Kernel based methods for volatility modelling: the problem of bandwidth selection

Stimatori kernel per la stima della volatilità: il problema della scelta della ampiezza di banda

Francesco Giordano and Maria Lucia Parrella [1]
Dipartimento di Scienze Economiche e Statistiche,
Università di Salerno, e-mail: giordano@unisa.it, mparrella@unisa.it

Riassunto: La volatilità dei rendimenti è considerata una misura di rischio nei mercati finanziari. Sono stati proposti in letteratura diversi e sofisticati modelli parametrici per l'analisi della volatilità, in grado di cogliere i principali e più importanti aspetti strutturali dei mercati finanziari. Come noto, questi modelli sono soggetti al rischio della errata specificazione. I metodi nonparametrici di tipo kernel, d'altra parte, svincolano lo stimatore della volatilità dalle specifiche ipotesi del modello, e sono utili sia ai fini di stima, sia ai fini della scelta e della verifica del modello parametrico. In cambio, la difficoltà tecnica di questi stimatori è rappresentata dalla selezione del parametro di smoothing, che determina in modo cruciale la consistenza dei risultati. In questo lavoro noi descriviamo gli aspetti pratici di tale questione, e confrontiamo il classico metodo plug-in proposto in letteratura per la scelta automatica del parametro di smoothing con un nuovo approccio basato sull'uso delle reti neurali. Presentiamo, infine, i risultati di uno studio di simulazione.

Keywords: Volatility modelling, kernel estimators, bandwidth selection, dependent data, neural networks.

1. Introduction

Since the ARCH model of Engle (1982), many sophisticated parametric volatility models have been proposed in the literature, which are able to capture the salient features of the underlying economic structures. See, for example, the TARCH model of Zakoïan (1994), the QTARCH model of Gouriéroux and Monfort (1992), the GARCH model of Bollerslev (1986), and so on. Nevertheless, there is always a danger that misspecification of a model leads to erroneous valuation and forecasts. On the other hand, nonparametric estimators of the volatility functions are not subject to the constraints related to the specific models assumed, and they adapt naturally even to highly nonlinear structures of the underlying unknown functions. The utility of nonparametric estimators is twofold: first of all, they give consistent estimations of the volatility functions (see Härdle and Tsybakov, 1997; Fan and Yao, 1998; Franke and Diagne, 2006); secondly, they can be used for finding simple and appropriate parametric models or even for testing them rigorously (Kreiss and Neumann, 1999; Fan and Huang, 2001; Chen *et al.*, 2003).
In particular, kernel based regression estimators have good properties, provided that they are correctly implemented (Masry and Fan, 1997, Masry and Tjøstheim, 1995). The most

---

appealing are the local polynomial estimators, which include the Nadaraya-Watson estimator as a particular case. The main difficulty with these kind of estimators is represented by the selection of the smoothing parameter, the *bandwidth* of the *kernel function*, which sensibly affects the consistency of the results.

In this paper, we study the problem of the selection of the optimal bandwidth in kernel based estimation of the volatility function. There are several approaches for the estimation of the optimal value of the bandwidth. The methods proposed so far in the literature may be divided into two broad categories: *cross-validation* methods and *plug-in* methods (for a dependent data context, see for example Härdle and Vieu, 1992, Hall *et al.*, 1995, Hart, 1996, Kim and Cox, 1997, Sköld, 2000, Masry and Fan, 1997). The plug-in approach is based on an analytical optimization: the unknown functionals that appear in the expression of the asymptotically optimal bandwidth are substituted by kernel estimates. The cross-validation methods are based on a numerical optimization: the mean squared prediction error $CV(h)$ is estimated through the leave-p-out estimator and than the optimal bandwidth is chosen numerically as the minimizer of $CV(h)$. The two approaches are based on different optimality criteria, so it is difficult to compare them analitically. Nevertheless, there is concordance in acknowledging a substantial superiority of plug-in procedures over cross-validation ones (see, for example, Chiu, 1991, Hall *et al.*, 1991, Park and Turlach, 1992, Ruppert *et al.*, 1995, Loader, 1999).

Let $\{X_t; t = 1, \ldots, n\}$ be a realization of length n from a real valued stationary stochastic process, $\{X_t; t \in \mathbb{N}\}$. In this paper, we consider the following nonlinear autoregressive model of order one

$$X_t = m\left(X_{t-1}\right) + s\left(X_{t-1}\right)\varepsilon_t, \tag{1}$$

where $m(\cdot)$ and $s(\cdot)$ are real valued functions defined on $\mathbb{R}$, with $s(\cdot) > 0$. The errors $\{\varepsilon_t\}$ are *i.i.d.* random variables with unknown density function f_ε. Model (1) is useful to analyze financial and econometric time series, which are generally characterized by nonlinear structures of the functions m and s (see, for example, Tjø stheim, 1994). Note that, for $x \in \mathbb{R}$, the function $m(x)$ represents the conditional mean function of the process, which may be interpreted as the predicted value based on the past information, while $s^2(x)$ is the volatility function, which measures the risk associated with this prediction. As a consequence, both estimations and forecasts of such quantities play a key role in the area of asset pricing, portfolio selection and risk management.

