Nonlinear and Nonparametric Methods for Analyzing Financial Time Series

Jürgen Franke, University of Kaiserslautern

Summary: We consider nonparametric generalization of various well-known financial time series models and study estimates of the trend and volatility functions and forecasts based on kernel smoothers as well as on neural networks.

1. A Nonparametric Black-Scholes model in discrete time

During the last decade, a lot of parametric models for financial time series have been studied. To decide which of these models is appropriate for a particular data set, nonparametric methods may be used analogously to the situation in regression analysis (Härdle, 1990). Using the latter, one may derive confidence bands for trend and volatility functions, develop tests of particular parametric hypotheses against quite general alternatives or construct nonparametric forecasts which do not depend on specific and perhaps restrictive assumptions on the structure of the time series. We illustrate this approach with three examples, starting from a rather straightforward nonparametric ARCH-model which may also be interpreted as a generalization of the Black-Scholes model in discrete time, then proceed to a nonparametric stochastic volatility model and finally consider forecasting a autoregressive time series with exogeneous components using a neural network-based approach for estimating the prediction function nonparametrically.

The Black-Scholes approach to option pricing is based on modelling the price $s(t)$ of the underlying asset at real-valued time t as a geometric Wiener process, i.e. as a solution of the stochastic differential equation

$$ds(t) = \mu\, s(t)dt + \sigma\, s(t)dw(t)$$

where, $w(t)$ is a standard Wiener process, and the unknown parameters μ and σ are interpreted as the expected return per time unit and as the volatility of the price process. A discrete-time version of this model is (compare, e.g., Hull, 1997, ch. 10.4)

$$S_{t+1} = (1+\mu)S_t + \sigma\, S_t \varepsilon_{t+1} \tag{1}$$

where ε_t, t integer, are independent standard normal random variables. Over a reasonable period of time, this type of model provides a useful approximation to, e.g., stock prices or foreign exchange rates. However, under particular circumstances, assumptions like a constant volatility independent of time and price level are violated. We therefore consider a generalization of (1):

$$S_{t+1} = m(S_t) + \sigma(S_t)\varepsilon_{t+1} \tag{2}$$

where ε_t are independent identically distributed (i.i.d.) random variables with mean 0 and variance 1. $m(x)$, $\sigma(x)$ are arbitrary functions up to some regularity assumptions. For the particular choice

$$m(x) = 0 \quad \text{and} \quad \sigma^2(x) = \omega + \alpha x^2\,,$$

(2) becomes an ARCH(1)-process (Engle, 1982), the ancestor of the popular ARCH-GARCH-family of financial time series models (Bollerslev et al., 1992). (2), therefore, may be seen as a nonparametric generalization of an ARCH(1)-model, too.

Nonparametric estimates of the trend function m and the volatility function σ have been the topic of many papers starting with the work of Robinson (1983). A popular choice of estimates are kernel estimates constructed in an analogous manner to the Nadaraya-Watson estimates in regression analysis (Härdle, 1990) from a sample $S_0, \ldots, S_T$:

$$\hat{m}(x,h) = \frac{1}{Th}\sum_{t=1}^{T} K\Big(\frac{x - S_{t-1}}{h}\Big) S_t / \hat{p}(x,h) \tag{3}$$

$$\hat{\sigma}^2(x,h) = \frac{1}{Th}\sum_{t=1}^{T} K\Big(\frac{x - S_{t-1}}{h}\Big) S_t^2 / \hat{p}(x,h) - \hat{m}^2(x,h) \tag{4}$$

where a kernel estimate of the stationary marginal density $p(x)$ of the S_t is used:

$$\hat{p}(x,h) = \frac{1}{Th}\sum_{t=1}^{T} K\Big(\frac{x - S_t}{h}\Big). \tag{5}$$

$K(u)$ is a given kernel function satisfying some standard regularity assumptions, e.g.:

(A1) $0 \leq K(u) \leq C_K$ for some constant C_K, $\int uK(u)du = 0$, $\int u^2 K(u)du < \infty$, $\int K(u)du = 1$.

The bandwidth h controls the smoothness of the function estimates $\hat{m}, \hat{\sigma}, \hat{p}$. For simplicity of notation, we have chosen the same bandwidth for all three estimates. From a theoretical viewpoint, this is not necessarily so, and in applications, bandwidths for estimating trend and volatility may differ.

The asymptotic properties of the estimates $\hat{m}, \hat{\sigma}$ and $\hat{p}$ are now well understood. Robinson (1983) has shown asymptotic normality, and Collomb and Härdle (1986) have proven uniform strong consistency. Asymptotic expansions of bias and variance have been derived by Auestad and Tjostheim (1990) and Masry and Tjostheim (1994). However, asymptotic normality provides reasonable approximations to, e.g., the distribution of $\hat{m}$ only for quite large sample sizes as has been illustrated by Franke and Wendel (1992) in a simulation study. For practical purposes like calculating confidence bands or testing parametric hypotheses against nonparametric alternatives, other approximations are of interest like, e.g., the bootstrap which, since Efron's (1979) seminal paper, has become a standard tool in a statistical analysis. In nonparametric situations, however, the bootstrap has to be designed with some care as straightforward applications of the general ideas do not always work.

