Local Maximum Likelihood Estimation of Volatility Functions

Panagiotis Avramidis

Department of Statistics, London School of Economics,

Houghton Street, WC2A 2AE

March, 15, 2002

Abstract

Estimation of the variance function has become increasingly popular mainly due to applications in financial risk management. In non-parametric context, local polynomial fitting using kernel function has been widely used. In this work, we introduce local-Linear Maximum Likelihood Estimation (LMLE) which makes use of the error density function. We derive the bias and asymptotic variance and establish asymptotic normality. Comparison with existing methods shows that when the error distribution departures from Gaussian, LMLE has a smaller asymptotic mean square error. Further, conditions for adaptiveness of the mean function are identified. Two simulation examples are presented.

KEY WORDS: Non-parametric, likelihood function, kernel, local polynomial fitting

[0]I would like to thank my supervisor Dr Q. Yao for his helpful comments and guidanse. The author was supported by an ESRC grant.

1 Introduction

The term volatility refers to the characteristic of the non-constant variation, also called conditional heteroskedasticity, observed in many financial data sets. In non-parametric theory various methods of estimation of the volatility function have appeared. Härdle and Tsybakov (1997) propose an estimator based on the simple decomposition $\sigma^2(x) = E(Y^2|X = x) - (E(Y|X = x))^2$ while Fan and Yao (1996) for time series and Ruppert *et al* (1997) for iid, study a residual-based estimator in conjunction with local kernel smoothing.

If the distribution of the error term is assumed to be known, an alternative approach is local maximum likelihood estimation. The idea is not new, Simonoff (1996) uses local likelihood in the density estimation context while Staniswalis (1989) derives the asymptotic properties of a kernel-based estimate of a regression function based on likelihood for the i.i.d case. Further, Linton and Xiao (2001) propose an improved local likelihood based regression estimator that adapts to the error distribution. Bandwidth selection and confidence intervals are discussed by Fan, Farmen and Gijbels (1998). However, little appears in the literature on the use of local-likelihood estimation for the volatility function.

In this work, we adopt a local linear approximation of the log-variance function and use the local likelihood function to estimate the parameters. A direct comparison of local Linear Maximum Likelihood Estimator (LMLE) and local Linear Least Square Estimator (LLSE) is based on the comparison of the asymptotic mean square error (AMSE). In fact, we argue that, although in the Gaussian case efficiency is the same for the two estimates, in a non-Gaussian case, the difference of the two AMSE tends to be significant. Using additional information provided from the error distribution, LMLE is more efficient when the errors are non Gaussian. In practice, the error dis-

tribution is unknown. Nevertheless, this does not reduce the value of the results and they can be used as a benchmark to prove adaptiveness of the error distribution.

Another implication is the introduction of the log transformation in the local polynomial fitting. The variance function should be positive and log transformation ensures this property without any restrictions. However, the results suggest that squared bias may increase, equivalently AMSE will increase, if the second derivative of the variance function is negative. Therefore, it is likely that some gain in asymptotic variance may be overshadowed by an increase of the squared bias. However, this is solely due to the introduction of the transformation and not because of the likelihood function.

The paper is organized as follows. In section 2 we introduce the model and discuss the linear fitting of the log standard deviation. In section 3 and 4 we study the asymptotic properties of the local maximum likelihood estimator for variance when the mean function is assumed to be known and unknown. Moreover, in section 4 we derive a sufficient condition under which local maximum likelihood estimator is asymptotically adaptive to unknown mean function $m(.)$. In this sense, we can estimate the variance function as well as if $m(.)$ was known. Further, the gain in efficiency using LMLE is illustrated with examples in section 5. Regularity conditions and proofs are included in the Appendix.

2 The model and the local likelihood function

The model that we consider is a non-parametric autoregressive conditional heteroscedastic time series model. Let $\{Y_t\}$ be a strictly stationary process and let $m(y) = E(Y_t|Y_{t-1} = y)$ and $\sigma^2(y) = \text{Var}(Y_t|Y_{t-1} = y) \neq 0$ the conditional mean and variance which are known as the mean and the variance

function. Let

$$Y_t = m(Y_{t-1}) + \sigma(Y_{t-1})\varepsilon_t \tag{1}$$

then $E(\varepsilon_t|Y_{t-1}) = 0$ and $\mathrm{Var}(\varepsilon_t|Y_{t-1}) = 1$. Assume ε_t are i.i.d. with the common known density function $f(.)$. We will estimate the variance function both where the mean function is known and for the unknown mean case. The conditional distribution can be expressed as

$$l_n = \sum_{t=1}^{n} \log f\left(\frac{Y_t - m(Y_{t-1})}{\sigma(Y_{t-1})}\right) - \sum_{t=1}^{n} \log \sigma(Y_{t-1}) \tag{2}$$

For the local polynomial fitting a minimum degree of smoothness is required in order to apply Taylor expansion, thus suppose that $\sigma(y)$ has at least a continuous third derivative. Without any loss of generality, we look at the log of standard deviation function, $s(y) = \log \sigma(y)$. The first order Taylor expansion in a small neighborhood around y gives

$$s(y) = \alpha + \beta(Y_t - y)$$

from which we get,

$$\sigma(Y_t) = e^{\alpha+\beta(Y_t-y)} \tag{3}$$

where $\alpha = s(y)$ and $\beta = s'(y)$.

Using equations (2) and (3) the conditional local log-likelihood function is given by

$$l_n(\alpha, \beta) = \sum_{t=1}^{n} \Big(\log f\Big(\frac{Y_t - m(Y_{t-1})}{e^{\alpha+\beta(Y_{t-1}-y)}}\Big) - \Big(\alpha+\beta(Y_{t-1}-y)\Big)\Big) K_h(Y_{t-1}-y) \tag{4}$$

where $m(.)$ is the mean function, $K_h(u)$ is the kernel function that assigns weights to the data points Y_t according to their distance from y and h is the bandwidth which defines the size of the neighborhood. We usually write $K_h(u) = K(\frac{u}{h})\frac{1}{h}$ where $K(.)$ is a density function. Having defined the likelihood function, the local maximum likelihood estimate, $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$, is the solution of the likelihood equation $\nabla l_n(\theta) = 0$.

