London School of Economics and Political Science
Department of Statistics

Estimation of the Volatility Function: Non-parametric and Semiparametric Approaches

Panagiotis Avramidis

May 2004

Thesis submitted to the University of London in partial fulfillment of the requirements for the degree of Doctor of Philosophy

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

We investigate two problems in modelling time series data that exhibit conditional heteroscedasticity. The first part deals with the local maximum likelihood estimation of volatility functions which are in the form of conditional variance functions. The existing estimation procedures yield plausible results. Yet, they often fail to take into account special features of the data at the cost of reduced accuracy of prediction. More precisely, many of the parametric and nonparametric conditional variance models ignore the fact that the error distribution departs significantly from gaussian distribution. We propose a novel nonparametric estimation procedure that replaces popular local least squares method with local maximum likelihood estimation. Intuitively, using information from the error distribution improves the estimators and therefore increases the accuracy in prediction. This conclusion is proved theoretically and illustrated by numerical examples. In addition, we show that the proposed estimator adapts asymptotically to the error distribution as well as to the mean regression function. Applications with real data examples demonstrate the potential use of the adaptive maximum likelihood estimator in financial risk management.

The second part deals with the variable selection for a particular class of semiparametric models known as the partial linear models. The existing selection methods are computationally demanding. The proposed selection procedure is computationally more efficient. In particular, if P and Q are the number of linear and nonparametric candidate regressors, respectively, then the proposed procedure reduces the order of the number of variable subsets to be investigated from 2^{Q+P} to $2^Q + 2^P$. At the same time, it maintains all the good properties of existing methods, such as consistency. The latter is proven theoretically and confirmed numerically by simulated examples. The results are presented for the mean regression function while the generalization to the conditional variance function is discussed separately.

Acknowledgement

I am deeply grateful to my supervisor Professor Qiwei Yao for being a limitless source of knowledge and for all his inspiration. I would like to thank my MSc tutor Dr Jeremy Penzer for helping me build the necessary foundations and all the people from the Department of Statistics for their ongoing support. The financial sponsorship from ESRC is gratefully acknowledged. Moreover, I want to thank my physics professor Mr Dimitris Giannakos for his encouragement from the early stages of my studies.

Furthermore, I would like to thank my office mates whose presence made this process a joyful and enriching journey. Special thanks to Jaya, Yorghos and Diego with whom I started this journey and who paved its way. I am also very grateful to all my friends here in London and those back home, for sharing with me the good and the bad moments during these years and giving me the courage to succeed.

Finally, I would like to thank my uncle Iakovos, and my brother Yiannis, for their unconditional love and support. Last here but first in my heart, two people who have defined my life so far: my father for being a stimulating source of wisdom and harmony, and my mother for her unmeasurable kindness, care and love.

To my parents, Konstantinos and Eirini,
my greatest teachers of life,
for their endless love.

In loving memory of my grandmother,
whom I adored.

Πρόσθεσις

Αν ευτυχής ή δύστυχής είμαι δεν εξετάζω.
Πλην ένα πράγμα με χαράν στον νου μου πάντα βάζω-
που στην μεγάλη πρόσθεσι (την πρόσθεσί των που μισώ)
που έχει τόσους αριθμούς, δεν είμαι εγώ εκεί
απ' τες πολλές μονάδες μια. Μες στ' ολικό ποσό
δεν αριθμήθηκα. Κι αυτή η χαρά μ' αρκεί.

Κωνσταντίνος Π. Καβάφης (1897)

Addition

I do not question whether I am happy or unhappy.
Yet there is one thing that I keep gladly in mind -
that in the great addition (their addition that I abhor)
that has so many numbers, I am not one
of the many units there. In the final sum
I have not been calculated. And this joy suffices me.

Constantine P. Cavafy (1897)

Contents

List of Tables

List of Figures

Chapter 1

Introduction in volatility modelling

There is a wide range of time series data sets where the sample variation changes over time, a phenomenon known as heteroscedasticity. For example, in financial markets, large movements tend to be followed by large movements and the same pattern applies for the small movements. The fluctuating behavior of the finance market is referred to as the "*volatility*". Volatility is typically characterized by the conditional variance or standard deviation. Given the extended number of applications, modelling, estimating and predicting the volatility, in the form of conditional variance, have attracted much of the attention in the recent research work. As a result, a large number of volatility models and estimation methods have been developed.

Undoubtedly, using conditional arguments, the most important parametric models are the Auto-Regressive Conditional Heteroscedastic (ARCH) model (Engle 1982) and the Generalized Auto-Regressive Conditional Heteroscedastic (GARCH) model (Bollerslev 1986). Due to the practicality and relatively good performance, GARCH remains one of the most frequently employed conditional variance models in finance. In consequence of its success, variations of GARCH appeared in literature including the popular exponential-GARCH (Nelson 1991) and the GARCH-in-Mean (Engle, Lilien, and Robins 1987). See also Gouriéroux (1997), Hamilton (1994) and Bollerslev, Chu, and Kroner (1994) for an extensive review of the GARCH-type models.

Various non-parametric procedures for estimating conditional variance functions, have been proposed. Härdle and Tsybakov (1997) introduce a kernel estimator based on the decomposition $\sigma^2(x) = \mathrm{E}(Y^2|X = x) - (\mathrm{E}(Y|X = x))^2$. However, the proposed estimator may not be positive and it also creates large bias. Ruppert, Wand, Holst, and Hössjer (1997) for i.i.d. and Fan and Yao (1998) for time series, study a residual-based estimator in conjunction with local kernel smoothing. The resulting estimator is mean regression adaptive[1]. Likewise, Müller and Stadtmüller (1987) obtain the uniform convergence rates for an alternative mean regression adaptive estimator, the kernel-smoothed local variance estimator.

In addition, many of the models introduced in the more general setting of mean regression function can be implemented for conditional variance function after some modification. Recall the well-known Nadaraya-Watson estimator (Nadaraya 1964 and Watson 1964) and the boundary Gasser-Müller kernel regression estimator (Gasser and Müller 1984). However, both estimators suffer from drawbacks. Particularly, the Nadaraya-Watson estimator results in an increase in the bias while Gasser-Müller estimator yields large asymptotic variance. On the other hand, the combination of local polynomial approximation and least squares estimation leads to the local polynomial kernel estimator (Stone 1977 and more recently, Fan 1992, Ruppert and Wand 1994). The local polynomial kernel estimator adapts automatically to estimation at the boundaries. Equivalently, it does not suffer from the lack of sufficient observations at the boundaries, a phenomenon known as "*boundary effects*". Moreover, Fan (1993) showed that it achieves the highest "*linear minimax efficiency*", in the sense of minimum possible supremum of Asymptotic Mean Square Error, among the class of linear smoothers. A detailed picture about the properties and advantages of the local polynomial smoother can be found in Wand and Jones (1995) and Fan and Gijbels (1996). See also Fan and Yao (2003) for an overview of the nonparametric estimation methods including the local polynomial estimator.

[1]see below for definition of adaptiveness

An alternative approach to least squares method is the maximum likelihood estimation. Our motivation stems from the parametric theory and particularly from the fact that maximum likelihood estimator achieves asymptotically the Cramèr Rao bound of the variance of the unbiased estimators. It is understood that using information on the error distribution improves the performance of the estimator and increases the accuracy of the prediction. Although not very frequently employed, likelihood function in nonparametric framework is not totally unknown. Simonoff (1996) and Hjort and Jones (1996) use local likelihood in the context of density function estimation while Staniswalis (1989) derives the asymptotic properties of a kernel likelihood-based estimator of a regression function for the case of i.i.d. data. Further, Linton and Xiao (2001) propose a local polynomial, likelihood-based, estimator of the mean function that adapts to the error distribution. Bandwidth selection and confidence intervals are discussed by Fan, Farmen, and Gijbels (1998). On the other hand, little appears in the literature on the use of likelihood estimation for the volatility function. Yu and Jones (2004) introduce a maximum likelihood estimator for the variance component. However, their approach is a pseudo-likelihood estimation since they use gaussian distribution instead of the unknown error distribution.

We propose a novel, general approach to the use of the likelihood function in the estimation of the conditional variance function. In Chapter 2 we assume that the error distribution is known and we introduce the local linear Maximum Likelihood estimator of the conditional variance function. Specifically, the estimator is a local linear approximation of the log-standard deviation function combined with the likelihood function as the minimizing estimation function. The introduction of the log-transformation in the local polynomial fitting is also pivotal. The estimator of the variance function should be positive, a property implied by the log-transformation with no need for further restrictions. However, the asymptotic results of the estimator suggest that the effect of the log-transformation on the squared bias depends on the properties of the derivatives of the variance function. Therefore, it is likely that

any gain in asymptotic variance may be overshadowed by an increase of the squared bias, see Yu and Jones (2004) for a similar conclusion. In an attempt to quantify the gain due to the use of likelihood function, we perform a direct theoretical comparison based on the Asymptotic Mean Square Error, between the likelihood estimator and existing estimators. The comparison applies for large n as it involves the asymptotic properties of the estimators. The results reflect the initial impression that use of the information from the error distribution improves the performance of the estimator especially when there is significant departure from the assumption of gaussian errors.

Although the nonparametric conditional heteroscedastic model includes directly the conditional mean function, the above results were derived assuming it is known. In order to extend these results for the more realistic case of unknown mean function, we continue in Chapter 2 with the investigation of a joint maximum likelihood estimator for both the mean and variance function. By establishing the asymptotic properties of the joint estimator, we identify sufficient conditions for the adaptiveness of the variance function estimator with respect to the mean function estimator. The term "*adaptiveness*" refers to the characteristic that without knowing the mean function $m(.)$, we can estimate the variance function asymptotically as well as if $m(.)$ was known. Equivalently, adaptiveness implies that the use of the estimated mean function instead of the true mean function has no effect on the first order asymptotic properties of the variance estimator. The identified condition for adaptiveness is the symmetry of the error distribution. It is a well known requirement for similar conclusion about location and scale parameters within the context of parametric regression, see Severini (2000). Furthermore, at the end of Chapter 2, we present two numerical examples that reinforce the conclusions drawn from the direct theoretical comparison.

There is no doubt that the condition of known error distribution is fairly restrictive. For this reason the estimator from Chapter 2 is referred to as the infeasible-Maximum Likelihood estimator. It is therefore understood that this condition needs to be relaxed in order to ensure that ML-estimator can be implemented in practice.

In Chapter 3, we propose a new, likelihood-based estimator that requires no prior knowledge of the error distribution. More precisely, we replace the error density and its derivatives by the nonparametric kernel estimators to obtain an estimate for the score function (the first derivative of the likelihood function) and the Hessian matrix (the second derivative of the likelihood function). Then, the new estimator is the one step Newton Raphson likelihood estimator calculated using the estimated score function and Hessian matrix. It is proven that it shares the same asymptotic properties with the infeasible Maximum Likelihood estimator. The latter implies adaptiveness with respect to the error distribution. Hence, we call this estimator the adaptive Maximum Likelihood estimator. Note that by requiring no particular form for the error density, the results apply for any density function $f(.)$ that satisfies the imposed regularity conditions. This makes the estimator more flexible especially when the error distribution departs significantly from gaussian distribution. Chapter 3 concludes with simulated examples in order to evaluate numerically the performance of the adaptive estimator in comparison with the infeasible estimator as well as other estimators e.g. the Least Squares estimator.

Another important issue in modelling conditional variance function is the choice of the variables used as regressors in the model. The nonparametric model introduced in Chapter 2 assumed a fixed, d-dimensional, set of variables. However, there is little probability that this set of variables is known a priori. More often, we have a restricted number of candidate variables with some of them having no significant effect on the dependent variable. In that case, we need to include only the significant predictors. The reason is that the convergence rate and consequently the performance of the nonparametric estimator decreases as the number of regressors increases, a phenomenon known as "*the curse of dimensionality*". It is therefore critical that the included regressors have significant effect on the dependent variable. The selection of the regressors is usually based on the minimization of a loss function, the choice of which defines the selection criterion. Some of the more frequently used criteria in

nonparametric context include, among others, the Cross-Validation criterion (Cheng and Tong 1992, 1993) and the equivalent of the Final Prediction Error criterion (Tjøstheim and Auestad 1994). Yao and Tong (1994) establish asymptotic results for the Cross-Validation criterion based on kernel estimation. Their approach includes time series. Correspondingly, the developments in variable selection in linear models have been substantial. The Akaike information criterion (Akaike 1974, Shibata 1981), the Cross-Validation (Stone 1974, Shao 1993) and the Generalized-Cross Validation criterion (Craven and Wahba 1979) are the most frequently used model selection procedures. See also Wei (1992) for an overview on the problem of variables selection for linear models. Here, in parallel to the issue of variables selection, we focus on the form of the conditional variance function. It is well known that nonparametric estimators converge more slowly than parametric estimators, see Robinson (1988) and Fan (1993). This means that if the true model contains a linear component then nonparametric estimators will not be as efficient as parametric estimators. In that case a more flexible model form needs to be introduced. This new class of models is semiparametric and is known as partially linear models.

Consequently, in Chapter 4, we deal with the problem of the variable selection for the semiparametric class of partially linear regression models. Following an idea introduced by Gao and Tong (2002), we propose a novel, computationally efficient, variable selection procedure for the class of partially linear models. Particularly, a two step selection procedure is proposed. At the first step, we select the nonparametric regressors based on a Cross-Validation criterion applied to the residuals from linear regression with all the candidate regressors. Then, given the selected nonparametric variables, we use a parametric Cross-Validation criterion to remove any unnecessary linear regressors. It is proven that the proposed selection procedure is consistent. The innovation here is that when calculating the residuals at the first step we include all the linear regressors (even those that are proven insignificant at the second step) and then apply the nonparametric criterion. In this way, we reduce the number of combinations

that should be considered which effectively leads to a significant reduction in the computations. We should mention here that, though presented for a mean regression model, the generalization of the proposed selection method to the variance function is straightforward and is discussed in a separate section. Simulation examples are presented at the end of Chapter 4 to illustrate numerically the consistency as well as the reduction in the computations.

In finance, volatility is often linked to the concept of risk. Within the risk theory, one of the most frequently employed measures is the "*Value at Risk*" (VaR). Hence, in Chapter 5, we implement the proposed nonparametric method for estimating conditional heteroscedastic models in connection with the prediction of the VaR. Using real data, we calculate the VaR along with other performance tests and deviation measures in order to compare the adaptive ML-estimator with existing parametric and nonparametric estimators. We use three different types of financial data sets i.e. stock-indices, stocks and exchange rates. They were chosen on the grounds that financial data sets often exhibit heteroscedasticity while at the same time are heavy tailed. The latter is important because as we shall see in the following chapters the improvement attributed to the proposed estimator becomes more apparent when analyzing heavy tailed data sets. Chapter 5 ends with a summary of the main conclusions concerning the adaptive estimator, drawn from the analysis of the real data.

Chapter 2

Local Linear Maximum Likelihood Estimator

2.1 Model and conditional likelihood function

The model that we are going to study is a non-parametric regression conditional heteroscedastic model. Let $\{Y_t, \mathbf{X}_t\}$ be a strictly stationary process with scalar Y_t and d-dimensional $\mathbf{X}_t^T = (X_{t,1}, \ldots, X_{t,d})$. Denote by $m : \mathbb{R}^d \to \mathbb{R}$ the conditional mean function, $m(\mathbf{x}) = \mathrm{E}(Y_t|\mathbf{X}_t = \mathbf{x})$, and $\sigma^2 : \mathbb{R}^d \to \mathbb{R}^+$ the conditional variance function $\sigma^2(\mathbf{x}) = \mathrm{Var}(Y_t|\mathbf{X}_t = \mathbf{x}) > 0$. Define

$$\epsilon_t = \frac{Y_t - m(\mathbf{X}_t)}{\sigma(\mathbf{X}_t)}. \tag{2.1}$$

It is easy to see that $\mathrm{E}(\epsilon_t|\mathbf{X}_t) = 0$ and $\mathrm{Var}(\epsilon_t|\mathbf{X}_t) = 1$. Assume that ϵ_t are i.i.d. and call $f(.)$ the error density function. At this stage, we assume that the error density is known. We are interested in estimating the variance function and we are going to do this for both cases of known and unknown mean function. From (2.1) we get

$$Y_t = m(\mathbf{X}_t) + \sigma(\mathbf{X}_t)\epsilon_t. \tag{2.2}$$

Note here that time series is included as a special case: if we define $\mathbf{X}_t = (X_{t,1}, \ldots, X_{t,d})$

with $X_{t,i} = Y_{t-i}$ for $i = 1, \ldots, d$, then model (2.2) is an autoregressive conditional heteroskedastic non-linear time series model. Hence, we will not assume independence between Y_t and $\mathbf{X}_t$, though we will impose some mixing conditions necessary for the derivation of the asymptotic properties. Using equation (2.1) we have that the conditional density function of $Y_t|\mathbf{X}_t$ and the error density are related through

$$f_{Y|\mathbf{X}}(y|\mathbf{X}_t = \mathbf{x}) = f\Big(\frac{y - m(\mathbf{x})}{\sigma(\mathbf{x})}\Big)\frac{1}{\sigma(\mathbf{x})}.$$

Consequently, the conditional log-likelihood function is defined

$$l_n(\mathbf{Y}|\mathbf{X}) = \sum_{t=1}^{n} \log f(\frac{Y_t - m(\mathbf{X}_t)}{\sigma(\mathbf{X}_t)}) - \sum_{t=1}^{n} \log \sigma(\mathbf{X}_t) \tag{2.3}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)$.

2.2 Local polynomial fitting

The local polynomial fitting is a useful tool for the estimation of unknown functions. Wand and Jones (1995), Fan and Gijbels (1996), among others, establish the asymptotic properties of local polynomial estimators for the mean regression function. The main idea is that we treat the unknown function as a polynomial in a small neighborhood and by estimating the coefficients of this polynomial we obtain an approximation of the function. The choice of the order of the approximation has an effect on the estimator. As Fan and Gijbels (1996) point out, "*odd order polynomials are preferable to even order polynomials fits*" a conclusion drawn also by Hjort and Jones (1996). The reason is that the constant term of the asymptotic variance increases when moving from an odd order polynomial to the next even order polynomial. Furthermore, higher order polynomial reduces the bias though this reduction is not that crucial given that the bias is controlled by the bandwidth. In the light of these remarks and for the sake of parsimony, we choose the first order polynomial approximation also known as linear approximation. We look at the logarithm of the standard deviation function,

$s(\mathbf{x}) = \log\sigma(\mathbf{x})$ at point $\mathbf{x}$. It is necessary that a minimum degree of smoothness is required in order to apply Taylor expansion, thus for the chosen polynomial order we assume that $\sigma^2(.)$ and hence $s(.)$, has continuous third partial derivatives. The first order Taylor expansion of $s(\mathbf{X}_t) = \log\sigma(\mathbf{X}_t)$ in a small neighborhood around $\mathbf{x}$ is

$$s(\mathbf{X}_t) = s(\mathbf{x}) + \sum_{i=1}^{d} \frac{\partial}{\partial x_i} s(\mathbf{x})(X_{t,i} - x_i) + R(\mathbf{X}_t - \mathbf{x})$$

where the remainder $R(\mathbf{X}_t - \mathbf{x})$ includes the higher order terms. In the above expansion, if we ignore the higher order terms the log-standard deviation function is approximated by a first order polynomial i.e.

$$s(\mathbf{X}_t) = \mathbf{Z}_t^T \boldsymbol{\theta} \tag{2.4}$$

where $\mathbf{Z}_t^T = (1, X_{t,1} - x_1, \ldots, X_{t,d} - x_d)$, $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_d)^T$. Denote with $\boldsymbol{\theta}^0$ the vector of the true values, that is

$$\theta_0^0 = s(\mathbf{x}) \text{ and } \theta_i^0 = \frac{\partial}{\partial x_i} s(\mathbf{x}) \equiv \dot{s}_i(\mathbf{x}).$$

The use of the log transformation ensures that the variance function is always positive, a necessary condition, without having to pose any further restrictions. Moreover, it is proven that under certain conditions, see discussion in section 2.6, the use of log-transformation reduces the bias of the estimator.

2.3 The local linear maximum likelihood estimator

We substitute equation (2.4) into (2.3) to calculate the conditional log-likelihood function for the local linear approximation

$$l_n(\boldsymbol{\theta}; \mathbf{Y}|\mathbf{X}) = \sum_{t=1}^{n} \{\log f(\frac{Y_t - m(\mathbf{X}_t)}{e^{\mathbf{Z}_t^T\boldsymbol{\theta}}}) - \mathbf{Z}_t^T\boldsymbol{\theta}\} K_h(\mathbf{X}_t - \mathbf{x}) \tag{2.5}$$

where $K_h(.)$ is d-dimensional kernel function that assigns weights to the data points Y_t according to the distance of $\mathbf{X}_t$ from the fixed point $\mathbf{x}$ and h is the bandwidth

which defines the size of the neighborhood. Although some minimum conditions are required, the results apply for a wide range of kernel functions. We usually write $K_h(.) = 1/h^d K(./h)$ and let $K(\mathbf{x}) = \prod_{r=1}^{d} k(x_r)$ where $k(.)$ a univariate density function. At this point, we claim that the mean function $m(.)$ in (2.2) is known while the study of the case of unknown mean function is postponed to a later section. Hence, without loss of generality assume that $\mathrm{E}(Y_t|\mathbf{X}_t) = 0 \Rightarrow m(\mathbf{X}_t) = 0$. The local linear Maximum Likelihood estimator (MLE) is obtained by maximizing $l_n(\boldsymbol{\theta}; \mathbf{Y}|\mathbf{X})$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$ where $\boldsymbol{\Theta}$ is a compact set. For notational convenience, call $e_t(\boldsymbol{\theta}) = Y_t e^{-\mathbf{Z}_t^T \boldsymbol{\theta}}$ and particularly $e_t^0 = e_t(\boldsymbol{\theta}^0)$ while note that $\epsilon_t = Y_t e^{-s(\mathbf{X}_t)}$. Then, ML-estimator is defined as the solution to the nonlinear system of equations:

$$S_n(\boldsymbol{\theta}) \equiv -\frac{\partial l_n}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}; \mathbf{Y}|\mathbf{X}) = \mathbf{0} \quad \Rightarrow$$

$$S_n(\boldsymbol{\theta}) \equiv \sum_{t=1}^{n} \{\Psi(e_t(\boldsymbol{\theta}))e_t(\boldsymbol{\theta}) + 1\}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}) = \mathbf{0}, \tag{2.6}$$

where $\Psi(y) = f'(y)/f(y)$. Further, we calculate the Hessian matrix at $\boldsymbol{\theta} \in \Theta$:

$$\mathcal{H}_n(\boldsymbol{\theta}) \equiv \frac{\partial^2 l_n}{\partial \boldsymbol{\theta} \boldsymbol{\theta}^T}(\boldsymbol{\theta}; \mathbf{Y}|\mathbf{X}) = \sum_{t=1}^{n} e_t(\boldsymbol{\theta})\Omega(e_t(\boldsymbol{\theta}))\mathbf{Z}_t\mathbf{Z}_t^T K_h(\mathbf{X}_t - \mathbf{x})$$

where

$$\Omega(y) = \frac{d}{dy}(\Psi(y)y + 1) = \frac{yf''(y)f(y) + f'(y)f(y) - yf'(y)^2}{f(y)^2}.$$

The general theory of likelihood-based statistical inference requires some regularity conditions on the model under consideration. Severini (2000) states the properties of a regular likelihood function (see R1-R3, page 80-81). In particular, property R3 involves the interchanging of integral and differentiation. Consequently, the identities that a regular likelihood function $l(\boldsymbol{\theta})$ should follow are:

$$\mathrm{E}[l'(\boldsymbol{\theta}^0)] = 0$$

$$\mathrm{E}[l''(\boldsymbol{\theta}^0)] = -\mathrm{E}[(l'(\boldsymbol{\theta}^0))(l'(\boldsymbol{\theta}^0))^T]. \tag{2.7}$$

These equations are also known as Bartlett identities. Without loss of generality we calculate identities (2.7) standardized by $n^{-1}H^{-1}$ (see below for definition of the matrix H). Then, the first Bartlett identity yields

$$\mathrm{E}(\frac{1}{n}H^{-1}S_n(\boldsymbol{\theta}^0)) = \mathrm{E}((\Psi(e_t^0)e_t^0 - \Psi(\epsilon_t)\epsilon_t)H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}))$$

$$+\mathrm{E}((\Psi(\epsilon_t)\epsilon_t + 1)H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}))) = \mathbf{0}$$

but Taylor expansion of $\Psi(e_t^0)e_t^0$ around ϵ_t yields

$$\Psi(e_t^0)e_t^0 = \Psi(\epsilon_t)\epsilon_t + \Omega(\epsilon_t)(e_t^0 - \epsilon_t) + o(e_t^0 - \epsilon_t) \tag{2.8}$$

where

$$e_t^0 - \epsilon_t = Y_t(e^{-\mathbf{z}_t^T\boldsymbol{\theta}^0} - e^{-s(\mathbf{X}_t)}) = \epsilon_t(e^{s(\mathbf{X}_t)-\mathbf{z}_t^T\boldsymbol{\theta}^0} - 1). \tag{2.9}$$

Second order Taylor expansion of the log-standard deviation function around $\mathbf{x}$ implies

$$s(\mathbf{X}_t) - \mathbf{Z}_t^T\boldsymbol{\theta}^0 = s(\mathbf{X}_t) - s(\mathbf{x}) - \sum_{i=1}^{d}\dot{s}_i(\mathbf{x})(X_{t,i} - x_i) = \frac{1}{2}\sum_{i,j=1}^{d}\ddot{s}_{i,j}(\mathbf{x}')(X_{t,i} - x_i)(X_{t,j} - x_j) \tag{2.10}$$

with $\mathbf{x}'$ lying between $\mathbf{X}_t, \mathbf{x}$. Substitution of (2.10) to (2.9) yields

$$e_t^0 - \epsilon_t = \frac{1}{2}\epsilon_t\sum_{i,j=1}^{d}\ddot{s}_{i,j}(\mathbf{x}')(X_{t,i} - x_i)(X_{t,j} - x_j) \tag{2.11}$$

which combined with (2.8) leads to

$$\mathrm{E}((\Psi(e_t^0)e_t^0 - \Psi(\epsilon_t)\epsilon_t)H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x})) =$$

$$\mathrm{E}(\Omega(\epsilon_t)\epsilon_t)\mathrm{E}(\frac{1}{2}\sum_{i,j}\ddot{s}_{i,j}(\mathbf{x}')(X_{t,i} - x_i)(X_{t,j} - x_j)H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x})) + o(h^2).$$

It is easy to see under conditions C1-C4 below

$$\mathrm{E}(\frac{1}{2}\sum_{i,j}\ddot{s}_{i,j}(\mathbf{x}')(X_{t,i} - x_i)(X_{t,j} - x_j)H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x})) = O(h^2)$$

which implies that $\mathrm{E}((\Psi(e_t^0)e_t^0 - \Psi(\epsilon_t)\epsilon_t)H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x})) = O(h^2) = o(1)$ since $h \to 0$. Call $p(.)$ the density function of $\mathbf{X}_t$ then,

$$\mathrm{E}((\Psi(\epsilon_t)\epsilon_t+1)H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t-\mathbf{x})) = p(\mathbf{x}) \int (\Psi(\epsilon)\epsilon+1)f(\epsilon)d\epsilon \int (1, \mathbf{u}^T)^T K(\mathbf{u})d\mathbf{u}+O(h)$$

and from assumptions C1-C4, we conclude that Bartlett's first identity yields

$$\int \{\Psi(\epsilon)\epsilon + 1\}f(\epsilon)d\epsilon = 0 \Rightarrow \int \epsilon f'(\epsilon)d\epsilon = -1. \tag{2.12}$$

Similarly, the second identity yields

$$\int \{\Psi(\epsilon)\epsilon + 1\}^2 f(\epsilon)d\epsilon = -\int \epsilon\Omega(\epsilon)f(\epsilon)d\epsilon \Rightarrow \int \epsilon^2 f''(\epsilon)d\epsilon = 2. \tag{2.13}$$

2.4 Asymptotic properties of the ML-estimator

The following regularity conditions are sufficient for the derivation of the asymptotic properties. In many cases, these conditions can be altered at the cost of lengthier proofs. We define $H = \mathrm{diag}\{1, h, \ldots, h\}$ the $(d+1) \times (d+1)$ bandwidth matrix and let $0 < C < \infty$ a generic constant that may take different values at different places.

