Chapter 5

LOCAL POLYNOMIAL PREDICTION AND VOLATILITY ESTIMATION IN FINANCIAL TIME SERIES

Zhan-Qian Lu
National Institute of Standards and Technology, USA
e-mail: john.lu@nist.gov.

Abstract Chaos and nonlinear theory has significant impact on the analysis of economic and financial time series. Nonlinearity plays an important role in explaining the empirical features of asymmetric business cycles, clustered volatility, and regime switching in finance data. In this Chapter, we will focus the popular local polynomial prediction method and its applications to chaotic time series prediction and financial volatility estimation. Volatility and conditional covariance estimation is important in many aspects of modern finance theory. We introduce a nonparametric volatility model, called local ARCH, and propose a weighted least square method for goodness of fit. The statistical theory is based on a martingale regression framework developed in Lu (1999a,b), which includes a wide variety of nonlinear time series models, such as nonlinear autoregression, ARCH, and nonlinear vector autoregressive models. The daily AOL stock data is used as an example to illustrate the developed techniques. First, we apply the nonlinear regression procedure to model the spread-volume relationship—We find a nice power-law relationship in all appropriate periods after discovering that the spurious nonlinearity in the overall data is due to nonstationarity. We also find a vastly changing structure in GARCH models fitted to different parts of the return rate series based on closing prices. We apply the developed local ARCH theory to a stationary subseries of the return series, and find some encouraging results.

Keywords: Nonlinear prediction, volatility modeling, interval prediction, local ARCH, martingale regression, chaotic time series, regime switching, nonparametric technique.

1. Introduction

Chaos theory has significant impact on the analysis of economic and financial time series. With the explosion of high-frequency and real-time ticker

data, nonlinearity theory will play an increasingly important role in understanding the complex dynamics and stylized facts in microstructure finance data. Asymmetric business cycles, clustering variability, and regime shift are just a few glimpses of nonlinearity in action. In this Chapter, we will focus on an important nonlinear prediction procedure using local polynomial fits and will discuss applications to chaotic time series prediction and financial volatility estimation. Local polynomial prediction method is a popular approach to nonlinear deterministic modeling and chaos prediction (Farmer and Sidorowich 1988). Recently, Lu (1999a) has developed an interesting statistical explanation of why local method is effective for low-dimensional modeling and chaos prediction, even in the presence of small noises and when the embedding dimension is very high. In the statistics literature, local polynomial regression is an established approach to nonparametric regression estimation (Fan and Gijbel 1996, Lu 1996a). Lu (1999b) extends the current statistical theory to stochastic nonlinear models which include most nonlinear time series models of Tong (1990) and Priestley (1988). In the following Chapter, we first review recent statistical theory on local polynomial regression, and then local prediction theory for both one-step and multi-step prediction, as well as non-mean prediction methods such as conditional distribution function and quantile estimation for prediction intervals.

Nonlinearity occurs in finance data under various guises (Franses and Dijk 2000). Regime switching is an important mechanism of introducing nonlinearity and flexibility in traditional models. Many macroeconomic and microeconomic data contain asymmetric cyclic components and Markov switching models have been a great success story (Hamilton 1994, Krolzig 1997). Recently, regime switching models have also been applied to model many empirical finance phenomena such as clustered volatility (Franses and Dijk 2000). While various ARCH models have been proposed for modeling volatility in finance time series (Gouriéroux 1997), it is hard to justify any of the proposed parametric models. Subsequently, there is an ever increasing literature on nonparametric volatility estimation (e.g. Fan and Yao 1998). We introduce yet another nonparametric volatility model, called local ARCH in this Chapter, and propose a weighted least square method as a more conservative measure of goodness of fit of volatility function.

Stationarity is a basic assumption in time series modeling. However, economic or financial time series data typically contain time-varying statistical properties such as trend and business cycles, and may contain stochastic nonstationary process such as unit roots and cointegration (e.g. Dhrymes 1997). To deal with nonstationarity, we propose a moving window approach for local modeling and focus on analysis using *local time models*. This approach seems to be effective in dealing with temporal changes and regime shifts in large data sets. We use the daily AOL stock data as an example to illustrate this idea. A nice power-law in spread-volume relationship is found in separate periods of the AOL data after discovering that the spu-

rious nonlinearity in overall data is due to nonstationarity. We also find a vastly changing structure in GARCH models fitted to different parts of the return rate series based on closing prices. We apply the local ARCH model to a stationary segment of the return series, and obtain some encouraging results.

2. Local polynomial method

Local polynomial method has been in use under various disguises in time series analysis for a long time. For example, local polynomial fitting generalizes *moving average* and *exponential smoothing*. Our focus is different from standard applications in that we are applying local polynomial fit to the *state space* and the goal is state-space based prediction. In the nonparametric regression literature, local polynomial regression generalizes the kernel and nearest neighbor method, and has some superior theoretical properties over standard methods in stochastic regression models(Fan and Gijbel 1996). In the multivariate case, Lu (1996a) develops bias and mean squared error formula for local polynomial regression, including the popular *local linear* and *local quadratic* regression. These results are useful for contructing confidence and prediction intervals. In the multivariate case, the use of bandwidth matrix (Lu, 1996a), has the similar effect to that of the *radial basis* function method in the neural network literature. In the nonlinear time series literature, Priestley (1988)'s *state dependent* models and Tong (1990)'s *threshold* models are closely related to the nonparametric method discussed here.

We will embed time series data in the state space regression form

$$\{(X_0, Y_1), (X_1, Y_2), \ldots, (X_{n-1}, Y_n)\} \tag{5.1}$$

where X_{i-1} is the reconstructed state vector at time $i-1$ and Y_i is the time series response at time i (or $i+T-1$ for $T-$step prediction). (Some general technical assumptions on how the data are generated from multivariate time series observations are given in Section 3). The local polynomial regression, also called the locally weighted regression, computes an estimate of $m(\mathbf{x}) = E(Y_{i+1}|X_i = \mathbf{x})$ at every state point $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$, by the *weighted* least square method based on the local regression model

$$Y = \mathbf{X}\boldsymbol{\beta} + E, \text{ where } E \sim N(0, \nu(\mathbf{x})W^{-1}), \tag{5.2}$$

where the design matrix $\mathbf{X}$ consists of row vectors of polynomial expansion terms up to certain degree around $\mathbf{x}$ evaluated at X_{i-1}, and W is a diagonal weight matrix determining the influence of each data point on $\mathbf{x}$ (controlled by a kernel function K and bandwidth parameter h), defined by

$$W = \text{diag}\{K(X_1, h), K(X_2, h), \ldots, K(X_n, h)\}.$$

Usually a *radial basis* kernel function is chosen, $K(X_i, h) = k(\|X_i - \mathbf{x}\|/h)$, where k is some univariate function on $[0, \infty]$, and $\|\cdot\|$ is the Euclidean

norm. More generally, one may employ the smoothing bandwidth matrix for multivarite data as $K(X_i, H) = k(\|H^{-1/2}(X_i - \mathbf{x})\|)$, where H is any positive definite matrix. The kernel k is said to have finite support if $k(x) = 0$ for $x > 1$. Use of finite support kernels has computational advantages over the Gaussian kernel in that only data points in the local neighborhood are used.