3 Asymptotic Properties when mean function is known

First concentrate on the case of known mean function. Knowing $m(.)$ we can assume that $E(Y_t|Y_{t-1}) = 0 \Rightarrow m(Y_{t-1}) = 0$ without loss of generality. We introduce the notation. Define,

$$\tilde{g}(t) = \frac{f'(t)}{f(t)}t + 1, \qquad \tilde{s}(t) = \frac{t^2 f''(t)f(t) + tf'(t)f(t) - t^2 f'(t)^2}{(f(t))^2}$$

We use $p(.)$ to denote the marginal density function of Y_t. Let $S^* = [\nu_{j+l}]_{0\leq j,l\leq 2}$ with $\nu_j = \int u^j K^2(u)du$, $S = [\mu_{j+l}]_{0\leq j,l\leq 2}$ with $\mu_j = \int u^j K(u)du$ and $X_t = (1, Y_{t-1} - y^*)^T$. We can now state the first theorem in which we present the bias and asymptotic variance of the estimator.

Theorem 1 *Suppose that conditions C1-C4 in Appendix hold. Then for the local maximum likelihood estimator $\hat{\theta} = (\hat{\alpha}, \hat{\beta})^T$ it holds that*

$$\begin{pmatrix} \sqrt{nh}(\hat{\alpha} - \alpha - h^2 b_1) \\ \sqrt{nh^3}(\hat{\beta} - \beta - h^2 b_2) \end{pmatrix} \xrightarrow{D} N(0, \Sigma)$$

where

$$\Sigma = \frac{1}{p(y)} \frac{\int \tilde{g}^2(y)\ f(y)\ dy}{\Big(\int \tilde{s}(y)\ f(y)\ dy\Big)^2}\ S^{-1}\ S^*\ S^{-1}$$

and

$$b_1 = \frac{1}{2}s''(y)\frac{\mu_2}{\mu_0}, \qquad b_2 = \frac{1}{6}s'''(y)\frac{\mu_4}{\mu_2}$$

where $s''(y) = \{\frac{\partial^2}{\partial x^2}\log\sigma(x)\}|_{x=y}$ and $s'''(y) = \{\frac{\partial^3}{\partial x^3}\log\sigma(x)\}|_{x=y}$

The above results are for estimating $\hat{\alpha} = \hat{s}(y)$ and $\hat{\beta} = \hat{s}'(y)$ of the log-standard deviation. Simple algebra calculation and using Taylor expansion yields the following results for the estimator $\hat{\sigma}^2(y)$,

Corollary 1 *Under conditions C1-C4 in Appendix, the estimate of the log-standard deviation function* $\hat{\alpha} = \log \hat{\sigma}(y)$ *is asymptotically normally distributed i.e.*

$$\sqrt{nh}(\hat{\alpha} - \alpha - h^2 b_1) \rightarrow N(0, v^2)$$

Therefore, the local maximum likelihood estimate of the variance function is asymptotically normally distributed with

$$\sqrt{nh}(\hat{\sigma}^2(y) - \sigma^2(y) - h^2 b_1') \rightarrow N(0, 4\sigma^4(y)v^2)$$

where

$$b_1' = \frac{1}{2}\mu_2(\sigma^{2\prime\prime}(y) - \frac{(\sigma^{2\prime}(y))^2}{\sigma^2(y)}) \text{ and } v^2 = \frac{\nu_0}{{\mu_0}^2}\frac{1}{p(y)}\frac{\int \tilde{g}^2(y)\ f(y)\ dy}{\Big(\int \tilde{s}(y)\ f(y)\ dy\Big)^2}$$

Note, we see that the error density appears in the formula for the asymptotic variance but not in the bias. In fact, the bias term involves the kernel, the estimated function and its derivatives. This is reasonable because bias is based on the properties of the estimated function rather than on the stochastic properties of the data. Nevertheless the log-transformation has affected the bias and the gain or loss in efficiency depends on the second derivative of variance function.

To calculate the asymptotic variance for the case of normally distributed errors is of interest. Suppose $\varepsilon \sim N(0,1)$ so $f(y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}$. To calculate

$$4\sigma^4 v^2 = 4\sigma^4 \nu_0 \frac{1}{p(y)}\frac{\int \tilde{g}^2(y)\ f(y)\ dy}{\Big(\int \tilde{s}(y)\ f(y)\ dy\Big)^2}$$

we first note that

$$\int \tilde{s}(y)f(y)dy = \int y^2 f''(y)dy + \int y f'(y)dy - \int y^2 \frac{(f'(y))^2}{f(y)}dy = I_1 + I_2 - I_3$$

while the numerator can be written,

$$\int \tilde{g}^2(y)f(y)dy = \int y^2 \frac{(f'(y))^2}{f(y)}dy + \int 2y f'(y)dy + \int f(y)dy = I_3 + 2I_2 + 1$$

Simple calculation lead to $I_1 = 2$, $I_2 = -1$, $I_3 = 3$. Hence,

$$\frac{\int \tilde{g}^2(y)f(y)dy}{\Big(\int \tilde{s}(y)f(y)dy\Big)^2} = \frac{1}{2}$$

Consequently, under the assumption of normally distributed errors the bias and asymptotic variance of the estimate $\hat{\sigma}^2$ are

$$B_{LMLE} = \frac{h^2}{2}\mu_2(\sigma^{2''}(y) - \frac{(\sigma^{2'}(y))^2}{\sigma^2(y)}) \quad AV_{LMLE} = 2\nu_0\frac{\sigma^4(y)}{p(y)}.$$