Now we present the assumptions needed in the paper. They refer to the mixing properties of the process, to the structure of the errors ε_t and of the functions m and s, and to the usual requirements concerning the kernel estimator and the neural network estimator.

A1 The errors have continuous and positive density function f_ε. Moreover, we have

$$E(\varepsilon_t^2) = 1, \quad E(\varepsilon_t) = E(\varepsilon_t^3) = 0, \quad \text{and} \quad E|\varepsilon_t|^\delta < \infty, \text{ with } \delta > 4.$$

A2 The function $m(\cdot)$ is supposed to have a continuous second order derivative on $\mathbb{R}$.

A3 The function $s(x)$ is positive on $\mathbb{R}$, and it has continuous second order derivative.

A4 There exist constants $C_1 > 0$ and $C_2 > 0$ such that, for $y \in \mathbb{R}$ we have

$$|m(y)| \le C_1(1 + |y|), \quad |s(y)| \le C_2(1 + |y|), \quad C_1 + C_2\left(E|\varepsilon_t|^\delta\right)^{1/\delta} < 1.$$

A5 The density function $f_X(\cdot)$ of the stationary distribution exists, is bounded, continuous and strictly positive in $\mathbb{R}$.

A6 The kernel function K is a density function defined on a compact set, say $[-1, 1]$.

A7 The bandwidth of the kernel h is of order $O(n^{-1/5})$.

A8 The number of nodes d in the hidden layer of the neural network estimator is such that $d = d(n) = O\left(\sqrt{n/\log n}\right)$.

A9 The weigth function w is symmetric, positive and it is such that $\int w(u)du = 1$ and $\int u^4 w(u)du < \infty$.

Remark 1: Under the conditions (A1) and (A5), the process $\{X_t\}$ is geometrically ergodic and exponentially β-mixing (Ango Nze, 1992; Doukhan, 1994).

The paper is organized as follows. In section 2 we describe the kernel based estimator of the volatility function. In section 3 we explain what is the role played by the bandwidth. In section 4, we present a new plug-in bandwidth selector based on the use of the neural networks. Finally, in section 5, we present the results of a simulation study which show the performances of different *plug-in* bandwidth estimators.

2. Kernel based estimators of the volatility function

Various nonparametric estimators of the functions m and s have been proposed in the literature. This section gives a brief overview on the nonparametric techniques based on the use of the kernel functions. In the context of estimating the volatility function of a process, kernel methods have been studied by many authors in the past. For example, Masry and Tjøstheim (1995) obtained strong convergence rates and asymptotic normality for kernel estimators of the functions m and s under α-mixing conditions. Härdle and Tsybakov (1997) treated the problem of estimating the volatility function directly, by using local polynomial regression methods. Independently, Fan and Yao (1998) apply the local linear technique to estimate the volatility function, using a slightly different estimator. Ziegelmann (2002) presents the local exponential estimator of the volatility function. Yang *et al.* (1999) have considered the joint estimation of both additive and multiplicative volatility. Laïb (2005) have established the strong uniform consistency and asymptotic normality of the kernel estimators of the functions m and s^2, when the process is stationary and ergodic but it does not satisfy any mixing condition. Franke *et al.* (2004) analize the bootstrap distribution of the kernel estimator of the volatility function.

Let $K_i(\cdot)$ be a *kernel function*, satisfying assumption (A6). Consider the column vectors $\mathbf{u} = (1, 0, \ldots, 0)^T$ and $\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_p\}^T$, both of length $p+1$. The local polynomial estimator of order p of a regression function $g(x) = E\{g(X_t)|X_{t-1} = x\}$ is equal to $\hat{g}(x; h) = \mathbf{u}^T \hat{\boldsymbol{\beta}}$, derived by solving the following weighted least squares problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{t=2}^{n} \left\{ g(X_t) - \sum_{j=0}^{p} \beta_j (X_{t-1} - x)^j \right\}^2 K\left(\frac{X_{t-1} - x}{h}\right). \qquad (2)$$

The positive real number h represents the *bandwidth* of the kernel estimator. For simplicity of notation, we will not indicate explicity the dependence of the bandwidth on the length of the series, as implied by assumption (A7). The (2) includes several different setups by considering different values of p and $g(\cdot)$. For example, for $p = 0$ we have the Nadaraya-Watson estimator, which is usually considered as the classic kernel regression estimator. For $p = 1$ we have the local linear estimator and for $p = 2$ the local quadratic

estimator. The local polynomial estimator of order p, with $p \geq 1$, is also useful for deriving an estimate of the derivatives of the function $g(x)$, up to the order p. By considering particular functions of $g(\cdot)$, we obtain estimators of the conditional moments of the process, as well as of the conditional distribution of the process. The asymptotic properties of the local polynomial estimator reported in (2) have been derived by several researchers. For mixing processes, see for example Masry and Fan (1997).