C1 (i) For fixed $\mathbf{x}$, $p(\mathbf{x}) > 0$ with continuous first derivative and $f(.)$ has up to four continuous derivatives. Further, it holds that the function $y\Psi(y)$, $y \in \mathbb{R}$, is twice continuously differentiable with $\Omega(y) = (d/dy)(y\Psi(y) + 1)$ and $R(y) = (d/dy)\Omega(y)$.

(ii) For the i.i.d. error term ϵ_t, there is $\delta > 2$ such that

$$\mathrm{E}|\Psi(\epsilon)\epsilon + 1|^{2\delta-2} < \infty, \ \mathrm{E}|\epsilon\Omega(\epsilon)|^2 < \infty \text{ and } \mathrm{E}|\epsilon^2 R(\epsilon)| < \infty.$$

(iii) The log-standard deviation function $s(\mathbf{x})$ has up to three continuous derivatives and it holds that for fixed $\mathbf{x}$, $|\ddot{s}_{ij}(\mathbf{x}')| < \infty$ for $||\mathbf{x}' - \mathbf{x}|| \leq C$ and $i, j = 1, \ldots, d$.

C2 The kernel function defined earlier is a continuous and symmetric density function with a bounded support. Further, we assume that for $\mathbf{u} = (u_1, \ldots, u_d)^T$ we have that

$$\mu_0 = \int K(\mathbf{u})d\mathbf{u} = 1, \quad \int \mathbf{u}K(\mathbf{u})d\mathbf{u} = 0 \text{ and } \int \mathbf{u}^T\mathbf{u}K(\mathbf{u})d\mathbf{u} = \mu_2\mathbf{I}$$

with $\mu_2 = \int u_i^2 K(\mathbf{u})d\mathbf{u}$ independent of i. Note also that

$$\int u_i u_j u_k K(\mathbf{u})d\mathbf{u} = 0 \; \forall \; i, j, k$$

$$\int u_i u_j u_k u_l K(\mathbf{u})d\mathbf{u} = \begin{cases} \int u_i^2 u_k^2 K(\mathbf{u})d\mathbf{u} = \mu_2^2, & i = j, k = l \\ 0, & \text{otherwise.} \end{cases}$$

C3 The strictly stationary process $(Y_t, \mathbf{X}_t)$ is strongly mixing, i.e.

$$\alpha(t) \equiv \sup_{A \in \Im_{-\infty}^0, B \in \Im_t^\infty} |P(A)P(B) - P(AB)| \to 0 \quad \text{as } t \to \infty$$

where with $\Im_{t_1}^{t_2}$ we denote the σ-field generated by $\{(Y_t, \mathbf{X}_t) : t = t_1, \ldots, t_2\}$. Further we assume that for the same $\delta > 2$ given in C1

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

C4 As $n \to \infty$, $h \to 0$ and $nh^d \to \infty$.

The conditions in C1 are a minimum requirement to ensure the convergence in probability of the first, second and third order derivatives of the likelihood function. They are also important in the use of the Central Limit Theorem in the derivation of the asymptotic distribution. The conditions for the kernel are self-explanatory and rather common within this context. Note that the bounded support can be relaxed but it requires further conditions. Condition C3 determines the mixing properties of the process under investigation. Condition C4 involving the rate of the bandwidth is standard in nonparametric theory.

2.4.1 Consistency

We proceed to the asymptotic properties of the ML-estimator. Particularly, in the following proposition we show that it is a consistent estimator.

Proposition 2.1 *Suppose that conditions C1-C4 hold. Then there exists at least one solution* $\hat{\boldsymbol{\theta}}_n$ *of the likelihood equations (2.6) that is consistent, i.e.* $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}^0$ *as* $n \to \infty$.

Proof of Proposition 2.1 Call $D_n(\boldsymbol{\theta}) = -1/nH^{-1}S_n(\boldsymbol{\theta})$. Then $\hat{\boldsymbol{\theta}}_n : l_n'(\hat{\boldsymbol{\theta}}_n) = 0 \Leftrightarrow D_n(\hat{\boldsymbol{\theta}}_n) = 0$. Further, let $D(\boldsymbol{\theta}; h) = \mathrm{E}(D_n(\boldsymbol{\theta}))$. We claim that in a compact set $\boldsymbol{\Theta}' \subset \boldsymbol{\Theta}$ that contains $\boldsymbol{\theta}^0$

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}'} \parallel D_n(\boldsymbol{\theta}) - D(\boldsymbol{\theta}; h) \parallel \to 0 \text{ as } n \to \infty \tag{2.14}$$

proof of which is given in Lemma 2.1 below. Moreover, Bartlett identity in (2.7) yields

$$D(\boldsymbol{\theta}^0; h) = 0. \tag{2.15}$$

Now, suppose that none of the solutions of the likelihood function converges in probability to $\boldsymbol{\theta}^0$. Consequently, if $\hat{\boldsymbol{\theta}}_n$ is a solution then there exists subsequence $\hat{\boldsymbol{\theta}}_{kn}$, such that $P\left(\parallel \hat{\boldsymbol{\theta}}_{kn} - \boldsymbol{\theta}^0 \parallel > \delta\right) > \epsilon$ which implies that $P\Big(\inf_{\parallel \boldsymbol{\theta} - \boldsymbol{\theta}^0 \parallel \leq \delta} \parallel D_n(\boldsymbol{\theta}) \parallel > \eta\Big) > \epsilon$ for a sufficiently large n. Equivalently

$$\inf_{\parallel \boldsymbol{\theta} - \boldsymbol{\theta}^0 \parallel \leq \delta} \parallel D_n(\boldsymbol{\theta}) \parallel \xrightarrow{P} 0. \tag{2.16}$$

Since,

$$\inf_{\parallel \boldsymbol{\theta} - \boldsymbol{\theta}^0 \parallel \leq \delta} \parallel D(\boldsymbol{\theta}; h) \parallel \geq \inf_{\parallel \boldsymbol{\theta} - \boldsymbol{\theta}^0 \parallel \leq \delta} \parallel D_n(\boldsymbol{\theta}) \parallel - \sup_{\parallel \boldsymbol{\theta} - \boldsymbol{\theta}^0 \parallel \leq \delta} \parallel D_n(\boldsymbol{\theta}) - D(\boldsymbol{\theta}; h) \parallel$$

from (2.14) and (2.16) we have that $\inf_{\parallel \boldsymbol{\theta} - \boldsymbol{\theta}^0 \parallel \leq \delta} \parallel D(\boldsymbol{\theta}; h) \parallel \xrightarrow{P} 0$ which contradicts with (2.15). Therefore we have proven consistency.

Lemma 2.1 *Under conditions C1-C4 it holds that for any compact set Θ' that includes $\boldsymbol{\theta}^0$*

$$\sup_{\boldsymbol{\theta}\in\Theta'} \parallel D_n(\boldsymbol{\theta}) - D(\boldsymbol{\theta};h) \parallel \to 0 \text{ as } n \to \infty.$$

Proof of Lemma 2.1 Note that

$$D_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n}\{\Psi(e_t(\boldsymbol{\theta}))e_t(\boldsymbol{\theta}) + 1\}H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}) = \frac{1}{n}\sum_{t=1}^{n} V_t(\boldsymbol{\theta}).$$

For fixed $\boldsymbol{\theta}$, the process $V_t(\boldsymbol{\theta})$ is strictly stationary as a function of the strictly stationary process $(Y_t, \mathbf{X}_t)$. Moreover we have that

$$\mathrm{E}||V_t(\boldsymbol{\theta})|| = \mathrm{E}(|\Psi(e_t(\boldsymbol{\theta}))e_t(\boldsymbol{\theta}) + 1| \; ||H^{-1}\mathbf{Z}_t|| \; K_h(\mathbf{X}_t - \mathbf{x}))$$

but $\Psi(e_t(\boldsymbol{\theta}))e_t(\boldsymbol{\theta}) + 1 = (\Psi(\epsilon_t)\epsilon_t + 1) + \Omega(\epsilon_t)\epsilon_t(e^{s(\mathbf{X}_t) - \mathbf{Z}_t^T\boldsymbol{\theta}^0 + \mathbf{Z}_t^T(\boldsymbol{\theta}^0 - \boldsymbol{\theta})} - 1)$ and using expansion (2.10)

$$\mathrm{E}||V_t(\boldsymbol{\theta})|| \leq \mathrm{E}(|\Psi(\epsilon_t)\epsilon_t + 1| \; ||H^{-1}\mathbf{Z}_t|| \; K_h(\mathbf{X}_t - \mathbf{x}))$$

$$+\mathrm{E}(|\Omega(\epsilon_t)\epsilon_t| \; |\frac{1}{2}\sum_{i,j=1}^{d} \ddot{s}_{i,j}(\mathbf{x}')(X_{t,i} - x_i)(X_{t,j} - x_j)| \; ||H^{-1}\mathbf{Z}_t|| \; K_h(\mathbf{X}_t - \mathbf{x}))$$

$$+\mathrm{E}(|\Omega(\epsilon_t)\epsilon_t| \; ||\mathbf{Z}_t^T(\boldsymbol{\theta} - \boldsymbol{\theta}^0)H^{-1}\mathbf{Z}_t|| \; K_h(\mathbf{X}_t - \mathbf{x})).$$

Conditions C1-C4 and the fact that $||\boldsymbol{\theta} - \boldsymbol{\theta}^0|| \leq M$, since Θ' is a compact neighborhood of $\boldsymbol{\theta}^0$, yield that $\mathrm{E}||V_t(\boldsymbol{\theta})|| < \infty$. Consequently Proposition 2.8 (Fan and Yao 2003), which from now on we refer to as the ergodic theorem, yields

$$\frac{1}{n}\sum_{t=1}^{n}\{V_t(\boldsymbol{\theta}) - \mathrm{E}(V_t(\boldsymbol{\theta}))\} \xrightarrow{a.s.} 0 \;\; \Rightarrow \;\; D_n(\boldsymbol{\theta}) - D(\boldsymbol{\theta};h) \xrightarrow{a.s.} 0$$

as $n \to \infty$ while the uniform convergence is implied from the almost sure convergence over the compact set Θ'.

2.4.2 Asymptotic normality

Statistical inferences drawn from confidence intervals or hypothesis tests as well as the bandwidth selection require the distribution of the estimator as Fan, Farmen, and Gijbels (1998) pointed out. Henceforth, we derive the asymptotic distribution of the ML-estimator. We establish asymptotic normality using a Central Limit Theorem and we calculate the asymptotic variance. However, like all the nonparametric estimators, the asymptotic mean square error of MLE includes a bias term along with the asymptotic variance. The first order Taylor expansion of the derivative of the likelihood around the true value $\boldsymbol{\theta}^0$ yields:

$$l_n'(\hat{\boldsymbol{\theta}}_n) = l_n'(\boldsymbol{\theta}^0) + l_n''(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0) \tag{2.17}$$

where $\boldsymbol{\theta}^*$ lies within $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^0$. By definition, $l_n'(\hat{\boldsymbol{\theta}}_n) = 0$ thus from (2.17) we write

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0 = \mathcal{H}_n^{-1}(\boldsymbol{\theta}^*) S_n(\boldsymbol{\theta}^0). \tag{2.18}$$

Next, we present a number of lemmas that involve the calculation of the asymptotic Hessian matrix, the bias term and the asymptotic variance before we proceed to the main theorem.

Lemma 2.2 *Suppose conditions C1-C4 hold. For the Hessian matrix it holds that*

$$\frac{1}{n} H^{-1}\Big(\mathcal{H}_n(\boldsymbol{\theta}^*) - \mathcal{H}_n(\boldsymbol{\theta}^0)\Big) H^{-1} \xrightarrow{P} 0 \qquad as\ n \to \infty.$$

Proof of Lemma 2.2 Note that $n^{-1}H^{-1}\Big(\mathcal{H}_n(\boldsymbol{\theta}^*) - \mathcal{H}_n(\boldsymbol{\theta}^0)\Big)H^{-1} =$

$$\frac{1}{n}\sum_{t=1}^{n}\Big(e_t(\boldsymbol{\theta}^*)\Omega(e_t(\boldsymbol{\theta}^*) - e_t(\boldsymbol{\theta}^0)\Omega(e_t(\boldsymbol{\theta}^0)))\Big) H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1} K_h(\mathbf{X}_t - \mathbf{x}).$$

Since, $\boldsymbol{\theta}^*$ lies within $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^0$, we have that $||\boldsymbol{\theta}^* - \boldsymbol{\theta}^0|| \leq ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||$. But in Proposition 2.1 we proved that $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0|| \xrightarrow{P} 0 \Rightarrow ||\boldsymbol{\theta}^* - \boldsymbol{\theta}^0|| \xrightarrow{P} 0$ as $n \to \infty$. Note that $\Omega(.)$ is a continuous function, equivalently $\mathcal{H}_n(.)$ is continuous, in respect to $\boldsymbol{\theta}$ thus Slutsky's Theorem entails the required result.

Denote with $I(f) = \int (\Psi(\epsilon)\epsilon + 1)^2 f(\epsilon) d\epsilon$ the Fisher information for the error density f. Let $\mathbf{S}_K$ be a $(d+1) \times (d+1)$ diagonal matrix, $\mathbf{S}_K = \text{diag}(\mu_0, \mu_2, \ldots, \mu_2)$, with $\mu_0 = \int K(\mathbf{u}) d\mathbf{u} = 1$ and $\mu_2 = \int u_i^2 K(\mathbf{u}) d\mathbf{u}$. In the following lemma we study the convergence in probability of the Hessian matrix $\mathcal{H}_n(\boldsymbol{\theta}^0)$.

Lemma 2.3 *Suppose conditions C1-C4 hold. The negative Hessian matrix calculated at the true value $\boldsymbol{\theta}^0$ converges in probability to a positive definite $(d+1) \times (d+1)$ matrix $\mathcal{I} \equiv p(\mathbf{x}) I(f)\, \mathbf{S}_K$, i.e.*

$$-\frac{1}{n} H^{-1} \mathcal{H}_n(\boldsymbol{\theta}^0) H^{-1} \xrightarrow{P} \mathcal{I}.$$

Proof of Lemma 2.3 By definition, $-n^{-1} H^{-1} \mathcal{H}_n(\boldsymbol{\theta}^0) H^{-1} = T_1 + T_2$ where

$$T_1 = -\frac{1}{n} \sum_{t=1}^{n} (e_t^0 \Omega(e_t^0) - \epsilon_t \Omega(\epsilon_t)) H^{-1} \mathbf{Z}_t \mathbf{Z}_t^T H^{-1} K_h(\mathbf{X}_t - \mathbf{x})$$

$$T_2 = -\frac{1}{n} \sum_{t=1}^{n} \epsilon_t \Omega(\epsilon_t) H^{-1} \mathbf{Z}_t \mathbf{Z}_t^T H^{-1} K_h(\mathbf{X}_t - \mathbf{x}).$$

Taylor expansion of $e_t^0 \Omega(e_t^0)$ around ϵ_t yields

$$e_t^0 \Omega(e_t^0) - \epsilon_t \Omega(\epsilon_t) = (\epsilon_t R(\epsilon_t) + \Omega(\epsilon_t))(e_t^0 - \epsilon_t) + O((e_t^0 - \epsilon_t)).$$

We substitute expansion (2.11) to obtain

$$e_t^0 \Omega(e_t^0) - \epsilon_t \Omega(\epsilon_t) = (\epsilon_t^2 R(\epsilon_t) + \epsilon_t \Omega(\epsilon_t)) \sum_{i,j} \ddot{s}_{i,j}(\mathbf{x}')(X_{t,i} - x_i)(X_{t,j} - x_j) + O((e_t^0 - \epsilon_t))$$

hence, the first term T_1 is equal to

$$-\frac{1}{n} \sum_{t=1}^{n} (\epsilon_t^2 R(\epsilon_t) + \epsilon_t \Omega(\epsilon_t)) \sum_{i,j} \ddot{s}_{i,j}(\mathbf{x}')(X_{t,i} - x_i)(X_{t,j} - x_j) H^{-1} \mathbf{Z}_t \mathbf{Z}_t^T H^{-1} K_h(\mathbf{X}_t - \mathbf{x}) + o_p(1).$$

Under C1(ii) $\mathrm{E}|\epsilon_t^2 R(\epsilon_t) + \epsilon_t \Omega(\epsilon_t)| < \infty$, C2 and bounded $\ddot{s}_{ij}(\mathbf{x}')$ it is easy to see that

$$\mathrm{E}(||(\epsilon_t^2 R(\epsilon_t) + \epsilon_t \Omega(\epsilon_t)) \sum_{i,j} \ddot{s}_{i,j}(\mathbf{x}')(X_{t,i} - x_i)(X_{t,j} - x_j) H^{-1} \mathbf{Z}_t \mathbf{Z}_t^T H^{-1} K_h(\mathbf{X}_t - \mathbf{x})||) = O(h^2)$$

and therefore ergodic theorem yields $T_1 = o_p(1)$ since $h \to 0$.

For T_2, define the strictly stationary process $R_t = \epsilon_t \Omega(\epsilon_t) H^{-1} \mathbf{Z}_t \mathbf{Z}_t^T H^{-1} K_h(\mathbf{X}_t - \mathbf{x})$. It holds that

$$\mathrm{E}||R_t|| = \mathrm{E}|\epsilon_t \Omega(\epsilon_t)| E\Big(||H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}||K_h(\mathbf{X}_t - \mathbf{x})\Big)$$

$$= \int |\epsilon_t \Omega(\epsilon_t)| f(\epsilon_t) d\epsilon_t \int ||(1, \mathbf{u}^T)^T (1, \mathbf{u}^T)||K(\mathbf{u}) p(\mathbf{x} + h\mathbf{u}) d\mathbf{u}$$

and from C1 and first order Taylor expansion of $p(\mathbf{x} + h\mathbf{u})$ around $\mathbf{x}$, we have

$$\mathrm{E}||R_t|| \leq Cp(\mathbf{x}) \int ||(1, \mathbf{u}^T)^T (1, \mathbf{u}^T)||K(\mathbf{u}) d\mathbf{u} + O(h)$$

which along with C2 and C4 imply that $\mathrm{E}||R_t|| < \infty$. Hence, application of the ergodic theorem for the process R_t yields

$$\frac{1}{n} \sum_{t=1}^{n} \{R_t - \mathrm{E}(R_t)\} \xrightarrow{a.s.} 0. \tag{2.19}$$

But,

$$\mathrm{E}(R_t) \to p(\mathbf{x}) \int (1, \mathbf{u}^T)^T (1, \mathbf{u}^T) K(\mathbf{u}) d\mathbf{u} \int \epsilon \Omega(\epsilon) f(\epsilon) d\epsilon \quad \text{as} \quad n \to \infty$$

which combined with (2.19) yields

$$T_2 = -\frac{1}{n} \sum_{t=1}^{n} R_t \xrightarrow{P} - \int \epsilon \Omega(\epsilon) f(\epsilon) d\epsilon \; p(\mathbf{x}) \; \mathbf{S}_K$$

from

$$\int (1, \mathbf{u}^T)^T (1, \mathbf{u}^T) K(\mathbf{u}) d\mathbf{u} = \mathbf{S}_K.$$

Substituting (2.13) we conclude that $T_2 \xrightarrow{P} p(\mathbf{x}) I(f) \; \mathbf{S}_K$ that completes the proof.

Note that for a symmetrical kernel density function the *information matrix* $\mathcal{I}$ is a diagonal, positive definite matrix since $I(f) = \int (\Psi(\epsilon)\epsilon + 1)^2 f(\epsilon) d\epsilon > 0$, $p(\mathbf{x}) > 0$ and $\mu_0, \mu_2 > 0$ i.e. the Hessian matrix is negative definite which is a requirement for

the existence of local maximum of the likelihood function. The next lemma is used to assess the bias. Define the $(d+1) \times (d+1)$ and $(d+1) \times 1$-matrices,

$$\mathcal{M}_{K,1} = \begin{pmatrix} \mu_2 & 0 & \dots & 0 \\ 0 & \mu_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mu_2^2 \end{pmatrix}, \ \mathbf{H}_s = \begin{pmatrix} \frac{1}{2}\sum_{j=1}^{d} \ddot{s}_{jj}(\mathbf{x}) \\ \frac{1}{6}\sum_{j=1}^{d} \dddot{s}_{1jj}(\mathbf{x}) \\ \vdots \\ \frac{1}{6}\sum_{j=1}^{d} \dddot{s}_{djj}(\mathbf{x}) \end{pmatrix} \tag{2.20}$$

where $\mu_2 = \int u_i^2 K(\mathbf{u}) d\mathbf{u}$.