Fan and Yao (1996) showed using local Linear Least Square Estimation (LLSE) the bias and asymptotic variance of the estimate are given by

$$B_{LLSE} = \frac{h^2}{2}\mu_2\sigma^{2''}(y) \quad AV_{LLSE} = \nu_0\frac{\sigma^4(y)}{p(y)}(E(\varepsilon_t^4) - 1) = 2\nu_0\frac{\sigma^4(y)}{p(y)}$$

with the latter equality holding when Gaussian errors are assumed. Further Ziegelmann (2001) has shown that using local LLS-estimation together with the exponential transformation (called here ELLSE) would give bias and asymptotic variance:

$$B_{ELLSE} = \frac{h^2}{2}\mu_2(\sigma^{2''}(y) - \frac{(\sigma^{2'}(y))^2}{\sigma^2(y)}) \quad AV_{ELLSE} = 2\nu_0\frac{\sigma^4(y)}{p(y)}.$$

again assuming Gaussian errors. Apparently $AV_{LMLE} = AV_{LLSE} = AV_{ELLSE}$ i.e. all estimates share the same asymptotic variance for the case of Gaussian error term. The bias is the same for the LLME and ELLSE. This is because the log transformation is used in both cases while in LLSE it is not. Note that any comparison is made under the assumption that we have same kernel function and bandwidth for both LMLE and LLSE. In summary, for Gaussian errors, local linear maximum likelihood is equally asymptotically efficient with the local linear least squares and this is not surprising, since the two methods are known to be asymptotically equivalent.

A discussion of the effeciency of the LMLE and ELLSE for non-normal errors follows. Let errors follow a t-distribution with k-degrees of freedom, namely that $\varepsilon_t \sim f(y)$ where $f(y) = C(1 + \frac{y^2}{k-2})^{-\frac{k+1}{2}}$ with $C =$

$(k-2)^{-\frac{1}{2}}(B(\frac{1}{2},\frac{k}{2}))^{-1}$. Note that we have standardized the distribution to ensure that the variance of the ε_t is one. Similarly to the previous example, we write

$$\int \tilde{s}(y)f(y)dy = I_1 + I_2 - I_3 \qquad \int \tilde{g}^2(y)f(y)dy = I_3 + 2I_2 + 1$$

and from $I_1 = 2$, $I_2 = -1$, $I_3 = \frac{3(k+1)}{k+3}$ we end up with

$$\frac{\int \tilde{g}^2(y)f(y)dy}{\Big(\int \tilde{s}(y)f(y)dy\Big)^2} = \frac{k+3}{2k}.$$

Consequently, the asymptotic variance of the LML-estimate is now given by

$$AV_{LMLE} = 4\sigma_0^4\nu_0\frac{1}{p(y)}\frac{k+3}{2k}.$$

Compare the result again with the common asymptotic variance of LLSE and ELLSE which is

$$AV_{LLSE} = \sigma_0^4(E(\varepsilon_t^4)-1)\nu_0\frac{1}{p(y)} \quad \text{with} \quad E(\varepsilon_t^4) = (k-2)^2\frac{B(\frac{5}{2},\frac{k-4}{2})}{B(\frac{1}{2},\frac{k}{2})}.$$

In order that forth moments exist we require $k > 4$. In table 1 we summarize the results taking as degrees of freedom k=6 and k=8. The ratio of the asymptotic variances

$$Eff \equiv \frac{AV_{LMLE}}{AV_{LLSE}}$$

is 0.6 for 6 degrees of freedom and 0.78 for 8 degrees of freedom and it can be viewed as a measure of the efficiency parallel to the asymptotic efficiency in the parametric theory. This indicates that local maximum likelihood estimator is asymptotically more efficient than the estimator using local linear least square estimation when the error distribution departure from Gaussian.

Using additional information provided from the error distribution, we obtain a more accurate estimate comparing to that of the ELLSE that uses no information from the error distribution. Particularly, when the distribution

	λ	$E(\varepsilon_t^4)$	Eff
k=6	0.75	6	0.6
k=8	0.69	3.5	0.78

Table 1: Efficiency for different degrees of freedom

is heavy tailed (like the t-distribution with small degrees of freedom) the reduction in asymptotic variance is even more significant.

For the bandwidth selection, we know from the theory that there is a trade off between bias and variance in the choice of h. A natural data driven optimal bandwidth selection would be the value of h that minimizes the variance and the squared bias of the estimator. We choose h_{opt} as the value in which the asymptotic mean square error AMSE attains its minimum. From $AMSE(y,h) = B^2(y,h) + AV(y,h)$, we find h_{opt} by minimizing the function of h:

$$AMSE(y,h) = h^4(b_1')^2 + 4\sigma_0^4 v^2 = h^4(b_1')^2 + 4\sigma_0^4 \frac{1}{nh}\frac{\nu_0}{\mu_0^2}\frac{1}{p(y)}\frac{\int \tilde{g}^2(y)f(y)dy}{\Big(\int \tilde{s}(y)f(y)dy\Big)^2}.$$

Following simple derivative calculations we obtain

$$h_{opt} = \Big(\frac{4}{p(y)}\frac{\nu_0}{\mu_2^2}\frac{1}{(\sigma^{2\prime\prime}(y) - \frac{(\sigma^{2\prime}(y))^2}{\sigma^2(y)})^2}\frac{\int \tilde{g}^2(y)f(y)dy}{\Big(\int \tilde{s}(y)f(y)dy\Big)^2}\Big)^{\frac{1}{5}} n^{-\frac{1}{5}}$$

as the optimal bandwidth.

4 Asymptotic Properties when mean function unknown

Throughout the previous sections we assume that the mean function $m(y)$ is known. In practice $m(y)$ is unknown and it has to be estimated. We now

use the local linear maximum likelihood approach, not only for the variance function, but also for the mean function. Recall, the approximation of the variance function in (3). In the same way, we assume that the mean function is locally linear. Of course, some minimum degree of smoothness is required for both the mean and the variance function. We assume that both have at least third continuous derivatives on an open set containing y. In a small neighborhood of y we can write

$$\sigma(Y_t) = e^{\alpha+\beta(Y_t-y)}, \qquad m(Y_t) = \gamma + \zeta(Y_t - y)$$

The conditional local log-likelihood function is now,

$$l_n(\alpha,\beta,\gamma,\zeta) = \sum_{t=1}^{n} \Big(log f\Big(\frac{y_t - \gamma - \zeta(Y_{t-1}-y)}{e^{\alpha+\beta(Y_{t-1}-y)}}\Big) - \alpha - \beta(Y_{t-1}-y)\Big) K_h(Y_{t-1}-y) \tag{5}$$