For kernel estimates of the autoregression function m and the volatility function σ, various bootstrap procedures have been studied recently by Franke et al. (1997). For sake of illustration, we consider only one of them, and apply it to get confidence bands for m and σ based on some stock index series. To begin with, we impose some assumptions on the model (2):

(A2) m and σ are Lipschitz continuous functions with Lipschitz constants L_m, L_σ resp.

(A3) $\sigma_0 = \inf_{x \in \mathbb{R}} \sigma(x) > 0$.

(A4) The innovations ε_t are i.i.d. random variables with mean 0, variance 1 and probability density p_ε satisfying $\inf_{x \in C} p_\varepsilon(x) > 0$ for all compact sets C.
(A5) $L_m + L_\sigma \, \mathcal{E}|\varepsilon_1| < 1$.

These assumptions are stronger than what is really needed but, in this form, easy to formulate. They guarantee that there exists a stationary and geometric ergodic solution S_t, $-\infty < t < \infty$, of (2). Geometric ergodicity implies, in particular, the strong or α-mixing property with exponentially decreasing mixing coefficients. Therefore, the time series $\{S_t\}$ shares many properties with sequences of i.i.d. random variables like a central limit theorem for the mean, Bernstein inequalities etc.

If $\sigma \equiv 1$, (2) is a nonlinear autoregression of order 1, and for general σ it also defines a Markov process. Therefore, we mimick the familiar residual-based bootstrap for autoregressions (compare, e.g., Kreiß and Franke, 1992.). As a first step, we calculate sample residuals

$$\hat{\varepsilon}_t = \frac{S_t - \hat{m}(S_{t-1}, g)}{\hat{\sigma}(S_{t-1}, g)},$$

i.e. we solve (2) for ε_t and replace the unknown functions m, σ by some initial estimates of kernel type, but with arbitrary bandwidth g. Some of these sample residuals are rather crude approximations of their true counterpart ε_t. $\hat{m}(x, g)$ is formed by a local smoothing operation and, therefore, is an unreliable estimate of $m(x)$ if only few data S_k, $1 \le k \le T$, happen to be close to x. Due to our above assumptions, with large probability this will only happen if $|x| > \gamma_T$ for some constant γ_T depending on T. Therefore, we retain only the reliable sample residuals $\hat{\varepsilon}_t$, $t \in A$, where $A = \{t = 1, \ldots, T;\ |S_{t-1}| \le \gamma_T\}$. Then, the remaining residuals are centered:

$$\tilde{\varepsilon}_t = \hat{\varepsilon}_t - \frac{1}{|A|} \sum_{k \in A} \hat{\varepsilon}_k$$

such that the empirical distribution $\hat{F}_T$ of the $\tilde{\varepsilon}_t$, $t \in A$, has mean 0. For theoretical purposes only, we have to smooth $\hat{F}_T$ by convoluting it with $H_b(u) = \frac{1}{b} H(\frac{u}{b})$ where H is a probability density with mean 0 and variance 1, e.g. the standard normal density. The smoothed empirical law $\hat{F}_{T,b} = \hat{F}_T * H_b$ then has a density. Now, we draw i.i.d. bootstrap residuals ε_t^*, $t = 1, \ldots, T$, from the distribution $\hat{F}_{T,b}$, and we get the bootstrap sample $S_1^*, \ldots, S_T^*$ by

$$S_t^* = \hat{m}(S_{t-1}^*, g) + \hat{\sigma}(S_{t-1}^*, g) \varepsilon_t^* \ , \ \ t = 1, \ldots, T,$$

with, for sake of simplicity, $S_0^* = S_0$. Now, we can calculate bootstrap versions $\hat{m}^*(x, h)$, $\hat{\sigma}^*(x, h)$ and $\hat{p}^*(x, h)$ by replacing the original data in (3) - (5) by the bootstrap resample $S_0^*, \ldots, S_T^*$.

It can be shown, that this type of bootstrap works in the sense that the distributions of the adequately centered and scaled estimates $\hat{m}^*(x, h)$, $\hat{\sigma}^*(x, h)$ in the bootstrap world provide good approximations of the distributions of $\hat{m}(x, h)$ and $\hat{\sigma}(x, h)$ in the real world. To formulate a corresponding theorem we need some technical assumptions:

(A6) $m, \sigma, p_\varepsilon, H$ are twice continuously differentiable with bounded derivatives. $\mathcal{E}\varepsilon_t^6 < \infty$, and $\sup_{x \in \mathbf{R}} |x \ p_\varepsilon'(x)| < \infty$. $\int v^4 H(v) dv < \infty$, $\int v^2 |H'(v)| dv < \infty$. K has compact support $[-1, +1]$, is symmetric and three times continuously differentiable.

(A7) $b \to 0$ and $g/b^{12} \to 0$ as $T \to \infty, \gamma_T \to \infty$, $\gamma_T / \log T$ is bounded and $\inf_{|x| \leq 2\gamma_T/\sigma_0} p_\varepsilon(x) \geq (g \log T)^2$ as $T \to \infty$.