Specifically, the local linear fit at a given point $\mathbf{x} \in \mathbb{R}^p$ is defined by minimizing the weighted sum of squares

$$\sum_{i=1}^{n} \{Y_i - a - \mathbf{b}^T(X_{i-1} - \mathbf{x})\}^2 K(X_i, h), \tag{5.3}$$

over a: a real number, and $\mathbf{b}$: a $p-$dimensional vector. The parameter estimates $\hat{a}, \hat{\mathbf{b}}$ correspond respectively to the regression and partial derivative estimates of m at the point $\mathbf{x}$. The design matrix $\mathbf{X}$ in (5.2) consists of row vectors $(1, (X_i - \mathbf{x})), i = 0, 1, \ldots, n-1$.

The local quadratic fit at any given point $\mathbf{x} \in \mathbb{R}^p$ is based on minimizing the weighted sum of squares

$$\sum_{i=1}^{n} \{Y_i - a - \mathbf{b}^T(X_{i-1} - \mathbf{x}) - (X_{i-1} - \mathbf{x})^T L(X_{i-1} - \mathbf{x})\}^2 K(X_i, h), \tag{5.4}$$

where a is a real number, $\mathbf{b}$ is a $p-$dimensional vector, and L is a $p \times p$ matrix which is restricted to be a lower triangular matrix for identifiability. The solution from minimizing (5.4) corresponds to $\widehat{a} = \hat{m}(\mathbf{x})$, an estimate of regression function at $\mathbf{x}$, of $\hat{\mathbf{b}} = \hat{D}_m(\mathbf{x})$ which corresponds to an estimate of $D_m(\mathbf{x}) = (\partial m(\mathbf{x})/\partial x_1, \cdots, \partial m(\mathbf{x})/\partial x_p)^T$ at $\mathbf{x}$, and of $\hat{L}$ which corresponds to estimates of elements in the Hessian matrix of $H_m(\mathbf{x}) = (\partial^2 m(\mathbf{x})/\partial x_i \partial x_j)$ at $\mathbf{x}$. That is, $L(x) = (l_{ij})$ satisfies $l_{ij} = h_{ij}$ if $i > j$ and $l_{ij} = h_{ii}/2$ if $i = j$, where $H_m(\mathbf{x}) = (h_{ij})$ is the Hessian. Define the local design matrix $\mathbf{X}$ consisting of rows $(1, (X_i - \mathbf{x}), \text{vech}^T[(X_i - \mathbf{x})(X_i - \mathbf{x})])$ where vech is an operator of stacking *distinct* elements in a symmetric matrix into a row vector (for example, discarding elements above the diagonal).

The vector of local polynomial estimates such as $\mathbf{c} = (\hat{a}, \hat{\mathbf{b}}^T)^T$ from local linear fitting in (5.3) can be solved through

$$(\mathbf{X}^T W \mathbf{X})\mathbf{c} = \mathbf{X}^T W Y \tag{5.5}$$

and when $(\mathbf{X}^T W \mathbf{X})^{-1}$ is available,

$$\mathbf{c} = (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W Y. \tag{5.6}$$

Computational concerns. By smoothness assumption, local polynomial fit may be computed at some uniformly selected points in the state space and values at other points are then interpolated based on nearby points, an idea cleverly used in Cleveland et al (1992). To see how this idea reduces

computation dramatically, consider the situation when the n design points $x_0, x_1, \ldots, x_{n-1}$ lie uniformly on $[0,1]$, and the kernel function is symmetric and has finite support, and h corresponds to the radius of ℓ nearest neighbor, then the coefficient matrix $P_{\mathbf{x}} = (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W$ is the same for all $\mathbf{x}$ at interior point, and there are only ℓ different coefficient matrices at the boundary, $P_0, P_1, \ldots, P_{\ell-1}$. That is, the following correspondence:

$$\begin{array}{lccccccccc} \mathbf{x} = & x_0, & x_1, & \ldots, & x_{\ell-1}, & x_\ell, & \ldots, & x_{n-\ell}, & x_{n-\ell+1}, & \ldots, & x_{n-1} \\ \text{coeff} & P_0, & P_1, & \ldots, & P_{\ell-1}, & P_{\ell-1}, & \ldots, & P_{\ell-1}, & P_{\ell-1}, & \ldots, & P_0. \end{array}$$

Furthermore, the uniform design and symmetric kernel give rise to an orthogonal matrix $(\tilde{\mathbf{X}}^T W \tilde{\mathbf{X}})$ at interior points; That is, it is diagonal, and so the solution is given as

$$b_i = \frac{\tilde{\mathbf{x}}_i W Y}{(\tilde{\mathbf{x}}_i^T W \tilde{\mathbf{x}}_i)} = \frac{\sum k_j \tilde{x}_{ij} y_j}{\sum k_j \tilde{x}_{ij}^2}, \ i = 1, \ldots, q.$$

where $\tilde{\mathbf{x}}_i$ is the ith column of $\tilde{\mathbf{X}}$ and $k_i = K(X_i, h)$. Orthogonality is an important virtue in function approximation, and one may seek orthogonality by using the orthogonal polynomial basis which has been constructed for a given kernel function. The *lowess* or *loess* function (up to $p = 2$), which is available in SPLUS, is an implementation of some of these fast computational ideas for robust locally weighted polynomial fits (Cleveland et al 1992).

Prior information and qualitative knowledge may also be incorporated in locally weighted fit, in the same way that linear constraints are used in linear regression. In particular, one may impose restriction such as $a \geq 0$ for obtaining nonnegative regression estimates, or $b > 0$ for monotone function fit, or setting higher-order coefficients such as higher-order mixed derivatives to zero in additive and low-order models. If the least square estimate is in the feasible region, then the restricted estimate is the same as the unrestricted estimate; Otherwise, when one or more restrictions are in effect, the affected components assume the equality constraint values, and the rest of components are then modified accordingly.

For example, consider the linear constraints $A\beta = c$ for some known matrix A and vector c. Then, the constrained weighted least square estimate is given by

$$b_c = b + (\mathbf{X}^T W \mathbf{X})^{-1} A^T [A(\mathbf{X}^T W \mathbf{X})^{-1} A^T]^{-1} (c - Ab) \qquad (5.7)$$

(cf. Seber 1977). Note that $Ab_c = c$ so the linear constraint is satisfied, while other components of b is modified by adding the second term to reflect (update) this prior information.

3. Technical setup for statistical theory

The martingale nonlinear regression model, first discussed in Lu (1999b), is a natural and general setup for discussion of statistical theory for local

polynomial prediction in time series data. We consider the regression-type model:

$$Y_t = m(X_t) + \nu^{1/2}(X_t)\varepsilon_t \tag{5.8}$$

where $m : \mathbb{R}^p \to \mathbb{R}$ is some nonlinear function, $\nu \geq 0$ is a variance function. We further assume that

(**A**) $\{\varepsilon_i\}$ is a sequence of martingale differences with respect to a sequence of increasing $\sigma-$fields $\{\mathcal{F}_i\}$ such that $X_0 \in \mathcal{F}_0, X_i, \varepsilon_i \in \mathcal{F}_i$ for all $i \geq 1$ and $\mathrm{E}\{\varepsilon_i|\mathcal{F}_{i-1}\} = 0, \mathrm{E}\{\varepsilon_i^2|\mathcal{F}_{i-1}\} = 1$.