Call $\theta = (\alpha,\beta,\gamma,\zeta)^T$ the parameter vector. As before, we first introduce some notation. In addition to $\tilde{g}$, and $\tilde{s}$ introduced in the previous section, define

$$\tilde{q}(t) = f'(t)(f(t))^{-1} \qquad \tilde{z}(t) = \frac{f''(t)f(t) - (f'(t))^2}{f^2(t)}$$

$$\text{and} \quad \tilde{v}(t) = \frac{tf''(t)f(t) + f'(t)f(t) - t(f'(t))^2}{f^2(t)}$$

Call

$$X_{1,t} = (1, Y_{t-1} - y, 0, 0)^T, \qquad X_{2,t} = (0, 0, 1, Y_{t-1} - y)^T.$$

Moreover let

$$H = \begin{pmatrix} 1 & 0 \\ 0 & h \end{pmatrix}, \qquad G = \begin{pmatrix} H & 0 \\ 0 & H \end{pmatrix}$$

a 2x2 and a 4x4 diagonal matrix respectively. Let

$$\lambda_1 = \int \tilde{s}(y)f(y)dy, \quad \lambda_2 = \frac{1}{\sigma(y)}\int \tilde{v}(y)f(y)dy, \quad \lambda_3 = \frac{1}{\sigma^2(y)}\int \tilde{z}(y)f(y)dy$$

and

$$\omega_1 = \int \tilde{g}^2(y)f(y)dy, \quad \omega_2 = \frac{1}{\sigma(y)}\int \tilde{g}(y)\ \tilde{q}(y)f(y)dy, \quad \omega_3 = \frac{1}{\sigma^2(y)}\int \tilde{q}^2(y)f(y)dy.$$

We can now establish the asymptotic properties of the joint mean and variance estimator. It holds that

Theorem 2 *Suppose conditions C_1-C_4 in Appendix hold. The vector of the joint estimator of the mean and variance function converges in distribution to a normal distribution. Particularly,*

$$\begin{pmatrix} \sqrt{nh}(\hat{\alpha} - \alpha - h^2 b_1) \\ \sqrt{nh^3}(\hat{\beta} - \beta - h^2 b_2) \\ \sqrt{nh}(\hat{\gamma} - \gamma - h^2 b_3) \\ \sqrt{nh^3}(\hat{\zeta} - \zeta - h^2 b_4) \end{pmatrix} \xrightarrow{D} N(0, \Sigma)$$

where

$$\Sigma = \frac{1}{p(y)} \frac{1}{(\lambda_1 \lambda_3 - {\lambda_2}^2)^2} \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{1,2}^T & \Sigma_{2,2} \end{pmatrix}$$

with

$$\Sigma_{1,1} = ({\lambda_3}^2 \omega_1 - 2\lambda_3 \lambda_2 \omega_2 + {\lambda_2}^2 \omega_3) S^{-1} S^* S^{-1}$$

$$\Sigma_{1,2} = (-\lambda_2 \lambda_3 \omega_1 + \lambda_2^2 \omega_2 + \lambda_1 \lambda_3 \omega_2 - \lambda_1 \lambda_2 \omega_3) S^{-1} S^* S^{-1}$$

$$\Sigma_{2,2} = (\lambda_2^2 \omega_1 - 2\lambda_1 \lambda_2 \omega_2 + \lambda_1^2 \omega_3) S^{-1} S^* S^{-1}$$

and

$$b_1 = \frac{1}{2}\mu_2 s''(y), \quad b_2 = \frac{1}{6}\frac{\mu_4}{\mu_2} s'''(y), \quad b_3 = \frac{1}{2}\mu_2 m''(y), \quad b_4 = \frac{1}{6}\frac{\mu_4}{\mu_2} m'''(y).$$

Apart from deriving the asymptotic results for the local maximum likelihood estimate, our aim is to check whether our estimate is fully regression adaptive. In a regression adaptive model, without knowing the mean function $m(.)$ we can estimate the conditional variance $\sigma^2(.)$ asymptotically as well as if $m(.)$ was known.

Proposition 1 *The local maximum likelihood estimator of the variance function is adaptive with respect to the mean function, if the Hessian matrix is*

diagonal or equivalently $\lambda_2 = 0$. *A sufficient condition would be that the error density function is symmetric.*

Indeed, substituting $\lambda_2 = 0$ in the asymptotic variance of the log-standard deviation estimator (given by $\Sigma_{1,1}$) would lead to the same asymptotic variance derived in theorem 1 where the mean function was assumed to be known. Moreover a symmetric error density would mean that $\tilde{v}(y)$ is antisymmetric in which case $\lambda_2 = 0$ so adaptiveness.

5 Numerical Examples

We give two examples. We want to make a comparison of the estimators and therefore need to know the true values of the variance function for this reason, both examples use simulated data. In both instances we look at the variance function while the mean function has been ignored. The reason is that since the simulated error distributions are the normal and the t-distribution and both are symmetric, we have adaptiveness and therefore, we can treat it as known.

Example 1

We simulate 200 random samples of size n=200 from the model

$$Y_t = \sigma(Y_{t-1})\varepsilon_t$$

where $\varepsilon \sim N(0,1)$, assuming the standard deviation function to be

$$\sigma(x) = \frac{1}{\sqrt{1+x^2}} + log(1+x^2).$$

The performance of the estimator is evaluated by the mean absolute deviation error

$$\varepsilon_{MAD} = \frac{1}{n_{grid}} \sum_{i=1}^{n_{grid}} |\hat{\sigma}^2(x_i) - \sigma^2(x_i)|,$$

where $\{x_i, i = 1, \ldots, n_{grid}\}$ are grid points on $[-2, 2)$ with $n_{grid} = 40$. Moreover, as a kernel, we have chosen the Epanechnikov kernel and the bandwidth is selected by minimizing the asymptotic mean square error.

As expected, for Gaussian errors, there is no significant difference in the performance of the three estimators; figure 1.1 (last page, first in the first row). Further, figure 1.2 (last page, second in the first row) displays the estimated curve and the true variance function and we can see from there, the quality of the estimation.