Theorem 1 (Franke et al., 1997): *Assume (A1)-(A7). Let* $g, h \to 0$ *for* $T \to \infty$ *such that* $Th^5 \longrightarrow c \geq 0$ *and* $g \sim T^{-\alpha}$ *with* $0 < \alpha \leq \frac{2}{15}$. *Then, for all* $x \in \mathbb{R}$:

$$d_K(\sqrt{Th}\{\hat{m}^*(x,h) - \hat{m}(x,g)\},\ \sqrt{Th}\{\hat{m}(x,h) - m(x)\}) \underset{p}{\longrightarrow} 0$$

$$d_K(\sqrt{Th}\{\hat{\sigma}^{*2}(x,h) - \hat{\sigma}^2(x,g)\},\ \sqrt{Th}\{\hat{\sigma}^2(x,h) - \sigma^2(x)\}) \underset{p}{\longrightarrow} 0$$

Here, for real random variables X, Y with distribution functions F, G resp.,

$$d_K(X,Y) \equiv d_K(F,G) \equiv \sup_{u \in \mathbf{R}} |F(u) - G(u)|$$

denotes the Kolmogorov distance between the distributions of X and Y.

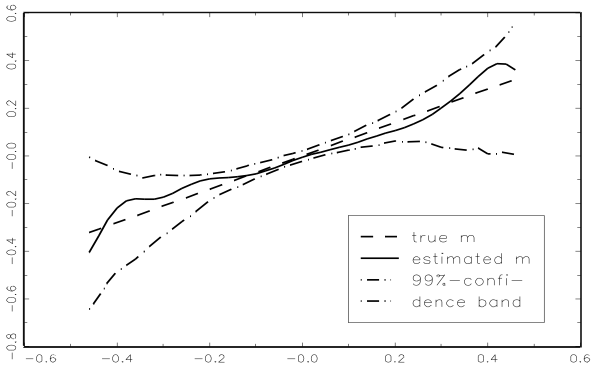

Figure 1a: true and estimated m

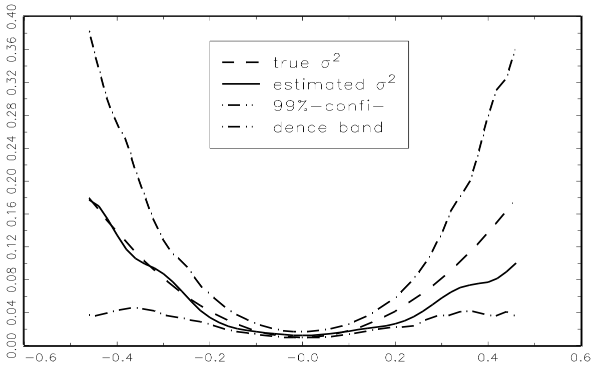

Figure 2a: True and estimated σ^2

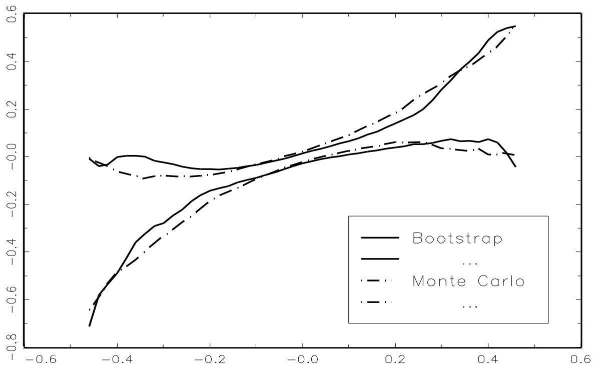

Figure 1b: 99%-confidence bands for m

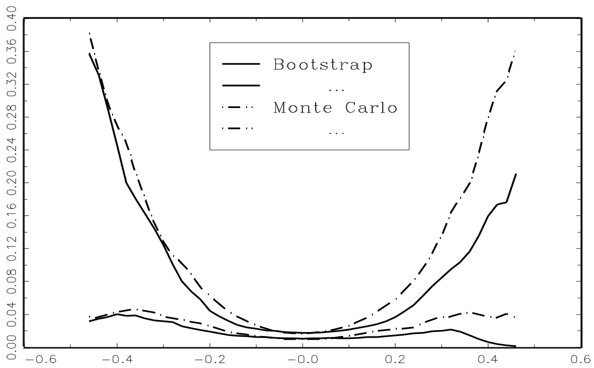

Figure 2b: 99%-confidence bands for σ^2

To illustrate the performance of the bootstrap, Figures 1 and 2 show the results of a small Monte Carlo study for a time series S_t of type (2) with linear trend $m(x) = 0.7x$, ARCH-like volatility $\sigma^2(x) = 0.01 + 0.8x^2$ and standard normal innovations ε_t . Figure 1a shows the true function m, a kernel estimate $\hat{m}(x,h)$ with $h = 0.05$ and a set of 99%-confidence intervals for $m(x)$ for an equidistant grid of values of x , joined to form a pointwise confidence band. The latter is based on a Monte Carlo-simulation of size 500. Figure 1b shows again the true 99%-confidence band of

Figure 1a together with a bootstrap approximation based on the same sample as the kernel estimate of Figure 1a. For calculating the bootstrap band, 500 resamples were used where as the bandwidth of the reference kernel estimate $g = 0.06$ and as the cut-off point $\gamma_T = 10 \log T$ have been chosen. Figure 2 shows the same type of picture for σ^2 instead of m using the same bandwidths h and g . In both cases, the bootstrap confidence bands approximate the true confidence quite well, in particular in the range of values x close to 0 where the process S_t stays most of the time.