Lemma 2.4 *Under conditions C1-C2, it holds that*

$$\mathrm{E}(\frac{1}{n} H^{-1} S_n(\boldsymbol{\theta}^0)) = -h^2 p(\mathbf{x}) I(f)\ H\ \mathcal{M}_{K,1}\ \mathbf{H}_s + (o(h^2), o(h^3), \dots, o(h^3))^T.$$

Proof of Lemma 2.4. Note that from stationarity

$$\mathrm{E}(\frac{1}{n} S_{n,0}(\boldsymbol{\theta}^0)) = \int\int (\Psi(e_t^0)e_t^0 + 1)\frac{1}{h^d} K(\frac{\mathbf{x}_t - \mathbf{x}}{h}) f_{Y|\mathbf{X}}(y_t|\mathbf{x}_t) p(\mathbf{x}_t) dy_t d\mathbf{x}_t \tag{2.21}$$

and for $r = 1, \dots, d$

$$\mathrm{E}(\frac{1}{nh} S_{n,r}(\boldsymbol{\theta}^0)) = \int\int (\Psi(e_t^0)e_t^0 + 1)\frac{x_{t,r} - x_r}{h}\frac{1}{h^d} K(\frac{\mathbf{x}_t - \mathbf{x}}{h}) f_{Y|\mathbf{X}}(y_t|\mathbf{x}_t) p(\mathbf{x}_t) dy_t d\mathbf{x}_t. \tag{2.22}$$

Concentrate on (2.21). Recall Taylor expansion in (2.8) and (2.11) where

$$(\Psi(e_t^0)e_t^0 + 1) = (\Psi(\epsilon_t)\epsilon_t + 1) +$$

$$\epsilon_t \Omega(\epsilon_t)\frac{1}{2}\sum_{i,j=1}^{d} \ddot{s}_{ij}(\mathbf{x})(X_{t,i} - x_i)(X_{t,j} - x_j) + o((X_{t,i} - x_i)(X_{t,j} - x_j)). \tag{2.23}$$

Substitution of (2.23) to (2.21) yields

$$\mathrm{E}(\frac{1}{n} S_{n,0}(\boldsymbol{\theta}^0)) = \int (\Psi(\epsilon_t)\epsilon_t + 1) f(\epsilon_t) d\epsilon_t \int \frac{1}{h^d} K(\frac{\mathbf{x}_t - \mathbf{x}}{h}) p(\mathbf{x}_t) d\mathbf{x}_t + \int \epsilon_t \Omega(\epsilon_t) f(\epsilon_t) d\epsilon_t$$

$$\int\int \Big(\frac{1}{2}\sum_{i,j=1}^{d} \ddot{s}_{ij}(\mathbf{x})(x_{t,i} - x_i)(x_{t,j} - x_j) + o((x_{t,i} - x_i)(x_{t,j} - x_j))\Big)\frac{1}{h^d} K(\frac{\mathbf{x}_t - \mathbf{x}}{h}) p(\mathbf{x}_t) d\mathbf{x}_t.$$

It is easy to see that from (2.12), the first term is

$$\int (\Psi(\epsilon_t)\epsilon_t + 1) f(\epsilon_t) d\epsilon_t \int \frac{1}{h^d} K(\frac{\mathbf{x}_t - \mathbf{x}}{h}) p(\mathbf{x}_t) d\mathbf{x}_t = 0.$$

For the second integral, using the transformation $\mathbf{x}_t - \mathbf{x} = h\mathbf{u}$ we have that

$$\int \epsilon_t \Omega(\epsilon_t) f(\epsilon_t) d\epsilon_t \int \left[\frac{1}{2} \sum_{i,j=1}^{d} \ddot{s}_{ij}(\mathbf{x}) h^2 u_i u_j + o(h^2) \right] K(\mathbf{u}) p(\mathbf{x} + h\mathbf{u}) d\mathbf{u}$$

$$= h^2 p(\mathbf{x}) \int \epsilon_t \Omega(\epsilon_t) f(\epsilon_t) d\epsilon_t \frac{1}{2} \sum_{i,j=1}^{d} \ddot{s}_{ij}(\mathbf{x}) \int u_i u_j K(\mathbf{u}) d\mathbf{u} + o(h^2)$$

$$= -\frac{1}{2} h^2 p(\mathbf{x}) I(f) \sum_{i,j=1}^{d} \ddot{s}_{ij}(\mathbf{x}) \mu_{i,j} + o(h^2)$$

using (2.13). Thus, condition C2 yields

$$\mathrm{E}(\frac{1}{n} S_{n,0}(\boldsymbol{\theta}^0)) = -\frac{1}{2} h^2 \mu_2 p(\mathbf{x}) I(f) \sum_{j=1}^{d} \ddot{s}_{jj}(\mathbf{x}) + o(h^2). \tag{2.24}$$

Wand and Jones (1995) and Fan and Gijbels (1996) showed that the calculation of the bias of the derivative estimator requires a third order Taylor expansion of the log-standard deviation function assuming that the kernel is a symmetric function. Therefore we extend (2.10) to

$$s(\mathbf{x}_t) - \mathbf{z}_t^T \theta^0 = \frac{1}{2} \sum_{i,j=1}^{d} \ddot{s}_{ij}(\mathbf{x})(x_{t,i} - x_i)(x_{t,j} - x_j)$$

$$+ \frac{1}{6} \sum_{i,j,k=1}^{d} \dddot{s}_{ijk}(\mathbf{x})(x_{t,i} - x_i)(x_{t,j} - x_j)(x_{t,k} - x_k) + o((x_{t,i} - x_i)(x_{t,j} - x_j)(x_{t,k} - x_k))$$

and hence we obtain

$$\mathrm{E}(\frac{1}{nh} S_{n,r}(\boldsymbol{\theta}^0)) = \int (\Psi(\epsilon_t)\epsilon_t + 1) f(\epsilon_t) d\epsilon_t \int \frac{x_{t,r} - x_r}{h} \frac{1}{h^d} K(\frac{\mathbf{x}_t - \mathbf{x}}{h}) p(\mathbf{x}_t) d\mathbf{x}_t +$$

$$\int \epsilon_t \Omega(\epsilon_t) f(\epsilon_t) d\epsilon_t \int \frac{x_{t,r} - x_r}{h} \frac{1}{2} \sum_{i,j=1}^{d} \ddot{s}_{ij}(\mathbf{x})(x_{t,i} - x_i)(x_{t,j} - x_j) \frac{1}{h^d} K(\frac{\mathbf{x}_t - \mathbf{x}}{h}) p(\mathbf{x}_t) d\mathbf{x}_t$$

$$+\int \epsilon_t \Omega(\epsilon_t) f(\epsilon_t) d\epsilon_t \int \frac{x_{t,r}-x_r}{h} \frac{1}{6}\Big[\sum_{i,j,k=1}^{d} \dddot{s}_{ijk}(\mathbf{x})(x_{t,i}-x_i)(x_{t,j}-x_j)(x_{t,k}-x_k)$$

$$+o((x_{t,i}-x_i)(x_{t,j}-x_j)(x_{t,k}-x_k))\Big]\frac{1}{h^d}K(\frac{\mathbf{x}_t-\mathbf{x}}{h})p(\mathbf{x}_t)d\mathbf{x}_t.$$

It has already been shown that the first integral is zero while the second is also equal to zero from condition C2, $\int u_i u_j u_r K(\mathbf{u})d\mathbf{u} = 0$ for all $i, j, r = 1, \ldots, d$. Using similar arguments as above, we write the third integral as

$$= \frac{1}{6}h^3 p(\mathbf{x}) \int \epsilon_t \Omega(\epsilon_t) f(\epsilon_t) d\epsilon_t \sum_{i,j,k=1}^{d} \dddot{s}_{ijk}(\mathbf{x}) \int u_i u_k u_j u_r K(\mathbf{u})d\mathbf{u} + o(h^3)$$

and from C2, $\int u_i u_k u_j u_r K(\mathbf{u})d\mathbf{u} = \int u_r^2 u_j^2 K(\mathbf{u})d\mathbf{u} = \mu_2^2$ if $i = r$, $j = k$ and zero otherwise, it follows that for $i = 1, \ldots, d$

$$\mathrm{E}(\frac{1}{nh}S_{n,r}(\boldsymbol{\theta}^0)) = -\frac{1}{6}h^3\mu_2^2 p(\mathbf{x}) I(f) \sum_{j=1}^{d} \ddot{s}_{rjj}(\mathbf{x}) + o(h^3) \tag{2.25}$$

and the proof is complete.

The following lemma will be used to assess the asymptotic variance of the estimator, see Cai, Fan, and Yao (2000) for a similar idea. Define the processes $U_t = (V_t - \mathrm{E}(V_t))$ and $Q_n = n^{-1}\sum_{t=1}^{n} U_t$ where $V_t = (\Psi(e_t^0)e_t^0 + 1)H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x})$. Call $\mathbf{S}_K^* = [\nu_{i,j}]_{0\le i,j\le d}$ where $\nu_{i,j} = \int u_i u_j K^2(\mathbf{u})d\mathbf{u}$ for $i, j = 1, \ldots, d$, $\nu_{0,j} = \int u_j K^2(\mathbf{u})d\mathbf{u} = 0$ for $j = 1, \ldots, d$ due to symmetry and $\nu_{0,0} = \int K^2(\mathbf{u})d\mathbf{u}$.

Lemma 2.5 *Under conditions C1-C4 the following propositions hold*

(a) $h^d \mathrm{Var}(U_t) \to p(\mathbf{x}) I(f)\, \mathbf{S}_K^*$.

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

(c) $nh^d \mathrm{Var}(Q_n) \to p(\mathbf{x}) I(f)\, \mathbf{S}_K^*$.

Proof of Lemma 2.5 (a) We have that $\mathrm{Var}(U_t) = \mathrm{E}(U_t U_t^T)$ where

$$\mathrm{E}(U_t U_t^T) = \mathrm{E}(V_t V_t^T) - \mathrm{E}(V_t)\mathrm{E}(V_t)^T$$

with $\mathrm{E}(V_t V_t^T) = E\left((\Psi(e_t^0)e_t^0 + 1)^2 H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1} K_h^2(\mathbf{X}_t - \mathbf{x})\right)$. The first order Taylor expansion of $(\Psi(e_t^0)e_t^0 + 1)^2$ around ϵ_t along with (2.11) yields

$$(\Psi(e_t^0)e_t^0 + 1)^2 = (\Psi(\epsilon_t)\epsilon_t + 1)^2 + (\Psi(\epsilon_t)\epsilon_t + 1)\epsilon_t\Omega(\epsilon_t)\sum_{i,j}\ddot{s}_{i,j}(\mathbf{x}')(X_{t,i} - x_i)(X_{t,j} - x_j).$$

Cauchy-Schwartz inequality and C1 ensure that $\mathrm{E}|(\Psi(\epsilon_t)\epsilon_t + 1)\epsilon_t\Omega(\epsilon_t)| < \infty$, which combined with C1(iii), C2 and $h \to 0$ implies that

$$\mathrm{E}(V_t V_t^T) = \int (\Psi(\epsilon_t)\epsilon_t + 1)^2 f(\epsilon_t)d\epsilon_t \int H^{-1}\mathbf{z}_t\mathbf{z}_t^T H^{-1} K_h^2(\mathbf{x}_t - \mathbf{x})p(\mathbf{x}_t)d\mathbf{x}_t + o(1)$$

and using again the transformation $\mathbf{x}_t - \mathbf{x} = h\mathbf{u}$,

$$= h^{-d}\int (\Psi(\epsilon_t)\epsilon_t + 1)^2 f(\epsilon_t)d\epsilon_t \int (1, \mathbf{u}^T)^T(1, \mathbf{u}^T)K^2(\mathbf{u})p(\mathbf{x} + h\mathbf{u})d\mathbf{u} + o(1)$$

consequently, $h^d\mathrm{E}(V_t V_t^T) \to p(\mathbf{x})I(f)\ \mathbf{S}_K^*$ as $n \to \infty$. Moreover, we proved in Lemma 2.4 that $\mathrm{E}(V_t) = O(h^2) = o(1)$. Therefore, we conclude

$$h^d\mathrm{Var}(U_t) \to p(\mathbf{x})I(f)\ \mathbf{S}_K^*. \tag{2.26}$$

Further, it holds that

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

hence, by stationarity, statement (c) follows easily from (a) and (b) as a result of the dominated convergence theorem and C4, $nh^d \to \infty$. Thus, it remains to prove part (b). Let $d_n \to \infty$ be a sequence of positive integers such that $h^d d_n \to 0$. Define $J_1 = \sum_{t=1}^{d_n}|\mathrm{Cov}(U_1, U_{t+1})|$ and $J_2 = \sum_{t=d_n}^{n-1}|\mathrm{Cov}(U_1, U_{t+1})|$. We have that for all $t \geq 1$

$$||\mathrm{Cov}(U_1, U_{t+1})|| = ||\mathrm{E}(U_1 U_{t+1}^T)|| \leq ||\mathrm{E}(V_1 V_{t+1}^T)|| + ||\mathrm{E}(V_1)\mathrm{E}(V_{t+1})^T||. \tag{2.27}$$

But, note that

$$||\mathrm{E}(V_1V_{t+1}^T)|| \leq \mathrm{E}||(\Psi(e_{t+1}^0)e_{t+1}^0+1)(\Psi(e_1^0)e_1^0+1)H^{-1}\mathbf{Z}_1\mathbf{Z}_{t+1}^TH^{-1}K_h(\mathbf{X}_1-\mathbf{x})K_h(\mathbf{X}_{t+1}-\mathbf{x})||$$

$$= \mathrm{E}|(\Psi(\epsilon_{t+1})\epsilon_{t+1}+1)(\Psi(\epsilon_1)\epsilon_1+1)|\ E\Big(||H^{-1}\mathbf{Z}_1\mathbf{Z}_{t+1}^TH^{-1}||K_h(\mathbf{X}_1-\mathbf{x})K_h(\mathbf{X}_{t+1}-\mathbf{x})\Big)+o(1)$$

using similar expansion arguments as in part (a). Condition C1(ii) implies that: $\mathrm{E}|(\Psi(\epsilon_{t+1})\epsilon_{t+1}+1)(\Psi(\epsilon_1)\epsilon_1+1)| < \infty$ hence $||\mathrm{E}(V_1V_{t+1}^T)|| < \infty$. Further, $\mathrm{E}(V_t) = o(1)$, thus from inequality (2.27):

$$||\mathrm{Cov}(U_1, U_{t+1})|| = O(1) \text{ for all } t \geq 1 \tag{2.28}$$

and therefore $J_1 \leq Cd_n \Rightarrow J_1 = O(d_n)$. From the choice of d_n we conclude that $h^dJ_1 = o(1)$. Next we consider the upper bound of J_2. By using Davydov's inequality (Bosq 1998 Corollary 1.1), for $\delta > 2$ given in C1 and C3, we obtain

$$||\mathrm{Cov}(U_1, U_{t+1})|| \leq C\{\alpha(t)\}^{1-\frac{2}{\delta}}\Big(\mathrm{E}||U_1||^\delta\Big)^{\frac{1}{\delta}}\Big(\mathrm{E}||U_{t+1}||^\delta\Big)^{\frac{1}{\delta}} \tag{2.29}$$

where $\alpha(t)$ is the mixing coefficient of the process $(Y_t, \mathbf{X}_t)$ defined in C3. Note that $\mathrm{E}||V_t||^\delta = \mathrm{E}\Big(|\Psi(e_t^0)e_t^0 + 1|^\delta||H^{-1}\mathbf{Z}_t||^\delta K_h^\delta(\mathbf{X}_t - \mathbf{x})\Big)$. Taylor expansion yields

$$(\Psi(e_t^0)e_t^0+1)^\delta = (\Psi(\epsilon_t)\epsilon_t+1)^\delta+\delta(\Psi(\epsilon_t)\epsilon_t+1)^{\delta-1}\Omega(\epsilon_t)\epsilon_t\frac{1}{2}\sum_{i,j}\ddot{s}_{i,j}(\mathbf{x}')(X_{t,i}-x_i)(X_{t,j}-x_j)$$

and from Cauchy-Schwartz inequality

$$(\mathrm{E}|(\Psi(\epsilon_t)\epsilon_t + 1)^{\delta-1}\Omega(\epsilon_t)\epsilon_t|)^2 \leq \mathrm{E}|\Psi(\epsilon_t)\epsilon_t + 1|^{2\delta-2}\mathrm{E}|\Omega(\epsilon_t)\epsilon_t|^2 < \infty$$

bounded, from C1(ii). Hence under C1(iii) and C2 it holds that

$$\mathrm{E}||V_t||^\delta \leq \mathrm{E}|\Psi(\epsilon_t)\epsilon_t + 1|^\delta\mathrm{E}\Big(||H^{-1}\mathbf{Z}_t||^\delta K_h^\delta(\mathbf{X}_t - \mathbf{x})\Big) + O(h^2).$$

Equivalently, $\mathrm{E}||V_t||^\delta \leq Ch^{(1-\delta)d}\int||(1, \mathbf{u}^T)||^\delta K^\delta(\mathbf{u})d\mathbf{u}$ from $\mathrm{E}|\Psi(\epsilon_t)\epsilon_t + 1|^\delta < \infty$, see C1(ii). Moreover, $\int||(1, \mathbf{u}^T)||^\delta K^\delta(\mathbf{u})d\mathbf{u} < \infty$ implying that $\mathrm{E}||V_t||^\delta \leq Ch^{(1-\delta)d}$ and therefore

$$\mathrm{E}||U_t||^\delta = O(h^{(1-\delta)d}). \tag{2.30}$$

The combination of (2.29) and (2.30) leads to

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

hence from C3, $J_2 = o(h^{(2/\delta-2)d} d_n^{-2})$. Therefore, if we define $d_n = Ch^{(2/\delta-1)d/2}$ then $d_n \to \infty$ since $\delta > 2$, $h^d d_n \to 0$ as required for J_1, and $J_2 = o(h^{-d})$, so conclude.

We can now proceed to the main theorem that entails the asymptotic distribution of the ML-estimator.

Theorem 2.1 *Suppose that conditions C1-C4 hold. Then for the Maximum Likelihood estimator we have that*

$$\sqrt{nh^d} H(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0 - \mathbf{b}) \xrightarrow{d} N(0, \mathcal{I}^{-1}\boldsymbol{\Sigma}\mathcal{I}^{-1})$$

where

$$\mathcal{I}^{-1}\boldsymbol{\Sigma}\mathcal{I}^{-1} = p^{-1}(\mathbf{x}) I^{-1}(f)\, \mathbf{S}_K^{-1}\, \mathbf{S}_K^{*}\, \mathbf{S}_K^{-1}$$

and bias

$$\mathbf{b} = h^2 \mathbf{S}_K^{-1}\, \mathcal{M}_{K,1}\, \mathbf{H}_s + (o(h^2), \ldots, o(h^2))^T.$$

Proof of Theorem 2.1 From (2.18) we have that $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0 = I_1 + I_2$ where

$$I_1 = H^{-1}\Big(H^{-1}\mathcal{H}_n(\boldsymbol{\theta}^*)H^{-1}\Big)^{-1} H^{-1}\left(S_n(\boldsymbol{\theta}^0) - \mathrm{E}(S_n(\boldsymbol{\theta}^0))\right)$$

and

$$I_2 = H^{-1}\Big(-H^{-1}\mathcal{H}_n(\boldsymbol{\theta}^*)H^{-1}\Big)^{-1} H^{-1}\mathrm{E}(-S_n(\boldsymbol{\theta}^0)).$$

It is easy to see that the theorem follows from statements:

(a) $\sqrt{nh^d} H I_1 \xrightarrow{d} N(0, \mathcal{I}^{-1}\boldsymbol{\Sigma}\mathcal{I}^{-1})$.

(b) $I_2 = h^2 \mathbf{S}_K^{-1}\, \mathcal{M}_{K,1}\, \mathbf{H}_s + (o(h^2), \ldots, o(h^2))^T$.

We begin with the proof for (a). Recall the process U_t defined in Lemma 2.5 and note that $H^{-1}\{S_n(\boldsymbol{\theta}^0) - \mathrm{E}(S_n(\boldsymbol{\theta}^0))\} = \sum_{t=1}^n U_t$. Thus, we write

$$\sqrt{nh^d}HI_1 = \Big(\frac{1}{n}H^{-1}\mathcal{H}_n(\boldsymbol{\theta}^*)H^{-1}\Big)^{-1}\frac{1}{\sqrt{nh^d}}\sum_{t=1}^n h^d U_t. \tag{2.31}$$

The process h^dU_t is a zero mean, strictly stationary process and using (2.30), there is $\delta > 2$ such that $\mathrm{E}||h^dU_t||^\delta = O(h^{d-d\delta+d\delta}) = O(h^d)$. Further, if we call $\tilde{\alpha}(j)$ the mixing coefficient of U_t then, since U_t is a function of $(Y_t, \mathbf{X}_t)$, it holds from the properties of strong mixing conditions (Bradley 1985) that $\tilde{\alpha}(j) < \alpha(j)$. Therefore using C3

$$\sum_{j\geq 1}\tilde{\alpha}(j)^{1-\frac{2}{\delta}} < \sum_{j\geq 1}\alpha(j)^{1-\frac{2}{\delta}} < \infty. \tag{2.32}$$

Application of the Central Limit Theorem 2.21 (Fan and Yao 2003) yields

$$\frac{1}{\sqrt{nh^d}}\sum_{t=1}^n h^dU_t \xrightarrow{d} N(0,\boldsymbol{\Sigma}_n) \text{ with } \boldsymbol{\Sigma}_n \equiv \mathrm{Var}(\frac{1}{\sqrt{nh^d}}\sum_{t=1}^n h^dU_t) = nh^d\mathrm{Var}(Q_n).$$

Using Lemma 2.5 (c), it follows that

$$\boldsymbol{\Sigma}_n \to \boldsymbol{\Sigma} = p(\mathbf{x})I(f)\,\mathbf{S}_K^* \tag{2.33}$$

and therefore we have shown that

$$\frac{1}{\sqrt{nh^d}}\sum_{t=1}^n h^dU_t \xrightarrow{d} N(0,\boldsymbol{\Sigma}). \tag{2.34}$$

From Lemma 2.2 and 2.3, $n^{-1}H^{-1}\mathcal{H}_n(\boldsymbol{\theta}^*)H^{-1} \xrightarrow{P} -\mathcal{I}$ with $\mathcal{I} = p(\mathbf{x})I(f)\,\mathbf{S}_K$ and the latter along with (2.31), (2.34) yields

$$\sqrt{nh}HI_1 \xrightarrow{d} N(0,\mathcal{I}^{-1}\boldsymbol{\Sigma}\mathcal{I}^{-1}) \tag{2.35}$$

where

$$\mathcal{I}^{-1}\boldsymbol{\Sigma}\mathcal{I}^{-1} = p^{-1}(\mathbf{x})\,I^{-1}(f)\,\mathbf{S}_K^{-1}\,\mathbf{S}_K^*\,\mathbf{S}_K^{-1}. \tag{2.36}$$

We now prove statement (b). Recall that

$$I_2 = H^{-1}\Big(-\frac{1}{n} H^{-1}\mathcal{H}_n(\boldsymbol{\theta}^*)H^{-1}\Big)^{-1} \mathrm{E}(-\frac{1}{n} H^{-1} S_n(\boldsymbol{\theta}^0)).$$

From Lemma 2.4, $\mathrm{E}(-n^{-1}H^{-1}S_n(\boldsymbol{\theta}^0)) = h^2 p(\mathbf{x}) I(f)\ H\ \mathcal{M}_{K,1}\ \mathbf{H}_s + (o(h^2), \ldots, o(h^3))$. Hence $I_2 = H^{-1}(\mathcal{I}^{-1} + o_p(1))(h^2 p(\mathbf{x}) I(f)\ H\ \mathcal{M}_{K,1}\ \mathbf{H}_s + (o(h^2), \ldots, o(h^3)))$. Substitute $\mathcal{I}^{-1}$ and after some algebraic calculations conclude that

$$I_2 = h^2 \mathbf{S}_K^{-1}\ \mathcal{M}_{K,1}\ \mathbf{H}_s + (o(h^2), \ldots, o(h^2))^T$$

which completes the proof of the theorem.

Theorem 2.1 contains the asymptotic distribution of the vector of estimators of the log-standard deviation and its derivatives. The calculation of the univariate asymptotic distribution of the log-standard deviation and equivalently the asymptotic distribution of the variance function itself is straightforward. The next corollary summarizes these results.

Corollary 2.1 *Under conditions C1-C4, the ML-estimator of the log-standard deviation function $\hat{\theta}_0$ is asymptotically normally distributed i.e.*

$$\sqrt{nh^d}(\hat{\theta}_0 - \theta_0^0 - b_0) \xrightarrow{d} N(0, v^2)$$

where

$$b_0 = \frac{h^2}{2}\mu_2 \sum_{j=1}^{d} \ddot{s}_{jj}(\mathbf{x}) \ \textit{and}\ \ v^2 = p^{-1}(\mathbf{x}) I^{-1}(f) \int K^2(\mathbf{u}) d\mathbf{u}.$$

Therefore, the local linear Maximum Likelihood estimator of the variance function $\hat{\sigma}^2(\mathbf{x}) = \exp(2\hat{\theta}_0)$ is asymptotically normally distributed with

$$\sqrt{nh^d}(\hat{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x}) - b) \xrightarrow{d} N(0, 4\sigma^4(\mathbf{x}) v^2)$$

where

$$b = \frac{h^2}{2}\mu_2 \left\{ \sum_{j=1}^{d} \ddot{\sigma}_j^2(\mathbf{x}) - \sum_{j=1}^{d} \Big(\dot{\sigma}_j^2(\mathbf{x})\Big)^2 / \sigma^2(\mathbf{x}) \right\}.$$

Proof of Corollary 2.1 First we derive the asymptotic distribution for the estimator of the log-standard deviation function. Note that $(\hat{\theta}_0 - \theta_0^0 - b_0) = {e_1}^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 - \mathbf{b})$. Thus,

$$v^2 = \mathrm{Var}(\sqrt{nh^d}(\hat{\theta}_0 - \theta_0^0 - b_0)) = {e_1}^T \mathrm{Var}(\sqrt{nh^d}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 - \mathbf{b}))e_1 \Rightarrow$$

$$\mathrm{Var}(\sqrt{nh^d}(\hat{\theta}_0 - \theta_0^0 - b_0)) = \frac{1}{p(\mathbf{x})} I^{-1}(f)\, {e_1}^T\, \mathbf{S}_K^{-1}\, \mathbf{S}_K^*\, \mathbf{S}_K^{-1}\, e_1$$

and substitute $\mu_0 = 1$ and $\nu_{0,0} = \int K^2(\mathbf{u})d\mathbf{u}$ to find v^2. For the asymptotic distribution of the variance estimator we use, for convenience, the notation of $\hat{\sigma}^2 \equiv \hat{\sigma}^2(\mathbf{x})$ and $\sigma^2 \equiv \sigma^2(\mathbf{x})$. From $\hat{\sigma}^2 = e^{2\hat{\theta}_0} \Rightarrow \hat{\sigma}^2/\sigma^2 = e^{2(\hat{\theta}_0 - \theta_0^0)}$ we can equivalently write $(\hat{\sigma}^2 - \sigma^2)/\sigma^2 = e^{2(\hat{\theta}_0 - \theta_0^0)} - 1$. But from Taylor expansion, we have that $e^{2(\hat{\theta}_0 - \theta_0^0)} - 1 = 2(\hat{\theta}_0 - \theta_0^0) + o_p(\hat{\theta}_0 - \theta_0^0)$ hence we conclude that,

$$\sqrt{nh^d}\Big(\frac{\hat{\sigma}^2 - \sigma^2}{\sigma^2}\Big) = 2\sqrt{nh^d}(\hat{\theta}_0 - \theta_0^0) + o_p(\sqrt{nh^d}(\hat{\theta}_0 - \theta_0^0)).$$

Since $\sqrt{nh^d}(\hat{\theta}_0 - \theta_0^0)$ converges in distribution then it is bounded in probability, $\sqrt{nh^d}(\hat{\theta}_0 - \theta_0^0) = O_p(1)$. Therefore, we have that

$$\sqrt{nh^d}\Big(\frac{\hat{\sigma}^2 - \sigma^2}{\sigma^2}\Big) = 2\sqrt{nh^d}(\hat{\theta}_0 - \theta_0^0) + o_p(1). \tag{2.37}$$

From (2.37) it holds that $\sqrt{nh^d}(\hat{\sigma}^2 - \sigma^2)/\sigma^2$ and $2\sqrt{nh^2}(\hat{\theta}_0 - \theta_0^0)$ follow asymptotically the same distribution, i.e.

$$\sqrt{nh^d}(\hat{\sigma}^2 - \sigma^2 - 2b_0\sigma^2) \xrightarrow{d} N(0, 4\sigma^4 v^2).$$

For the bias term, note that $b = 2b_0\sigma^2 = h^2\mu_2\sigma^2 \sum_{j=1}^d (\partial^2/\partial x_j^2) \log\sigma$ where

$$2\sigma^2 \frac{\partial^2}{\partial x_j^2} \log\sigma = \left\{ \sigma^2 \frac{\partial^2}{\partial x_j^2}\sigma^2 - \Big(\frac{\partial}{\partial x_j}\sigma^2\Big)^2 \right\} / \sigma^2$$

and the proof concludes here.