It should be pointed out that assumption (A) is very natural in the context of financial and economic time series (e.g. Bollerslev et al 1994). Indeed, broadly the martingale assumption follows from the finite-dimensionality or Markovian assumption of the underlying process. (A) is more general in the sense that only the first and second-order moments of the predictive distribution are assumed to be functions of a finite-number of past observations (as represented in vector X), whereas the usual Markovian property imposes this finite-dimensionality property for the *whole* predictive distribution.

Several familiar models are just special cases of this general setup, as explained below.

Model I. Nonlinear autoregression (NAR): Consider a scalar time series $\{y_t\}$, for integer p, we can always write

$$y_t = m(y_{t-1}, \ldots, y_{t-p}) + \nu^{1/2}(y_{t-1}, \ldots, y_{t-p})\varepsilon_t \tag{5.9}$$

where $m(y_{t-1}, \ldots, y_{t-p}) = E[y_t|y_{t-1}, \ldots, y_{t-p}]$ and $\nu(y_{t-1}, \ldots, y_{t-p}) = \mathrm{Var}(y_t|y_{t-1}, \ldots, y_{t-p}]$ and ε_t has zero mean, unit variance.

In the time series literature, the residual $\{\varepsilon_t\}$ is usually assumed independent for some big enough embedding dimension p. Under this assumption, (5.9) satisfies condition (A) with

$$Y_t = y_t, \;\; X_t = (Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}). \tag{5.10}$$

It is noted that, the embedded state vectors $\{X_t\}$ in the autoregression case is actually a Markov chain in $\mathbb{R}^p$, or equivalently $\{Y_t\}$ is an Markov chain of order p. When m is linear and ν is constant, this is the familiar $AR(p)$ model. When m is nonlinear and ν is constant, it is the classic nonlinear autoregression (NAR) model. Tong (1990) gave an extensive discussion of NAR models in nonlinear time series analysis. When both m is nonlinear and ν is time-varying (so called heteroscedastic variance), it is a hybrid of NAR and ARCH model, and is termed NAR-ARCH model.

Model II. Nonlinear multivariate models: Vector time series models are important in modeling the interdependence among two or more time series processes. There is very little written on nonlinear multivariate time series models. We argue that why a model like (5.8)

is relevant to nonlinear vector time series modeling. For simplicity we consider a bivariate time series given by $\{(y_{1t}, y_{2t})^T\}$, which is assumed to follow a Markov model or vector autoregression model given by

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} m_1(y_{1\ t-1}, y_{2\ t-1}) \\ m_2(y_{1\ t-1}, y_{2\ t-1}) \end{pmatrix} + \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} \tag{5.11}$$

where $\{(e_{1t}, e_{2t})^T\}$ is sequence of independent random vectors with zero mean and covariance structure, say $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$. Suppose our interest is in prediction, and Σ is usually unknown, for all practical purpose we can usually ignore the inter-dependence expressed in Σ, and then joint estimation of m_1 or m_2 is equivalent to the marginal estimation approach, in that m_1 and m_2 are estimated separately. The component model such as

$$y_{1t} = m_1(y_{1\ t-1}, y_{2\ t-1}) + \sigma_{11}^{1/2}\varepsilon_{1t}$$

satisfies the stochastic regression model (5.8) with $Y_t = y_{1t}, X_t = (y_{1\ t-1}, y_{2\ t-1})^T$.

Written in vector notation, let Y_t be a vector in $\mathbb{R}^p$, and $M = (m_1, \ldots, m_p)$ be a map $\mathbb{R}^p \to \mathbb{R}^p$, and $E_t = (e_{1t}, \ldots, e_{pt})$ is iid random vector with zero mean and covariance matrix Σ (where $p = 2$ in (5.11)). Then, a general vector NAR or noisy dynamical system model can be written as

$$Y_t = M(Y_{t-1}) + E_t. \tag{5.12}$$

Thus, model like (5.8) is relevant to time series prediction involving covariates.

Model III. Volatility models: Traditional time series usually assume a homoscedastic model such as in (5.11). However, in economic or financial time series, it is more realistic to assume that the conditional variance or volatility function is *time-varying*. In particular, we assume an *autoregressive* model for volatility, in which we assume that $\sigma_t = \mathrm{Var}(y_t|y_{t-1}, \ldots, y_{t-p})$ (in the multivariate case, the conditional covariance matrix $\Sigma_t = \mathrm{Cov}(Y_t|Y_{t-1})$) is a function of past variables only.

For general discussion, let $\mathcal{F}_t$ denote all relevant information at time t. For example, in the case of an pth-order Markov chain such as (5.9), $\mathcal{F}_t$ consists of $y_t, \ldots, y_{t-p+1}$. We write the residual process $e_t = y_t - \mathrm{E}(y_t|\mathcal{F}_{t-1})$ as

$$e_t = \sigma_t \eta_t, \tag{5.13}$$

where σ_t is a time-varying and measurable function of information $\mathcal{F}_{t-1}$ at time $t-1$, and η_t is iid and $E(\eta_t) = 0, \mathrm{Var}(\eta_t) = 1$.

Assuming that σ_t has finite dimensionality, we can write $\sigma_t^2 = \nu(e_1^2, \ldots, e_{t-p}^2)$ for some positive function ν and integer p. Then, define $y_t = e_t^2$, and rewrite (5.13) as

$$y_t = \sigma_t^2 + \sigma_t^2(\eta_t^2 - 1) = \nu(y_{t-1}, \ldots, y_{t-p}) + \nu(y_{t-1}, \ldots, y_{t-p})(\eta_t^2 - 1), \tag{5.14}$$

which is seen to have the form of NAR model (5.9). A special model is the popular ARCH model which refers to the parametric case when m is constant, and ν is a *linear* function

$$\nu(y_1, \ldots, y_p) = a + b_1 y_1 + \ldots + b_p y_p. \tag{5.15}$$

(e.g. Bollerslev et al 1994, Gourieroux 1997).

In summary, the martingale regression model (5.8) is a very general model. We should mention that modeling of noisy chaos can be studied in the framework (5.8) (Lu 1999a).

Two technical assumptions. For the asymptotic statistical results to be stated later, the following assumptions are used as in Lu (1999b). Noticeable is the strict stationarity assumption of $\{X_i\}$ in that any finite distribution of $(X_{t+i_1}, \ldots, X_{t+i_k})$ is independent of time t for any integers $i_1, \ldots, i_k$ and $k = 1, 2, \ldots$. In particular, we assume the following *strong mixing* condition for a stationary sequence:

(**B**) The vector sequence $\{X_i\}$ is strictly stationary and satisfies the short-range dependence condition: let $f_j(\cdot, \cdot)$ denote the joint density of X_1, X_{j+1} and $f(\cdot)$ denote the marginal density, then

$$\sup_{\mathbf{u}, \mathbf{v} \in \mathbb{R}^p} \sum_{j=1}^{\infty} |f_j(\mathbf{u}, \mathbf{v}) - f(\mathbf{u}) f(\mathbf{v})| < \infty. \tag{5.16}$$

Note that (5.16) can be extended to situations of chaotic time series when the design density does not exist (Lu 1999a). In addition, a minor technicality condition on ε_t is also used:

(**C**) $\sup_{i \geq 1} \mathrm{E}\{|\varepsilon_i|^{2+\delta} | \mathcal{F}_{i-1}\} < \infty$ for some $\delta > 0$.