2. General nonparametric stochastic volatility models

A common approach to improve the Black-Scholes model allows the volatility to be not only time-dependent or - as in chapter 1 above - a function of past prices but to be a stochastic process $\sigma(t)$ in its own right. Various stochastic volatility models have been proposed which differ with respect to the stochastic differential equation prescribing the evolution of the volatility process. To apply these models some parameters have to be estimated, and for this purpose time-discrete versions are considered. As a result of such a discretization process starting from a continuous-time stochastic volatility model investigated by Scott (1987) and Wiggins (1987), Taylor (1994) studies the so-called lagged autoregressive random variance (LARV) model for the return $R_t = \log(S_t/S_{t-1})$ of an asset with price S_t at time t:

$$\begin{aligned} R_t &= \mu + \sigma_{t-1}\varepsilon_t \\ \log \sigma_t &= \omega + \beta\{\log \sigma_{t-1} - \omega\} + \vartheta\, \varepsilon_t^\sigma \end{aligned}$$

If we assume, for sake of simplicity, that the mean return μ is known and if we use the abbreviations

$$\begin{aligned} X_t &= \frac{1}{2}\log(R_t - \mu)^2 - \mu_\eta, \\ \eta_t &= \frac{1}{2}\log \varepsilon_t^2 - \mu_\eta, \\ \xi_t &= \log \sigma_t \quad \text{and} \quad \tilde{\varepsilon}_t = \vartheta\, \varepsilon_t^\sigma, \end{aligned}$$

where μ_η is chosen such that $\mathcal{E}\eta_t = 0$, the LARV model can be written as

$$X_t = \xi_{t-1} + \eta_t, \qquad \xi_t = m(\xi_{t-1}) + \tilde{\varepsilon}_t \tag{6}$$

with $m(u) = \omega + \beta\{u - \omega\}$. For general function m instead, (6) models the log-volatility ξ_t as a non-parametric autoregression of order 1 as discussed in chapter 1. However, ξ_t is not observed directly. For estimating m, only the transformed returns X_t are available. (6) formulates the relation between observables X_t and non-observables ξ_t as an errors-in-variables model, and as a common fact related to such situations, we have to assume that the distribution of the "errors" η_t is known to have any chance to infer structural properties of the process $\{\xi_t\}$ from the observation $\{X_t\}$. If we retain the normality assumption of the LARV for the ε_t then the corresponding η_t would be distributed like $\frac{1}{2}$ times the logarithm of a χ_1^2-random variable centered such that $\mathcal{E}\eta_t = 0$. As in chapter 1, we impose some assumptions on m and on the laws of the $\tilde{\varepsilon}_t, \eta_t$ which, among other things, guarantee the existence of stationary solution of (6):

(B1) m is twice continuously differentiable

(B2) For some $\vartheta > 0$ $\tilde{\varepsilon}_t$ are i.i.d. $\mathcal{N}(0, \vartheta^2)$-random variables, and there exist i.i.d. $\mathcal{N}(0,1)$-random variables ε_t such that $\eta_t = \frac{1}{2}\log \varepsilon_t^2 - \mu_\eta$ and $(\varepsilon_t, \tilde{\varepsilon}_t)$ are i.i.d. bivariate Gaussian variables with $\mathrm{corr}(\varepsilon_t, \tilde{\varepsilon}_t) = \rho$ for some $|\rho| < 1$.

As an immediate consequence of (B1), (B2), the stationary density p_ξ of the log volatilities ξ_t is twice continuously differentiable. If the ξ_t would be i.i.d. themselves, it would be possible to estimate p_ξ nonparametrically from observations $X_1, \dots, X_T$ by a so-called deconvolution density estimate as described by Carroll and Hall (1988). The basic idea is the following: If $\xi_1, \dots, \xi_T$ would be known, we could estimate p_ξ directly by a kernel estimate as in (5). By (6), the density p_x of the X_t is the convolution of p_ξ and the known density p_η of the η_t : $p_x = p_\xi * p_\eta$. Now, we estimate the Fourier transform ϕ_x of p_x from $X_1, \dots, X_T$ by the sample characteristic function

$$\hat{\phi}_x(\omega) = \frac{1}{T} \sum_{t=1}^{T} e^{i\omega X_t}.$$

Then, we deconvolute, i.e. in the Fourier domain we divide by the known Fourier transform ϕ_η of p_η, and, simultaneously, we smooth which, in the Fourier domain, becomes a multiplication with the scaled Fourier transform ϕ_K of some kernel function K. Finally, we apply an inverse Fourier transformation and get the desired estimate $\hat{p}(x,h)$ of $p_\xi(x)$:

$$\hat{p}(x,h) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\omega x} \phi_K(\omega h) \frac{\hat{\phi}_x(\omega)}{\phi_\eta(\omega)} \, d\omega. \tag{7}$$