2.5 Implementation and bandwidth selection

It is understood that the performance of the estimator depends critically on the bandwidth h. Although small value for h reduces the bias, the variance of the estimator will be large since there will be fewer data points within the local neighborhood. On the other hand, large value will decrease the variance but it will increase the bias. In other words, there is a trade off between the variance and the bias regarding the bandwidth selection. Consequently, we require a good compromise between these two terms. Due to the importance of the bandwidth parameter, an extensive number of procedures appear in the literature with most of them based on the idea of minimization of a loss function. More specifically, Ruppert, Sheather, and Wand (1995) propose a direct plug-in global bandwidth. They calculate the minimizer of the conditional asymptotic Mean Integrated Squared Error (MISE) and substitute the unknown quantities by their estimates. Härdle, Hall, and Marron (1988), on the other hand, study the asymptotic behavior of a bandwidth selected using a weighted cross validation function as the loss function.

Following these ideas, we propose a direct plug in algorithm. Although we choose the Asymptotic Mean Square Error (AMSE) as the selection criterion, it is understood that other criteria, like cross validation, can be implemented after some modifications. Note here that the proposed algorithm generates a local bandwidth. However, a global bandwidth criterion, equivalent to the asymptotic MISE, could also be considered. Based on the decomposition

$$\mathrm{AMSE}(\mathbf{x};h) = \mathrm{B}^2(\mathbf{x};h) + \mathrm{AV}(\mathbf{x};h)$$

$$\mathrm{AMSE}(\mathbf{x};h) = \frac{h^4}{4}\mu_2^2 \left(\sum_{j=1}^{d} \ddot{\sigma}_{jj}^2(\mathbf{x}) - \sum_{j=1}^{d} \left(\dot{\sigma}_j^2(\mathbf{x})\right)^2 / \sigma^2(\mathbf{x}) \right)^2 + \frac{1}{nh^d} 4\sigma^4(\mathbf{x}) p^{-1}(\mathbf{x}) I^{-1}(f) \int K^2(\mathbf{u}) d\mathbf{u} \tag{2.38}$$

we calculate h_{opt} by minimizing (2.38). Following simple derivative calculations we obtain

$$h_{opt} = C_d(K)C(f)\left(\sum_{j=1}^{d}\ddot{\sigma}_j^2(\mathbf{x})/\sigma^2(\mathbf{x}) - \sum_{j=1}^{d}(\dot{\sigma}_j^2(\mathbf{x}))^2/\sigma^4(\mathbf{x})\right)^{-2/(d+4)} n^{-1/(d+4)} \quad (2.39)$$

where $C_d(K) = (4d\nu_{0,0}\mu_2^{-2})^{1/(d+4)}$ and $C(f) = (p(\mathbf{x})I(f))^{-1/(d+4)}$. Clearly the optimal bandwidth consists of unknown quantities such as the variance function and its derivatives. Further, the error density f in $C(f)$ is assumed to be known but as we have already argued this is not the case in reality so it needs to be estimated as well. The only quantity that can be directly calculated is the kernel-related constant $C_d(K)$. Substitution of the unknown quantities by some pilot estimators leads to the following practical data driven algorithm:

1. Start with the bandwidth $h_0 = c_0 n^{-1/(d+4)}$ with e.g. $c_0 = 0.5$.

2. In the j-th iteration calculate using the bandwidth h_{j-1} the estimators $\hat{\sigma}^2$, $\hat{\dot{\sigma}}_j^2$ and $\hat{\ddot{\sigma}}_{jj}^2$ for $j = 1, \ldots, d$.

3. Improve h_{j-1} by

$$h_j = C_d(K)C(f)\left(\hat{\sigma}^2\sum_{j=1}^{d}\hat{\ddot{\sigma}}_{jj}^2 - \sum_{j=1}^{d}(\hat{\dot{\sigma}}_j^2)^2\right)^{-2/(d+4)} n^{-1/(d+4)}$$

   and repeat from step 2 until convergence is reached or until a specified number of steps has been carried out.

Note here that the proposed algorithm assumes the error density known $f(.)$ which is a condition in this section. However, the generalization to the unknown error density case is straightforward and requires the estimation of $f(.)$. The above algorithm is used in the numerical examples. Hence numerical evaluation of the performance of the bandwidth algorithm is postponed to following section.

2.6 Comparison of MLE with existing estimators

For notational convenience, we study the univariate case $d = 1$ while the generalization for $d \geq 2$ is straightforward. The direct theoretical comparison of the variance function estimators is based on their asymptotic properties. Numerical investigation based on small samples follows in later section. From Corollary 2.1, the bias and the asymptotic variance of the estimator $\hat{\sigma}^2(x)$ are: $b = (h^2/2)\mu_2 \{\ddot{\sigma}^2(x) - (\dot{\sigma}^2(x))^2/\sigma^2(x)\}$ and $4n^{-1}h^{-1}\sigma^4(x)\nu_{0,0}p^{-1}(x)I^{-1}(f)$. Hence, from (2.38), the Asymptotic Mean Square Error is:

$$\text{AMSE}_{\text{MLE}}(x;h) = \frac{4}{nh}\nu_{0,0}\frac{\sigma^4(x)}{p(x)}I^{-1}(f) + \frac{h^4}{4}\mu_2^2\left\{\ddot{\sigma}^2(x) - (\dot{\sigma}^2(x)^2/\sigma^2(x)\right\}^2. \quad (2.40)$$

For the univariate case $d = 1$, Fan and Yao (1998) showed that for the local linear Least Squares estimator it holds that

$$\text{AMSE}_{\text{LSE}_1}(x;h) = \frac{1}{nh}\nu_{0,0}\frac{\sigma^4(x)}{p(x)}(\text{E}(\epsilon_t^4) - 1) + \frac{h^4}{4}\mu_2^2\left\{\ddot{\sigma}^2(x)\right\}^2 \quad (2.41)$$

while Ziegelmann (2002) has shown that use of the log-transformation along with least squares leads to

$$\text{AMSE}_{\text{LSE}_2}(x;h) = \frac{1}{nh}\nu_{0,0}\frac{\sigma^4(x)}{p(x)}(\text{E}(\epsilon_t^4)-1)+\frac{h^4}{4}\mu_2^2\left\{\ddot{\sigma}^2(x) - (\dot{\sigma}^2(x))^2/\sigma^2(x)\right\}^2. \quad (2.42)$$

A direct comparison yields that a reduction in bias can be achieved using the log-transformation under the assumption that $(\dot{\sigma}^2(x))^2 \leq 2\sigma^2(x)\ddot{\sigma}^2(x)$, see Yu and Jones (2004) and Hall and Tao (2002) for a similar conclusion. Consequently, the effect of the log-transformation on the AMSE depends on the properties of the variance function. Furthermore, it is evident that the bias depends solely on deterministic quantities as it is not related to the stochastic error term. We should point out that the comparison involves the constant terms of the AMSE and is made under the assumption of common kernel function.

On the other hand, asymptotic variance is closely related to the stochastic term. Note that the ratio of the asymptotic variances of ML-estimator and LS-estimator is

$$\frac{\mathrm{AV_{MLE}}}{\mathrm{AV_{LSE}}} = 4I^{-1}(f)/(\mathrm{E}(\epsilon_t^4) - 1).$$

Consequently the likelihood estimator has smaller asymptotic variance than the Least Squares estimator if

$$4I^{-1}(f) \leq (\mathrm{E}(\epsilon_t^4) - 1). \tag{2.43}$$

Inequality (2.43) is similar to a Cramèr-Rao type lower bound of the variance of the estimator. In fact, Cramèr-Rao inequality says that

$$\mathrm{Var}_\theta(W(\mathbf{Y})) \geq \frac{\left(\frac{d}{d\theta}\mathrm{E}_\theta(W(\mathbf{Y}))\right)^2}{\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{Y};\theta)\right)^2}$$

where $W(\mathbf{Y})$ is an estimator of θ. The equality holds for the case where the error density function is of the form: $(\partial/\partial\theta)\log f(\mathbf{Y};\theta) = g(\theta)(W(\mathbf{Y})-\theta)$ for some function $g(\theta)$. For instance, gaussian errors satisfy the above condition and hence, it is expected that ML-estimator and LS-estimator have equal asymptotic variance. The latter is shown explicitly below. However, the inequality does not say anything about the actual gain in efficiency. In an attempt to quantify any possible gain in efficiency from the use of likelihood function and have an idea of how "much better" ML-estimator performs asymptotically, we explore analytically two different cases. First, we derive the Asymptotic Mean Square Error for normally distributed errors. Suppose $\varepsilon \sim N(0,1)$ i.e. $f(\epsilon) = (2\pi)^{-1/2}e^{-\epsilon^2/2}$. Note that

$$\int \epsilon\Omega(\epsilon)f(\epsilon)d\epsilon = \int \epsilon^2 f''(\epsilon)d\epsilon + \int \epsilon f'(\epsilon)d\epsilon - \int \epsilon^2\frac{(f'(\epsilon))^2}{f(\epsilon)}d\epsilon = I_1 + I_2 - I_3$$

and that Bartlett identities (2.12),(2.13) yield $I_1 = 2$ and $I_2 = -1$. Using the transformation $\epsilon^2/2 = x \Rightarrow \epsilon d\epsilon = dx$ we have that

$$I_3 = \frac{2}{\sqrt{2\pi}}\int_0^\infty \epsilon^4 e^{-\frac{\epsilon^2}{2}}d\epsilon = \frac{4}{\sqrt{\pi}}\int_0^\infty x^{\frac{3}{2}}e^{-x}dx = 3.$$

Hence, $I(f) = -(I_1 + I_2 - I_3) = 2$. Since ϵ_t are gaussian, then $\mathrm{E}(\epsilon_t^4) = 3$ and therefore (2.43) holds as an equality. In other words, for the case of gaussian error term, likelihood and least squares yield estimators with the same asymptotic variance, equivalently equal AMSE, in terms of constants, given that the log-transformation is employed for both estimators. This result is in line with the well known property that likelihood and least squares are equivalent methods when errors are gaussian. Consider now the case where the error term follows a t-distribution with degrees of freedom $k > 2$, i.e. $f(\epsilon) = C(1 + \epsilon^2/(k-2))^{-(k+1)/2}$ with $C = (k-2)^{-\frac{1}{2}}(B(\frac{1}{2}, \frac{k}{2}))^{-1}$ where $B(.,.)$ the beta function. Note that we have standardized the distribution to ensure that the variance of the ϵ_t is one, otherwise we may have identifiability problems. Similar to the gaussian case, we write $\int \epsilon \Omega(\epsilon) f(\epsilon) d\epsilon = I_1 + I_2 - I_3$ with $I_1 = 2$ and $I_2 = -1$ independent of the distribution. Hence, concentrate on I_3:

$$I_3 = \int \epsilon^2 \frac{(f'(\epsilon))^2}{f(\epsilon)} d\epsilon = 2C \frac{(k+1)^2}{(k-2)^2} \int_0^\infty \epsilon^4 (1 + \frac{\epsilon^2}{k-2})^{-\frac{k+5}{2}} d\epsilon \Rightarrow$$

$$I_3 = (k+1)^2 \frac{B(\frac{5}{2}, \frac{k}{2})}{B(\frac{1}{2}, \frac{k}{2})} = \frac{3(k+1)}{k+3}.$$

Therefore, $I(f) = I_3 - I_1 - I_2 = 2k/(k+3)$. Further, $\mathrm{E}(\varepsilon_t^4) = (k-2)^2 B(\frac{5}{2}, \frac{k-4}{2})/B(\frac{1}{2}, \frac{k}{2})$ with $k > 4$, so that fourth moments exist. The ratio of the asymptotic variances is

$$\frac{\mathrm{AV_{MLE}}}{\mathrm{AV_{LSE}}} = \frac{4(k+3)}{2k\big((k-2)^2 B(\frac{5}{2}, \frac{k-4}{2})/B(\frac{1}{2}, \frac{k}{2}) - 1\big)}$$

and using the properties of the beta $B(x,y)$ and gamma $\Gamma(x)$ functions, we have that

an

$$\frac{\mathrm{AV_{MLE}}}{\mathrm{AV_{LSE}}} = \frac{8(k+3)}{3k(k-2)^2}\Big(\frac{\Gamma((k-4)/2)}{\Gamma(k/2)} - \frac{4}{3(k-2)^2}\Big)^{-1}.$$

$$\frac{\mathrm{AV_{MLE}}}{\mathrm{AV_{LSE}}} = \frac{8(k+3)}{3k(k-2)^2}\Big(\frac{\Gamma((k-4)/2)}{\Gamma(k/2)} - \frac{4}{3(k-2)^2}\Big)^{-1}.$$

The property $\Gamma(x) = (x-1)\Gamma(x-1)$ of gamma function, yields $\Gamma((k-4)/2)/\Gamma(k/2) = 4/((k-4)(k-2))$. Consequently, the ratio is simplified to

$$\frac{\mathrm{AV_{MLE}}}{\mathrm{AV_{LSE}}} = \frac{(k+3)(k-4)}{k(k-1)}. \tag{2.44}$$

Table 2.1: Efficiency for t-distribution with k degrees of freedom

	$I^{-1}(f)$	$\mathrm{E}(\varepsilon_t^4)$	$\mathrm{AV_{MLE}}/\mathrm{AV_{LSE}}$
k=5	0.8	9.0	0.4
k=6	0.750	6.0	0.6
k=8	0.690	4.5	0.780
k=20	0.575	3.375	0.968
k=100	0.515	3.062	0.999

From (2.44), it is easy to see that $\mathrm{AV_{MLE}}/\mathrm{AV_{LSE}} < 1$ for $k > 4$ and $\mathrm{AV_{MLE}}/\mathrm{AV_{LSE}} \to 1$ as $k \to \infty$. Consequently, when the error term follows a t_k-distribution with $k > 4$ degrees of freedom, ML-estimator is more efficient compared to the LS-estimator. In Table 2.1, we summarize the results for different degrees of freedom k in order to quantify the reduction in asymptotic variance. The ratio of the asymptotic variances is 0.4 for 5 degrees of freedom, 0.78 for 8 degrees of freedom going up to 0.999 for $k = 100$. This means that if the error distribution is a t-distribution with 5 degrees of freedom, the asymptotic variance of the estimator based on the likelihood function, is reduced up to 60% compared to the asymptotic variance of the Least Squares estimator. The reduction decreases as the degrees of freedom of t_k-distribution increase and it is only 0.032% for $k = 20$ while it reaches 0.001% for the case of $k = 100$. When k becomes large enough, t_k-distribution approximates the normal distribution therefore it is reasonable that for large k the asymptotic properties of the two estimators do not differ significantly.

Summing up, we conclude that using additional information provided from the error distribution, yields a more accurate estimator, in the sense of smaller AMSE, compared to the LS-estimator that uses no information from the error distribution. Particularly, when the distribution is heavy tailed (like the t_k-distribution with small degrees of freedom) the gain in AMSE is even more significant. However, we should point out that all these conclusions are based on direct comparison of asymptotic

properties and therefore hold, in principle, for n large. The choice of AMSE as the measure of comparison is common in nonparametric theory. However, in the numerical examples presented in later sections we employ alternative measures in order to ensure that any gain in efficiency is independent of the performance measure that has been used.

2.7 Simultaneous estimation of the mean and variance function

Throughout the previous sections we assumed that the mean function $m(.)$ is known. However, in practice, the mean function is unknown and has to be estimated as well. Therefore, we extend the use of likelihood estimation to include the mean function. Since both variance and mean function are approximated with first order polynomials, it is understood that some minimum degree of smoothness is required. Thus, we assume that both have at least continuous third derivative around $\mathbf{x}$. Hence, in a small neighborhood of $\mathbf{x}$ we can write

$$m(\mathbf{X}_t) = \mathbf{Z}_t^T \boldsymbol{\gamma} \tag{2.45}$$

along with the linear approximation of the log-standard deviation given in (2.4): $s(\mathbf{X}_t) = \mathbf{Z}_t^T \boldsymbol{\theta}$ where $\mathbf{Z}_t^T = (1, X_{t,1} - x_1, \ldots, X_{t,d} - x_d)$, $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_d)^T$ and $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_d)^T$. Further let $\boldsymbol{\gamma}^0 = (\gamma_0^0, \ldots, \gamma_d^0)^T$ with $\gamma_0^0 = m(\mathbf{x})$ and $\gamma_j^0 = \dot{m}_j(\mathbf{x})$ the true values similar to $\boldsymbol{\theta}^0$. The conditional local linear log-likelihood function is now given by

$$l_n(\boldsymbol{\theta}, \boldsymbol{\gamma}; \mathbf{Y}|\mathbf{X}) = \sum_{t=1}^{n} \{\log f\Big(\frac{Y_t - \mathbf{Z}_t^T \boldsymbol{\gamma}}{e^{\mathbf{Z}_t^T \boldsymbol{\theta}}}\Big) - \mathbf{Z}_t^T \boldsymbol{\theta}\} K_h(\mathbf{X}_t - \mathbf{x}).$$

Note here that we choose to work with the one step estimation. However, one could adopt a two step estimation procedure where the mean function is estimated first

and then the calculated estimator of the mean is plugged in the likelihood function to estimate the variance function. It is understood that, in the two step estimation approach, a different kernel could be introduce to account for the mean function along with the existing kernel of the variance function. This would also imply that a different bandwidth is used.

Similar to $e_t(\boldsymbol{\theta}) = Y_t e^{-\mathbf{Z}_t^T\boldsymbol{\theta}}$, we define $e_t(\boldsymbol{\gamma},\boldsymbol{\theta}) = (Y_t - \mathbf{Z}_t^T\boldsymbol{\gamma})e^{-\mathbf{Z}_t^T\boldsymbol{\theta}}$. The joint, local linear Maximum Likelihood estimator is the solution of the following system of equations

$$S_n^{(1)}(\boldsymbol{\theta},\boldsymbol{\gamma}) \equiv -\frac{\partial l_n}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\gamma};\mathbf{Y}|\mathbf{X}) = \sum_{t=1}^{n}\{\Psi(e_t(\boldsymbol{\gamma},\boldsymbol{\theta}))e_t(\boldsymbol{\gamma},\boldsymbol{\theta}) + 1\}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}) = \mathbf{0},$$

$$S_n^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma}) \equiv -\frac{\partial l_n}{\partial \boldsymbol{\gamma}}(\boldsymbol{\theta},\boldsymbol{\gamma};\mathbf{Y}|\mathbf{X}) = \sum_{t=1}^{n}\frac{1}{e^{\mathbf{Z}_t^T\boldsymbol{\theta}}}\Psi(e_t(\boldsymbol{\gamma},\boldsymbol{\theta}))\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}) = \mathbf{0}.$$

For the second derivative of the likelihood function $l_n''(\boldsymbol{\theta},\boldsymbol{\gamma})$, call

$$\mathcal{H}_n^{(1)}(\boldsymbol{\theta},\boldsymbol{\gamma}) = \sum_{t=1}^{n} e_t(\boldsymbol{\gamma},\boldsymbol{\theta})\Omega_1(e_t(\boldsymbol{\gamma},\boldsymbol{\theta}))\mathbf{Z}_t\mathbf{Z}_t^T K_h(\mathbf{X}_t - \mathbf{x}),$$

$$\mathcal{H}_n^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma}) = \sum_{t=1}^{n}\frac{1}{e^{\mathbf{Z}_t^T\boldsymbol{\theta}}}\Omega_1(e_t(\boldsymbol{\gamma},\boldsymbol{\theta}))\mathbf{Z}_t\mathbf{Z}_t^T K_h(\mathbf{X}_t - \mathbf{x}),$$

$$\mathcal{H}_n^{(3)}(\boldsymbol{\theta},\boldsymbol{\gamma}) = \sum_{t=1}^{n}\frac{1}{e^{2\mathbf{Z}_t^T\boldsymbol{\theta}}}\Omega_2(e_t(\boldsymbol{\gamma},\boldsymbol{\theta}))\mathbf{Z}_t\mathbf{Z}_t^T K_h(\mathbf{X}_t - \mathbf{x}),$$

where $\Omega_1(y)$[1]$= (d/dy)(\Psi(y)y + 1)$ and $\Omega_2(y) = (d/dy)\Psi(y)$. The $2(d+1) \times 2(d+1)$ Hessian matrix is given by

$$\mathcal{H}_n(\boldsymbol{\theta},\boldsymbol{\gamma}) = \begin{pmatrix} \mathcal{H}_n^{(1)}(\boldsymbol{\theta},\boldsymbol{\gamma}) & \mathcal{H}_n^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma}) \\ \mathcal{H}_n^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma}) & \mathcal{H}_n^{(3)}(\boldsymbol{\theta},\boldsymbol{\gamma}) \end{pmatrix}.$$

Due to the continuity of the 2nd partial derivatives of the likelihood function that allow the interchange of the partial derivatives the Hessian matrix is symmetric.

[1]Note here that for notational convenience we rename Ω from earlier sections as Ω_1.

Under the conditions listed below, it can be proven that the Bartlett identities yield the following:

$$\int \Psi(\epsilon) f(\epsilon) d\epsilon = 0 \Rightarrow \int f'(\epsilon) d\epsilon = 0 \tag{2.46}$$

$$\int \Psi^2(\epsilon) f(\epsilon) d\epsilon = -\int \Omega_2(\epsilon) f(\epsilon) d\epsilon \Rightarrow \int f''(\epsilon) d\epsilon = 0 \tag{2.47}$$

$$\int (\Psi(\epsilon)\epsilon + 1)\Psi(\epsilon) f(\epsilon) d\epsilon = -\int \Omega_1(\epsilon) f(\epsilon) d\epsilon \Rightarrow \int \epsilon f''(\epsilon) d\epsilon = 0 \tag{2.48}$$

along with (2.12) and (2.13) derived earlier. Call $G_1(y) = y\Psi(y)+1$ and $G_2(y) = \Psi(y)$ and let $0 < C < \infty$ be a generic constant that can take different values at different places. Moreover recall $H = \text{diag}\{1, h, \ldots, h\}$ the $(d+1)\times(d+1)$ diagonal bandwidth matrix and define

$$G = \begin{pmatrix} H & 0 \\ 0 & H \end{pmatrix}$$

the augmented $2(d+1)\times 2(d+1)$ diagonal bandwidth matrix. The following conditions are sufficient for the derivation of the asymptotic properties.

C1′ (i) For fixed $\mathbf{x}$, $p(\mathbf{x}) > 0$ with continuous first derivative and $f(.)$ has up to 4th continuous derivatives. Further, it holds that $G_i(y), i = 1, 2$ are twice continuously differentiable with $R_i(y) = (d/dy)\Omega_i(y)$ for $i = 1, 2$.

(ii) For the i.i.d. error term ϵ_t, there are $\delta_i, > 2$, $i = 1, 2$ such that

$$\mathrm{E}|G_i(\epsilon)|^{2\delta_i - 2} < \infty,\ \mathrm{E}|\Omega_i(\epsilon)|^2 < \infty,\ \mathrm{E}|\Omega_i(\epsilon)\epsilon|^2 < \infty$$

$$\mathrm{E}|R_i(\epsilon)| < \infty\ \ \mathrm{E}|R_i(\epsilon)\epsilon| < \infty \text{ for } i = 1, 2 \text{ and } \mathrm{E}|R_1(\epsilon)\epsilon^2| < \infty.$$

(iii) The mean function $m(\mathbf{x})$ and the log-standard deviation function $s(\mathbf{x})$ have up to 3rd continuous derivatives and it holds that for fixed $\mathbf{x}$: $|\ddot{s}_{ij}(\mathbf{x}')| < \infty$ and $|\ddot{m}_{ij}(\mathbf{x}')| < \infty$ for $||\mathbf{x}' - \mathbf{x}|| \leq C$ and $i, j = 1, \ldots, d$.

C2′ The kernel function is a continuous and symmetric density function with bounded support. Further, we assume that for $\mathbf{u} = (u_1, \ldots, u_d)^T$ we have that

$$\mu_0 = \int K(\mathbf{u}) d\mathbf{u} = 1,\ \int \mathbf{u} K(\mathbf{u}) d\mathbf{u} = 0 \text{ and } \int \mathbf{u}^T \mathbf{u} K(\mathbf{u}) d\mathbf{u} = \mu_2 \mathbf{I}$$

with $\mu_2 = \int u_i^2 K(\mathbf{u}) d\mathbf{u}$ independent of i. Note further that

$$\int u_i u_j u_k K(\mathbf{u}) d\mathbf{u} = 0 \; \forall \, i, j, k$$

$$\int u_i u_j u_k u_l K(\mathbf{u}) d\mathbf{u} = \begin{cases} \int u_i^2 u_k^2 K(\mathbf{u}) d\mathbf{u} = \mu_2^2, & i = j, k = l \\ 0, & \text{otherwise.} \end{cases}$$

C3′ The strictly stationary process $(Y_t, \mathbf{X}_t)$ is strongly mixing, i.e.

$$\alpha(t) \equiv \sup_{A \in \Im_{-\infty}^0, B \in \Im_t^\infty} |P(A)P(B) - P(AB)| \to 0 \text{ as } t \to \infty$$

where with $\Im_{t_1}^{t_2}$ we denote the σ-field generated by $\{(Y_t, \mathbf{X}_t) : t = t_1, \ldots, t_2\}$. Further we assume that for $\delta = \min\{\delta_1, \delta_2\} > 2$, with δ_i $i = 1, 2$ given in C1′

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

C4′ As $n \to \infty$, $h \to 0$ and $nh^d \to \infty$.

Again, we should point out that the conditions are not the weakest possible. Note that these conditions are a generalization of conditions C1-C4 to include the analysis of the mean function. Hence, all the remarks for C1-C4 still hold. Further, the rate of the bandwidth h remains the same and this is promising for our aim of adaptiveness.

2.7.1 Consistency of the joint estimator

The following proposition proves the consistency of the joint likelihood estimator.

Proposition 2.2 *Under conditions* C1′-C4′, *there exists at least one consistent solution of the likelihood equation. Equivalently, there is* $(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\gamma}}_n)$, *a solution of the likelihood equation, such that* $(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\gamma}}_n) \to (\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)$, *as* $n \to \infty$.