Asymptotic bias and variance. Assumptions (A), (B), (C) together with some condition on m (such as twice or third-degree differentiable) and the bandwidth $h = h_n$ (such as $h_n \to 0$ and $nh_n^p \to \infty$) guarantee consistency, and optimal convergence rate of local polynomial estimators (Lu 1996a, 1999b). Furthermore, asymptotic normality can also be proved, and the following proxy for asymptotic bias and variance can be used for most purposes:

$$\begin{array}{ll} \text{Bias:} & (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W (M - \mathbf{X}\beta), \\ \text{Variance:} & \nu(\mathbf{x}) (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W^2 \mathbf{X} (\mathbf{X}^T W \mathbf{X})^{-1}, \end{array}$$

where $M = (m(X_1), \ldots, m(X_n))^T$. By using Taylor expansion on m near $\mathbf{x}$, the bias expression can be shown to depend on the higher-order derivatives of m at $\mathbf{x}$ only.

By plug-in any consistent estimate of ν and higher-order derivatives of m, one can obtain estimates for the asymptotic bias and asymptotic variance. This in turn gives rise to an approach for computing pointwise confidence interval and prediction intervals for the conditional mean predictor. For chaotic and nonlinear deterministic modelings, Lu (1999a) has established a convergence rate in terms of the fractal dimension for the well-known Nadaraya-Watson kernel estimator (which corresponds to local constant fit), thus confirming a conjecture of Farmer and Sidorowich (1988). I also show that the choice of embedding dimension is not very crucial as long as it is large enough. It is demonstrated there that the local polynomial method is in general a very effective prediction procedure for systems whose true fractal dimension is not too high.

4. Prediction methods

In this section, we concentrate on nonlinear prediction theory of time series, with emphasis on nonparametric techniques. The nonlinear regression technique in Section 2 can be applied directly to provide one-step and multi-step point predictors, namely for developing a point predictor through estimating the conditional mean, which is the best mean square predictor. The one-step predictor $\hat{m}_t$ of y_{t+1} based on $y_t, \ldots, y_1$ is the same as the regression estimator based on the embedded data

$$\{[y_i, (y_{i-1}, \ldots, y_{i-p})], p+1 \le i \le t\},$$

such as the case in **Model I** (5.9).

Let $\mathbf{X}_t, W_t, Y_t$ denote the corresponding design matrix, weight matrix, and data vector at time t, and $e_1 = (1, 0, \ldots, 0)$, and let $\mathbf{b}_t$ denote the estimation vector using data up to time t only. Then

$$\mathbf{b}_t = (\mathbf{X}_t^T W_t \mathbf{X}_t)^{-1} \mathbf{X}_t W_t Y_t = P_t Y_t. \tag{5.17}$$

where $P_t = (\mathbf{X}_t^T W_t \mathbf{X}_t)^{-1} \mathbf{X}_t W_t$ is the coefficient matrix at time t.

On-line prediction algorithm. There is a nice recursive formula for updating $\mathbf{b}_t$ when a new data point, say y_{t+1} is available. Let $\mathbf{r}_{t+1}^T$ denote the *extra* row vector of design matrix at time $t+1$, and k_{t+1} denote the weight corresponding to the time $t+1$ vector (that is, $k_{t+1} = K(X_{t+1}, h)$). Then,

$$\begin{aligned} \mathbf{b}_{t+1} &= \mathbf{b}_t - k_{t+1} g_{t+1} (\mathbf{X}_t^T W_t \mathbf{X}_t)^{-1} \mathbf{r}_{t+1} \mathbf{r}_{t+1}^T \mathbf{b}_t \\ &\quad - k_{t+1} g_{t+1} c_{t+1} y_{t+1} (\mathbf{X}_t^T W_t \mathbf{X}_t)^{-1} \mathbf{r}_{t+1}, \end{aligned} \tag{5.18}$$

where $c_{t+1} = k_{t+1} \mathbf{r}_{t+1} (\mathbf{X}_t^T W_t \mathbf{X}_t)^{-1} \mathbf{r}_{t+1}$, and $g_{t+1} = (1 + c_{t+1})^{-1}$. Obviously, if there are multiple data points available, (5.18) can be applied multiple times to obtain the desired estimates. To start our sequential algorithm, one can either apply the batch algorithm for some historical data, or some prior guess of b and its covariance matrix. Note that there is close connection

between (5.18) and the extended Kalman filtering algorithm. The on-line algorithm should be useful for predicting nonstationary systems and for data filtering.

Mean squared prediction error: Note that the important difference from the regression problem is that the mean squared prediction error is the sum of mean squared estimation error plus the intrinsic noise variance, of which the latter cannot be removed. Asymptotic mean squared prediction error

$$\mathrm{E}(\hat{m}_t - y_{t+1})^2 \approx \mathrm{E}(\hat{m}_t - m_t)^2 + \nu_t,$$

which consists of the decomposition of

$$\text{asymptotic bias}^2 + \text{ asymptotic variance} + \text{conditional variance},$$

where the asymptotic bias and asymptotic variance refer to those of $\hat{m}_t$ and $\nu_t = \nu(y_t, \ldots, y_{t-p+1})$. Any consistent estimates can used to substitute any unknown quantities such as ν in above formula. This approach applies to estimation of prediction error for any nonparametric prediction procedure.

Multi-step prediction: We discuss some procedures for constructing multi-step predictor. Suppose we want to predict $y_{t+\tau}$ based on data up to time t, and denote an τ-step predictor by $\hat{m}_{\tau|t}$. Since $m_{\tau|t} = \mathrm{E}[y_{t+\tau}|y_t, \ldots, y_{t+p-1}]$, one obvious predictor is the *direct* predictor, which is the regression estimate based on data

$$\{[y_{i+\tau}, X_i = (y_i, \ldots, y_{i-p+1})], p \le i \le t - \tau\}. \tag{5.19}$$