Example 2: We simulate 250 random samples of size n=200 from the model

$$Y_t = \sigma(Y_{t-1})\varepsilon_t$$

where $\varepsilon \sim t_d$, with $d = 6$ and $d = 2$ degrees of freedom respectively. The standard deviation function is of the form

$$\sigma(x) = 0.8x + 0.6e^{-0.8x^2}.$$

Again we look at the mean absolute deviation error

$$\varepsilon_{MAD} = \frac{1}{n_{grid}} \sum_{i=1}^{n_{grid}} |\hat{\sigma}^2(x_i) - \sigma^2(x_i)|,$$

where $\{x_i, i = 1, \ldots, n_{grid}\}$ are grid points on $[-0.8, 0.7]$ with $n_{grid} = 38$. The boxplot for these two cases are given in figure 1. From figure 2.1 (last page, first in the second row) we can see all the estimators perform rather well with a slight preference for the LMLE although the difference is not significant. This is because we have 6 degrees of freedom and therefore the distribution is closer to the normal case. Another explanation could be that the use of the log-transformation may has increased the bias due to negativity of the second derivative of the variance function at some grid points (all grid points bigger than 0.1 have negative second derivative). We reduce the degrees of freedom to two and re-run the simulation. The result is quite different. Figure 2.2

(last page, second in the second row) indicates what we have already noticed theoretically. Local LMLE outperforms both LLSE and ELLSE. Therefore, we can claim that when the data appears to be heavy tailed, local linear maximum likelihood estimator is more appropriate.

Appendix

Regularity conditions

We introduce some regularity conditions and notation. We use C to denote a generic constant that may be different at different places. Note that C1(ii) is used in the proof of theorem 1 while C1(ii)$'$ is for theorem 2.

C1 (i) For fixed x, $p(x) > 0$ and is differentiable with continuous third derivatives. Moreover, for the error-density function $f(.)$ we assume that $f(y) > 0$ for all $y \in \Re$ and has continuous 3rd derivatives. Further, any interchange of integral and derivative is justified for $f(.)$.
(ii)Call

$$g(y_1, x, \theta) = \tilde{g}(\frac{y_1}{e^{\alpha+\beta(x-y)}}), \quad s(y_1, x, \theta) = \tilde{s}(\frac{y_1}{e^{\alpha+\beta(x-y)}})$$

$$r_1(y_1, x) = g(y_1, x, \theta) f(y_1|x) \quad \text{and} \quad r_2(y_1, x) = s(y_1, x, \theta) f(y_1|x)$$

and $G \subset \Re$ is a compact set with $y \in G$, then we assume that there are functions $H_j(.)$ $j = 1, 2$ such that for $i = 0, 1, 2, 3$ and $\forall x \in G$

$$|\frac{\partial^i}{\partial x^i} r_j(y_1, x)| \leq H_j(y_1)$$

with

$$\int H_1^{\delta}(y_1) dy_1 < \infty, \quad \int H_2(y_1) dy_1 < \infty \quad \text{for some } \delta > 2.$$

(ii)$'$ Call

$$g(y_1, x, \theta) = \tilde{g}(\tfrac{y_1-\gamma-\zeta(x-y)}{e^{\alpha+\beta(x-y)}}), \quad q(y_1, x, \theta) = \tilde{q}(\tfrac{y_1-\gamma-\zeta(x-y)}{e^{\alpha+\beta(x-y)}})$$

$$s(y_1, x, \theta) = \tilde{s}(\tfrac{y_1-\gamma-\zeta(x-y)}{e^{\alpha+\beta(x-y)}}), \quad z(y_1, x, \theta) = \tilde{z}(\tfrac{y_1-\gamma-\zeta(x-y)}{e^{\alpha+\beta(x-y)}})$$

$$v(y_1, x, \theta) = \tilde{v}(\tfrac{y_1-\gamma-\zeta(x-y)}{e^{\alpha+\beta(x-y)}}) \text{ and } r_1(y_1, x) = g(y_1, x, \theta) f(y_1|x),$$

$$r_2(y_1, x) = q(y_1, x, \theta) f(y_1|x),\ r_3(y_1, x) = s(y_1, x, \theta) f(y_1|x),$$

$$r_4(y_1,x)=z(y_1,x,\theta)f(y_1|x), r_5(y_1,x)=v(y_1,x,\theta)f(y_1|x)$$

and let $G \subset \Re$ a compact set with $y \in G$, then there are $H_j(.)$ $j = 1,...,5$ and some $\delta_1 > 2$ and $\delta_2 > 2$ such that for $i = 0,1,2,3$ and $\forall x \in G$,

$$|\frac{\partial^i}{\partial x^i} r_j(y_1,x)| \leq H_j(y_1) \text{ with } \int H_1^{\delta_1}(y_1)dy_1 < \infty,$$

$$\int H_2^{\delta_2}(y_1)dy_1 < \infty \text{ and } \int H_j(y)dy_1 < \infty \text{ for } j = 3,4,5.$$

C2 The kernel function is a continuous and symmetric density function with a bounded support.

C3 The strictly stationary process $\{(Y_t, Y_{t-1})\}$ is strongly mixing, i.e.

$$\alpha(j) \equiv \sup_{A \in \Im_{-\infty}^0, B \in \Im_j^\infty} |P(A)P(B) - P(AB)| \to 0$$

as $j \to \infty$ where $\Im_i^j$ we denote the σ-field generated by $\{(Y_t, Y_{t-1}) : t = i,\ldots,j\}$. Further, we assume that for the $\delta > 2$ given in (ii) i.e. for $\delta = min\{\delta_1, \delta_2\} > 2$ where δ_1,δ_2 given in (ii)$'$, it holds that

$$\sum_{j=1}^{\infty} j^2 \alpha(j)^{1-\frac{2}{\delta}} < \infty$$

C4 As $n \to \infty$, $h \to 0$ and $\liminf_{n\to\infty} nh^4 > 0$.

Remarks The conditions are not the weakest possible and can be altered at the cost of lengthier proofs. In many cases where the assumption of interchanging derivative with integral is used, symmetry of the error density would have been sufficient. Note that C4 implies that $nh \to \infty$. The stronger convergence rate of h is used in the asymptotic distribution.