Proof of Proposition 2.2 Define $D_n^{(k)}(\boldsymbol{\theta},\boldsymbol{\gamma}) = n^{-1}H^{-1}S_n^{(k)}(\boldsymbol{\theta},\boldsymbol{\gamma})$, $k = 1,2$ then, we write

$$D_n(\boldsymbol{\theta},\boldsymbol{\gamma}) \equiv \frac{1}{n}G^{-1}\nabla l_n(\boldsymbol{\theta},\boldsymbol{\gamma}) = (D_n^{(1)}(\boldsymbol{\theta},\boldsymbol{\gamma}), \mathbf{0}_{d+1})^T + (\mathbf{0}_{d+1}, D_n^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma}))^T$$

where $\mathbf{0}_{d+1}$ is the $d+1$-dimensional zero vector. Call $D^{(k)}(\boldsymbol{\theta},\boldsymbol{\gamma};h) = \mathrm{E}(D_n^{(k)}(\boldsymbol{\theta},\boldsymbol{\gamma}))$, $k = 1,2$ and

By definition, $(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}})^T : \nabla l_n(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}}) = \mathbf{0} \Leftrightarrow D_n(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}}) = \mathbf{0}$. Moreover it is easy to see that equations (2.12) and (2.46) yield

By

$$D^{(k)}(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0;h) = \mathbf{0}, \quad k = 1,2 \Rightarrow D(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0;h) = \mathbf{0}. \tag{2.49}$$

that equations (2.12) and (2.46) yield

$$D^{(k)}(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0;h) = \mathbf{0}, \quad k = 1,2 \Rightarrow D(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0;h) = \mathbf{0}. \tag{2.49}$$

Now, suppose that none of the solutions of the likelihood equation converges in probability to $(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0)$. Then if $(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\gamma}}_n)$ is a solution then there exists a subsequence $(\hat{\boldsymbol{\theta}}_{kn},\hat{\boldsymbol{\gamma}}_{kn})$, such that for $\delta',\varepsilon > 0$,

$$P\Big(\parallel (\hat{\boldsymbol{\theta}}_{kn},\hat{\boldsymbol{\gamma}}_{kn}) - (\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0) \parallel > \delta'\Big) > \varepsilon \Leftrightarrow P\Big(\parallel (\hat{\boldsymbol{\theta}}_{kn},\hat{\boldsymbol{\gamma}}_{kn}) - (\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0) \parallel \leq \delta'\Big) < 1-\varepsilon$$

which implies that

$$P\Big(\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq\delta_1,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\|\leq\delta_2} \parallel D_n(\boldsymbol{\theta},\boldsymbol{\gamma}) \parallel > \eta\Big) > \epsilon$$

for a sufficiently large n and for some $\delta_i > 0$, $\eta > 0$. Equivalently

$$\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq\delta_1,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\|\leq\delta_2} \parallel D_n(\boldsymbol{\theta},\boldsymbol{\gamma}) \parallel \xrightarrow{P} 0. \tag{2.50}$$

Since

$$\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq\delta_1,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\|\leq\delta_2} \parallel D(\boldsymbol{\theta},\boldsymbol{\gamma};h) \parallel \geq$$

$$\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq\delta_1,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\|\leq\delta_2} \parallel D_n(\boldsymbol{\theta},\boldsymbol{\gamma}) \parallel - \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq\delta_1,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\|\leq\delta_2} \parallel D_n(\boldsymbol{\theta},\boldsymbol{\gamma}) - D(\boldsymbol{\theta},\boldsymbol{\gamma};h) \parallel$$

from (2.50) and Lemma 2.6 below, we have that

$$\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq\delta_1,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\|\leq\delta_2} \| D(\boldsymbol{\theta},\boldsymbol{\gamma};h) \| \xrightarrow{P} 0$$

which contradicts with (2.49) hence conclude.

Lemma 2.6 *Under conditions* C1'-C4' *we have that*

$$\sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\boldsymbol{\Theta}'\times\boldsymbol{\Gamma}'} \| D_n(\boldsymbol{\theta},\boldsymbol{\gamma}) - D(\boldsymbol{\theta},\boldsymbol{\gamma};h) \| \to 0$$

as $n \to \infty$ *where* $\boldsymbol{\Theta}' \times \boldsymbol{\Gamma}'$, *a compact set that contains* $(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)$.

Proof of Lemma 2.6 It holds that

$$\sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\boldsymbol{\Theta}'\times\boldsymbol{\Gamma}'} \| D_n(\boldsymbol{\theta},\boldsymbol{\gamma}) - D(\boldsymbol{\theta},\boldsymbol{\gamma};h) \| \leq$$

$$\sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\boldsymbol{\Theta}'\times\boldsymbol{\Gamma}'} \| D_n^{(1)}(\boldsymbol{\theta},\boldsymbol{\gamma}) - D^{(1)}(\boldsymbol{\theta},\boldsymbol{\gamma};h) \| + \sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\boldsymbol{\Theta}'\times\boldsymbol{\Gamma}'} \| D_n^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma}) - D^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma};h) \|.$$

The proof is similar to Lemma 2.1 and we only present the result for $D_n^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma})$. Let

$$D_n^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma}) = \frac{1}{n}\sum_{t=1}^{n} \frac{1}{e^{\mathbf{Z}_t^T\boldsymbol{\theta}}} \Psi(e_t(\boldsymbol{\theta},\boldsymbol{\gamma}))H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}) = \frac{1}{n}\sum_{t=1}^{n} V_t^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma})$$

then note that $V_t^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma})$ is a strictly stationary process with

$$\mathrm{E}||V_t^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma})|| = E\left(\mathrm{E}(|e^{-\mathbf{Z}_t^T\boldsymbol{\theta}}\Psi(e_t(\boldsymbol{\theta},\boldsymbol{\gamma}))|\ |\mathbf{X}_t)\ ||H^{-1}\mathbf{Z}_t||K_h(\mathbf{X}_t - \mathbf{x})\right).$$

Taylor expansion of $\Psi(e_t(\boldsymbol{\theta},\boldsymbol{\gamma}))$ around ϵ_t yields

$$\Psi(e_t(\boldsymbol{\theta},\boldsymbol{\gamma})) = \Psi(\epsilon_t) + \Omega_2(\epsilon_t)(e_t(\boldsymbol{\theta},\boldsymbol{\gamma}) - \epsilon_t) + O((e_t(\boldsymbol{\theta},\boldsymbol{\gamma}) - \epsilon_t)^2). \tag{2.51}$$

Note that

$$(e_t(\boldsymbol{\theta},\boldsymbol{\gamma}) - \epsilon_t) = (m(\mathbf{X}_t) - \mathbf{Z}_t^T\boldsymbol{\gamma})e^{-\mathbf{Z}_t^T\boldsymbol{\theta}} + \epsilon_t(s(\mathbf{X}_t) - \mathbf{Z}_t^T\boldsymbol{\theta}) =$$

from (2.50) and Lemma 2.6 below, we have that

$$\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq\delta_1,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\|\leq\delta_2} \| D(\boldsymbol{\theta},\boldsymbol{\gamma};h) \| \xrightarrow{P} 0$$

which contradicts with (2.49) hence conclude.

Lemma 2.6 *Under conditions* C1′-C4′ *we have that*

$$\sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\boldsymbol{\Theta}'\times\boldsymbol{\Gamma}'} \| D_n(\boldsymbol{\theta},\boldsymbol{\gamma}) - D(\boldsymbol{\theta},\boldsymbol{\gamma};h) \| \to 0$$

as $n \to \infty$ *where* $\boldsymbol{\Theta}' \times \boldsymbol{\Gamma}'$, *a compact set that contains* $(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)$.

Proof of Lemma 2.6 It holds that

$$\sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\boldsymbol{\Theta}'\times\boldsymbol{\Gamma}'} \| D_n(\boldsymbol{\theta},\boldsymbol{\gamma}) - D(\boldsymbol{\theta},\boldsymbol{\gamma};h) \| \leq$$

$$\sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\boldsymbol{\Theta}'\times\boldsymbol{\Gamma}'} \| D_n^{(1)}(\boldsymbol{\theta},\boldsymbol{\gamma}) - D^{(1)}(\boldsymbol{\theta},\boldsymbol{\gamma};h) \| + \sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\boldsymbol{\Theta}'\times\boldsymbol{\Gamma}'} \| D_n^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma}) - D^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma};h) \|.$$

The proof is similar to Lemma 2.1 and we only present the result for $D_n^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma})$. Let

$$D_n^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma}) = \frac{1}{n}\sum_{t=1}^{n} \frac{1}{e^{\mathbf{Z}_t^T\boldsymbol{\theta}}} \Psi(e_t(\boldsymbol{\theta},\boldsymbol{\gamma}))H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}) = \frac{1}{n}\sum_{t=1}^{n} V_t^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma})$$

then note that $V_t^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma})$ is a strictly stationary process with

$$\mathrm{E}||V_t^{(2)}(\boldsymbol{\theta},\boldsymbol{\gamma})|| = E\left(\mathrm{E}(|e^{-\mathbf{Z}_t^T\boldsymbol{\theta}}\Psi(e_t(\boldsymbol{\theta},\boldsymbol{\gamma}))|\ |\mathbf{X}_t)\ ||H^{-1}\mathbf{Z}_t||K_h(\mathbf{X}_t - \mathbf{x})\right).$$

Taylor expansion of $\Psi(e_t(\boldsymbol{\theta},\boldsymbol{\gamma}))$ around ϵ_t yields

$$\Psi(e_t(\boldsymbol{\theta},\boldsymbol{\gamma})) = \Psi(\epsilon_t) + \Omega_2(\epsilon_t)(e_t(\boldsymbol{\theta},\boldsymbol{\gamma}) - \epsilon_t) + O((e_t(\boldsymbol{\theta},\boldsymbol{\gamma}) - \epsilon_t)^2). \tag{2.51}$$

Note that

$$(e_t(\boldsymbol{\theta},\boldsymbol{\gamma}) - \epsilon_t) = (m(\mathbf{X}_t) - \mathbf{Z}_t^T\boldsymbol{\gamma})e^{-\mathbf{Z}_t^T\boldsymbol{\theta}} + \epsilon_t(s(\mathbf{X}_t) - \mathbf{Z}_t^T\boldsymbol{\theta}) =$$

2.7.2 Asymptotic normality of the joint estimator

We proceed to the asymptotic distribution of the joint estimator of the mean and variance function. The first order Taylor expansion of the derivative of the likelihood function around the true value $(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)$ yields,

$$(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0, \hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}^0) = \mathcal{H}_n^{-1}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*) S_n(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0) \tag{2.52}$$

where $S_n(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0) = (S_n^{(1)}(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0), S_n^{(2)}(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0))$ and $(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)$ lies between $(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\gamma}}_n)$ and $(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)$. We continue as we did in section 2.4.2 and prove some preliminary results. To avoid unnecessary repetition, we refer to some of the earlier results. For instance, recall that

$$\frac{1}{n} G^{-1} \Big(\mathcal{H}_n(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*) - \mathcal{H}_n(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0) \Big) G^{-1} \xrightarrow{P} 0 \tag{2.53}$$

as a result of the Slutsky's Theorem and the continuity of the matrix as a function of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ while details of the proof can be found in Lemma 2.2. Further, denote with $\otimes$ the kronecker product of two matrices. In the following lemma, we prove that $-\mathcal{H}_n(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)$ converges in probability to the information matrix.

Lemma 2.7 *Under conditions* C1′- C4′ *we have that*

$$-\frac{1}{n} G^{-1} \mathcal{H}_n(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0) G^{-1} \xrightarrow{P} \mathcal{I}_2 = p(\mathbf{x}) \left(\mathbf{V}^{-1}(\mathbf{x}) \, \mathbf{I}(f) \, \mathbf{V}^{-1}(\mathbf{x}) \right) \otimes \mathbf{S}_K$$

where

$$\mathbf{I}(f) = \begin{pmatrix} I_1(f) & I_2(f) \\ I_2(f) & I_3(f) \end{pmatrix} \quad \textit{and} \quad \mathbf{V}(\mathbf{x}) = \begin{pmatrix} 1 & 0 \\ 0 & \sigma(\mathbf{x}) \end{pmatrix}$$

with

$$I_1(f) = \int (\Psi(\epsilon)\epsilon + 1)^2 f(\epsilon) d\epsilon, \quad I_2(f) = \int (\Psi(\epsilon)\epsilon + 1)\Psi(\epsilon) f(\epsilon) d\epsilon, \quad I_3(f) = \int \Psi^2(\epsilon) f(\epsilon) d\epsilon$$

and $\mathbf{S}_K$ *defined in Lemma 2.3.*

Proof of Lemma 2.7 It is sufficient to prove that: $-n^{-1} H^{-1} \mathcal{H}_n^{(i)}(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0) H^{-1} \xrightarrow{P} p(\mathbf{x}) \sigma^{i-1} I_i(f) \, \mathbf{S}_K$ for $i = 1, 2, 3$. We present the result for $i = 2$ and the remaining

cases are proven in the same way. Note that

$$-\frac{1}{n}H^{-1}\mathcal{H}_n^{(2)}(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0)H^{-1} = -\frac{1}{n}\sum_{t=1}^{n}\frac{1}{e^{\mathbf{z}_t^T\boldsymbol{\theta}^0}}\Omega_1(e_t(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0))H^{-1}\mathbf{Z}_t\mathbf{Z}_t^TH^{-1}K_h(\mathbf{X}_t-\mathbf{x}).$$

Call $R_t = e^{-\mathbf{z}_t^T\boldsymbol{\theta}^0}\Omega_1(e_t(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0))H^{-1}\mathbf{Z}_t\mathbf{Z}_t^TH^{-1}K_h(\mathbf{X}_t-\mathbf{x})$ which is a strictly stationary process. Using Taylor expansion of $\Omega_1(e_t(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0))$ around ϵ_t we have that

$$\Omega_1(e_t(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0)) = \Omega_1(\epsilon_t) + R_1(\epsilon_t)(e_t(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0)-\epsilon_t) + o(e_t(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0)-\epsilon_t)$$

but

$$e_t(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0)-\epsilon_t = (m(\mathbf{X}_t)-\mathbf{Z}_t^T\boldsymbol{\gamma}^0)e^{-\mathbf{z}_t^T\boldsymbol{\theta}^0} + \epsilon_t(s(\mathbf{X}_t)-\mathbf{Z}_t^T\boldsymbol{\theta}^0) \tag{2.54}$$

consequently

$$e_t(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0)-\epsilon_t = e^{-\mathbf{z}_t^T\boldsymbol{\theta}^0}\frac{1}{2}\sum_{i,j=1}^{d}\ddot{m}_{i,j}(\mathbf{x}')(X_{t,i}-x_i)(X_{t,j}-x_j)$$

$$+\epsilon_t\frac{1}{2}\sum_{i,j=1}^{d}\ddot{s}_{i,j}(\mathbf{x}')(X_{t,i}-x_i)(X_{t,j}-x_j). \tag{2.55}$$

Therefore, for R_t it holds that

$$R_t = \Omega_1(\epsilon_t)e^{-\mathbf{z}_t^T\boldsymbol{\theta}^0}H^{-1}\mathbf{Z}_t\mathbf{Z}_t^TH^{-1}K_h(\mathbf{X}_t-\mathbf{x})$$

$$+R_1(\epsilon_t)\Big(e^{-2\mathbf{z}_t^T\boldsymbol{\theta}^0}\frac{1}{2}\sum_{i,j=1}^{d}\ddot{m}_{i,j}(\mathbf{x}')(X_{t,i}-x_i)(X_{t,j}-x_j)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^TH^{-1}K_h(\mathbf{X}_t-\mathbf{x})\Big)$$

$$+\epsilon_tR_1(\epsilon_t)\Big(e^{-\mathbf{z}_t^T\boldsymbol{\theta}^0}\frac{1}{2}\sum_{i,j=1}^{d}\ddot{s}_{i,j}(\mathbf{x}')(X_{t,i}-x_i)(X_{t,j}-x_j)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^TH^{-1}K_h(\mathbf{X}_t-\mathbf{x})\Big).$$

Based on this decomposition, from C1′: $\mathrm{E}|\Omega_1(\epsilon)|<\infty$, $\mathrm{E}|R_1(\epsilon)|<\infty$, $\mathrm{E}|R_1(\epsilon)\epsilon|<\infty$, C2′ and bounded 2nd derivatives of the mean and variance function it follows that $\mathrm{E}||R_t||<\infty$. Moreover,

$$\mathrm{E}(R_t) = \int\Omega_1(\epsilon_t)f(\epsilon_t)d\epsilon_t\int e^{-\mathbf{z}_t^T\boldsymbol{\theta}^0}H^{-1}\mathbf{z}_t\mathbf{z}_t^TH^{-1}K_h(\mathbf{x}_t-\mathbf{x})p(\mathbf{x}_t)d\mathbf{x}_t + O(h^2)$$

and substitute $\mathbf{x}_t - \mathbf{x} = h\mathbf{u}$

$$= \int \Omega_1(\epsilon_t) f(\epsilon_t) d\epsilon_t \int e^{-(1,h\mathbf{u}^T)\theta^0} \begin{pmatrix} 1 & \mathbf{u}^T \\ \mathbf{u} & \mathbf{u}\mathbf{u}^T \end{pmatrix} K(\mathbf{u}) p(\mathbf{x} + h\mathbf{u}) d\mathbf{u} + O(h^2)$$

$$= p(\mathbf{x}) e^{-s(\mathbf{x})} \int \Omega_1(\epsilon_t) f(\epsilon_t) d\epsilon_t \int \begin{pmatrix} 1 & \mathbf{u}^T \\ \mathbf{u} & \mathbf{u}\mathbf{u}^T \end{pmatrix} K(\mathbf{u}) d\mathbf{u} + o(1)$$

but note that

$$\int \begin{pmatrix} 1 & \mathbf{u}^T \\ \mathbf{u} & \mathbf{u}\mathbf{u}^T \end{pmatrix} K(\mathbf{u}) d\mathbf{u} = \begin{pmatrix} \mu_0 & 0 \\ 0 & \mu_2 \mathbf{I}_d \end{pmatrix} = \mathbf{S}_K$$

so using (2.48) we conclude that $\mathrm{E}(-R_t) = p(\mathbf{x})\sigma^{-1}(\mathbf{x}) I_2(f)\ \mathbf{S}_K + o(1)$. Direct application of the ergodic theorem for the process $-R_t$ entails the required result.

The following lemma involves the bias of the joint likelihood estimator. Recall $\mathcal{M}_{K,1}$ and $\mathbf{H}_s$ defined in (2.20) and let $\mathbf{H}_m$ similar to $\mathbf{H}_s$ but substitute $s(.)$ for $m(.)$. Further, define the matrix operator vec : $\mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^{nm}$ which takes a $n \times m$ matrix and by stacking the columns underneath each other from left to right, it returns a nm-dimensional vector. Then

Lemma 2.8 *Under conditions* C1′-C4′ *we have that*

$$\mathrm{E}(\frac{1}{n} G^{-1} S_n(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)) = -h^2 p(\mathbf{x}) G \mathcal{M}_K \mathrm{vec}[\mathbf{H}_{s,m}(\mathbf{V}^{-1}(\mathbf{x}) \mathbf{I}(f) \mathbf{V}^{-1}(\mathbf{x}))] + o(h^2)$$

where

$$\mathcal{M}_K = \begin{pmatrix} \mathcal{M}_{K,1} & 0 \\ 0 & \mathcal{M}_{K,1} \end{pmatrix} \text{ and } \mathbf{H}_{s,m} = (\mathbf{H}_s, \mathbf{H}_m)\,.$$

Proof of Lemma 2.8 Note that $\mathrm{E}(n^{-1} G^{-1} S_n(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)) = (\mathrm{E}(n^{-1} H^{-1} S_n^{(1)}(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)$, $\mathrm{E}(n^{-1} H^{-1} S_n^{(2)}(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)))^T$ and from stationarity,

$$\mathrm{E}(\frac{1}{n} S_{n,0}^{(k)}(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)) = \int \int e^{-(k-1)\mathbf{z}_t^T \boldsymbol{\theta}^0} G_k(e_t(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)) K_h(\mathbf{x}_t - \mathbf{x}) f(y_t | \mathbf{x}_t) p(\mathbf{x}_t) dy_t d\mathbf{x}_t \tag{2.56}$$

$$\mathrm{E}((nh)^{-1}S_{n,r}^{(k)}(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0) =$$

$$= \int\int e^{-(k-1)\mathbf{z}_t^T\boldsymbol{\theta}^0} G_k(e_t(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0))\Big(\frac{x_{t,r}-x_r}{h}\Big)K_h(\mathbf{x}_t-\mathbf{x})f(y_t|\mathbf{x}_t)p(\mathbf{x}_t)dy_t d\mathbf{x}_t \quad (2.57)$$

for $k=1,2$. Focus on the case of $k=1$. First order Taylor expansion of $G_1(e_t(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0))$ around ϵ_t yields $G_1(e_t(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0)) = G_1(\epsilon_t) + \Omega_1(\epsilon_t)(e_t(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0)-\epsilon_t) + o(e_t(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0)-\epsilon_t)$. Using expansion (2.55), rearranged to include the remainder of the expansion, it follows that

$$G_1(e_t(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0)) = G_1(\epsilon_t) + \Omega_1(\epsilon_t)\Big(e^{-s(\mathbf{x}_t)}\frac{1}{2}\sum_{i,j=1}^{d}\ddot{m}_{ij}(\mathbf{x})(X_{t,i}-x_i)(X_{t,j}-x_j)$$

$$+\epsilon_t\frac{1}{2}\sum_{i,j=1}^{d}\ddot{s}_{ij}(\mathbf{x})(X_{t,i}-x_i)(X_{t,j}-x_j)\Big) + o((X_{t,i}-x_i)(X_{t,j}-x_j)). \quad (2.58)$$

Substitute (2.58) in (2.56) to obtain

$$\mathrm{E}(S_{n,0}^{(1)}(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0)) = \int G_1(\epsilon_t)f(\epsilon_t)d\epsilon_t\int K_h(\mathbf{x}_t-\mathbf{x})p(\mathbf{x}_t)d\mathbf{x}_t +$$

$$+\int\Omega_1(\epsilon_t)f(\epsilon_t)d\epsilon_t\int e^{-s(\mathbf{x}_t)}\frac{1}{2}\sum_{i,j=1}^{d}\ddot{m}_{ij}(\mathbf{x})(X_{t,i}-x_i)(X_{t,j}-x_j)K_h(\mathbf{x}_t-\mathbf{x})p(\mathbf{x}_t)d\mathbf{x}_t$$

$$+\int\Omega_1(\epsilon_t)\epsilon_t f(\epsilon_t)d\epsilon_t\int\frac{1}{2}\sum_{i,j=1}^{d}\ddot{s}_{ij}(\mathbf{x})(X_{t,i}-x_i)(X_{t,j}-x_j)K_h(\mathbf{x}_t-\mathbf{x})p(\mathbf{x}_t)d\mathbf{x}_t + o(h^2).$$

Bartlett identity in (2.12) yields that the first integral is zero. For the second integral, applying the transformation $\mathbf{x}_t-\mathbf{x}=h\mathbf{u}$ yields

$$\int\Omega_1(\epsilon_t)f(\epsilon_t)d\epsilon_t\int e^{-s(\mathbf{x}+h\mathbf{u})}\frac{1}{2}\sum_{i,j=1}^{d}\ddot{m}_{ij}(\mathbf{x})h^2u_iu_jK(\mathbf{u})p(\mathbf{x}+h\mathbf{u})d\mathbf{u}$$

but $\mu_{i,j}=0$ for $i\neq j$ and using Bartlett's identity in (2.48) we conclude

$$= -\frac{h^2}{2}\mu_2 p(\mathbf{x})e^{-s(\mathbf{x})}I_2(f)\sum_{j=1}^{d}\ddot{m}_{jj}(\mathbf{x}) + o(h^2).$$

Similarly, the third integral is equal to

$$-\frac{h^2}{2}\mu_2 p(\mathbf{x}) I_1(f) \sum_{j=1}^{d} \ddot{s}_{jj}(\mathbf{x}) + o(h^2).$$

Summing up, it follows that

$$\mathrm{E}(\frac{1}{n} S_{n,0}^{(1)}(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)) = -\frac{h^2}{2}\mu_2 p(\mathbf{x}) \Big(\sigma^{-1}(\mathbf{x}) I_2(f) \sum_{j=1}^{d} \ddot{m}_{jj}(\mathbf{x}) + I_1(f) \sum_{j=1}^{d} \ddot{s}_{jj}(\mathbf{x}) \Big) + o(h^2)$$

and note that using the same arguments, we obtain $\mathrm{E}(n^{-1} S_{n,0}^{(2)}(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)) =$

$$= -\frac{h^2}{2}\mu_2 p(\mathbf{x}) \Big(\sigma^{-2}(\mathbf{x}) I_3(f) \sum_{j=1}^{d} \ddot{m}_{jj}(\mathbf{x}) + \sigma^{-1}(\mathbf{x}) I_2(f) \sum_{j=1}^{d} \ddot{s}_{jj}(\mathbf{x}) \Big) + o(h^2).$$