Note the data vector $Y_{\tau|t} = (y_{p+\tau}, \ldots, y_t)^T$, then we can compute local regression coefficient vector $\mathbf{b}_{\tau|t}$ by

$$\mathbf{b}_{\tau|t} = P_t Y_{\tau|t}. \tag{5.20}$$

Except for sharing the same coefficient matrix P_t (more or less, by ignoring a few earliest data points), the direct predictor does not use the one-step predictor explicitly. Now we discuss some multi-step procedures which exploit the *dynamical* structure of the underlying process. Consider the NAR process (5.9) $y_{t+1} = m(y_t, \ldots, y_{t-p+1}) + \nu_{t+1}\varepsilon_{t+1}$ and $\tau = 2$ for easy account. Note the following recursive relation:

$$\begin{aligned} m_{2|t} &= \mathrm{E}[m(y_{t+1}, y_t, \ldots, y_{t-p+2})|y_t, \ldots, y_2, y_1] \\ &= \mathrm{E}[m(m(X_t) + \nu(X_t)\varepsilon_{t+1}, y_t, \ldots, y_{t-p+2})], \end{aligned} \tag{5.21}$$

where the last expectation operator is over ε_{t+1} while $y_t, \ldots, y_{t-p+1}$ are kept fixed. Thus, if the distribution of ε_{t+1} is *known*, one can compute $m_{2|t}$ based on estimates of m and ν, using Monte Carlo method such as

$$\hat{m}_{2|t} = \frac{1}{R}\sum_{\ell=1}^{R} \hat{m}(\{\hat{m}(X_t) + \hat{\nu}_{t+1} \cdot e_{t+1}(\ell)\}, y_t, \ldots, y_{t-p+1}),$$

where $e_{t+1}(1), \ldots, e_{t+1}(R)$ are random samples from the distribution of ε_{t+1}, and $\hat{m}, \hat{\nu}_{t+1}$ are estimates of m, ν at Y_t.

When ε_{t+1} is small, or more precisely the probability of ε_{t+1} is concentrated on a small neighborhood near the origin, (5.21) is approximated by

$$m_{2|t} \approx m(m(Y_t), y_t, \ldots, y_{t-p+1}),$$

or $M_{2|t} = M^2$ in vector notation, where $Y_t = (y_t, \ldots, y_{t-p+1})^T$ and $M_{2|t}$ is the two-step conditional mean in the dynamical model and $M(Y_t) = (m(Y_t), y_t, \ldots, y_{t-p+2})^T$. The resulting two-step predictor

$$\hat{m}_{2|t} = \hat{m}(\hat{m}(Y_t), y_t, \ldots, y_{t-p+2}),$$

or $\hat{M}_{2|t} = \hat{M}^2$. Longer period multi-step prediction can be derived in similar fashion. Lu (1996b) discusses the advantages of iterative predictor, and demonstrates that the invariant measure on a chaotic attractor is better estimated using this approach.

Conditional distribution function: non-mean prediction. Economic and financial data often exhibit nonnormal, asymmetric, and heavy tail behaviors. Thus, it is likely that the conditional mean is not adequate as a lone predictor. Other likely choices such as conditional median, regression quantile may be more informative (e.g. Chaudhuri 1991). In particular, the extremes such as low and high predictive values, or the spread (high-low) will be very interesting. More formally we consider estimation of conditional distribution function (cdf) $F_{\tau|t}(y) = P(Y_{t+\tau} \leq y|$ data at time $t)$. A simple estimator is the Nadaraya-Watson estimator

$$\hat{F}_{\tau|t}(y) = \frac{\sum_{i=p+\tau}^{t-\tau} k_i 1_{\{y_i \leq y\}}}{\sum_{i=p+\tau}^{t-\tau} k_i} = \sum_{i=p+\tau}^{t-\tau} w_i 1_{\{y_i \leq y\}}, \tag{5.22}$$

where (X_i, y_i) as defined in (5.19), $k_i = K(X_{i-\tau}, h)$ is the weight associated with the ith data point and $w_i = k_i / \sum k_i$ is the probability mass assigned to y_i. Recall that K is the kernel function and may be chosen to have finite support, then w_i is just the local probability weight for prediction at time t. Predictive characteristics such as mean, spread, variance, and quantile can be computed based on the weighted empirical distribution function. Ensemble forecasting is another option, presenting simulated samples from the predictive distribution (5.22) as future scenarios. An important application is to interval prediction for a future value $y_{t+\tau}$. For any two numbers $a < b$, we have that the predictive probability that $y_{t+\tau}$ falls within $[a, b]$ conditional on information $\mathcal{F}_t$ available up to time t is estimated by

$$\hat{\Pr}(a \leq y_{t+\tau} \leq b|\mathcal{F}_t) = \hat{F}_{\tau|t}(b) - \hat{F}_{\tau|t}(a). \tag{5.23}$$

So the $100 \times (1-\alpha)\%$ prediction interval estimated by $[a, b]$ where a, b are chosen to be the $100 \times (\alpha/2)\%$th and $100 \times (1-\alpha/2)\%$th quantiles of $\hat{F}_{\tau|t}$.

Implementation details. Implementation of local polynomial prediction for a given time series $\{y_i\}$ consists of the following steps.

1 Define the embedding vector

$$X_i = (y_i, y_{i-d}, \ldots, y_{i-(p-1)d})^T, i = (p-1)d+1, \ldots, t,$$

the choice of embedding dimension p and time delay d.

2 Decide the norm to use in computing the interpoint distances d_{ij} in the embedded state space. Examples include
Euclidean $d_{ij} = \|X_i - X_j\|_2 = \sqrt{\sum_{\ell=1}^{p}(X_{i\ell} - X_{j\ell})^2}$,
L^1 norm: $d_{ij} = \|X_i - X_j\|_1 = \sum_{\ell=1}^{p} |X_{i\ell} - X_{j\ell}|$,
maxmimum absolute distance: $d_{ij} = \|X_i - X_j\|_0 = \max_{\ell=1}^{p} |X_{i\ell} - X_{j\ell}|$.

3 Given a choice of bandwidth h or nearest neighbor ℓ at a given state vector of interest, say last X_t, if the kernel function is chosen to be radial function of finite support, local computation can be excised in the following way: one decides the index of embedded state vectors such that d_{it} is less than or equal to h or the order of d_{it} is less than or equal to ℓ.

4 For $\tau-$step ahead prediction based on data up to time t, find the response vector $Y = (y_{(p-1)d+\tau}, \ldots, y_t)$.

Apply standard (weighted) least square procedure to local data consisting of submatrix of X, Y and weights $k(d_{it}/h)$ and obtain the local least square estimate. The fitted regression hyperplane evaluated at X_t is the predicted value of $y_{t+\tau}$.

5 Repeat Step 3-4 if prediction at other time points is needed, or repeat Step 4 only if prediction at other lead times are needed.

Choice of embedding parameters p and d is certainly important. More crucially, however, is the choice of bandwidth h or number of nearest neighbors ℓ. See Lu (1999a) for some interesting examples in the context of chaotic time series prediction.

5. Volatility estimation

Time-varying volatility function is a popular way of incorporating nonlinearity in financial time series models. The volatility is obviously predictable, considering that there may be more volatility in an equity price around the time of earning announcements, or releases of macroeconomics news. There are also (surprisingly) a lot of co-movement and common factors among different equity prices in the same sector.

Many models have been proposed to model volatility, e.g. Gouriermoux (1997). Since there is no basis for committing to a specific parametric form,

there is a lot of interest recently in the nonparametric approach. The goal of this section is to present a nonparametric approach to estimating volatility or conditional variance function using the local polynomial method.