Proof of theorem 1: Define the processes $Z_t = H^{-1}X_tK_h(Y_{t-1}-y)g(Y_t, Y_{t-1}, \theta_0)$ and $U_t = h(Z_t - E(Z_t))$ both strictly stationary as functions of the strictly

stationary process (Y_t, Y_{t-1}). We follow a similar approach to the one of Cai, Fan and Yao (2000) for estimating functional coefficient regression models. First we prove two lemmas that will be used at the proof of the theorem.

Lemma 1 *Under conditions C1(ii) and C2 we can show that*

$$H^{-1}E(-\frac{1}{n}l_n'(\theta_0)) = (h^2a_1, h^3a_2)^T$$

with

$$a_1 = \frac{1}{2}p(y)s''(y)\mu_2 \int y\ \tilde{g}'(y)f(y)dy + o(1)$$

$$a_2 = \frac{1}{6}p(y)s'''(y)\mu_4 \int y\ \tilde{g}'(y)f(y)dy + o(1)$$

where $s''(y) = \{\frac{\partial^2}{\partial x^2}\log\sigma(x)\}|_{x=y}$ *and* $s'''(y) = \{\frac{\partial^3}{\partial x^3}\log\sigma(x)\}|_{x=y}$

Proof. $H^{-1}E(-\frac{1}{n}l_n'(\theta_0)) = E(\frac{1}{n}\sum_{t=1}^n Z_t)$ where under C1-C2 $E|Z_t|^\delta < \infty$. Since, the first two moments exist, then it is weak stationary. Thus we can write $H^{-1}E(-\frac{1}{n}l_n'(\theta_0)) = E(Z_t)$ where

$$E(Z_t) = \int\int (1, \frac{x-y}{h})^T K(\frac{x-y}{h})\frac{1}{h}g(y_1, x, \theta_0)f(y_1|x)p(x)dy_1dx$$

But

$$\int g(y_1, x, \theta_0)f(y_1|x)dy_1 = \int \tilde{g}(\frac{y_1}{e^{\alpha+\beta(x-y)}})f(\frac{y_1}{e^{s(x)}})\frac{1}{e^{s(x)}}dy_1$$

$$= \int y_1f'(y_1)dy_1 + \int f(y_1)dy_1 + (e^{s(x)-\alpha-\beta(x-y)} - 1)\int y_1\tilde{g}'(y_1)f(y_1)dy_1$$

using Taylor expansion of $\tilde{g}$. Note that $\int y_1f'(y_1)dy_1 = -1$ thus, from the Taylor expansion

$$e^{s(x)-a-\beta(x-y)} - 1 = \frac{s''(y)}{2}(x-y)^2 + o((x-y)^2)$$

we get

$$= s''(y)\frac{1}{2}(x-y)^2\int y_1\tilde{g}'(y_1)f(y_1)dy_1 + o((x-y)^2)$$

and therefore

$$h^2 a_1 = h^2 \frac{1}{2} p(y) s''(y) \mu_2 \int y\ \tilde{g}'(y) f(y) dy + o(h^2)$$

For the second term (the derivative) $a_2 = 0$ since the kernel is symmetric ($\mu_3 = 0$), thus the second order Taylor expansion is not sufficient. Extending to include the third order we have:

$$\int g(y_1, x, \theta_0) f(y_1|x) dy_1 = (s''(y)\frac{1}{2}(x-y)^2 + s'''(y)\frac{1}{6}(x-y)^3) \int y_1\ \tilde{g}'(y_1) f(y_1) dy_1$$

from which we get

$$h^3 a_2 = h^3 \frac{1}{6} p(y) s'''(y) \mu_4 \int y\ \tilde{g}'(y) f(y) dy + o(h^3)$$

and the lemma has been proved.

Lemma 2 *Under conditions C1-C4 the following propositions hold*

a. $h^{-1} Var(U_t) \to p(y) \int \tilde{g}^2(y) f(y) dy\ S^*$

b. $h^{-1} \sum_{t=1}^{n-1} |\mathrm{Cov}(U_1, U_{t+1})| = o(1)$

c. $nh^{-1} Var(Q_n) \to p(y) \int \tilde{g}^2(y) f(y) dy\ S^*$

Proof a. Since $E(U_t) = 0$ we have that

$$Var(U_t) = E(U_t U_t^T) = h^2 \Big(E(Z_t Z_t^T) - E(Z_t) E(Z_t)^T \Big).$$

Using Taylor expansion in conjunction with Condition C1(ii) and C2, we have that for $0 \leq j \leq 2$

$$\{h^2 E(Z_t Z_t^T)\}_j = h \int \int u^j K^2(u) g^2(y_1, hu + y, \theta_0) f(y_1|hu + y) p(hu + y) du dy_1$$

$$= hp(y) \int u^j K^2(u) du \int \tilde{g}^2(y_1) f(y_1) dy_1 + O(h^2).$$

In addition, $E(Z_t) = O(h^2)$ (see lemma 1), therefore $E(Z_t)E(Z_t)^T = O(h^4)$. Thus,

$$h^{-1}Var(U_t) = p(y)\int u^j K^2(u)du \int \tilde{g}^2(y)f(y)dy + O(h^5)$$

and since $h \to 0$ when $n \to \infty$ we conclude a.