The functions $m(\cdot)$ and $s(\cdot)$ can be regarded as particular regression functions, since they represent (combinations of) conditional moments of the process. There are substantially two different approaches for estimating the volatility function s^2. We will describe separately each one. First of all, thanks to the stationarity of the process, the volatility function $v(x) = s^2(x)$ can be decomposed as follows

$$v(x) = m_2(x) - m^2(x),$$

where $m_2(x) = E(X_t^2|X_{t-1} = x)$ and $m(x) = E(X_t|X_{t-1} = x)$, so a direct estimator of the volatility is the following:

$$\hat{v}_1(x; h_1, h_2) = \hat{g}_2(x; h_2) - \left\{ \hat{g}_1(x; h_1) \right\}^2, \tag{3}$$

where $\hat{g}_1(x; h_1)$ and $\hat{g}_2(x; h_2)$ are respectively the kernel based estimators for $m(x)$ and $m_2(x)$. They are obtained by considering respectively $g_1(z) = z$ and $g_2(z) = z^2$ in (2). The estimator $\hat{v}_1(x; h_1, h_2)$ has been analyzed, for example, in Yao and Tong (1994), Härdle and Tsybakov (1997), Fan and Yao (1998). We refer, in particular, to the results derived in Härdle and Tsybakov (1997), which derive the asymptotic normality of such estimator. The inconvenience with this estimator is that it can have large bias and it can produce a negative estimate od the volatility function, especially if different smoothing parameters are used in estimating $m(x)$ and $m_2(x)$. Härdle and Tsybakov proposed an improved version of the estimator (3), by using a common bandwidth and a common kernel function. For this reason, we will refer later on in the paper to their estimator and we will denote it with $\hat{v}_1(x; h_2)$.

The alternative estimator of $v(x)$ is considered, for example, by Fan and Yao (1998), Franke *et al.* (2004), Kreiss and Neumann (1999), Hall and Carrol (1989) and Ziegelmann (2002). We refer in particular to the results reported in the first paper. It is known that

$$v(x) = E\left\{ \left[X_t - m(x) \right] | X_{t-1} = x \right\}^2.$$

Now consider the estimated squared residuals $\hat{r}(X_t; h_1) = \left\{ X_t - \hat{g}_1(X_{t-1}; h_1) \right\}^2$. The residual based estimator of the volatility function is equal to

$$\hat{v}_2(x; h_1, h_3) = \hat{g}_3(x; h_1, h_3),$$

derived by considering the function $g_3(z) = \hat{r}(z; h_1)$ in the (2), as follows

$$\hat{\boldsymbol{\beta}}_3 = \arg\min_{\boldsymbol{\beta}_3} \sum_{t=2}^{n} \left\{ \hat{r}(X_t; h_1) - \sum_{j=0}^{p} \beta_{3j}(X_{t-1} - x)^j \right\}^2 K_3\left(\frac{X_{t-1} - x}{h_3} \right). \tag{4}$$

Note that we use eventually a different kernel function K_3. Fan and Yao (1998) derived the asymptotic normality of the local linear estimator $\hat{v}_2(x; h_1, h_3)$, obtained by fixing $p = 1$ in the (4). One would expect some loss in convergence for this estimator, since the local regression in (4) is based on a previous kernel estimation $\hat{r}(X_t; h_1)$. This would require generally the necessity of undersmoothing the nonparametric link

estimated function in order to achieve a faster rate of consistency (see, for example, Kreiss and Neumann, 1999). In contrast, the estimator $\hat{v}_2(x; h_1, h_3)$ is regression adaptive, in the sense that the volatility function can be estimated asymptotically as well as if the conditional mean function $m(x)$ were known (Hall and Carrol, 1989, Fan and Yao, 1998).

3. The role played by the bandwidths

In the implementation of local polynomial regression estimators two parameters must be selected: the order of the local polynomial fit and the bandwidth of the kernel function. These parameters play a crucial role in the performance of the estimators, since they both influence the mean square error of the estimators. It is very difficult to select the optimal estimator by tuning simultaneously both these parameters. However, a good quality of approximation can be reached by choosing an appropriate bandwidth when using a fixed order of fit. In this paper we focus on the problem of selecting the optimal bandwidth for the local linear estimator (when the order of the polynomial p is fixed to 1).

Let us consider first the estimator $\hat{v}_1(x; h_2)$ of Härdle and Tsybakov (1997). The mean squared error is asymptotically equal to

$$
\begin{aligned}
\widetilde{MSE}\{\hat{v}_1(x; h_2)\} &= E\{\hat{v}_1(x; h_2) - v(x)\}^2 \\
&= \frac{v^2(x)\lambda_4}{n h_2 f_X(x)} \int K^2(u)du + \frac{\sigma_K^4 h_2^4}{4}\left[v''(x) + 2\{m'(x)\}^2\right]^2.
\end{aligned}
\tag{5}
$$