We start with **Model III** (5.14). Note that this is a regression model with special structure, namely the regression (mean) and standard deviation are *proportional*, resembling the Gamma distribution. Alternatively, we may say that the normalized quantity

$$\xi_t = \frac{y_t - \nu_t}{\nu_t} = \frac{e_t^2}{\sigma_t^2} - 1 \tag{5.24}$$

is a random variable with mean zero and constant variance (note $y_t = e_t^2$, the squared residuals). Thus, we may derive our estimators using an estimation equation type approach. The heuristic rational is that e_t^2 is a sufficient statistic for σ_t^2, and the *relative* scale e_t^2/σ_t^2 matters most. In practice, e_t can be effectively replaced by some *estimated* residual process as long as the unknown mean function is a consistently estimated (Fan and Yao 1998).

Subsequently, we work with quantity

$$\frac{y_t}{\sigma_t} - 1 = \varepsilon_t \tag{5.25}$$

where $y_t = |e_t|, \varepsilon_t = |\eta_t| - 1$, and propose to minimize

$$\sum_{t=1}^{n} k_t \frac{(y_t - \sigma_t)^2}{\sigma_t^2}, \tag{5.26}$$

under the constraint that $\sigma_t \geq 0$, where ν_t is substituted by its modeling form and k_t is the weighting function.

For example, for ARCH models of order p,

$$\nu_t = a + b_1 y_{t-1} + \ldots + b_p y_{t-p}$$

and the constraints are:

$$0 \leq a, b_1, \ldots, b_p; b_1 + \ldots + b_p < 1.$$

Note that this weighted least square (wls) method emphasizes the ratio y_t/σ_t, a more natural quantity for modeling scales, rather than the absolute differences in standard least square methods such as in $\sum_{t=1}^{n}(y_t - \sigma_t)^2$ or $\sum_{t=1}^{n}(y_t^2 - \sigma_t^2)^2$.

Now we define a *local* ARCH methodology as follows. At each point of interest, say $\mathbf{u} = (u_1, \ldots, u_p) \geq 0$, we estimate $\nu(u_1, \ldots, u_p)$ by minimizing (5.26) with

$$\sigma_t = a + b_1(y_{t-1} - u_1) + \ldots + b_p(y_{t-p} - u_p),$$

and $k_t = K(Y_t, h) = k(\|Y_t - \mathbf{u}\|/h)$ is the kernel weighting function. over $0 \leq a, \mathbf{b} = (b_1, \ldots, b_p)^T$. Note that we do not require the boundedness condition of b_i's with the local fitting approach.

Compared to standard least square or likelihood method as used e.g. in Martin et al (1996), the weighted least square method as given in (5.26) emphasizes and favors *large* volatility, while at the expense of yielding upward biased estimate. In general, the weighted least square (WLS) method gives more realistic variance estimate, in terms of comparable scale of fluctuation and variability. The WLS approach is similar to the QGLS method of Gouriermoux (1997).

Besides using ratio-type criterion (5.26) for model comparison, we can also use the ratio statistics

$$\hat{\xi}_t = \frac{\hat{y}_t}{\hat{\sigma}_t}, \tag{5.27}$$

which should center around 1 in ideal situations, as a model diagnostic tool.

Implementation details. The implementation of predicting and estimating the volatility function of a time series consist of the following steps.

1 Compute the residual of time series. This step could be as simple as differencing, and as complicated as fitting a nonparametric regression to the conditional mean.

2 Deciding on the state space and embedding dimension of the squared residuals.

3 Given bandwidth h and kernel K, compute the volatility estimate $\hat{a}$ using the nonnegative local linear fit.

4 Repeat step 3 at other data points until estimates at all desirable time points are computed.

5 Plot volatility estimates along with data, judging goodness of fit, and possibly modify the choice of embedding and bandwidth, and repeat Step 4.

6. Risk analysis of AOL stock

In this section we apply some of our methods developed earlier to the study of daily AOL stock price. We use only the data from Jan. 3, 1995 to Sept. 3, 1999 on stock prices (open, high, low, close) and total daily volumes. We focus on two applications, first the use of stochastic regression in modeling the relationship of spread and volume, and as a byproduct we also explore issues of stationarity by subsampling technique. Second application is for modeling volatility of the return rate based on closing prices. We apply the local ARCH method and the weighted least squares criterion developed in this chapter, and we also compare them with standard results using ARCH and GARCH models.

Spread-volume relationship

Spread in the daily price is an important volatility factor, and poses both arbitrage opportunity and potential risk for shareholders. It is well-known that volume is directly related to price fluctuation. Higher trading volume is clear reflection of increased public interest and attention in a given stock. In particular, the price swing and spread will be greater with higher trading volume. We empirically explore the relationship of spread and volume in the AOL stock based on the given data.

Figure 5.1 shows the scatter plot of daily spread versus trading volume on the log scale, where different symbols represent data from different periods. The (nonlinear) smoothing line (solid) is the lowess fit based on the overall data. Apparently, one might conclude that there is a nonlinear relationship based on the overall line fit. Fortunately, after talking with an economist, I tried separate fits with data from different periods, chosen in rather ad hoc way: Period 1 (day 1 to 200): points ($\cdot$), smoothing line (short dash); Period 2 (day 201 to 600): points ($+$), smoothing line (mediate dash); Period 3(day 601 to 800): points($-$), smoothing line (dots); Period 4 (day 801 to 1195): points ($\diamond$), smoothing line (long dash). One can see that there is a clear linear relationship when looked properly within each subperiod. ¿From this analysis, we may infer that there is probably nonstationarity or regime switching in this data set, and within each regime, there is a *power-law* relationship between spread and volume. A lesson for data analysis: checking for stationarity is a very important step, and always be aware of regime switching and time varying events with large data sets.

Volatility in return rate

In this subsection, we consider nonlinear modeling of the conditional variance function in the return rate series based on closing price, defined as $y_t = (p_t - p_{t-1})/p_{t-1}$. Figure 5.2 shows basic plots of this series, which include time plot (5.2 a), marginal histogram plot (the overlaid line is a smooth density estimate) (5.2 b), scatter plot of squared series (5.2 d).

First consider the traditional GARCH(1,1) model as defined in

$$\begin{aligned} y_t &= \mu + \sigma_t \varepsilon_t, \text{ where} \\ \sigma_t^2 &= a + a_1 \varepsilon_{t-1}^2 + b_1 \sigma_{t-1}^2, \end{aligned}$$

where μ, a are constant, a_1 is the ARCH coefficient, and b_1 is the GARCH coefficient, and $0 \le a, a_1, b_1, a_1 + b_1 < 1$. SPLUS GARCH module in Martin et al (1996) is used here for fitting such models.