Further, for $\mathrm{E}((nh)^{-1} S_{n,r}^{(1)}(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0))$, $r = 1, \ldots, d$ we argued that a third order Taylor expansion of $m(.)$ and $s(.)$ is required, see discussion in Lemma 2.4. Hence, we extend the expansion in (2.58) to include the third order terms:

$$G_1(e_t(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)) = G_1(\epsilon_t) + \Omega_1(\epsilon_t) \Big(e^{-s(\mathbf{x}_t)} \frac{1}{2} \sum_{i,j=1}^{d} \ddot{m}_{ij}(\mathbf{x})(X_{t,i} - x_i)(X_{t,j} - x_j)$$

$$+ \frac{\epsilon_t}{2} \sum_{i,j=1}^{d} \ddot{s}_{ij}(\mathbf{x})(X_{t,i} - x_i)(X_{t,j} - x_j) + e^{-s(\mathbf{x}_t)} \frac{1}{6} \sum_{i,j,k=1}^{d} \dddot{m}_{ijk}(\mathbf{x})(X_{t,i} - x_i)(X_{t,j} - x_j)(X_{t,k} - x_k)$$

$$+ \frac{\epsilon_t}{6} \sum_{i,j,k=1}^{d} \dddot{s}_{ijk}(\mathbf{x})(X_{t,i} - x_i)(X_{t,j} - x_j)(X_{t,k} - x_k) \Big) + o((X_{t,i} - x_i)(X_{t,j} - x_j)(X_{t,k} - x_k)) \tag{2.59}$$

then for $r = 1, \ldots, d$

$$\mathrm{E}(\frac{1}{nh} S_{n,r}^{(1)}(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)) = \int G_1(\epsilon_t) f(\epsilon_t) d\epsilon_t \int \frac{x_{t,r} - x_r}{h} K_h(\mathbf{x}_t - \mathbf{x}) p(\mathbf{x}_t) d\mathbf{x}_t +$$

$$+ \int \Omega_1(\epsilon_t) f(\epsilon_t) d\epsilon_t \int e^{-s(\mathbf{x}_t)} \frac{x_{t,r} - x_r}{h} K_h(\mathbf{x}_t - \mathbf{x}) \Big(\frac{1}{2} \sum_{i,j=1}^{d} \ddot{m}_{ij}(\mathbf{x})(x_{t,i} - x_i)(x_{t,j} - x_j)$$

$$+\frac{1}{6}\sum_{i,j,k=1}^{d}\dddot{m}_{ijk}(\mathbf{x})(x_{t,i}-x_i)(x_{t,j}-x_j)(x_{t,k}-x_k)\Big)p(\mathbf{x}_t)d\mathbf{x}_t+o(h^3)$$

$$+\int\Omega_1(\epsilon_t)\epsilon_t f(\epsilon_t)d\epsilon_t\int\frac{x_{t,r}-x_r}{h}K_h(\mathbf{x}_t-\mathbf{x})+\Big(\frac{1}{2}\sum_{i,j=1}^{d}\ddot{s}_{ij}(\mathbf{x})(x_{t,i}-x_i)(x_{t,j}-x_j)$$

$$+\frac{1}{6}\sum_{i,j,k=1}^{d}\dddot{s}_{ijk}(\mathbf{x})(x_{t,i}-x_i)(x_{t,j}-x_j)(x_{t,k}-x_k)\Big)p(\mathbf{x}_t)d\mathbf{x}_t+o(h^3).$$

From Bartlett identity, the first integral is zero while for the second integral, the transformation $\mathbf{x}_t-\mathbf{x}=h\mathbf{u}$ yields

$$-I_2(f)p(\mathbf{x})e^{-s(\mathbf{x})}\int u_rK(\mathbf{u})\Big(\frac{1}{2}h^2\sum_{i,j=1}^{d}\ddot{m}_{ij}(\mathbf{x})u_iu_j+\frac{1}{6}h^3\sum_{i,j,k=1}^{d}\dddot{m}_{ijk}(\mathbf{x})u_iu_ju_k\Big)d\mathbf{u}+o(h^3)$$

$$=-I_2(f)p(\mathbf{x})e^{-s(\mathbf{x})}\Big(\frac{1}{2}h^2\sum_{i,j=1}^{d}\ddot{m}_{ij}(\mathbf{x})\int u_ru_iu_jK(\mathbf{u})d\mathbf{u}$$

$$+\frac{1}{6}h^3\sum_{i,j,k=1}^{d}\dddot{m}_{ijk}(\mathbf{x})\int u_ru_ju_lu_iK(\mathbf{u})d\mathbf{u}\Big)+o(h^3)$$

but note that symmetry of kernel implies that $\int u_ru_ju_iK(\mathbf{u})d\mathbf{u}=0$ while

$$\int u_ru_iu_ju_kK(\mathbf{u})d\mathbf{u}=\begin{cases}\mu_2^2 & i=r,j=k\\ 0 & \text{otherwise}\end{cases}$$

hence the second integral is equal to

$$-\frac{h^3}{6}\mu_2^2I_2(f)\sigma^{-1}(\mathbf{x})p(\mathbf{x})\sum_{j=1}^{d}\dddot{m}_{rjj}(\mathbf{x})+o(h^3)$$

and following the same procedure, we prove that the third integral is

$$-\frac{h^3}{6}\mu_2^2I_1(f)p(\mathbf{x})\sum_{j=1}^{d}\dddot{s}_{rjj}(\mathbf{x})+o(h^3).$$

Summing up, for $r = 1, \ldots, d$

$$\mathrm{E}(\frac{1}{nh} S_{n,r}^{(1)}(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)) = -\frac{h^3}{6}\mu_2^2 p(\mathbf{x})\Big(\sigma^{-1}(\mathbf{x}) I_2(f) \sum_{j=1}^{d} \dddot{m}_{rjj}(\mathbf{x}) + I_1(f) \sum_{j=1}^{d} \dddot{s}_{rjj}(\mathbf{x})\Big) + o(h^3)$$

and in the same way it can be shown that $\mathrm{E}((nh)^{-1} S_{n,r}^{(2)}(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)) =$

$$-\frac{h^3}{6}\mu_2^2 p(\mathbf{x})\Big(\sigma^{-2}(\mathbf{x}) I_3(f) \sum_{j=1}^{d} \dddot{m}_{rjj}(\mathbf{x}) + \sigma^{-1}(\mathbf{x}) I_2(f) \sum_{j=1}^{d} \dddot{s}_{rjj}(\mathbf{x})\Big) + o(h^3).$$

Writing up the results using matrix notation, we conclude.
Define the processes $U_t^{(1)} = (V_t^{(1)} - \mathrm{E}(V_t^{(1)}), \mathbf{0}_{d+1})$, $U_t^{(2)} = (\mathbf{0}_{d+1}, V_t^{(2)} - \mathrm{E}(V_t^{(2)}))^T$ $Q_n = n^{-1} \sum_{t=1}^{n} \{U_t^{(1)} + U_t^{(2)}\}$ where $V_t^{(k)} = e^{-(k-1)\mathbf{z}_t^T \boldsymbol{\theta}^0} G_k(e_t(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)) H^{-1} \mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x})$.

Lemma 2.9 *Under conditions* C1′-C4′ *it follows that*

(a) $h^d \mathrm{Var}(U_t^{(1)} + U_t^{(2)}) \to p(\mathbf{x})(\mathbf{V}^{-1}(\mathbf{x})\mathbf{I}(f)\mathbf{V}^{-1}(\mathbf{x})) \otimes \mathbf{S}_K^*$.

(b) $h^d \sum_{t=1}^{n} ||\mathrm{Cov}(U_1^{(1)} + U_1^{(2)}, U_t^{(1)} + U_t^{(2)})|| = o(1)$.

(c) $nh^d \mathrm{Var}(Q_n) \to p(\mathbf{x})(\mathbf{V}^{-1}(\mathbf{x})\mathbf{I}(f)\mathbf{V}^{-1}(\mathbf{x})) \otimes \mathbf{S}_K^*$.

Proof of Lemma 2.9 (a). Note that $\mathrm{Var}(U_t^{(1)} + U_t^{(2)}) = \mathrm{Var}(U_t^{(1)}) + \mathrm{Var}(U_t^{(2)}) + \mathrm{E}(U_t^{(1)} U_t^{(2)^T}) + \mathrm{E}(U_t^{(2)} U_t^{(1)^T})$. It holds that

$$\mathrm{Var}(U_t^{(1)}) = \mathrm{E}(U_t^{(1)} U_t^{(1)^T}) = \begin{pmatrix} \mathrm{E}(V_t^{(1)} V_t^{(1)^T}) - \mathrm{E}(V_t^{(1)})\mathrm{E}(V_t^{(1)})^T & 0 \\ 0 & 0 \end{pmatrix}$$

with $\mathrm{E}(V_t^{(1)} V_t^{(1)^T}) = \mathrm{E}\left(G_1^2(e_t(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)) H^{-1} \mathbf{Z}_t \mathbf{Z}_t^T H^{-1} K_h^2(\mathbf{X}_t - \mathbf{x})\right)$. In the expansion:

$$G_1^2(e_t(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)) = G_1^2(\epsilon_t) + 2G_1(\epsilon_t)\Omega_1(\epsilon_t)(e_t(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0) - \epsilon_t)$$

substitute (2.55), then from $\mathrm{E}|G_1(\epsilon)|^2 < \infty$, $\mathrm{E}|G_1(\epsilon)\Omega_1(\epsilon)\epsilon| < \infty$, (Cauchy-Schwartz inequality and C1′) and C2′, C4′, we have that

$$h^d \mathrm{E}(V_t^{(1)} V_t^{(1)^T}) = p(\mathbf{x}) \int G_1^2(\epsilon_t) f(\epsilon_t) d\epsilon_t \int \begin{pmatrix} 1 & \mathbf{u}^T \\ \mathbf{u} & \mathbf{u}\mathbf{u}^T \end{pmatrix} K^2(\mathbf{u}) d\mathbf{u} + o(1)$$

equivalently $h^d \mathrm{E}(V_t^{(1)} V_t^{(1)^T}) \to -p(\mathbf{x}) I_1(f) \mathbf{S}_K^*$. Further, from Lemma 2.8 $\mathrm{E}(V_t^{(1)}) = O(h^2) = o(1)$, so we conclude that

$$h^d \mathrm{Var}(U_t^{(1)}) = -p(\mathbf{x}) I_1(f) \begin{pmatrix} \mathbf{S}_K^* & 0 \\ 0 & 0 \end{pmatrix} + o(1). \tag{2.60}$$

Similarly, it can be shown that

$$h^d \mathrm{Var}(U_t^{(2)}) = -p(\mathbf{x}) \sigma^{-2}(\mathbf{x}) I_3(f) \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{S}_K^* \end{pmatrix} + o(1) \tag{2.61}$$

$$h^d \mathrm{E}(U_t^{(1)} U_t^{(2)^T}) = h^d \mathrm{E}(U_t^{(2)} U_t^{(1)^T})^T = -p(\mathbf{x}) \sigma^{-1}(\mathbf{x}) I_2(f) \begin{pmatrix} 0 & \mathbf{S}_K^* \\ 0 & 0 \end{pmatrix} + o(1). \tag{2.62}$$

Combining (2.60), (2.61) and (2.62), we write

$$h^d \mathrm{Var}(U_t^{(1)} + U_t^{(2)}) \to p(\mathbf{x}) (\mathbf{V}^{-1}(\mathbf{x}) \mathbf{I}(f) \mathbf{V}^{-1}(\mathbf{x})) \otimes \mathbf{S}_K^*.$$

It follows from stationarity that (a) and (b) along with $nh^d \to \infty$ from C4′ and the dominance convergence theorem on

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

are sufficient for (c) to hold. To prove (b) let $d_n \to \infty$ be a sequence of positive integers such that $d_n h^d \to 0$. Define,

$$J_1 = \sum_{t=1}^{d_n - 1} ||\mathrm{Cov}(U_1^{(1)} + U_1^{(2)}, U_{t+1}^{(1)} + U_{t+1}^{(2)})||, \quad J_2 = \sum_{t=d_n}^{n-1} ||\mathrm{Cov}(U_1^{(1)} + U_1^{(2)}, U_{t+1}^{(1)} + U_{t+1}^{(2)})||.$$

It is sufficient to show that $J_k = o(h^{-d})$, $k = 1, 2$. Note that

$$||\mathrm{Cov}(U_1^{(1)} + U_1^{(2)}, U_{t+1}^{(1)} + U_{t+1}^{(2)})|| \leq$$

$$||\mathrm{E}(U_1^{(1)} U_{t+1}^{(1)^T})|| + ||\mathrm{E}(U_1^{(2)} U_{t+1}^{(1)^T})|| + ||\mathrm{E}(U_1^{(1)} U_{t+1}^{(2)^T})|| + ||\mathrm{E}(U_1^{(2)} U_{t+1}^{(2)^T})||.$$

By definition, $||\mathrm{E}(U_1^{(1)}{U_{t+1}^{(1)}}^T)|| = ||\mathrm{E}(V_1^{(1)}{V_{t+1}^{(1)}}^T) - \mathrm{E}(V_1^{(1)})\mathrm{E}(V_{t+1}^{(1)})^T||$ and

$$||\mathrm{E}(V_1^{(1)}{V_{t+1}^{(1)}}^T)|| \leq \mathrm{E}|G_1(\epsilon_1)G_1(\epsilon_{t+1})|\mathrm{E}\Big(||H^{-1}\mathbf{Z}_1\mathbf{Z}_{t+1}^T||K_h(\mathbf{X}_{t+1}-\mathbf{x})K_h(\mathbf{X}_1-\mathbf{x})\Big)+O(h^2)$$

the latter using expansion (2.58). Then, from C1′-C2′ we have that $||\mathrm{E}(V_1^{(1)}{V_{t+1}^{(1)}}^T)|| < \infty$ and since $\mathrm{E}(V_t^{(1)}) = o(1)$ for all t, we conclude that $||\mathrm{E}(U_1^{(1)}{U_{t+1}^{(1)}}^T)|| = O(1)$. Similar arguments for the remaining terms of the inequality ensure that for all $t \geq 1$

$$||\mathrm{Cov}(U_1^{(1)} + U_1^{(2)}, U_{t+1}^{(1)} + U_{t+1}^{(2)})|| = O(1).$$

Therefore, $h^d J_1 = O(h^d d_n) = o(1)$, by definition of d_n. Next we consider the upper bound of J_2. By using Davydov's inequality, see Bosq (1998) Corollary 1.1, for $\delta = \min\{\delta_1, \delta_2\} > 2$ where δ_1 and δ_2 are given in C1′(ii) we have that for all $t \geq 1$

$$||\mathrm{Cov}(U_1^{(1)}+U_1^{(2)}, U_{t+1}^{(1)}+U_{t+1}^{(2)})|| \leq C\{\alpha(t)\}^{1-\frac{2}{\delta}}\Big(\mathrm{E}||U_1^{(1)}+U_1^{(2)}||^\delta\Big)^{\frac{1}{\delta}}\Big(\mathrm{E}||U_{t+1}^{(1)}+U_{t+1}^{(2)}||^\delta\Big)^{\frac{1}{\delta}} \tag{2.63}$$

where $\alpha(t)$ is the mixing coefficient of the process $(Y_t, \mathbf{X}_t)$ given in C3′. Under conditions C1′-C2′, for all $t \geq 1$ and $k = 1, 2$ it holds that

$$\mathrm{E}||V_{t+1}^{(k)}||^\delta \leq C\mathrm{E}\Big(|G_k(e_t(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0))|^\delta e^{-(k-1)s(\mathbf{X}_t)}||H^{-1}\mathbf{Z}_{t+1}||^\delta K_h^\delta(\mathbf{X}_{t+1} - \mathbf{x})\Big) \leq$$

$$\leq C\mathrm{E}|G_k(\epsilon_t)|^\delta E\Big(||H^{-1}\mathbf{Z}_{t+1}||^\delta K_h^\delta(\mathbf{X}_{t+1} - \mathbf{x})\Big) + O(h^2)$$

the latter from Taylor expansion of $G_k^\delta(e_t(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0))$ around ϵ_t and condition C1′: $\mathrm{E}|G_k(\epsilon_t)|^{2\delta-2} < \infty$, $\mathrm{E}|\Omega_k(\epsilon_t)\epsilon_t| < \infty$ and $\sigma^2(\mathbf{x}) > 0$. Hence, for $k = 1, 2$

$$\mathrm{E}||U_{t+1}^{(k)}||^\delta \leq Ch^{(1-\delta)d}\int |G_k(\epsilon_t)|^\delta f(\epsilon_t)d\epsilon_t \int ||(1, \mathbf{u}^T)||^\delta K^\delta(\mathbf{u})d\mathbf{u}.$$

Since $\mathrm{E}|G_k(\epsilon_t)|^\delta < \infty$ and from C2′, it follows that $\mathrm{E}||U_{t+1}^{(k)}|| = O(h^{(1-\delta)d})$. Application of the Minkowski inequality $\Big(\mathrm{E}||U_{t+1}^{(1)} + U_{t+1}^{(2)}||^\delta\Big)^{\frac{1}{\delta}} \leq \Big(\mathrm{E}||U_{t+1}^{(1)}||^\delta\Big)^{\frac{1}{\delta}} + \Big(\mathrm{E}||U_{t+1}^{(2)}||^\delta\Big)^{\frac{1}{\delta}}$ yields

$$\Big(\mathrm{E}(||U_{t+1}^{(1)} + U_{t+1}^{(2)}||^\delta)\Big)^{\frac{1}{\delta}} \leq Ch^{(1/\delta-1)d}. \tag{2.64}$$

Substituting (2.64) into (2.63), we conclude

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

and by choosing $d_n = Ch^{(2/\delta-1)d/2}$, then $d_n \to \infty$, $h^d d_n \to 0$ and $J_2 = o(h^{-d})$ and the proof is complete.

We now proceed to the main theorem for the asymptotic distribution of the joint mean and variance ML-estimator.

Theorem 2.2 *Suppose conditions* C1′-C4′ *hold. Then, it follows that*

$$\sqrt{nh^d} G\Big((\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0, \hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}^0)^T - (\mathbf{b}_1, \mathbf{b}_2)^T\Big) \xrightarrow{d} N(0, \mathcal{I}_2^{-1} \boldsymbol{\Sigma}_2 \mathcal{I}_2^{-1})$$

where

$$\mathcal{I}_2^{-1} \boldsymbol{\Sigma}_2 \mathcal{I}_2^{-1} = p^{-1}(\mathbf{x})(\mathbf{V}(\mathbf{x})\mathbf{I}^{-1}(f)\mathbf{V}(\mathbf{x})) \otimes (\mathbf{S}_K^{-1} \mathbf{S}_K^* \mathbf{S}_K^{-1})$$

and

$$(\mathbf{b}_1, \mathbf{b}_2)^T = (h^2 \mathbf{S}_K^{-1} \mathcal{M}_{K,1} \mathbf{H}_s, h^2 \mathbf{S}_K^{-1} \mathcal{M}_{K,1} \mathbf{H}_m)^T + o(h^2).$$

Proof of Theorem 2.2 Recall expansion (2.52): $(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0, \hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}^0) =$

$$= \mathcal{H}_n^{-1}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)\{S_n(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0) - \mathrm{E}(S_n(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0))\} + \mathcal{H}_n^{-1}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)\mathrm{E}(S_n(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)) \Rightarrow$$

$$(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0, \hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}^0) = L_1 + L_2$$

with

$$L_1 = G^{-1} \left(G^{-1} \mathcal{H}_n(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*) G^{-1}\right)^{-1} G^{-1} \{S_n(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0) - \mathrm{E}(S_n(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0))\}$$

$$L_2 = G^{-1} \left(-G^{-1} \mathcal{H}_n(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*) G^{-1}\right)^{-1} \mathrm{E}(-G^{-1} S_n(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)).$$

Then the results of the theorem come directly from statements:

(a) $\sqrt{nh^d} G L_1 \xrightarrow{d} N(0, \mathcal{I}_2^{-1} \boldsymbol{\Sigma}_2 \mathcal{I}_2^{-1})$,

(b) $L_2 \equiv (\mathbf{b}_1, \mathbf{b}_2)^T = (h^2 \mathbf{S}_K^{-1} \mathcal{M}_{K,1} \mathbf{H}_s, h^2 \mathbf{S}_K^{-1} \mathcal{M}_{K,1} \mathbf{H}_m)^T + o(h^2)$.

Focus on (a) and note that $G^{-1}(S_n(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0)-\mathrm{E}(S_n(\boldsymbol{\theta}^0,\boldsymbol{\gamma}^0))) = \sum_{t=1}^n (U_t^{(1)}+UJ_t^{(2)})$ hence

$$\sqrt{nh^d}GL_1 = \left(\frac{1}{n}G^{-1}\mathcal{H}_n(\boldsymbol{\theta}^*,\boldsymbol{\gamma}^*)G^{-1}\right)^{-1}\frac{1}{\sqrt{nh^d}}\sum_{t=1}^n h^d(U_t^{(1)}+U_t^{(2)}). \tag{2.65}$$

The process $U_t^{(1)}+U_t^{(2)}$ is a zero mean, strictly stationary process. Further,, in (2.64), Lemma 2.9, we showed that $\mathrm{E}||h^d(U_{t+1}^{(1)}+U_{t+1}^{(2)})||^\delta \leq Ch^{(1-\delta)d}h^{d\delta} = Ch^d$ for $\delta = \min\{\delta_1,\delta_2\} > 2$. If we denote with $\tilde{\alpha}(j)$ the mixing coefficients of the process $U_t^{(1)}+U_t^{(2)}$ then from the properties of strong mixing conditions it holds that $\tilde{\alpha}(j) < \alpha(j)$ where $\alpha(j)$ is the mixing coefficient of $(Y_t,\mathbf{X}_t)^T$. Thus for the fixed $\delta > 2$, condition C3′ yields $\sum_{j=1}^\infty \tilde{\alpha}(j)^{1-\frac{2}{\delta}} < \sum_{j=1}^\infty \alpha(j)^{1-\frac{2}{\delta}} < \infty$. Consequently, using the Central Limit Theorem 2.21 (Fan and Yao 2003), we have that $1/\sqrt{nh^d}\sum_{t=1}^n h^d(U_t^{(1)}+U_t^{(2)})$ follows asymptotically normal distribution with mean zero and variance $\boldsymbol{\Sigma}_{2,n}$ where

$$\boldsymbol{\Sigma}_{2,n} = \mathrm{Var}\Big(\frac{1}{\sqrt{nh^d}}\sum_{t=1}^n h^d(U_t^{(1)}+U_t^{(2)})\Big) = nh^d\mathrm{Var}(Q_n).$$

Lemma 2.9 (c) yields $\boldsymbol{\Sigma}_{2,n} \to \boldsymbol{\Sigma}_2 = p(\mathbf{x})(\mathbf{V}^{-1}(\mathbf{x})\mathbf{I}(f)\mathbf{V}^{-1}(\mathbf{x}))\otimes\mathbf{S}_K^*$. Hence, it holds that

$$\frac{1}{\sqrt{nh^d}}\sum_{t=1}^n h^d(U_t^{(1)}+U_t^{(2)}) \xrightarrow{d} N(0,\boldsymbol{\Sigma}_2) \tag{2.66}$$

and using Lemma 2.7, we conclude

$$\sqrt{nh^d}GL_1 \xrightarrow{d} N(0,\mathcal{I}_2^{-1}\boldsymbol{\Sigma}_2\mathcal{I}_2^{-1})$$

where

$$\mathcal{I}_2^{-1}\boldsymbol{\Sigma}_2\mathcal{I}_2^{-1} = p^{-1}(\mathbf{x})\Big((\mathbf{V}(\mathbf{x})\,\mathbf{I}^{-1}(f)\,\mathbf{V}(\mathbf{x}))\otimes\mathbf{S}_K^{-1}\Big)$$
$$\Big((\mathbf{V}^{-1}(\mathbf{x})\,\mathbf{I}(f)\,\mathbf{V}^{-1}(\mathbf{x}))\otimes\mathbf{S}_K^*\Big)\Big((\mathbf{V}(\mathbf{x})\,\mathbf{I}^{-1}(f)\,\mathbf{V}(\mathbf{x}))\otimes\mathbf{S}_K^{-1}\Big).$$

From Kronecker product properties[2], it follows that

$$\mathcal{I}_2^{-1}\boldsymbol{\Sigma}_2\mathcal{I}_2^{-1} = p^{-1}(\mathbf{x})\Big(\mathbf{V}(\mathbf{x})\,\mathbf{I}^{-1}(f)\,\mathbf{V}(\mathbf{x})\Big)\otimes(\mathbf{S}_K^{-1}\,\mathbf{S}_K^*\,\mathbf{S}_K^{-1}).$$

[2] $(A\otimes B)(C\otimes D) = AC\otimes BD$

To prove (b) note that $L_2 = G^{-1}(\mathcal{I}_2^{-1} + o_p(1))\mathrm{E}(-n^{-1}G^{-1}S_n(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0))$ and Lemma 2.7 and 2.8 imply

$$L_2 = h^2\mathcal{M}_K\Big((\mathbf{V}(\mathbf{x})\ \mathbf{I}^{-1}(f)\ \mathbf{V}(\mathbf{x}))\otimes\mathbf{S}_K^{-1}\Big)\mathrm{vec}[\mathbf{H}_{s,m}(\mathbf{V}^{-1}(\mathbf{x})\mathbf{I}(f)\mathbf{V}^{-1}(\mathbf{x}))] + o(h^2).$$

For the vec-operator and the kronecker product it holds that

$$\mathrm{vec}[\mathbf{H}_{s,m}(\mathbf{V}^{-1}(\mathbf{x})\mathbf{I}(f)\mathbf{V}^{-1}(\mathbf{x}))] = \Big((\mathbf{V}^{-1}(\mathbf{x})\mathbf{I}(f)\mathbf{V}^{-1}(\mathbf{x}))\otimes\mathbf{I}_2\Big)\mathrm{vec}[\mathbf{H}_{s,m}]$$

where $\mathbf{I}_2$ is the 2×2 unit matrix, thus

$$L_2 = h^2\mathcal{M}_K\Big((\mathbf{V}(\mathbf{x})\ \mathbf{I}^{-1}(f)\ \mathbf{V}(\mathbf{x}))\otimes\mathbf{S}_K^{-1}\Big)\Big((\mathbf{V}^{-1}(\mathbf{x})\mathbf{I}(f)\mathbf{V}^{-1}(\mathbf{x}))\otimes\mathbf{I}_2\Big)\mathrm{vec}[\mathbf{H}_{s,m}]+o(h^2)$$

and properties of kronecker product yield $L_2 = h^2\mathcal{M}_K(\mathbf{I}_2\otimes\mathbf{S}_K^{-1})\mathrm{vec}[\mathbf{H}_{s,m}] + o(h^2)$ which after some algebraic calculations we write as

$$L_2 \equiv (\mathbf{b}_1, \mathbf{b}_2)^T = (h^2\mathbf{S}_K^{-1}\mathcal{M}_{K,1}\mathbf{H}_s, h^2\mathbf{S}_K^{-1}\mathcal{M}_{K,1}\mathbf{H}_m)^T + o(h^2)$$

and the proof is complete.