This can be rewritten as a common kernel density estimate

$$\hat{p}(x,h) = \frac{1}{Th} \sum_{t=1}^{T} K_h\Big(\frac{x - X_t}{h}\Big) \tag{8}$$

depending on the observables X_t only. The kernel K_h depends on the bandwidth h explicitly:

$$K_h(u) = \frac{1}{2\pi} \int e^{-i\omega u} \frac{\phi_K(\omega)}{\phi_\eta(\omega/h)} d\omega. \tag{9}$$

By analogy, Fan and Truong (1993) have proposed deconvolution kernel estimates for regression functions by replacing the fixed kernel K in the Nadaraya-Watson estimate by the variable kernel K_h of (9). Their approach, transferred to the problem of estimating the autoregressive function m of (6), provides the following estimate based on data $X_0, \dots, X_T$

$$\hat{m}(x,h) = \frac{1}{Th} \sum_{t=1}^{T} K_h(\frac{x - X_{t-1}}{h}) \, X_t \, / \, \hat{p}(x,h) \tag{10}$$

where K_h and $\hat{p}$ are given by (9), (8). It can be shown that $\hat{m}$ is really a consistent estimate of m :

Theorem 2 (Franke et al., 1998): *Let the Fourier transform ϕ_K of K have a compact support $[-\kappa, \kappa]$, assume (B1), (B2), and let h converge to 0 as $c/\log T$ for some $c > \kappa\pi/2$. Then,*

$$\hat{m}(x,h) = m(x) + O_p((\log T)^{-2}).$$

The slow speed of convergence is typical for such nonparametric errors-in-variables models and is not caused by the time series structure we have assumed. It may be improved if other laws of the ε_t are assumed instead of the Gaussian (compare Fan and Truong, 1993).

3. Neural network based time series models

One drawback of the kernel estimates of the previous chapters is their practical limitation to low-dimensional situations. A straightforward generalization of (2) to higher-order models like

$$S_{t+1} = m(S_t, S_{t-1}, ..., S_{t-\tau}) + \sigma(S_t, S_{t-1}, ..., S_{t-\tau})\varepsilon_{t+1} \tag{11}$$

can be handled with kernel estimates only for small τ or under specific restrictions of the functions m and σ. An example of the latter are the generalized additive models where $m(x_0, ..., x_\tau) = m_0(x_0) + ... + m_\tau(x_\tau)$ and only one-dimensional functions have to be estimated nonparametrically.

An approach which works quite well in practice and additionally allows for including exogeneous variables is based on neural networks which within the last decade have become a popular tool for solving classification, regression and forecasting problems in particular in finance (Anders, 1997). To give a flavour of basic ideas and typical theoretical results we first consider a regression model:

$$Y_t = m(X_t) + \varepsilon_t \tag{12}$$

where we assume

(C1) $X_1, X_2, \ldots$ are i.i.d. d-dimensional random vectors with density $p(x)$. $\varepsilon_1, \varepsilon_2, \ldots$ are independent real random variables and $\mathcal{E}\{\varepsilon_t | X_t = x\} = 0$, $\mathcal{E}\{\varepsilon_t^2 | X_t = x\} = \sigma_\varepsilon^2(x) < \infty$. There exist constants C_n such that $\mathcal{E}\{|\varepsilon_t|^n | X_t = x\} \leq C_n < \infty$ is satisfied for all x, n.
(C2) m is continuous and bounded, $\sigma_\varepsilon^2(x)$ is continuous and $0 < \delta \leq \sigma_\varepsilon^2(x) \leq \Delta < \infty$ for all x.

We consider only feedforward neural networks with one hidden layer for estimating the function m. The basic building block is the so-called neuron which provides a nonlinear transformation of a linear combination of the inputs $x = (x_1, ..., x_d)' : x \mapsto \psi(u_0 + u_1 x_1 + ... u_d x_d)$, where ψ is the fixed activation function, e.g. a function of sigmoid form like the logistic function

$$\psi(s) = \tfrac{1}{1+e^{-s}}.$$

If we combine H of these neurons with, of course, differing weights u_k, we get as the function represented by the network

$$f_H(x, \vartheta) = v_0 + \sum_{h=1}^{H} v_h \psi(\tilde{x}' w_h)$$

where $\vartheta = (w_1', \ldots, w_H', v_0, \ldots, v_H)'$ is the vector of network weights with $w_h' = (w_{0h}, \ldots, w_{dh})$, $h = 1, \ldots, H$, and $\tilde{x}' = (1, x')$ is the input vector augmented by a bias component 1. The parametrization of the network function is not unique, as certain simple symmetry operations applied to the weight vectors obviously do not change the value of $f_H(x, \vartheta)$. To avoid this ambiguity we restrict ϑ to some large compact set Θ_H where we do not have problems with identifiability.