Garch (1,1) fit to whole series gives

$$\mu = 0.0049863, a = 0.0001255, a_1 = 0.0907047, b_1 = 0.8397169.$$

The fitted conditional standard deviation σ_t is plotted in Figure 5.2 c. Next, we check for regime switching and nonstationarity. In order to fit a time-

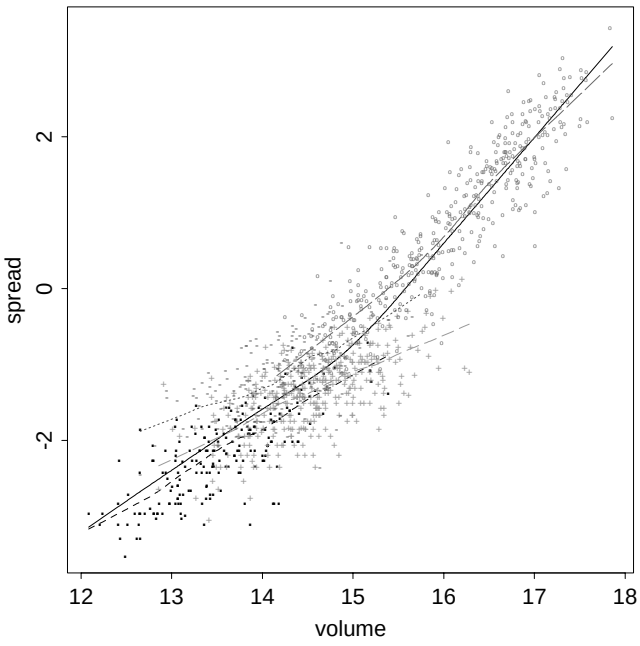

Figure 5.1. Power-law relation in spread-volume of AOL stock.
Scatter plot of daily price spread and trading volume of AOL data plotted on the log scale. The (nonlinear) smoothing line (solid) is the lowess fit based on the overall data. Probably due to nonstationarity or regime switching, one should put more emphasis on the information contained in subsample data sets in shorter time period: Period 1 (day 1 to 200): points ($\cdot$), smoothing line (short dash); Period 2 (day 201 to 600): points ($+$), smoothing line (mediate dash); Period 3(day 601 to 800): points($-$), smoothing line (dots); Period 4 (day 801 to 1195): points ($\diamond$), smoothing line (long dash). One can see that there is a clear linear relationship or power-law behavior on original scale.

varying GARCH model, we choose bandwidth=200 days, and consider moving average fit at a given time point to subsamples of within 200 days. That is, starting from the 201th day, at every 5th day i, we fit GARCH to data points in day $[-200+i, i+200]$. The time series plots of fitted coefficients are shown in Figure 5.3. It is seen that there is significant variation in the coefficients, indicating regime switching or time varying behavior in this process.

In light of previous analysis, we consider a subsample only, say data from day 801 to 1195, the end of data series, and so confounding effect of nonstationarity is avoided. Figure 5.4 shows the results of different fittings

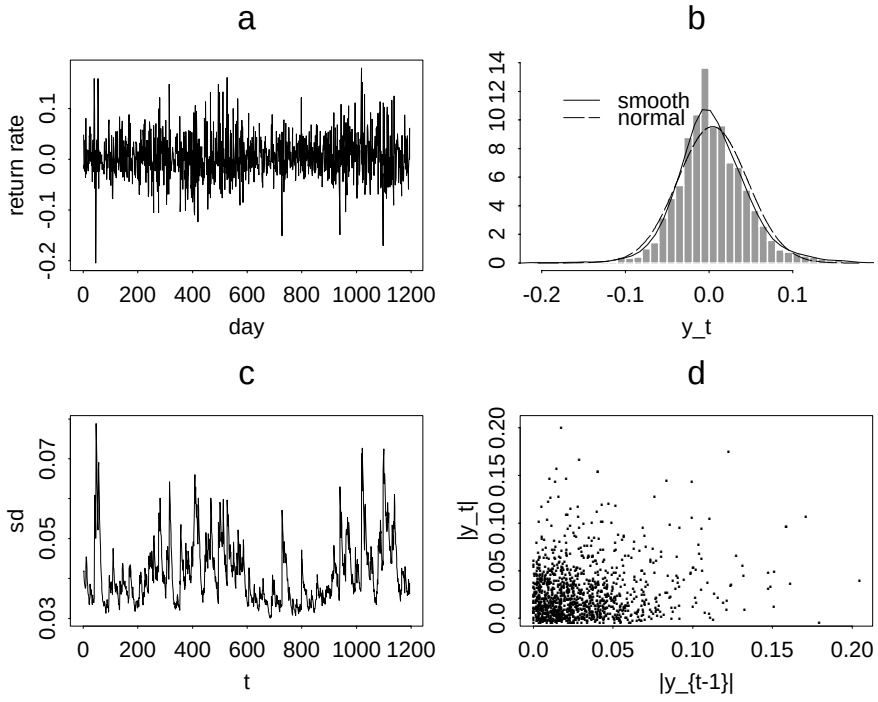

Figure 5.2. AOL closing price return rate series.
a. time plot; b. marginal histogram; c. fitted conditional standard deviation using standard GARCH(1,1) model; d. scatter plot of squared series.

Table 5.1. Comparison of GARCH and local ARCH models.

Models		*Comparison criterion* (5.26)
GARCH(1,1)		167.75
Local ARCH(1):	weighted fit	149.99
	unweighted loess fit	306.66

using GARCH(1,1) (Figure 5.4c,5.4d), and weighted or unweighted (loess) fits of local ARCH(1) (Figure 5.4a,5.4b), along with residual plots using ratio statistics (5.27). For the GARCH(1,1) fit, the value of criterion (5.26) is: 167.75. The local ARCH fit gives final value of (5.26) as 149.99. As a comparison, the loess fit using standard method gives 306.66. See also Table 5.1. Two points can be made from this analysis: the first is that the simple one-step local ARCH can do as well as or better than the GARCH model, which is known to depend heavily on the AR part; secondly, there is a great need in the use of weighted criterion such as (5.26) in order for the local ARCH

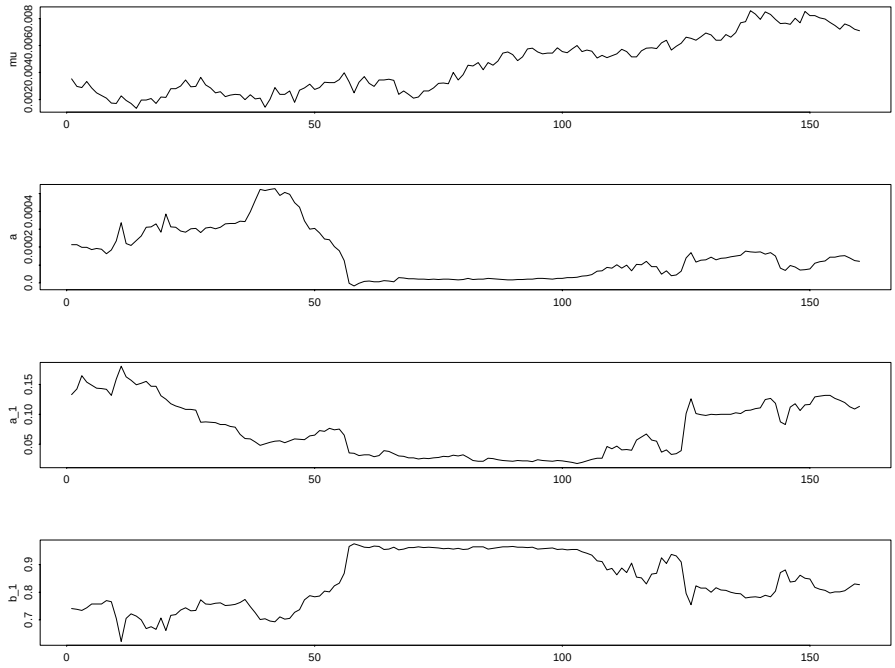

Figure 5.3. Moving GARCH fits of AOL return series.
Moving window GARCH (1,1) fit at every 5th day starting from day 201 to day 995, using data within 200 days. The four fitted coefficient series for μ, a, a_1, b_1 are plotted from top to bottom, respectively.

methodology to outperform a standard method. In summary, we think that the local ARCH model using the weighted criterion gives more comparable scale in large volatility values than the GARCH fit, and the ratio plot reflects both this fact as well as the weighted least square method's upward biasedness tendency. The conservative nature of our weighted least square approach will be welcomed in applying to prediction error bar estimation as well as in risk analysis and optimal portfolio design.