Here $\lambda_4 = E\{(\varepsilon_1^2 - 1)^2\}$, $f_X(x)$ is the density function of X_t and σ_K^4 represents the squared variance of the kernel function. The first term in the last expression of the (5) represents the variance, while the second term is the squared bias of the estimator. By evaluating the expression in (5), it can be seen what is the role played by the bandwidth h_2: it involves a trade-off between bias and variance, since a relatively large bandwidth brings to a reduction in the variance of the nonparametric estimator, whereas the bias of the estimator increases. Miminization of the (5) with respect to h_2 leads to the asymptotical optimal value of the local bandwidth

$$
h_2^{\mathrm{opt}}(x) = \left(\frac{C_K v^2(x)\lambda_4}{f_X(x)\left[v''(x) + 2\{m'(x)\}^2\right]^2}\right)^{1/5} n^{-1/5}.
\tag{6}
$$

The constant $C_K = \sigma_K^{-4}\int K^2(u)du$ is known, because it depends only on the kernel function. Note that the (6) represents a local optimal bandwidth, since it depends on the value of x. As expected, the local variability of the process and of the kernel function have a direct effect on the size of the bandwidth, while the local density of the process and the local smoothness of the function v determine an inverse effect on it. Therefore, a constant global bandwidth $h^{\mathrm{g.opt}}$ can be sufficient if the unknown volatility function has a high smoothness. This global bandwidth can be selected by minimizing a global measure of the estimation error, for example the asymptotical Mean Integrated Squared Error

$$
MISE(\hat{v}_1; h_2) = \int \widetilde{MSE}\{\hat{v}_1(x; h_2)\} w(x) f_X(x)dx.
\tag{7}
$$

For the sake of generality, we introduce the weigth function $w(x)$ in the construction of the MISE. This will enable us to generalize our neural network bandwidth selector

to different context, such as the estimation of a variable bandwidth. The asymptotically optimal global bandwidth is the bandwidth which minimizes the (7)

$$h_2^{\text{g.opt}} = \arg\min_{h_2} MISE(\hat{v}_1; h_2) = \left(\frac{C_K \lambda_4 R(v)}{R_f[v'' + 2(m')^2]} \right)^{1/5} n^{-1/5}. \tag{8}$$

We introduce the operators $R(\cdot)$ and $R_f(\cdot)$ to denote, respectively, the squared integral with respect to the Lesbeague measure and with respect to the measure $F_X(\cdot)$.
The functionals $R(v)$ and $R_f[v'' + 2(m')^2]$ are then equal to

$$R(v) = \int v^2(x)w(x)dx \tag{9}$$

$$R_f\left[v'' + 2(m')^2\right] = \int \left\{ v''(x) + 2[m'(x)]^2 \right\}^2 w(x)f_X(x)dx \tag{10}$$

If we consider the estimator $\hat{v}_2(x; h_1, h_3)$ of Fan and Yao (1998), the expression of the approximated mean squared error modifies in the following way

$$\widetilde{MSE}\{\hat{v}_2(x; h_1, h_3)\} = \frac{v^2(x)\lambda_4}{nh_3 f_X(x)} \int K^2(u)du + \frac{\sigma_K^4 h_3^4}{4} \left[v''(x)\right]^2.$$

Note that h_1 is the bandwidth used in the estimation of the function $m(x)$, in order to get the residuals $\hat{r}(X_t; h_1)$. This bandwidth does not affect the leading part of the *MSE*, as shown in Fan and Yao (1998). This simplifies the selection procedure of the smoothing parameter h_3, as it can be choosen independently from h_1. Nevertheless, it remains the fact that the use of the estimator $\hat{v}_2(x; h_1, h_3)$ implies the necessity of selecting two different bandwidths. By following the same arguments as before, we obtain the two asymptotical formulas for the bandwidths h_1 and h_3

$$h_1^{\text{g.opt}} = \arg\min_{h_1} MISE(\hat{m}; h_1) = \left(\frac{C_K R(s)}{R_f(m'')} \right)^{1/5} n^{-1/5}, \tag{11}$$

$$h_3^{\text{g.opt}} = \arg\min_{h_3} MISE(\hat{v}_2; h_3) = \left(\frac{C_K \lambda_4 R(v)}{R_f(v'')} \right)^{1/5} n^{-1/5}. \tag{12}$$

Here the functionals $R(s)$, $R_f(m'')$ and $R_f(v'')$ are defined as

$$R(s) = \int s^2(x)w(x)dx, \quad R_f(m'') = \int [m''(x)]^2 w(x)f_X(x)dx, \tag{13}$$

$$R_f(v'') = \int [v''(x)]^2 w(x)f_X(x)dx. \tag{14}$$

In conclusion, it is evident what is the importance of the correct selection of the smoothing parameter in kernel regression. All the optimal bandwidths (8), (11) and (12) are of order $O(n^{-1/5}) = Cn^{-1/5}$, as stated in assumption (A7). However, the constant C may assume very different values in the three cases and such values are very difficult to guess at a glance. An automatic data-driven bandwidth selection procedure should be then useful.