The result for c. follows easily from a. and b. along with

$$Var(Q_n) = \frac{1}{n}Var(U_t) + \frac{2}{n}\sum_{t=1}^{n-1}(1 - \frac{t}{n})\text{Cov}(U_1, U_{t+1})$$

and the fact that $nh \to \infty$ implied from C4. So, it remains to prove part b. Let $d_n \to \infty$ to be a sequence of positive integers such that $hd_n \to 0$. Define,

$$J_1 = \sum_{t=1}^{d_n} |\text{Cov}(U_1, U_{t+1})| \qquad \text{and} \qquad J_2 = \sum_{t=d_n}^{n-1} |\text{Cov}(U_1, U_{t+1})|$$

We should show that $J_1 = o(h)$ and $J_2 = o(h)$. We have that for all $t \geq 1$

$$|\text{Cov}(U_1, U_{t+1})| = |E(U_1 U_{t+1}^T)| \leq h^2|E(Z_1 Z_{t+1}^T)| + h^2|E(Z_1)E(Z_{t+1})^T|$$

But using the Markovian property and C1(ii)

$$|E(Z_1 Z_{t+1}^T)| \leq |C||E(H^{-1}X_1 X_{t+1}^T H^{-1}K_h(Y_0 - y)K_h(Y_t - y))| \overset{C2}{<} \infty.$$

Substitute and also use the fact that $E(Z_t) = O(h^2)$ to get, $|\text{Cov}(U_1, U_{t+1})| \leq h^2 C$. It follows that $h^{-1}J_1 \leq hd_n$ and from the choice of d_n we conclude that $J_1 = o(h)$. Next we consider the upper bound of J_2. By using Davydov's inequality, see corollary 1.1 Bosq (1998) , for $\delta > 2$ given in $C1$(ii) and C3, we obtain

$$|\text{Cov}(U_1, U_{t+1})| \leq C\{\alpha(t)\}^{1-\frac{2}{\delta}}\Big(E|U_1|^\delta\Big)^{\frac{1}{\delta}}\Big(E|U_{t+1}|^\delta\Big)^{\frac{1}{\delta}}$$

where $\alpha(t)$ is the mixing coefficient of the process (Y_t, Y_{t-1}) given in C3. Note that $E|U_t|^\delta \leq Ch^\delta E|Z_t|^\delta$ and

$$E|Z_t|^\delta = \int \{1 + u^2\}^{\frac{\delta}{2}} K^\delta(u)h^{1-\delta}p(hu + y)\int |r_1(y_1, hu + y)|^\delta dy_1 du < Ch^{1-\delta}$$

from C1(ii) and C2. Hence $E|U_t|^\delta \leq Ch$ which leads to

$$h^{-1}J_2 \leq Ch^{\frac{2}{\delta}-1}\sum_{t=d_n}^{\infty}\{\alpha(t)\}^{1-\frac{2}{\delta}} \leq Ch^{\frac{2}{\delta}-1}d_n^{-2}\sum_{t=d_n}^{\infty}t^2\{\alpha(t)\}^{1-\frac{2}{\delta}}$$

and by choosing d_n such that $d_n = Ch^{\frac{1}{\delta}-1}$ then the requirements $d_n \to \infty$ and $hd_n \to 0$ are met and moreover under $C3$ we ensure that $J_2 = o(h)$ therefore the lemma has been proved.

We can now conclude the theorem's proof. From the Taylor expansion of the derivative of the likelihood function (use the notation S_n for $l_n''(\theta_0)$)

$$l_n{}'(\hat{\theta}_n) = l_n'(\theta_0) + S_n(\hat{\theta}_n - \theta_0) \Rightarrow \hat{\theta}_n - \theta_0 = S_n^{-1}\{-l_n'(\theta_0)\} \Rightarrow$$

$$\hat{\theta}_n - \theta_0 = S_n^{-1}\{-l_n'(\theta_0) - E(-l_n'(\theta_0))\} + S_n^{-1}E(-l_n'(\theta_0)) \Rightarrow$$

$$\hat{\theta}_n - \theta_0 = I_1 + I_2$$

where

$$I_1 = H^{-1}\Big(H^{-1}S_nH^{-1}\Big)^{-1}H^{-1}\{-l_n'(\theta_0) - E(-l_n'(\theta_0))\}$$

$$I_2 = H^{-1}\Big(H^{-1}S_nH^{-1}\Big)^{-1}H^{-1}E(-l_n'(\theta_0))$$

It is easy to see that the theorem follows from statements a. and b.

a.

$$\sqrt{nh}HI_1 \xrightarrow{D} N(0,\Sigma)$$

b.

$$I_2 = h^2(b_1,b_2)^T + o_p(h^2)$$

Let first prove a: Note

$$H^{-1}\{-l_n'(\theta_0) - E(-l_n'(\theta_0))\} = \sum Z_t - nE(Z_t) \Rightarrow$$

$$\sqrt{nh}HI_1 = \Big(\frac{1}{n}H^{-1}S_nH^{-1}\Big)^{-1}\frac{1}{\sqrt{nh}}\sum_{t=1}^{n}U_t \tag{6}$$

Obviously $E(U_t) = 0$ and moreover in lemma 2 we showed $E|U_t|^\delta < Ch$ Now if we call $\tilde{\alpha}(j)$ the mixing coefficient of U_t then since U_t is a function of (Y_t, Y_{t-1}) it holds from the properties of strong mixing conditions Brandley (1985) , that $\tilde{\alpha}(j) < \alpha(j)$ therefore using C3

$$\sum_{j\geq 1} \tilde{\alpha}(j)^{1-\frac{2}{\delta}} < \sum_{j\geq 1} \alpha(j)^{1-\frac{2}{\delta}} < \infty$$

Then, from theorem 2.21 in Fan and Yao (2002) , we have that $Q_n = \frac{1}{\sqrt{nh}} \sum_{t=1}^{n} U_t$ is asymptotically normal with mean zero and variance

$$\sigma_n^2 \equiv Var(\frac{1}{\sqrt{nh}} \sum_{t=1}^{n} U_t) = nh^{-1} Var(Q_n).$$

Now, using the result of lemma 2

$$\sigma_n^2 \to \sigma^2 = p(y) \int \tilde{g}^2(y) f(y) dy \; S^*$$

we conclude that

$$\frac{1}{\sqrt{nh}} \sum_{t=1}^{n} U_t \xrightarrow{D} N(0, \sigma^2) \tag{7}$$

Further note that for the hessian matrix (under C1-C2):

$$\frac{1}{n} H^{-1} S_n H^{-1} \xrightarrow{P} R = p(y) \int \tilde{s}(y) f(y) dy \; S \tag{8}$$

using the ergodic theorem for $V_{t,j} = K_h(Y_{t-1} - y)(\frac{Y_{t-1}-y}{h})^j s(Y_t, Y_{t-1}, \theta_0)$ and the Taylor expansion

$$E(V_{t,j}) = p(y) \int u^j K(u) du \int s(y_1, y, \theta_0) f(y_1|y) dy_1 + O(h).$$