Network functions like $f_H(x, \vartheta)$ are universal approximators, i.e., for $H \to \infty$, they are able to approximate quite general functions m arbitrarily well (Hornik et al, 1989). Using a sample of noisy observations $(X_1,Y_1),...,(X_N,Y_N)$ from the model (12), for given H a network function is fitted to the data replacing the unknown ϑ by its nonlinear least squares estimate $\widehat{\vartheta}_N$, i.e. the solution of

$$\widehat{D}_N(\vartheta) \equiv \frac{1}{N}\sum_{t=1}^{N}(Y_t - f_H(X_t,\vartheta))^2 = \min_{\vartheta\in\Theta_H}!$$

For $N\to\infty$, $\widehat{\vartheta}_N$ converges to ϑ_0, the parameter vector for which the corresponding network function $f_H(x,\vartheta)$ provides the best approximation of the true regression function m , i.e. the solution of

$$\begin{aligned} D_0(\vartheta) &\equiv \mathcal{E}(Y_t - f_H(X_t,\vartheta))^2 \\ &= \int[(m(x) - f_H(x,\vartheta))^2 + \sigma_\varepsilon^2(x)]p(x)dx = \min_{\vartheta\in\Theta_H}! \end{aligned}$$

To show that $\widehat{\vartheta}_N$ is even asymptotically Gaussian, we impose some rather weak regularity conditions on the network. Here, ∇ and ∇^2 denote the gradient and the Hessian with respect to ϑ.

(C3) The activation function ψ is bounded and twice continuously differentiable with bounded derivatives.
(C4) $D_0(\vartheta)$ has a unique global minimum at ϑ_0 lying in the interior of Θ_H, and $\nabla^2 D_0(\vartheta_0)$ is positive definite.

To formulate the theorem, we introduce the sample version of $D_0(\vartheta)$:

$$D_N(\vartheta) = \frac{1}{N}\sum_{t=1}^{N}\left[(m(X_t) - f_H(X_t,\vartheta))^2 + \sigma_\varepsilon^2(X_t)\right],$$

and we define ϑ_N as the minimizer of $D_N(\vartheta)$. Instead of $\widehat{\vartheta}_N - \vartheta_0$, we consider its two components $\widehat{\vartheta}_N - \vartheta_N$ and $\vartheta_N - \vartheta_0$ separately which asymptotically are independent.

Theorem 3 (Franke and Neumann, 1998): *Suppose that (C1) - (C4) are satisfied. Then, for* $N\to\infty$*,*

$$\sqrt{N}\begin{pmatrix}\widehat{\vartheta}_N - \vartheta_N \\ \vartheta_N - \vartheta_0\end{pmatrix} \xrightarrow[d]{} \mathcal{N}\left(0, \begin{pmatrix}\Sigma_1 & 0 \\ 0 & \Sigma_2\end{pmatrix}\right),$$

i.e. $\sqrt{N}(\widehat{\vartheta}_N - \vartheta_N)$ *and* $\sqrt{N}(\vartheta_N - \vartheta_0)$ *are asymptotically independent normal random vectors with covariance matrices* Σ_1 *and* Σ_2*, respectively, where*

$$\begin{aligned} &\Sigma_1 = A(\vartheta_0)^{-1}B_1(\vartheta_0)A(\vartheta_0)^{-1},\;\; \Sigma_2 = A(\vartheta_0)^{-1}B_2(\vartheta_0)A(\vartheta_0)^{-1} \\ \textit{with}\quad &B_1(\vartheta) = 4\cdot\int\sigma_\varepsilon^2(x)\nabla f_H(x,\vartheta)\cdot\nabla f_H(x,\vartheta)'p(x)dx \\ &B_2(\vartheta) = 4\cdot\int(m(x) - f_H(x,\vartheta))^2\nabla f_H(x,\vartheta)\cdot\nabla f_H(x,\vartheta)'p(x)dx \end{aligned}$$

[1cm] and $A(\vartheta) = \nabla^2 D_0(\vartheta)$.

As an immediate consequence, $\sqrt{N}(\hat{\vartheta}_N - \vartheta_0)$ is asymptotically normal with mean 0 and covariance matrix $\Sigma_1+\Sigma_2$. Remark that in the correctly specified case, where $m(x) = f_H(x,\vartheta_0)$, Σ_2 is equal to the zero matrix, as then $\vartheta_N = \vartheta_0$.

As for the kernel estimates of chapter 1, asymptotic normality may be improved by the bootstrap where due to the heteroskedasticity of the errors ε_t of (12) the wild bootstrap has to be used (Franke and Neumann, 1998). In a series of papers, White (1989a, b; 1990) has not only shown asymptotic normality of $\widehat{\vartheta}_N$ for given H and the dependence of the asymptotic covariance matrix of $\sqrt{N}(\hat{\vartheta}_N - \vartheta_0)$ on the numerical procedure used for calculating the nonlinear least squares estimate, but he has also shown that the neural network function estimate $f_H(x, \widehat{\vartheta}_N)$ of m is consistent if the number H of neurons in the hidden layer increases with the sample size N with an appropriate rate. White does not only discuss the regression case with independent data $(X_1, Y_1), ..., (X_N, Y_N)$, but he allows (X_t, Y_t) to be generated by a stationary stochastic process satisfying appropriate mixing conditions. Therefore, an analogon of Theorem 3 holds in the time series context too under suitable conditions, e.g. for estimating the conditional expectation of S_{t+1} given the past in a model like

$$S_{t+1} = m(S_t, S_{t-1}, ..., S_{t-\tau}, X_t) + \varepsilon_{t+1} \tag{13}$$

where S_t is a real valued time series to be predicted, X_t is a d-dimensional time series of exogeneous variables and suitable transformations of past data $S_k, k \leq t$, and the ε_t are independent real random variables.