4. The plug-in method and the neural network bandwidth selector

The plug-in method for the selection of the optimal bandwidth starts with the analitical derivation of the asymptotical optimal bandwidth, that is the value of the bandwidth which minimizes the integrated mean squared error of the estimator. For the example of the conditional variance, such value is represented by the expression (8) for the estimator $\hat{v}_1$ and by the expression (12) for the estimator $\hat{v}_2$. The basic idea of the plug-in approach is to substitute the unknown functionals which appear in the expression of the asymptotical optimal bandwidth with some consistent estimates. As far as some derivative functions are involved, local polynomial estimators are usually (again) considered for the estimation of such quantities. Therefore, plug-in methods are generally implemented as multi-step procedures: in the preliminary stage, some pilot bandwidths must be selected for the auxiliary estimations of the unknown functionals in the (8) and (12) (one pilot bandwidth for each unknown function to estimate!); such estimates are then used in the final step (*plug-in*) to get the estimation of the optimal bandwidth.

The main criticism directed at plug-in methods is that they are heavily dependent on the correct specification of the pilot bandwidths in the first stage, being heavily biased when this specification is wrong (Loader, 1999, Giordano and Parrella, 2006b). Things run better if one starts with some consistent estimation of the "optimal" pilot bandwidths, but this would require the implementation of a specific "pilot" bandwidth selector, which usually requires the estimation of higher order derivatives of the unknown functions. Thus, the situation seems like that of a dog which tries to bite its tail.

In Giordano and Parrella (2006a)(2006b), we proposed a new plug-in method for the selection of the optimal bandwidth based on the use of the Neural Network technique. We showed the consistency of the selection method for the estimation of the global optimal bandwidth for nonparametric homoscedastic autoregressive models. Here we adapt the procedure to the nonparametric estimation of nonlinear heteroscedastic models.

The main advantage of our method lies in the fact that it does not depend on the selection of a preliminary pilot bandwidth, contrary to what happens with the traditional plug-in methods. The only tuning parameter to identify is the number of nodes in the hidden layer of the neural network function, but this can be easily done, thanks to the several automatic procedures available for this purpose. We refer to the paper of Giordano and Parrella (2006b) for a detailed description of the method, together with the theoretical justification and the computational performance of it, compared with the traditional plug-in approach. We describe first the neural network estimator. A Feedforward Neural Network, with one lagged variable and one hidden layer, is defined by the function

$$q(X_{t-1}; \boldsymbol{\eta}_i) = \sum_{k=1}^{d} c_{ik} \phi\left(a_{ik}, X_{t-1} + b_{ik}\right) + c_{i0}, \qquad i = 1, 2, \qquad (15)$$

where $\boldsymbol{\eta}_i = (c_{i0}, c_{i1}, \ldots, c_{id}, a_{i1}, \ldots, a_{id}, b_{i1} \ldots b_{id})$ is the vector of parameters to estimate, and d is the number of nodes of the hidden layer, satisfying assumption (A8). The quantities c_{i0} and $b_{i1} \ldots b_{id}$ are called bias terms, while $\phi(\cdot)$ is a sigmoidal activation function, *i.e.* a bounded measurable function on $\mathbb{R}$ with $\phi(u) \to 1$ as $u \to \infty$ and $\phi(u) \to 0$ as $u \to -\infty$, and $\phi \in C^\infty(\mathbb{R})$. The Neural Network estimator of a regression function $g_i(X_{t-1}) = E\{g_i(X_t)|X_{t-1}\}$ is equal to $q(X_{t-1}; \hat{\boldsymbol{\eta}}_i)$, given by solving

$$\hat{\boldsymbol{\eta}}_i = \arg\min_{\boldsymbol{\eta}_i} \sum_{t=2}^{n} \left[g_i(X_t) - q(X_{t-1}; \boldsymbol{\eta}_i)\right]^2, \qquad i = 1, 2, \qquad (16)$$

where $g_1(z) = z$ and $g_2(z) = z^2$, as before. Neural Network estimators have good approximation capabilities (Hornik, 1991). In this paper, we consider the logistic activation function $\phi(z) = [1 + \exp(-z)]^{-1}$. Moreover, we assume that the neural network estimator has the properties of the *approximate sieve extremum* estimator of Chen and White (1999).

The plug-in neural network bandwidth selector is based on the neural network estimation of the unknown functionals which appear in the formula of the asymptotically optimal bandwidth. Note that we can reformulate the (10) and the (14) in the following way

$$R_f\left[v'' + 2(m')^2\right] = \int \left\{m_2''(x) - 2m(x)m''(x)\right\}^2 w(x)f_X(x)dx$$

$$R_f[v''] = \int \left\{m_2''(x) - 2[m'(x)]^2 - 2m(x)m''(x)\right\}^2 w(x)f_X(x)dx.$$

Now using these last expressions, we propose the following neural network estimators of the functionals $R_f[v'' + 2(m')^2]$ and $R_f[v'']$.

$$\hat{R}(v) = \frac{1}{n^*}\sum_{t=1}^{n^*}\left[q(x_t; \hat{\boldsymbol{\eta}}_2) - q^2(x_t; \hat{\boldsymbol{\eta}}_1)\right]^2 w(x_t; \hat{\tau}), \tag{17}$$

$$\hat{R}_f[v'' + 2(m')^2] = \frac{1}{n-1}\sum_{t=2}^{n}\left\{q''(X_{t-1}; \hat{\boldsymbol{\eta}}_2) + \right. \tag{18}$$
$$\left. -2q(X_{t-1}; \hat{\boldsymbol{\eta}}_1)\,q''(X_{t-1}; \hat{\boldsymbol{\eta}}_1)\right\}^2 w(X_{t-1}; \hat{\tau}).$$