Combine (8) and (7) to get

$$\sqrt{nh} H I_1 \xrightarrow{D} N(0, R^{-1} \sigma^2 R^{-1}) \tag{9}$$

Substitute R and σ^2 in (9) and we have that the asymptotic variance is

$$\Sigma \equiv R^{-1} \sigma^2 R^{-1} = \frac{1}{p(y)} \frac{\int \tilde{g}^2(y) f(y) dy}{\Big(\int \tilde{s}(y) f(y) dy \Big)^2} \; S^{-1} \; S^* \; S^{-1} \tag{10}$$

We now show b. Recall
$I_2 = H^{-1}(H^{-1}S_nH^{-1})^{-1}H^{-1}E(-l_n'(\theta_0)) \Rightarrow I_2 = H^{-1}(R^{-1}+o_p(1))(h^2a_1, h^3a_2)^T$
substitute R^{-1} and a_1,a_2 from lemma 1 to get

$$I_2 = h^2\frac{\int y\ \tilde{g}'(y)f(y)dy}{\int \tilde{s}(y)f(y)dy}\ (\frac{1}{2}s''(y)\frac{\mu_2}{\mu_0}\ ,\ \frac{1}{6}s'''(y)\frac{\mu_4}{\mu_2})^T + o(h^2)$$

but note that

$$\int y\ \tilde{g}'(y)f(y)dy = \int \tilde{s}(y)f(y)dy$$

hence is reduced to

$$I_2 = h^2\ (\frac{1}{2}s''(y)\frac{\mu_2}{\mu_0}\ ,\ \frac{1}{6}s'''(y)\frac{\mu_4}{\mu_2})^T + o(h^2) \Rightarrow I_2 = h^2(b_1, b_2)^T + o(h^2)$$

and the theorem has been proved.

Proof of theorem 2 is straightforward by following the same steps as above. In that case, derivation of the bias involves second order Taylor expansion of the mean function too while the asymptotic normality is established using similar arguments.

References

Bosq, D. (1998). *Nonparametric statistics for stochastic processes: estimation and prediction*. New York: Springer.

Brandley, R. (1985). The basic properties of strongly mixing conditions. *Dependence in probability*.

Cai, Z., J. Fan, and Q. Yao (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association 95*, 941–956.

Fan, J., M. Farmen, and I. Gijbels (1998). Local maximum likelihood estimation and inference. *Journal of Royal Statist.Soc.B 60*, 591–608.

Fan, J. and Q. Yao (1996). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika 85*, 645–660.

Fan, J. and Q. Yao (2002). Nonlinear time series: nonparametric and parametric methods. To appear.

Hall, P., R. Wolff, and Q. Yao (1999). Methods of estimating a conditional distribution function. *Journal of the American Statistical Association 94*, 154–163.

Härdle, W. and A. Tsybakov (1997). Local polynomial estimator of the volatility function in non-parametric autoregression. *Journal of Econometrics 81*, 223–242.

Linton, O. and Z. Xiao (2001). A nonparametric regression estimator that adapts to error distribution of unknown form. Preprint.

Ruppert, D., M. Wand, U. Holst, and O. Hössjer (1997). Local polynomial variance function estimation. *Technometrics 39*, 262–73.

Simonoff, J. (1996). *Smoothing Methods in Statistics*. New York: Springer.

Staniswalis, J. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association 84*, 276–283.

Ziegelmann, F. (2002). Nonparametric estimation of volatility functions: the local exponential estimator. *Econometric Theory*. To appear.

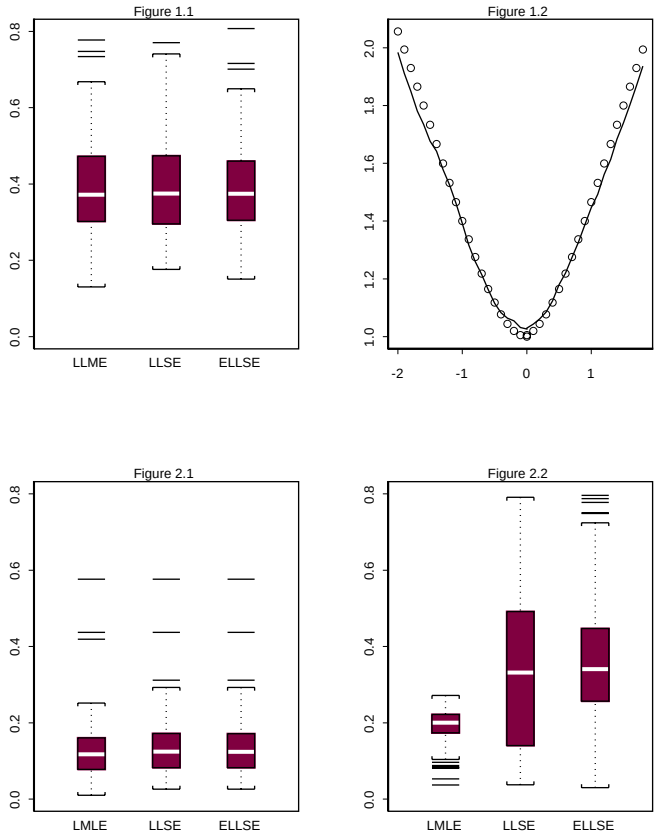

Figure 1: Fig 1.1: Boxplot of the mean absolute deviation error ε_{MAD} for LMLE, LLSE and ELLSE for ex1. Fig 1.2: Plot of true $\sigma(.)$ (doted curve) and estimated $\hat{\sigma}_{LMLE}(.)$ (solid curve) for ex1. Fig 2.1: Boxplot of the mean absolute deviation error ε_{MAD} of LMLE, LLSE and ELLSE for ex2, t-distribution with d=6 degrees of freedom. Fig 2.2: Boxplot of the mean absolute deviation error ε_{MAD} of LMLE,LLSE and ELLSE for ex2, t-distribution with d=2 degrees of freedom.