To illustrate the performance of neural network estimates for the autoregression function m of (13), we consider as an example the forecasting of stock prices. The data are daily closing prices of 28 major Dutch stocks which dominate the CBS index for the years 1993 - 1996. We use the data up to the end of 1995 for preliminary model selection, i.e. for deciding on the number H of neurons and on the number and type of inputs used for forecasting, and for estimation the model parameters. The 1996 data are used for the validation part of the procedure allowing to test the performance of various models on data not used for model building and parameter estimation. Additionally to the raw stock prices, we implement various transformations like moving averages, envelopes or average directional movement indicators commonly used in technical market analysis (Müller and Nietzer, 1993, Welcker, 1994) as well as several intermarket data like index prices, foreign exchange rates, international interest rates or the MG base metal price as inputs in our models. They correspond to the exogeneous variables X_t in (13).

The given data are used to design a procedure for calculating 3-months- (60 trading days-) forecasts of the stock prices which serve as the basis for developping a trading strategy. Point forecasts of time series over such a large lag are notoriously unreliable. Therefore, we concentrated our efforts on predicting the trend correctly, i.e. if the stock price will increase significantly (i.e. move upwards by more than 5% of its original value), decrease significantly (i.e. move downwards by more than 5% of its original value), or stay essentially at the same level (i.e. does not change by more than 5%).

We tested various neural network functions $f_H(S_t, S_{t-1}, ..., S_{t-\tau}, X_t, \vartheta)$ as approximations of m in (13) where the inputs were selected using knowledge of experts and statistical model selection procedures. As the model showing the best performance we finally decided for a feedforward network with $H = 3$ neurons using $d = 25$ input variables (stock prices, technical indicators and intermarket data) out of more than 60 candidates. Including a constant input, this model uses 82 network weights. The corresponding network function f_H could probably be simplified if some of the network connections between input layer and hidden layer would be eliminated, i.e. some of the weights w_{jh} are set to 0 in advance. As the software used during this project did not allow for fitting incompletely connected networks we did not test this option. The least-squares estimates $\hat{\vartheta}_N$ of these parameters ϑ were calculated using Jacobs enhanced backpropagation (Jacobs, 1988). Details of the model building including comparisons of various numerical least-squares procedures are discussed in (Decker, 1998).

To test the performance of the forecasting procedure we predicted the price trend of all the 28 stocks at the end of each quarter of 1996. The results are given in the following table. The second column shows the percentage of correct classifications of the future price trends. As we allow for three different qualitative predictions, a purely random forecast would have a chance of 33% to be correct. The neural network does better but the performance is not outstanding considering the number of misclassifications. However, the chance to predict a significant increase of the price correctly is usually much larger. This is reflected by looking at the performance of a network buy and hold trading strategy: buy those stocks whose prices are predicted to rise significantly, hold them up to the end of the 60-days prediction interval and sell them afterwards. This strategy is compared to a index based buy and hold strategy where all 28 of the CBS stocks considered in the study are bought and held for 60 days. The return of both trading schemes is given in columns 3 and 4 of the table. The network strategy outperforms the index benchmark considerable in each quarter. This is quite remarkable, as, except for the second quarter, the Dutch stock market experienced a period of rapid growth in 1996, and it is commonly acknowledged by practitioners that it is not easy to beat the index under those circumstances.

time	prob. of correct forecast	return network	return index
1. quarter	43.30%	15.41%	11.22%
2. quarter	46.82%	2.38%	-0.34%
3. quarter	63.42%	6.94%	5.75%
4. quarter	55.05%	13.17%	11.84%
annual average	52.46%	9.76%	7.22%

It is worth mentioning that all of the 28 stock price time series considered were forecasted using exactly the same neural network, and the weights were estimated using information from all 28 stocks simultaneously. Finally, we checked the stability of the forecasting procedure if new information becomes available. Our original network uses weights estimated from information available end of 1995. Now, we added successively the data from January 1996 up to June 1996 and reestimated the prediction parameters based on each of those growing data sets, getting 6 new networks. We used each of those networks to forecast the price trend in the last quarter of 1996. The following table shows again the percentage of correct forecasts and the returns of the network trading strategy compared to the index-based strategy for the last quarter. The first row of the table repeats the performance of the original network. Looking at the table, there is no systematic improvement in the performance of the network if additional information is included for parameter estimation. Therefore, it does not seem to be necessary to redesign the network frequently; the forecast procedure may be used unchanged for some time which is of some importance in practice as the construction of a suitable prediction network requires a considerable effort.

data up to end of	prob. of correct forecast	return network	return index
12/95	55.05%	13.17%	11.84%
1/96	49.38%	13.02%	11.84%
2/96	61.85%	15.61%	11.84%
3/96	51.36%	14.99%	11.84%
4/96	59.07%	13.52%	11.84%
5/96	62.93%	13.75%	11.84%
6/96	46.26%	14.11%	11.84%

Acknowledgement: Financial support was received by the Deutsche Forschungsgemeinschaft, SFB 373 "Quantifikation und Simulation Ökonomischer Prozesse" at Humboldt-University, Berlin. The application reported in chapter 3 is part of joined project with Commerzbank AG, Frankfurt am Main where the main work has been done by F. Decker of the University of Kaiserslautern and D. Oppermann of Commerzbank. The numerical analysis has been done using the software package ThinksPro.