Here $\{x_1, x_2, \ldots, x_{n^*}\}$ is a uniformly spaced values on a subset of $\mathbb{R}$, with $n^* << n$. The notation $q''(X_{t-1}; \hat{\boldsymbol{\eta}}_1)$ denotes the second derivative of the function $q(X_{t-1}; \hat{\boldsymbol{\eta}}_1)$, obtained by deriving the expression (15) in corrispondence of $\hat{\boldsymbol{\eta}}_1$. The weight function $w(x; \hat{\tau})$ is taken as the density of the normal distribution with mean zero and variance $\hat{\tau}^2$ equal to the estimated variance of X_t. Substituting the estimations (17) and (18) in the (8), we get an estimation of the optimal bandwidth for the volatility estimator of Härdle and Tsybakov. As concerning the bandwidths h_1 and h_3, we propose to use

$$\hat{R}(s) = \frac{1}{n^*}\sum_{t=1}^{n^*}\left[q(x_t; \hat{\boldsymbol{\eta}}_2) - q^2(x_t; \hat{\boldsymbol{\eta}}_1)\right] w(x_t; \hat{\tau}), \tag{19}$$

$$\hat{R}_f(m'') = \frac{1}{n-1}\sum_{t=2}^{n}\left[q''(X_{t-1}; \hat{\boldsymbol{\eta}}_1)\right]^2 w(X_{t-1}; \hat{\tau}), \tag{20}$$

$$\hat{R}_f(v'') = \frac{1}{n-1}\sum_{t=2}^{n}\left\{q''(X_{t-1}; \hat{\boldsymbol{\eta}}_2) - 2\left[q'(X_{t-1}; \hat{\boldsymbol{\eta}}_1)\right]^2 + \right.$$
$$\left. -2q(X_{t-1}; \hat{\boldsymbol{\eta}}_1)\,q''(X_{t-1}; \hat{\boldsymbol{\eta}}_1)\right\}^2 w(X_{t-1}; \hat{\tau}), \tag{21}$$

Note that the expressions from (17) to (21) are based only on two neural network estimations, *i.e.* $q(X_{t-1}; \hat{\boldsymbol{\eta}}_1)$ and $q(X_{t-1}; \hat{\boldsymbol{\eta}}_2)$. The derivatives which appear in the previous formulas are obtained directly by deriving appropriately the expression (15). On the other hand, a specific estimation for each derivative is generally required in the traditional approach.

Theorem: If the conditions (A1)-(A5) and (A8) hold, then

$$\int_{\mathbb{R}} \left[m(x) - q\left(x; \hat{\eta}_1\right) \right]^2 f_X(x) dx \xrightarrow{p} 0$$

$$\int_{\mathbb{R}} \left[s^2(x) + m^2(x) - q\left(x; \hat{\eta}_2\right) \right]^2 f_X(x) dx \xrightarrow{p} 0$$

Proof
Given the assumptions (A1)-(A5) and (A8), the result follows applying the Lemma 2 in Giordano and Parrella (2006b).

5. The computational performance of the bandwidth selectors

We made some simulations in order to test our bandwidth selection procedure and to compare it computationally with the traditional plug-in bandwidth selector based on the use of the local polynomial derivative estimation. In this section we present the results. We consider the following two models:

$$X_t = (0.3 + 0.3|X_{t-1}|)\,\varepsilon_t \tag{22}$$

$$X_t = \frac{0.7}{1 + \exp(-X_{t-1})} + \{\phi(X_{t-1} + 1.2) + 1.5\phi(X_{t-1} - 1.2)\}\,\varepsilon_t. \tag{23}$$

The errors ε_t are *i.i.d.* and normally distributed, with zero mean and unit variance. It can be easily shown that both the models are geometrically ergodic and exponentially β-mixing, since they satisfy the assumptions (A1)-(A5). Note that the first model has only one point for which the first derivative does not exist, and this does not have any effect on the estimation. Model (23) is similar to that considered in the paper of Härdle and Tsybakov (1997).

We considered two different time series lengths $n = (500, 1000)$. For each model, we generated 200 replications for a given time series length. We consider the problem of the bandwidth selection in the local linear estimation of the volatility function by using both the estimators described in section 2. Thus we have to estimate the five functionals reported in (9), (10), (13) and (14). We use two plug-in methods: the neural network plug-in approach described in section 4 and the traditional plug-in approach, based on the local polynomial estimation of the previous functionals. In all the simulations, we use the Epanechnikov kernel, which is defined as $K(u) = 0.75(1 - u^2)$, for $|u| \leq 1$.