Based on the results regarding the asymptotic distribution of the ML-estimator in Theorems 2.1 and 2.2, we identify sufficient conditions for adaptiveness in respect to the mean function. 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. This implies that the AMSE of the variance estimator when the mean function is unknown is equal to the AMSE of the variance estimator derived under the assumption of known mean function. Note here that the bias term in Theorem 2.2 for the variance estimator is equal to the bias term found in Theorem 2.1. However, the block of the asymptotic variance that corresponds to the variance estimator is not exactly the same as the asymptotic variance from Theorem 2.1. In the following proposition, we identify a sufficient condition for mean regression adaptiveness of the variance function estimator. Particularly, it holds that

Proposition 2.3 *The local linear Maximum Likelihood estimator of the variance function is asymptotically adaptive regarding the mean function, if the information matrix* $\mathbf{I}(f)$ *is diagonal. Equivalently, for the information of the joint variance and mean function estimator, it holds that* $I_2(f) = 0$. *A sufficient condition would be that the error density function is symmetric.*

Proof of Proposition 2.3 We concentrate on the variance term of the asymptotic distribution given that the bias term remained unchanged. Assume that the information matrix of the error density $\mathbf{I}(f)$ is diagonal i.e. $I_2(f) = 0$. Then for the asymptotic variance $\mathcal{I}_2^{-1}\Sigma_2\mathcal{I}_2^{-1} = p^{-1}(\mathbf{x})\Big(\mathbf{V}(\mathbf{x})\mathbf{I}^{-1}(f)\mathbf{V}(\mathbf{x})\Big) \otimes (\mathbf{S}_K^{-1}\mathbf{S}_K^*\mathbf{S}_K^{-1})$, found in Theorem 2.2, follows that

$$\mathcal{I}_2^{-1}\Sigma_2\mathcal{I}_2^{-1} = p^{-1}(\mathbf{x}) \begin{pmatrix} I_1^{-1}(f)\mathbf{S}_K^{-1}\mathbf{S}_K^*\mathbf{S}_K^{-1} & 0 \\ 0 & \sigma^2(\mathbf{x})I_3^{-1}(f)\mathbf{S}_K^{-1}\mathbf{S}_K^*\mathbf{S}_K^{-1} \end{pmatrix}$$

and the block of the asymptotic variance of the log-standard deviation estimator corresponds to $p^{-1}(\mathbf{x})I_1^{-1}(f)\mathbf{S}_K^{-1}\mathbf{S}_K^*\mathbf{S}_K^{-1}$ which is the same as the asymptotic variance of the log-standard deviation derived in Theorem 2.1 where the mean function $m(.)$ was assumed to be known. Therefore under the condition of $I_2(f) = 0$, we have proved that the estimator is mean regression adaptive. In addition, note that the assumption of symmetric density function means that the function $\Psi(y)$ is antisymmetric while $\Psi(y)y + 1$ is symmetric and therefore their product is antisymmetric implying that $I_2(f) = 0$. Therefore, the condition of symmetric error density, is sufficient for the ML-estimator to be mean regression adaptive.

Intuitively, the identified condition for adaptiveness is not surprising at all. Regression adaptiveness implies that the mean function $m(.)$ has no contribution to the estimation of the variance function $\sigma^2(.)$. The latter is also implied by the identified condition that the joint information for the mean and variance functions is zero i.e. $I_2(f) = 0$. Consequently, using the ML-estimator of the mean function has no cost at the performance of the variance function estimator in terms of Asymptotic Mean Square Error, assuming that the error distribution is symmetric.

2.8 Numerical applications

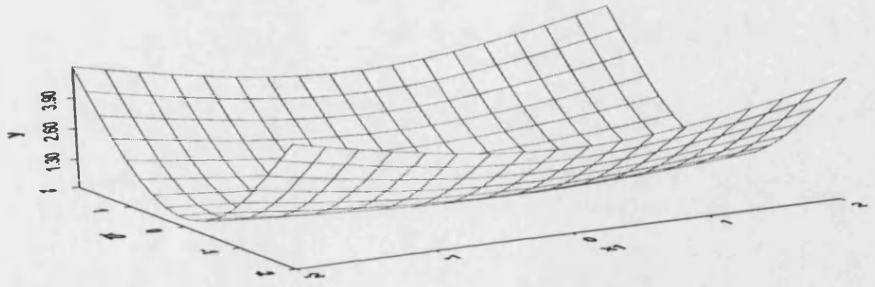
Theoretical findings, based on large sample properties, are in favor of ML-estimator. However, it remains to be seen if this holds when dealing with small sample data sets. In order to evaluate the performance of the ML-estimator, it would be better if we knew the true values of the variance function. For this reason, we present two examples based on simulated data. Real data analysis is postponed to Chapter 5. Note that our primary interest is the variance function and therefore the data is generated from conditional heteroscedastic models in the form of (2.2) with the mean function set equal to zero. The bandwidth is selected from the data driven algorithm described in earlier section with a slight modification. Specifically, instead of using the AMSE which would yield a local bandwidth, we calculate the bandwidth estimator by minimizing $\text{AMISE}(h) \equiv \sum_{i=1}^{n_{grid}} \text{AMSE}(\mathbf{x}_i, h)$ at given grid points $\mathbf{x}_i$, for the range of bandwidths $h = \hat{\sigma}\ C_K\ 1.2^k$ for $k = 0, \ldots, 15$. This function is an approximation to the Asymptotic Mean Integrated Square Error (AMISE) and it is based on the global performance of the estimator. As a result, the estimated bandwidth is a global bandwidth, independent of $\mathbf{x}_i$. The range of bandwidths has been suggested by Fan, Yao, and Cai (2003). The constant C_K depends on the kernel function and it is $C_K = 0.2$ for Epanichnikov kernel and $C_K = 1.2$ for Gaussian kernel, while $\hat{\sigma}$ is the sample standard deviation. Note that at each of the steps, we choose to fit a second order polynomial approximation in order to get estimates for the second derivative of the variance function. Furthermore, theoretical comparison was based on the AMSE. Instead, for the numerical evaluation of the performance of the estimator we employ an alternative measure, the Mean Absolute Deviation Error i.e. $\text{MADE} = n_{grid}^{-1} \sum_{i=1}^{n_{grid}} |\hat{\sigma}^2(\mathbf{x}_i) - \sigma^2(\mathbf{x}_i)|$, where $\{\mathbf{x}_i, i = 1, \ldots, n_{grid}\}$ are grid points in a given interval of the domain of the variance function. We choose a different measure from AMSE in order to ensure that the gain in efficiency is independent of the measure used in the derivation of the theoretical results.

2.8.1 Numerical example 2.1

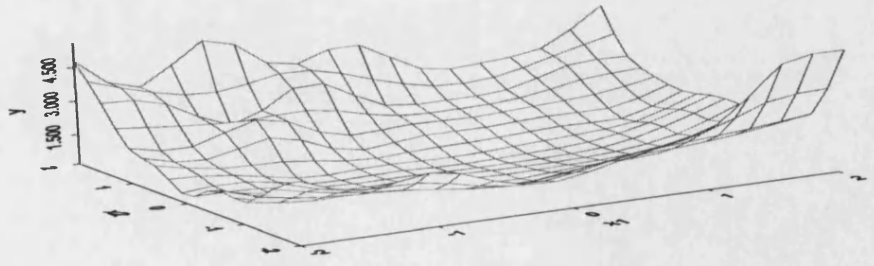
We simulate 100 random samples, each with size $n = 200$, from the model $Y_t = \sigma(Y_{t-1}, Y_{t-2})\varepsilon_t$ with $\sigma(x_1, x_2) = 0.3\sqrt{1 + x_1^2} + \log(1 + x_2^2)$ where the error distribution is assumed to be (i) standard normal, (ii) t-distribution with 6 degrees of freedom and (iii) t-distribution with 14 degrees of freedom (both standardized to ensure that $\mathrm{E}(\varepsilon^2) = 1$). The grid points $\mathbf{x}_i$ are equally spaced points on $[-2, 2] \times [-2, 2]$ with $n_{grid} = 9^2 = 81$. Moreover, we use the bivariate Epanechnikov kernel, $K(\mathbf{u}) = k(u_1)k(u_2)$ with $k(u) = 0.75(1 - u^2)_+$. The selected kernel is symmetric and has a bounded support as required in C2. Fan and Gijbels (1996) also proved that Epanechnikov kernel is the optimal kernel, in the sense of minimum MSE, over all nonnegative, symmetric and Lipschitz continuous functions.

In Figure 2.1 we present the plot: (a) the true variance function and (b) the Maximum Likelihood estimator of the variance function. Note that the MLE is, overall, a good approximation of the true function especially in the middle grid points but there seem to be a discrepancy at the boundary grid points, due to the lack of sufficient observations in neighborhoods closed to the boundaries. In Figure 2.2(i), we plot AMISE against the bandwidth. In plot (a) we have the AMISE based on the true values of the variance and its derivatives. The minimum is attained around $h_{\text{opt}} = 0.31$ which is the optimal bandwidth. In (b) we plot the estimated AMISE calculated using the ML-estimator of the variance and its derivatives used at the final step of the data-driven bandwidth selection algorithm described in section 2.5. Note that the curve is slightly shifted to the left and the estimated optimal global bandwidth is $\hat{h}_{\text{opt}} = 0.297$. This value lies relatively close to the true value. After 100 repetitions for all different samples we noticed that for the data driven bandwidth estimator it holds that $|\hat{h}_{\text{opt}} - h_{\text{opt}}| \leq 5 \times 10^{-2}$ except from 8 samples. In other words, the algorithm identified the optimal bandwidth with error margin less than 5×10^{-2} with probability 92%. The result implies that the estimated bandwidth from the proposed data driven converges in probability to the optimal global bandwidth.

Figure 2.1: Plot of standard deviation function $\sigma(x_1, x_2)$: (a) The true function (b) the ML-estimator for gaussian errors.



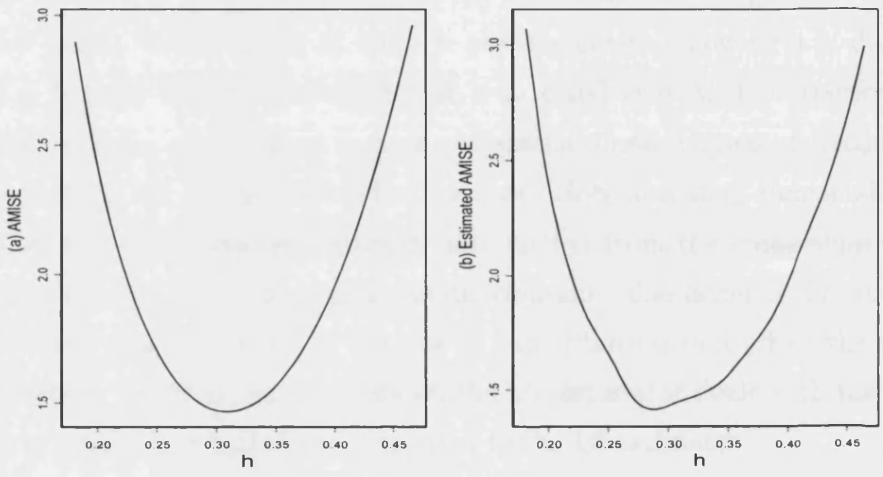
(a)



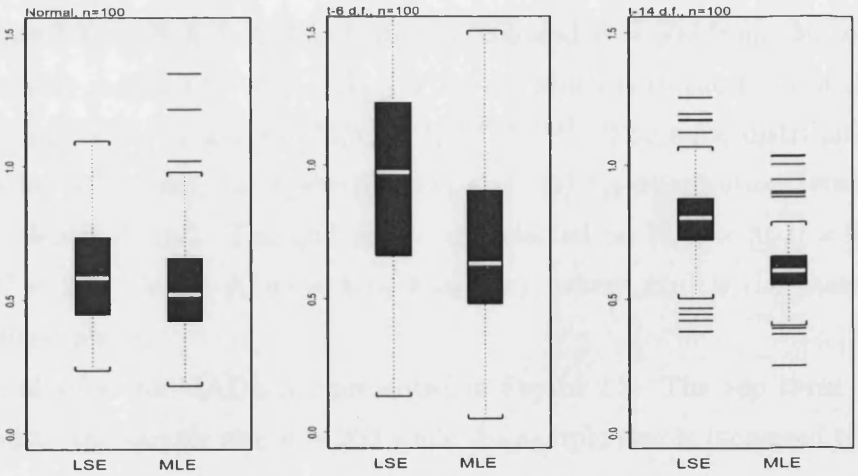
(b)

Figure 2.2: (i) Plot of AMISE vs bandwidth h using (a) the true value of $\sigma(.)$ and its derivatives (b) the ML-estimates. (ii) Box-Plot of MADE for the LSE and MLE for gaussian and t_k-distributed errors with $k = 6$ and $k = 14$ d.f.

(i)



(ii)



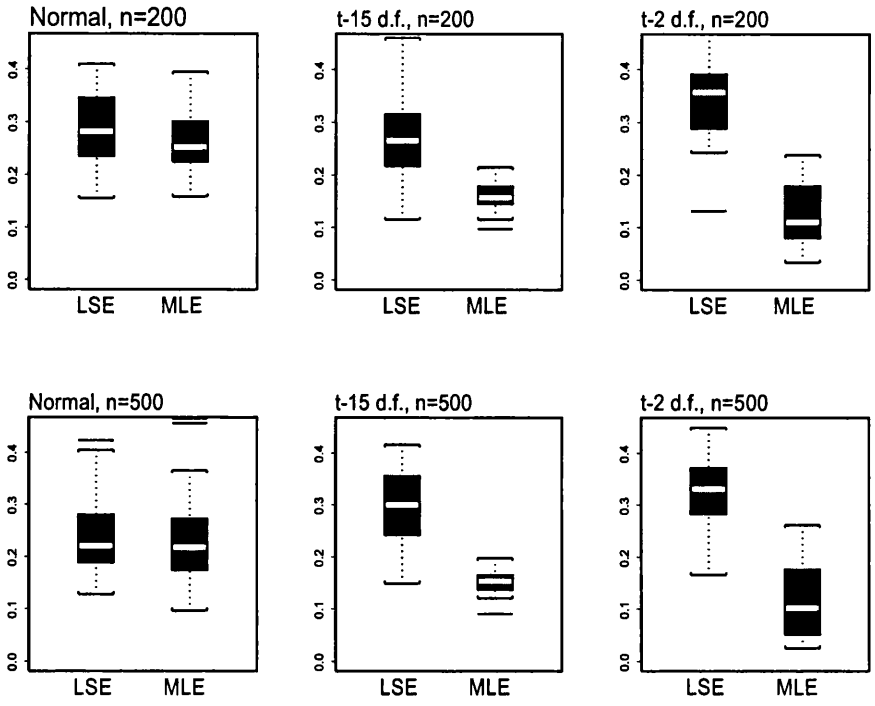
In Figure 2.2(ii), we present the box-plot of the Mean Absolute Deviation Error (MADE) defined above. Particularly, in the first plot where the error distribution is gaussian, the box plot shows that the two nonparametric estimators perform similarly, with a slight preference for the MLE but note also that MLE's mean absolute deviation takes same extreme values while LSE looks more robust. This is in line with our findings of equivalent AMSE of the two estimators for gaussian errors. On the other hand, when looking at the box plot for errors following a t_k-distribution with $k = 6$ (left) and $k = 14$ (right), it is revealed that MLE outperforms LSE. Their difference in performance is more evident for small degrees of freedom and it is reduced when the degrees of freedom increase. Note also that, independent of the estimation procedure, the estimators deviate further from the true values when the error distribution departs from gaussian distribution. The latter is due to the existence of many extreme values in the case of t-distribution that affect the quality of the estimation. However, as noted above, the ML-estimator deals with the existence of extreme values in a better way compared to the LS-estimator.

2.8.2 Numerical example 2.2

We simulate 100 random samples of size $n = 200$ and $n = 500$ from the conditional heteroscedastic model $Y_t = \sigma(Y_{t-1}, Y_{t-2}, Y_{t-3})\ \varepsilon_t$, which is in the form of (2.2) with zero mean and $\sigma(x_1, x_2, x_3) = 0.35(x_1^2 + 1)e^{0.25(x_2+x_3)}$. The error distribution is assumed to be (i) normal, (ii) t_2-distribution and (iii) t_{15}-distribution (standardized to ensure identifiability). The grid points are selected on $[0,1] \times [0,1] \times [0,1]$ with $n_{grid} = 3^3 = 27$. We use $K(\mathbf{u}) = k(u_1)k(u_2)k(u_3)$ where $k(u)$ is the Epanechnikov kernel defined above.

The results for the MADE are presented in Figure 2.3. The top three box-plots correspond to the sample size $n = 200$ while the sample size is increased to $n = 500$ for the box-plots at the bottom. It is clear that, independent of the sample size, the MLE outperforms the LSE. Their difference is more emphatic when the errors follow

Figure 2.3: Box-Plot of the MADE for the LSE and MLE for gaussian and t_k-distributed error with $k = 2$ and $k = 15$ d.f. and $n = 200,\ 500$.



a t-distribution with small degrees of freedom while the difference in the performance becomes less apparent when the degrees of freedom increase and it is almost indistinguishable for the gaussian errors. Note here that using a t_2-distribution violates our assumptions but it is revealing that, in practice, MLE performs better even when the error distribution has infinite variance. Further, comparing the box-plots for the different sample sizes, we can see that, though the increase in the sample size smooths out to some extent the differences of the two estimators, the improvement is still significant as implied from the theoretical results derived for large n.

Chapter 3

Adaptive Maximum Likelihood Estimator

3.1 Motivation and preliminary results

The reduction in the Asymptotic Mean Square Error and the performance of the Maximum Likelihood estimator in the numerical examples suggest that there is an alternative to the Least Squares estimator of the variance function within the non-parametric context. Note that in the implementation of the estimation procedure, we used the error distribution, assuming it is known. However, such an assumption raises serious concerns as there are few cases where we can claim that the error density is known a prior. Hence, though the likelihood estimation procedure yields plausible results, it cannot be implemented in practice when dealing with real data coming from unknown distribution.

The aim of this chapter is to propose a feasible, likelihood-based estimator for the variance function that will not require the error distribution to be known and will share the same asymptotic properties with the estimator introduced in Chapter 2. In other words, we introduce a likelihood-based estimator that is adaptive to the error

distribution. But first, we present some preliminary results that will play a key role in establishing the asymptotic properties of the proposed estimator. These results are known theorems in the context of nonparametric theory and involve rates of uniform convergence for kernel-based estimators of the error density as well as of a general regression function.

Recall that $(Y_t, \mathbf{X}_t)$ is a strictly stationary process with Y_t a scalar and $\mathbf{X}_t^T = (X_{t,1}, \ldots, X_{t,d})$ a d-dimensional random variable. For notational convenience denote with $f^{(0)}$ the density f and $f^{(i)}$ the i-th derivative of f and let $C > 0$ be a generic constant that takes different values at different places. We begin with the theorem that involves the optimal rate of uniform convergence of the kernel density estimator.

Theorem 3.1 *Let ϵ_t be an independent identically distributed process. Denote with $f(y)$ the density function of ϵ_t that has a compact support $S \subset \mathbb{R}$. Assume that the kernel function $W_{h_1}(.)$ has up to 2nd continuous derivatives that satisfy the Lipschitz condition. Then, for i=0,1 and $h_1 \sim n^{-1/(2i+5)}$*

$$\sup_{y \in S} |\tilde{f}^{(i)}(y) - f^{(i)}(y)| = O_p((\log n)^{1/2} n^{-2/(2i+5)})$$

where $\tilde{f}^{(0)} = \tilde{f}$ is the regular kernel density estimator and $\tilde{f}^{(1)}$ the estimator of the 1st-derivative.

Proof of Theorem 3.1 We prove the result for the derivative of the density function i.e. for $i = 1$. Proof for the density function itself can be found in Fan and Yao (2003). Note that $\tilde{f}^{(1)}(y) = n^{-1} \sum_{t=1}^{n} h_1^{-2} W'((\epsilon_t - y)/h_1)$. Take a partition of the compact set $S = \bigcup_{j=1}^{N} S_j$, where $N = [(n/h_1)^{1/2}]$. Then, if y_j is the center of S_j,

$$\sup_{y \in S} |\tilde{f}'(y) - \mathrm{E}(\tilde{f}'(y))| \leq \max_{1 \leq j \leq N} |\tilde{f}'(y_j) - \mathrm{E}(\tilde{f}'(y_j))| + C(nh_1^3)^{-1/2} \qquad (3.1)$$

the latter from the fact that $|\tilde{f}'(x) - \tilde{f}'(y)| \leq Ch_1^{-2}|x - y|$ and a similar result for the expectation. Define $U_t = h_1^{-2}\{W'((\epsilon_t - y)/h_1) - \mathrm{E}(W'((\epsilon_t - y)/h_1))\}$ then $\| U_t \|_\infty < Ch_1^{-2}$. Moreover, $\mathrm{E}(U_t) = 0$ and $\mathrm{Var}(U_t) = O(h_1^{-3})$. Call $\sigma = 1/n \sum_{t=1}^{n} \mathrm{Var}(U_t) =$

$O(h_1^{-3})$, hence Bernstein's inequality for independent, zero-mean, real valued processes, yields

$$P\Big(|\frac{1}{n}\sum_{t=1}^{n} U_t| > \epsilon\Big) \leq \exp(\frac{-n\epsilon^2}{2\sigma + Ch_1^{-2}\epsilon}) \leq \exp(\frac{-n\epsilon^2}{2h_1^{-3} + Ch_1^{-2}\epsilon})$$

and if we take $\epsilon^2 = \alpha n^{-1}h_1^{-3}\log n$ for sufficient large $\alpha > 0$ then

$$P\Big(|\frac{1}{n}\sum_{t=1}^{n} U_t| > \epsilon\Big) \leq \exp(\frac{-\alpha h_1^{-3}\log n}{2h_1^{-3} + \alpha C h_1^{-5} n^{-1}\log n}) \leq \exp(\frac{-\alpha}{C}\log n) = O(n^{-\frac{\alpha}{C}}).$$

Consequently

$$P\Big(\max_{1\leq j\leq N} |\tilde{f}'(y_j) - \mathrm{E}(\tilde{f}'(y_j))| > \epsilon\Big) \leq Nn^{-\alpha'} \to 0 \tag{3.2}$$

for large $\alpha' > 0$. From (3.1) and (3.2), it follows that

$$\sup_{y\in S} |\tilde{f}'(y) - \mathrm{E}(\tilde{f}'(y))| = O_p\Big((\log n)^{1/2}(nh_1^3)^{-1/2}\Big)$$

while standard kernel density theory implies that $\sup_{y\in S} |f'(y) - \mathrm{E}(\tilde{f}'(y))| = O(h_1^2)$. The last two uniform convergence rates with $h_1 = Cn^{-1/7}$ along with the inequality

$$\sup_{y\in S} |\tilde{f}'(y) - f'(y)| \leq \sup_{y\in S} |\tilde{f}'(y) - \mathrm{E}(\tilde{f}'(y))| + \sup_{y\in S} |f'(y) - \mathrm{E}(\tilde{f}'(y))|$$

entail the result for $i = 1$.

The assumption of i.i.d. can be relaxed to include time series. Indeed, the same rate of uniform convergence is found for a strictly stationary geometrically mixing process, see Bosq (1998) and Fan and Yao (2003). However, here the errors are assumed i.i.d. and hence the above result is sufficient for the error density estimator.

The second theorem involves the uniform rate of convergence of a kernel based estimator for the general regression function. Let $r(\mathbf{x}) = \mathrm{E}(q(Y_t)|\mathbf{X}_t = \mathbf{x})$ where $Y_t \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$ with density $p(\mathbf{x})$. Define $g(\mathbf{x}) = \int q(y)f(y,\mathbf{x})dy$, $\mathbf{x} \in \mathbb{R}^d$. The kernel estimators of g,p and r are then given by:

$$\tilde{g}(\mathbf{x}) = \frac{1}{nh^d}\sum_{t=1}^{n} q(Y_t)K(\frac{\mathbf{X}_t - \mathbf{x}}{h}) \text{ and } \tilde{p}(\mathbf{x}) = \frac{1}{nh^d}\sum_{t=1}^{n} K(\frac{\mathbf{X}_t - \mathbf{x}}{h})$$

and $\tilde{r}(\mathbf{x}) = \tilde{g}(\mathbf{x})/\tilde{p}(\mathbf{x})$ with $K_h(.) = 1/h^d K(./h)$ a d-dimensional kernel. Uniform convergence requires the density $p(\mathbf{x})$ to be bounded away from zero, equivalently that there is $\eta > 0$ such that $\inf_{\mathbf{x}\in B}\{p(\mathbf{x})\} > \eta$. In case of time series we have that $p(\mathbf{x})$ and the error density $f(\mathbf{x})$ are related, so this assumption precludes some interesting error distributions e.g. the beta distribution. Hence, we relax this assumption and we allow the density to converge to zero at the tails though we impose restrictions on the rate of convergence. Particularly,

Definition 3.1 *A sequence B_n of compact sets in $\mathbb{R}^d$ is called regular in respect to the density $p(\mathbf{x})$ if there exists a sequence of real numbers β_n and a constant $\gamma > 0$ such that for each n*

$$\inf_{\mathbf{x}\in B_n} p(\mathbf{x}) \geq \beta_n > 0 \text{ and } \delta(B_n) \leq n^{\gamma}$$

where $\delta(B_n)$ is the diameter of B_n.