Corresponding Author: Jürgen Franke
Department of Mathematics and ITWM, University of Kaiserslautern
P.O. Box 3049
67653 Kaiserslautern, Germany

References

[1] Anders, U. (1997). Statistische neuronale Netze. Vahlen, München.

[2] Auestad, B. and Tjostheim, D. (1990). Identification of nonlinear time series: First order characteristics and order determination. Biometrika **77**, 669-687.

[3] Bollerslev, T., Chou, R.Y. and Kroner, K.F. (1992). ARCH-modelling in finance. J. Econometrics **52**, 5-59.

[4] Carroll, R.J. and Hall, P. (1988). Optimal rates of convergence for deconvoluting a density. J. Amer. Statist. Assoc. **83**, 1184-1186.

[5] Collomb, G. and Härdle, W. (1986). Strong uniform convergence rates in robust nonparametric time series analysis and prediction. Stochastic Proc. and Appl. **23**, 77-89.

[6] Decker, F. (1998). Kombination markttechnischer Indikatoren und Intermarketdaten zur Kursprognose marktbreiter internationaler Einzelwerte mit Hilfe neuronaler Netze. Dipl. thesis. Univ. of Kaiserslautern.

[7] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. Ann. Statist. **7**, 1-26.

[8] Engle, RF. (1982). Autoregressive conditional heteroskedasticity with estimates of UK inflation. Econometrica **50**, 987-1008.

[9] Fan, J. and Truong, Y.K. (1993). Nonparametric regression with errors-in variables. Ann. Statist. **21**, 1900-1925.

[10] Franke, J., Härdle, W. and Kreiß, J.P. (1998). Nonparametric estimation in a stochastic volatility model. Report in Wirtschaftsmathematik **37**, University of Kaiserslautern.

[11] Franke, J., Kreiß, J.P. and Mammen, E. (1997). Bootstrap of kernel smoothing in nonlinear time series. Discussion paper **20**, SFB 373, Humboldt University, Berlin.

[12] Franke, J. and Neumann, M. (1998). Bootstrapping neural networks. Report in Wirtschaftsmathematik **38**, University of Kaiserslautern.

[13] Franke, J. and Wendel, M. (1992). A bootstrap approach for nonlinear autoregressions - some preliminary results. In Jöckel, K.H., Rothe, G. and Sendler, W. eds.: Bootstrapping and Related Techniques. Lecture Notes in Economics and Mathematical Systems **376**. Springer, New York.

[14] Härdle, W. (1990). Applied Nonparametric Regression. Cambridge University Press, Cambridge.

[15] Hull, J.C. (1997). Options, Futures and Other Derivatives. Prentice Hall, New York.

[16] Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks **2**, 359-366. bibitem17 Jacobs, R.A. (1988). Increased rates of convergence through learning rate adaptation. Neural Networks **1**, 295-307.

[17] Kreiß, J.P. and Franke, J. (1992). Bootstrapping stationary autoregressive moving average models. J. Time Ser. Anal. **13**, 297-317.

[18] Masry, E. and Tjostheim, D. (1994). Nonparametric estimation and identification of nonlinear ARCH time series: Strong convergence and asymptotic normality. Econometric Theory **11**, 258-289.

[19] Müller, Th. and Nietzer, H. (1993). Das große Buch der technischen Indikatoren. TM Börsenverlag.

[20] Neumann, M.H. and Kreiß, J.P. (1996). Bootstrap confidence bands for the autoregression function. Discussion paper **75**, SFB 373, Humboldt University, Berlin.

[21] Robinson, P.M. (1983). Nonparametric estimators for time series. J. Time Series Anal. **4**, 185-207.

[22] Scott, L.O. (1987). Option pricing when the variance changes randomly: Theory, estimation and application. J. Financial Quant. Anal. **22**, 419-438.

[23] Taylor, S.J. (1994). Modelling stochastic volatility: A review and comparative study. Mathem. Finance **4**, 183-204.

[24] Welcker, J. (1994) Technische Aktienanalyse. Verlag Moderne Industrie.

[25] White, H. (1989a). Some asymptotic results for learning in single hidden-layer feedforward network models. J. Amer. Statist. Assoc. **84**, 1008-1013.

[26] White, H. (1989b). Learning in artificial neural networks: a statistical perspective. Neural Computation **1**, 425-464.

[27] White, H. (1990). Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. Neural Networks **3**, 535-550.

[28] Wiggins, J.B. (1987). Option values under stochastic volatility: Theory and empirical estimates. J. Financial Econ. **19**, 351-372.