As a first step, we simulated the true values of the unknown functionals which appear in the asymptotical formulas of the bandwidths by Monte Carlo simulations, considering 5000 realizations of lenght $n = 1000$ for each model. These values can be used to derive the "true" optimal bandwidths. The results are reported in the table below

Figure 1 and 2 report the boxplots of the estimated functionals for models (22) and (23), using the neural network and the local polynomial approaches. In particular, the last procedure has been used with a pilot bandwidth near to the optimal value. The panel on the left considers time series of length $n = 500$, while the other panel reports the estimates for $n = 1000$. The first two boxplots of each panel refer to the estimates of the functional in (10) obtained respectively by our procedure (.NN) and by the traditional Local Polynomial estimator (.ker). The third and fourth boxplots concern the estimation of the functional in (9), while the last two refer to the functional $R_f(m'')$ in the (13). The

Table 1: *The true values for the functionals (9), (10), (13) and (14), derived by Monte Carlo simulations, for the two models (22) and (23).*

	$R(s)$	$R(v)$	$R_f(v'')$	$R_f(m'')$	$R_f[v'' + 2(m')^2]$
Model (22)	0.1064	0.0124	0.0231	0	0.0231
Model (23)	0.1617	0.0392	0.0839	0.0005	0.0646
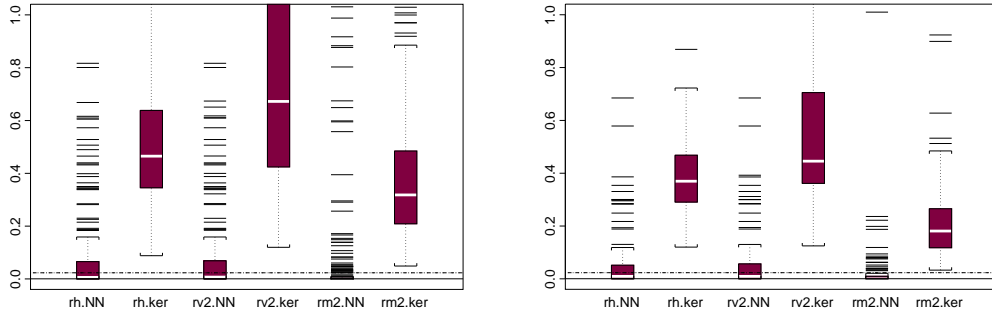
Figure 1: *Boxplots of the estimated functionals for model (22). The panel on the left considers time series of length $n = 500$, while the other panel reports the estimates for $n = 1000$. The first two boxplots of each panel refer to the estimates of the functional in the (10) obtained respectively by our procedure (.NN) and by the traditional Local Polynomial estimator (.ker). The third and fourth boxplots concern the estimation of the functional in the (9), while the last two refer to the functional $R_f(m'')$ in the (13).*

dashed line refers to the true value of the functional $R_f(v'')$, while the dotted line refers to the true value of the functional $R_f(v'' + 2(m')^2)$ (note that for model (22) we have that $R_f(v'') = R_f(v'' + 2(m')^2)$).

It is evident from the box-plots that both the procedures are consistent, but the neural networks seem to outperform the local polynomial estimators. For the sake of brevity, we do not report the results relative to the other functionals $R(v)$ and $R(s)$.

We should stress the fact that the implementation of our procedure is relatively simple, because we do not have to guess any value for the tuning parameter d (number of nodes of the hiden layer). We just used in our simulations a BIC selection algoritmh which automatically selects the optimal value of d.

References

Ango Nze P. (1992) Critéries d'ergodicité de quelques modéles à représentation markovienne, *C.R. Acad. Sci. Paris*, 315, 1301–1304.
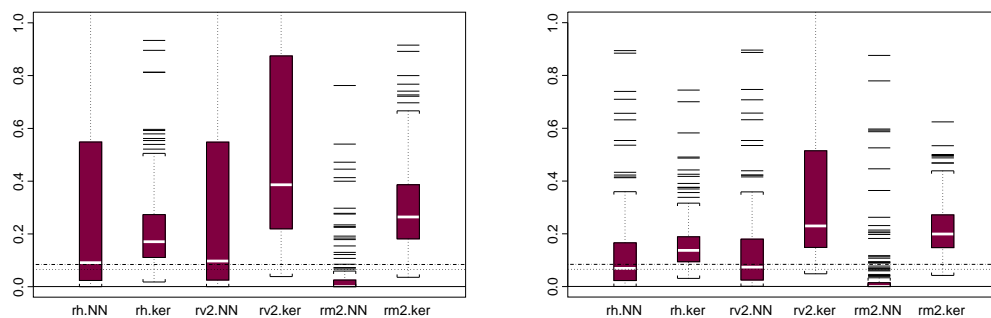
Figure 2: *Boxplots of the estimated functionals for model (23). The panel on the left considers time series of length $n = 500$, while the other panel reports the estimates for $n = 1000$. The first two boxplots of each panel refer to the estimates of the functional in (10) obtained respectively by our procedure (.NN) and by the traditional Local Polynomial estimator (.ker). The third and fourth boxplots concern the estimation of the functional in (9), while the last two refer to the functional $R_f(m'')$ in the (13).*

Bollerslev T. (1986) Generalized autoregressive conditional heteroscedasticity, *Econometrics*, 31, 307–327.