7. Concluding remarks

In this Chapter, we have surveyed recent developments in multivariate local polynomial fitting for time series, with emphasis on useful methods, prediction, and computational issues. We discuss predictors for multi-step prediction (which include short-range and mediate-range time horizon), quantile and non-mean predictors, and ensemble prediction. Variance estimation is important in economic and financial planning as well as in prediction (such

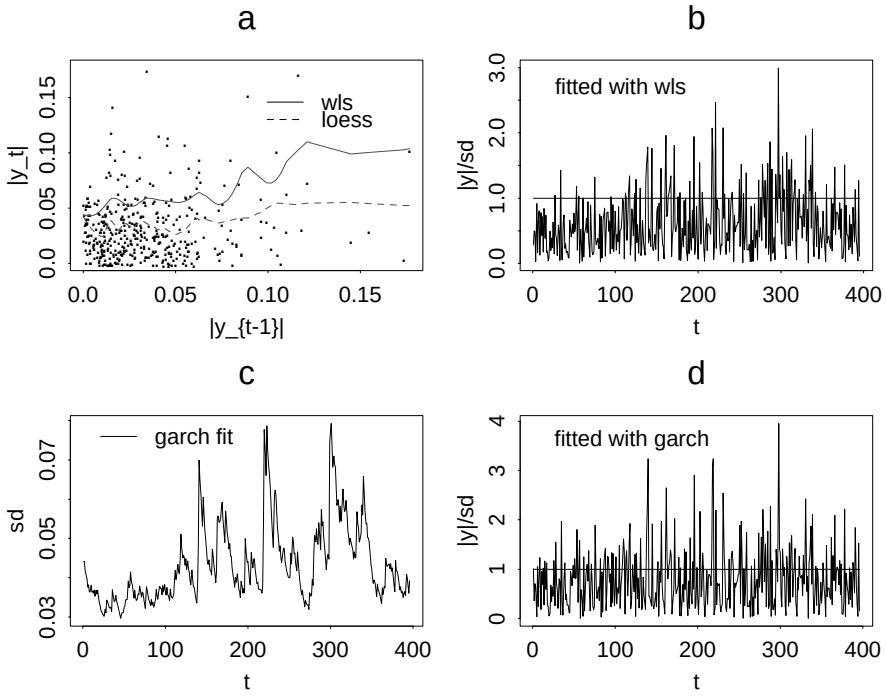

Figure 5.4. Comparison of local ARCH, GARCH, and loess fits.
a. local ARCH fit; b. ratio plot: residual/σ_t fitted with local ARCH; c. time series plot of GARCH(1,1) fit; d. ratio plot for GARCH fit.

as uncertainty estimate). We introduce a nonlinear volatility function, local ARCH, which compares favorably with standard approaches. The AOL stock data are used as a testbed for some of our methods, and we emphasize checking for stationarity, regime switching, and subsampling technique in the analysis of large time series data sets.

To end with a lighter note, we note the recent appearance of the book *The Predictors*, by Thomas Bass (1999), detailing chaotists Doyne Farmer and colleagues, in their new adventures in stock prediction and the study of financial markets. Now chaos, complexity, and predictors have come to the main street, the Wall Street. With easy access to ultra and high-frequency financial data or real-time ticker data, nonlinear techniques will find ample opportunities and exciting new challenges.

References

Bass, T.A. 1999. *The Predictors.* Henry Holt & Company.

Bollerslev,T.,Engle,R.F. and Nelson,D.B.(1994). ARCH models. Chapter 49 of *Handbook of Econometrics* **IV**, eds. Engle, R.F.. and McFadden, D.L., Elsevier Sciences.

Cleveland, W.S., Grosse,E. and Shyu, W.M.(1992). Local regression models. Chapter 8 of *Statistical Models in S*, eds. Chambers,J.M. and Hastie,T.J., CRC/Chapman and Hall, New York.

Chaudhuri, P.(1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *Annals of Statistics*, **19**, 760-777.

Cutler, C. and D. Kaplan (eds) (1997). *Nonlinear Dynamics and Time Series*. American Mathematical Society, Rode Island.

Dhrymes, P.(1997). *Time Series, Unit Roots, and Cointegration*. Academic Press.

Fan, J. and Gijbels, I. (1995). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.

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

Farmer, J.D. and Sidorowich, J.J.(1988). Exploiting chaos to predict the future and reduce noise. In *Evolution, Learning, and Cognition* (ed. Y. C. Lee), pp. 277-330. World Scientific, Singapore

Franses, P.H. and D.V. Dijk (2000). *Nonlinear Time Series Models in Empirical Finance*. Cambridge University Press, Cambridge.

Gouriermoux, C.(1997). *ARCH Models and Financial Applications*. Springer, New York.

Hamilton,J.D.(1994). *Time Series Analysis*. Princeton University Press, Princeton.

Krolzig, H.-M.(1997). *Markov Switching Vector Autoregressions, Modeling, Statistical Inference, and Applications to Business Cycle Analysis*. Lecture Notes in Economics and Mathematical Systems, Vol.454. Springer, Berlin.

Lu, Z. Q. (1996a). Multivariate locally weighted polynomial fitting and partial derivative estimation. *J. of Multivariate Anal.*, **59**, 187-205.

Lu, Z. Q. (1996b). Predicting chaotic time series. In *Business and Economic Statistics: Proceedings of Joint Statistical Meetings in Chicago, 1996*, pp.182-187. American Statistical Association, Alexandria, Virginia, USA.

Lu, Z.Q.(1999a). Nonparametric regression with singular design. *Journal of Multivariate Analysis*, **70**, 177-201.

Lu, Z.Q. (1999b). Multivariate local polynomial fitting for nonlinear martingale models. *Annals of the Institute of Statistical Mathematics*, **51**, No.4, 691-706.

Z. Q. Lu, L. M. Berliner (1999). Markov switching time series models with application to a daily runoff series. *Water Resources Research*, Vol. 35, No. 2, 523-534.

Martin, R.D., Gao, H.Y., Zhan,Y. and Ding, Z.(1996). *S+GARCH: User's Manual.* Data Analysis Products Division, MathSoft, Seattle.

Priestley, M. B. (1988). *Non-linear and Non-stationary Time Series Analysis.* London, Academic Press.

Seber, G.A.F.(1977).*Linear Regression Analysis.* John Wiley, New York.

Tong, H. (1990). *Nonlinear Time Series: a Dynamical System Approach.* Oxford, Oxford University Press.