Chen S.X., Härdle W. and Li M. (2003) An empirical likelihood goosness-of-fit test for time series, *Journal of the Royal Statistical Society, series B*, 65, 663–678.

Chen X. and White H. (1999) Improved rates and asymptotic normality for nonparametric neural network estimators, *IEEE Transactions on Information Theory*, 45, 682–691.

Chiu S.T. (1991) Bandwidth selection for kernel density estimation, *Annals of Statistics*, 19, 1883–1905.

Doukhan P. (1994) *Mixing: properties and examples*, Lecture Notes in Statistics 85. Springer, New York.

Engle R. (1982) Autoregressive conditional heteroscedasticity with estimates of variance of u.k. inflation, *Econometrica*, 4, 987–1008.

Fan J. and Huang L. (2001) Goodness-of-fit test for parametric regression models, *Journal of the American Statistical Association*, 96, 640–652.

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

Franke J. and Diagne M. (2006) Estimating market risk with neural networks, *Statistics & Decision*, 24, 233–253.

Franke J., Neumann M.H. and Stockis J.P. (2004) Bootstrapping nonparametric estimators of the volatility function, *Jounal of Econometrics*, 118, 189–218.

Giordano F. and Parrella M.L. (2006a) Local polynomial vs neural networks: some empirical evidences, *Proceedings of the National Congress on Time Series Research, April 18-19, Rome*.

Giordano F. and Parrella M.L. (2006b) Neural networks for bandwidth selection in local linear regression of time series, *accepted for the pubblication on Computational*

Statistics and Data Analysis.

Gouriéroux C. and Monfort A. (1992) Qualitative threshold arch models, *Journal of Econometrics*, 52, 159–199.

Hall P. and Carrol R.J. (1989) Variance function estimation in regression: the effect of estimating the mean, *Journal of the Royal Statistical Society, Series B*, 51, 3–14.

Hall P., Lahiri S.L. and Truong Y.K. (1995) On bandwidth choice for density estimation with dependent data, *Annals of Statistics*, 23, 2241–2263.

Hall P., Sheather S.J., Jones M.C. and Marron J.S. (1991) On optimal data-based bandwidth selection in kernel density estimation, *Biometrika*, 78, 263–270.

Härdle W. and Tsybakov A. (1997) Local polynomial estimation of the volatility function in nonparametric autoregression, *Econometrica*, 81, 223–242.

Härdle W. and Vieu P. (1992) Kernel regression smoothing of time series, *Journal of Time Series Analysis*, 13, 209–232.

Hart J. (1996) Some automated methods of smoothing time-dependent data, *Journal of Nonparametric Statistics*, 6, 115–142.

Hornik K. (1991) Approximation capabilities of multilayer feedforward networks, *Neural Networks*, 4, 251–257.

Kim T.Y. and Cox D.D. (1997) Bandwidth selection in kernel smoothing of time series, *Journal of Time Series Analysis*, 17, 49–63.

Kreiss J.P. and Neumann M. (1999) *Bootstrap test for parametric volatility structure in nonparametric autoregression*, in: Prob. Theory and Math. Statistics (B. Grigelionis et al. (eds.)).

Laïb N. (2005) Kernel estimates of the mean and the volatility functions in a nonlinear autoregressive model with arch errors, *Journal of Statistical Planning and Inference*, 134, 116–139.

Loader C.R. (1999) Bandwidth selection: classical or plug-in?, *Annals of Statistics*, 27, 415–438.

Masry E. and Fan J. (1997) Local polynomial estimation of regression functions for mixing processes, *Scandinavian Journal of Statistics*, 24, 165–179.

Masry E. and Tjøstheim D. (1995) Nonparametric estimation and identification of nonlinear arch time series: strong convergence and asymptotic normality, *Econometric Theory*, 11, 258–289.

Park B.U. and Turlach B.A. (1992) Practical performance of several data driven bandwidth selectors, *Computational Statistics*, 7, 251–270.

Ruppert D., Sheather S.J. and Wand M.P. (1995) An effective bandwidth selector for local least squares regression, *Journal of the American Statistical Association*, 90.

Sköld M. (2000) A bias correction for cross-validation bandwidth selection when a kernel estimate id based on dependent data, *Journal of Time Series Analysis*, 22, 493–503.

Tjø stheim D. (1994) Non-linear time series: a selective review, *Scandinavian Journal of Statistics*, 21, 97–130.

Yang L., Härdle W. and Nielsen J. (1999) Nonparametric autoregression with multiplicative volatility and additive mean, *Journal of time series analysis*, 20, 579–604.

Yao Q. and Tong H. (1994) Quantifying the influence of initial values on nonlinear prediction, *Journal of Royal Statistical Society*, 56, 701–725.

Zakoïan J.M. (1994) Threshold heteroscedastic models, *Journal of Economic Dynamics and Control*, 18, 931–955.

Ziegelmann F.A. (2002) Nonparametric estimation of volatility functions: the local exponential estimator, *Econometric Theory*, 18, 985–991.