Then, it can be shown that,

Theorem 3.2 *For $Y_t \in \mathbb{R}$ and $\mathbf{X}_t \in B \subset \mathbb{R}^d$ let $(Y_t, \mathbf{X}_t)$ be a strictly stationary geometrically mixing process and let $\alpha > 0$ such that $\mathrm{E}(e^{\alpha|q(Y_0)|}) < \infty$. Suppose that $g(\mathbf{x})$ and $p(\mathbf{x})$ are bounded and also have bounded second derivatives. Let $K_h(.) = 1/h^d K(./h)$ with $K(.)$ a continuous density function that satisfies the Lipschitz condition. If δ_n is a sequence of real numbers such that there exists a regular, in respect to the density p, sequence of compact sets B_n satisfying*

$$\frac{\delta_n (\log n)^{3/2-1/(d+4)}}{\beta_n n^{2/(d+4)}} \to 0$$

then for $h \simeq (\log n)^{1/(d+4)} n^{-1/(d+4)}$ it holds that

$$\delta_n \sup_{\mathbf{x}\in B_n} |\tilde{r}(\mathbf{x}) - r(\mathbf{x})| \xrightarrow{a.s.} 0.$$

The proof of Theorem 3.2 can be found in Bosq (1998), so is omitted. Note that for $q(u) = u^2$ the estimator $\tilde{r}(\mathbf{x}) = \tilde{g}(\mathbf{x})/\tilde{p}(\mathbf{x})$ is the Nadaraya-Watson estimator for the variance function.

3.2 The adaptive ML-estimator

For the ML-estimator introduced in Chapter 2, which from now on we refer to as the infeasible ML-estimator, the conditional local log-likelihood function $l_n(\boldsymbol{\theta}; \mathbf{Y}|\mathbf{X})$ in (2.5) and consequently the score function $S_n(\boldsymbol{\theta})$ introduced in (2.6), were defined under the assumption of known error density. The following likelihood-based estimator does not require such an assumption. Instead, we estimate $f(.)$ and its derivative $f'(.)$ to find an estimator for $\Psi(.)$ which we plug in to the score function. Then, using the estimated score function and information matrix, we calculate the new likelihood-based estimator. A detailed description of the proposed algorithm is:

Step 1. Let $\tilde{\sigma}_t^2 = \tilde{g}_t(\mathbf{X}_t)/\tilde{p}_t(\mathbf{X}_t)$ the Nadaraya-Watson initial estimators where

$$\tilde{g}_t(\mathbf{x}) = \frac{1}{nh^d} \sum_{s=1, s\neq t}^{n} Y_s^2 K(\frac{\mathbf{X}_s - \mathbf{x}}{h}) \text{ and } \tilde{p}_t(\mathbf{x}) = \frac{1}{nh^d} \sum_{s=1, s\neq t}^{n} K(\frac{\mathbf{X}_s - \mathbf{x}}{h}).$$

Use the initial estimators $\tilde{s}_t = \log \tilde{\sigma}_t$ to find the error estimates $e_t = Y_t e^{-\tilde{s}_t}$.

Step 2. Define

$$\hat{f}^{(i)}(y) = \frac{1}{nh_1^{i+1}} \sum_{s=1}^{n} W^{(i)}((e_s - y)/h_1), \quad \hat{f}_t^{(i)}(y) = \frac{1}{nh_1^{i+1}} \sum_{s=1, s\neq t}^{n} W^{(i)}((e_s - y)/h_1)$$

the standard kernel and the leave-one-out density and derivative estimators based on the estimated errors from step 1.

Step 3. Use $\hat{\Psi}_t(y) = \hat{f}_t'(y)/\hat{f}_t(y)$ to obtain an estimator for the score function:

$$\hat{S}_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \{\hat{\Psi}_t(e_t(\boldsymbol{\theta}))e_t(\boldsymbol{\theta}) + 1\}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}).$$

Further, the estimator for the $(d+1) \times (d+1)$ information matrix is

$$\hat{\mathcal{I}}_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \{\hat{\Psi}_t(e_t(\boldsymbol{\theta}))e_t(\boldsymbol{\theta}) + 1\}^2 \mathbf{Z}_t \mathbf{Z}_t^T K_h(\mathbf{X}_t - \mathbf{x}).$$

Step 4. Define $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_0, \ldots, \tilde{\theta}_d)^T$ to be an initial LS-estimator, then the proposed likelihood-based estimator is the one step Newton-Raphson estimator call $\hat{\boldsymbol{\theta}}_{NR}$ calculated from:

$$\hat{\boldsymbol{\theta}}_{NR} = \tilde{\boldsymbol{\theta}} - \hat{\mathcal{I}}_n^{-1}(\tilde{\boldsymbol{\theta}})\hat{S}_n(\tilde{\boldsymbol{\theta}}). \tag{3.3}$$

Note that the denominators of the estimated score function and information matrix, $\hat{S}_n(\boldsymbol{\theta})$ and $\hat{\mathcal{I}}_n(\boldsymbol{\theta})$ respectively, can be small enough to cause problems. At this point we follow Linton and Xiao (2001) and introduce a trimming function. A smooth trimming function that they propose is

$$G_b(x) = \begin{cases} 0, & x < b \\ \int_{-\infty}^{x} g_b(z)dz, & b \leq x \leq 2b \\ 1, & x > 2b \end{cases} \quad \text{with } g_b(x) = \frac{1}{b} g(\frac{x}{b} - 1)$$

where $g(.)$ is a density function with support $[0, 1]$, $g(0) = g(1) = 0$ and $b > 0$ the trimming parameter. Hence, we use the trimmed score function estimator

$$\hat{S}_n(\boldsymbol{\theta}) \equiv \sum_{t=1}^{n} \{\hat{\Psi}_t(e_t(\boldsymbol{\theta}))e_t(\boldsymbol{\theta}) + 1\}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}) G_b(\hat{f}_t(e_t(\boldsymbol{\theta})))$$

and the trimmed information matrix estimator

$$\hat{\mathcal{I}}_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \{\hat{\Psi}_t(e_t(\boldsymbol{\theta}))e_t(\boldsymbol{\theta}) + 1\}^2 \mathbf{Z}_t \mathbf{Z}_t^T K_h(\mathbf{X}_t - \mathbf{x}) G_b(\hat{f}_t(e_t(\boldsymbol{\theta})))$$

in (3.3) to calculate $\hat{\boldsymbol{\theta}}_{NR}$.

It is understood that the proposed estimator $\hat{\boldsymbol{\theta}}_{NR}$ requires the calculation of a number of initial estimators. We start by calculating the LS-estimator $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_0, \ldots, \tilde{\theta}_d)$ using a first order polynomial approximation. A necessary requirement is that $\tilde{\boldsymbol{\theta}}$ is a consistent estimator. The latter is a well-known result, proof of which can be found in many nonparametric textbooks, e.g. Härdle (1990) among others. Further, the Nadaraya-Watson estimators $\tilde{\sigma}_t^2$ are easily calculated based on standard kernel theory. The following corollary is a direct application of Theorem 3.2 and it entails the uniform convergence rate of the initial estimators $\tilde{\sigma}_t^2$.

Corollary 3.1 *For $Y_t \in \mathbb{R}$ and $\mathbf{X}_t \in B \subset \mathbb{R}^d$ bounded, let $(Y_t, \mathbf{X}_t)$ be a strictly stationary geometrically mixing process and assume that there exists $\alpha > 0$ such that $\mathrm{E}(e^{\alpha Y_0^2}) < \infty$. Suppose $\sigma^2(\mathbf{x}) = \mathrm{E}(Y_t^2|\mathbf{X}_t = \mathbf{x})$ and $p(\mathbf{x})$ the density of $\mathbf{X}_t$ are bounded and also have bounded second derivatives. For the kernel it holds that $K_h(.) = 1/h^d K(./h)$ where $K(.)$ is a continuous density function that satisfies the Lipschitz condition. If there exists a regular in respect to the density $p(.)$ sequence of compact sets B_n, where $\beta_n = n^{-\beta}\log n$ with $0 < \beta < 2/(d+4)$ then for $h \simeq (\log n)^{1/(d+4)} n^{-1/(d+4)}$ it holds that*

$$\sup_{0\le t\le n} |\tilde{\sigma}_t^2 - \sigma_t^2| = o_p(n^{-\delta}(\log n)^{1/2}) \quad for \quad \delta = \frac{2}{d+4} - \beta > 0.$$

Proof of Corollary 3.1 For bounded $\mathbf{X}_t$, Definition 3.1 implies that for each n, we can find B_n, independent of t, such that $\mathbf{X}_t \in B_n$ for all $0 \le t \le n$ and hence we have

$$\sup_{0\le t\le n} |\tilde{\sigma}_t^2 - \sigma_t^2| = \sup_{\mathbf{X}_t \in B_n} |\tilde{\sigma}^2(\mathbf{X}_t) - \sigma^2(\mathbf{X}_t)|. \tag{3.4}$$

We need to write our model in the same form as the model in Theorem 3.2 and prove that all assumptions hold. Indeed $Y_t = \sigma(\mathbf{X}_t)\epsilon_t$ can be written as follows: $Y_t^2 = \sigma^2(\mathbf{X}_t) + \sigma^2(\mathbf{X}_t)\xi_t$ with $\mathrm{E}(\xi_t|\mathbf{X}_t) = 0$, as $\xi_t = \epsilon_t^2 - 1$. Hence $\mathrm{E}(Y_t^2|\mathbf{X}_t) = \sigma^2(\mathbf{X}_t)$ which is in the form of $\mathrm{E}(q(Y_t)|\mathbf{X}_t) = r(\mathbf{X}_t)$ for $q(u) = u^2$ and $r(u) = \sigma^2(u)$. Note that $\beta_n = n^{-\beta}\log n$ with $0 < \beta < 2/(d+4)$ therefore, for $\delta_n = n^{-\delta}(\log n)^{1/2}$ with $\delta = 2/(d+4) - \beta > 0$, it holds that $\delta_n (\log n)^{3/2-1/(d+4)} \beta_n^{-1} n^{-2/(d+4)} \to 0$ thus, the rate of uniform convergence of the estimator is

$$\sup_{\mathbf{X}_t \in B_n} |\tilde{\sigma}^2(\mathbf{X}_t) - \sigma^2(\mathbf{X}_t)| = o_p(n^{-\delta}(\log n)^{1/2}) \tag{3.5}$$

the latter as a direct application of Theorem 3.2. Hence from equation (3.4) and (3.5) we conclude.

It is understood that the calculated rate of uniform convergence for the initial estimates is attained for the specific choice of bandwidth $h \simeq (\log n)^{1/(d+4)} n^{-1/(d+4)}$. This rate is close to the rate of the optimal theoretical bandwidth h_0 calculated by

minimizing the AMSE, (Wand and Jones 1995 and Simonoff 1996). In practice, we use the "plug in" bandwidth $\hat{h}_0$, found by minimizing the estimated AMSE. The numerical examples below reveal that the calculated bandwidth lies close to the required rate i.e. $\hat{h}_0 \sim n^{-1/(d+4)}$ and hence uniform rates from Corollary 3.1 can be achieved by using the "plug in" bandwidth. Nevertheless, the conclusion is not general since it only applies for the particular examples. At this point we argue that the required uniform rates, though important for the derivation of the asymptotic properties, are in practice as crucial as the constant terms especially for large n. If there is any significant departure from the assumptions, it will be revealed in the overall performance of the final estimator.

Apart from the initial variance estimators, we calculate an estimate for the error density and its derivative. Theorem 3.1 yields the uniform convergence rate of the regular density and derivative estimators:

$$\tilde{f}^{(i)}(y) = \frac{1}{nh_1^{i+1}} \sum_{s=1}^{n} W^{(i)}((\epsilon_s - y)/h_1), \; \tilde{f}_t^{(i)}(y) = \frac{1}{nh_1^{i+1}} \sum_{s=1, s\neq t}^{n} W^{(i)}((\epsilon_s - y)/h_1)$$

that are based on the true error values. However, the estimators defined in step 2 are based on error estimates rather than the true values hence Theorem 3.1 is not directly applicable. Instead, we prove the following proposition as a result of Corollary 3.1 and Theorem 3.1. There, we derive the uniform rate of convergence for the density and derivative estimators calculated using the error estimates e_t.

Proposition 3.1 *Let assumptions A1-A4 (below) hold. Then for all t and for i=0,1*

1. $\sup_{y\in S} |\hat{f}_t^{(i)}(y) - \tilde{f}_t^{(i)}(y)| = o_p(n^{-2/(d+4)+\beta}(\log n)^{1/2})$
2. $\sup_{y\in S} |\tilde{f}_t^{(i)}(y) - f^{(i)}(y)| = o_p(n^{-2/(2i+5)}(\log n)^{1/2})$.

Proof of Proposition 3.1 We claim that if $\inf_t \sigma_t^2 > 0$, then $\lim_{n\to\infty} \inf_t \tilde{\sigma}_t^2 > 0$. Indeed, from Corollary 3.1 we have that $\sup_t |\tilde{\sigma}_t^2 - \sigma_t^2| \to 0$ so for each $\varepsilon > 0$ there is $n_0 \in \mathbb{N}$ such that $|\tilde{\sigma}_t^2| > |\sigma_t^2| - \varepsilon$ for $n \geq n_0$ and all t. Therefore $\inf_t |\tilde{\sigma}_t^2| \geq \inf_t |\sigma_t^2| - \varepsilon$.

Choose ε such that $0 < \varepsilon < \inf_t |\sigma_t^2|$ to conclude that $\lim_{n\to\infty} \inf_t |\tilde{\sigma}_t^2| > 0$. Further, first order Taylor expansion of $W^{(i)}(e_s - y/h_1)$ around $(\epsilon_s - y)/h_1$ yields

$$|\hat{f}_t^{(i)}(y) - \tilde{f}_t^{(i)}(y)| \leq \frac{1}{nh_1^{i+1}} \sum_{s=1, s\neq t}^{n} |W^{(i)}(\frac{e_s - y}{h_1}) - W^{(i)}(\frac{\epsilon_s - y}{h_1})| \leq$$

$$\frac{1}{nh_1^{i+2}} \sum_{s=1, s\neq t}^{n} W^{(i+1)}(\frac{\epsilon_s^* - y^*}{h_1})|e_s - \epsilon_s|$$

with ϵ_s^* lying between e_s and ϵ_s. Assumptions A1-A4 ensure that Corollary 3.1 holds, implying that for the initial estimates calculated with the standard kernel method, the rate of uniform convergence is $\sup_{0\leq s\leq n} |\tilde{\sigma}_s^2 - \sigma_s^2| = o_p(n^{-2/(d+4)+\beta}(\log n)^{1/2})$ equivalently, $\sup_{0\leq s\leq n} |e_s - \epsilon_s| = o_p(n^{-2/(d+4)+\beta}(\log n)^{1/2})$ from A3 and $\lim_{n\to\infty} \inf_t \tilde{\sigma}_t^2 > 0$ proved above. Since $f^{(i+1)}(y)$ for $i = 0, 1$ is bounded, from Theorem 3.1 we conclude that

$$\sup_{y\in S} |\frac{1}{nh_1^{i+2}} \sum_{s=1, s\neq t}^{n} W^{(i+1)}(\frac{\epsilon_s^* - y}{h_1})| = O_p(1).$$

Hence, the right hand side of the inequality above is of order $o_p(n^{-2/(d+4)+\beta}(\log n)^{1/2})$. Thus we conclude that for all t:

$$\sup_{y\in S} |\hat{f}_t^{(i)}(y) - \tilde{f}_t^{(i)}(y)| = o_p(n^{-2/(d+4)+\beta}(\log n)^{1/2}) \text{ for } i = 0, 1.$$

Further,

$$|\tilde{f}_t^{(i)}(y) - f^{(i)}(y)| \leq |\tilde{f}^{(i)}(y) - f^{(i)}(y)| + \frac{1}{nh_1^{i+1}} W^{(i)}((\epsilon_t - y)/h_1).$$

From Assumptions A1-A4, application of Theorem 3.1 yields that

$$\sup_{y\in S} |\tilde{f}_t^{(i)}(y) - f^{(i)}(y)| = o_p(n^{-2/(2i+5)}(\log n)^{1/2})$$

for $h_1 = Cn^{-1/(2i+5)}$ and the proof is complete.

Based on the dimension d of the unknown variance function $\sigma^2(.)$ and the constant β that depends on the tails of the error density, Proposition 3.1 yields the following possible cases:

a. If $2/(d+4) \leq 2/7+\beta$ then for $i=0,1$,

$$\sup_{y\in S}|\hat{f}_t^{(i)}(y)-f^{(i)}(y)| = o_p(n^{-2/(d+4)+\beta}(\log n)^{1/2}) \text{ for all } t. \tag{3.6}$$

b. If $2/7+\beta \leq 2/(d+4) \leq 2/5+\beta$ then

$$\sup_{y\in S}|\hat{f}_t(y)-f(y)| = o_p(n^{-2/(d+4)+\beta}(\log n)^{1/2})$$

and

$$\sup_{y\in S}|\hat{f}_t'(y)-f'(y)| = o_p(n^{-2/7}(\log n)^{1/2}) \text{ for all } t. \tag{3.7}$$

It is understood that the lowest uniform convergence rate for the density and the first derivative occurs for d,β such that $2/(d+4) < 2/7+\beta$. Hence, the following results are derived for the particular case that yields the lowest rate and we claim that the same holds for the omitted cases of d,β that yield faster rates. Next we state the regularity conditions that are sufficient to derive the asymptotic properties of the proposed estimator. These are not the weakest possible and can be altered at the cost of lengthier proof.

A1 Let ϵ_t be i.i.d. with $f(y)$ the error density function with compact support S. Assume that f has up to 4th continuous and bounded derivatives with f and f' satisfying Lipschitz condition. Further, recall $\Omega(y)=(d/dy)(y\Psi(y)+1)$ and $R(y)=(d/dy)\Omega(y)$, then it holds that $\mathrm{E}|\Psi(\epsilon)\epsilon+1|^6<\infty$, $\mathrm{E}|\Omega(\epsilon)|^3<\infty$, $\mathrm{E}|R(\epsilon)|^2<\infty$ and $\mathrm{E}|\Omega(\epsilon)\Psi(\epsilon)\epsilon^2|^3<\infty$.

A2 For the d-dimensional symmetric kernel $K_h(.)=1/h^dK(./h)$ assume that it has compact support and satisfies the Lipschitz condition. Similarly, for the univariate kernel $W_{h_1}(.)=h_1^{-1}W(./h_1)$ defined on a compact support, we assume that it has up to 2nd continuous derivatives that satisfy Lipschitz condition.

A3 Let $(Y_t,\mathbf{X}_t)$ be a bounded, strictly stationary, absolutely regular process with the ß-mixing coefficient satisfying $ß(k)\leq c_0\rho^k$ for $c_0>0$ and $0<\rho<1$. Call

$p(\mathbf{x}) > 0$ the marginal density of $\mathbf{X}_t$ with bounded 2nd derivatives. Further, we assume that there exists a regular, in respect to $p(\mathbf{x})$, sequence called B_n with $\beta_n = n^{-\beta}\log n$ with $0 < \beta < 2/(d+4)$, see Definition 3.1.

A4 Assume that $\sigma^2(\mathbf{x}) = \mathrm{E}(Y_t^2|\mathbf{X} = \mathbf{x})$ has up to third continuous and bounded derivatives. Moreover, it holds that $\inf_t \sigma_t^2 > 0$.

A5 For the trimming parameter b assume that $b \sim n^{-a}$ with $0 < a < 1/(d+4) - \beta/2$. Further, $nh^{2d} \to \infty$ and $h \to 0$ as $n \to \infty$.

Note that assumptions A1-A5 ensure that conditions C1-C4 in Chapter 2 hold. Particularly, C1 holds for $\delta = 4$. It also follows that $(Y_t, \mathbf{X}_t)$ is a Geometrically-Strongly-Mixing process. Indeed, from A3 and properties of mixing coefficients we have that $\alpha(k) \leq ß(k) \leq c_0\rho^k$. The property of GSM is necessary for the application of Theorem 3.2. It is easy to see that conditions of Proposition 3.1 above, are implied from A1-A4.

3.2.1 Asymptotic properties of the adaptive ML-estimator

We can now proceed to the asymptotic properties of the proposed estimator $\hat{\boldsymbol{\theta}}_{NR}$. Denote with $\bar{\boldsymbol{\theta}}_n$ the infeasible estimator of Chapter 2, then from Theorem 2.1:

$$\sqrt{nh^d}H(\bar{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0 - \mathbf{b}) \xrightarrow{d} N(0, \mathcal{I}^{-1}\boldsymbol{\Sigma}\mathcal{I}^{-1}) \tag{3.8}$$

while from Lemma 2.3

$$-H^{-1}\mathcal{H}_n(\boldsymbol{\theta}^0)H^{-1} \xrightarrow{P} \mathcal{I} = p(\mathbf{x})I(f)\mathbf{S}_K. \tag{3.9}$$

Theorem 3.3 *Under A1-A5, we have that*

$$\sqrt{nh^d}H(\hat{\boldsymbol{\theta}}_{NR} - \boldsymbol{\theta}^0 - \mathbf{b}) \xrightarrow{d} N(0, \mathcal{I}^{-1}\boldsymbol{\Sigma}\mathcal{I}^{-1})$$

with $\mathcal{I}, \boldsymbol{\Sigma}$ and $\mathbf{b}$ defined in Theorem 2.1. Equivalently, the infeasible ML-estimator and the proposed estimator follow asymptotically the same distribution.

Proof of Theorem 3.3 From (3.3) it holds that

$$(\hat{\boldsymbol{\theta}}_{NR} - \boldsymbol{\theta}^0) = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) - \hat{\mathcal{I}}_n^{-1}(\tilde{\boldsymbol{\theta}})\hat{S}_n(\tilde{\boldsymbol{\theta}}). \tag{3.10}$$

Taylor expansion of $\hat{S}_n(\tilde{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}^0$ yields $\hat{S}_n(\tilde{\boldsymbol{\theta}}) = \hat{S}_n(\boldsymbol{\theta}^0) - \hat{\mathcal{H}}_n(\boldsymbol{\theta}^0)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + o_p(1)$ since $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 \xrightarrow{P} 0$ from consistency. Substitution in (3.10) yields

$$(\hat{\boldsymbol{\theta}}_{NR} - \boldsymbol{\theta}^0) = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + \hat{\mathcal{I}}_n^{-1}(\tilde{\boldsymbol{\theta}})\hat{\mathcal{H}}_n(\boldsymbol{\theta}^0)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) - \hat{\mathcal{I}}_n^{-1}(\tilde{\boldsymbol{\theta}})\hat{S}_n(\boldsymbol{\theta}^0) + o_p(1).$$

Note that the result of the theorem follows from

(i) $1/nH^{-1}(\hat{\mathcal{I}}_n(\tilde{\boldsymbol{\theta}}) - \mathcal{I}_n(\boldsymbol{\theta}^0))H^{-1} \xrightarrow{P} 0.$

(ii) $\sqrt{nh^d}(1/nH^{-1}\hat{S}_n(\boldsymbol{\theta}^0) + H\mathcal{I}\mathbf{b}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}).$

Indeed if (i) holds then from $n^{-1}H^{-1}\mathcal{I}_n(\boldsymbol{\theta}^0)H^{-1} \xrightarrow{P} \mathcal{I} = p(\mathbf{x})I(f)\mathbf{S}_K$ we conclude that

$$\frac{1}{n}H^{-1}\hat{\mathcal{I}}_n(\tilde{\boldsymbol{\theta}})H^{-1} \xrightarrow{P} \mathcal{I} = p(\mathbf{x})I(f)\mathbf{S}_K. \tag{3.11}$$

The latter along with (3.9) yields that $-\hat{\mathcal{I}}_n(\tilde{\boldsymbol{\theta}})^{-1}\hat{\mathcal{H}}_n(\boldsymbol{\theta}^0) \xrightarrow{P} \mathbf{I}_{d+1}$, the unit matrix, consequently

$$(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + \hat{\mathcal{I}}_n(\tilde{\boldsymbol{\theta}})^{-1}\hat{\mathcal{H}}_n(\boldsymbol{\theta}^0)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{P} 0$$

while (ii) along with (3.11) and Lemma 2.4 ensure that

$$\sqrt{nh^d}H\hat{\mathcal{I}}_n(\tilde{\boldsymbol{\theta}})^{-1}\Big(\hat{S}_n(\boldsymbol{\theta}^0) - \mathrm{E}(S_n(\boldsymbol{\theta}^0))\Big) \xrightarrow{d} N(0, \mathcal{I}^{-1}\boldsymbol{\Sigma}\mathcal{I}^{-1}).$$

Focus on $I = 1/nH^{-1}(\hat{\mathcal{I}}_n(\tilde{\boldsymbol{\theta}}) - \mathcal{I}_n(\boldsymbol{\theta}^0))H^{-1}$. For notational convenience, we denote $\tilde{e}_t = Y_t e^{-\mathbf{Z}_t^T\tilde{\boldsymbol{\theta}}}$ and $\bar{e}_t = Y_t e^{-\mathbf{Z}_t^T\boldsymbol{\theta}^0}$ while recall that $e_t = Y_t e^{-\tilde{s}_t}$ and $\epsilon_t = Y_t e^{-s_t}$ with $\tilde{s}_t = 1/2\log\tilde{\sigma}_t^2$ and $s_t = 1/2\log\sigma_t^2$. Then, we write

$$I = \frac{1}{n}H^{-1}(\hat{\mathcal{I}}_n(\tilde{\boldsymbol{\theta}}) - \mathcal{I}_n(\tilde{\boldsymbol{\theta}}))H^{-1} + \frac{1}{n}H^{-1}(\mathcal{I}_n(\tilde{\boldsymbol{\theta}}) - \mathcal{I}_n(\boldsymbol{\theta}^0))H^{-1} = I_1 + I_2.$$

Furthermore, I_1 is decomposed to:

$$I_1 = \frac{1}{n}\sum_{t=1}^{n}\Big(\{\hat{\Psi}_t(\tilde{e}_t)\tilde{e}_t + 1\}^2 - \{\Psi(\tilde{e}_t)\tilde{e}_t + 1\}^2\Big)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x})G_b(\hat{f}_t(\tilde{e}_t))$$

$$+\frac{1}{n}\sum_{t=1}^{n}\{\Psi(\tilde{e}_t)\tilde{e}_t+1\}^2 H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t-\mathbf{x})\Big(G_b(\hat{f}_t(\tilde{e}_t))-1\Big)=I_{11}+I_{12}.$$

From $\hat{\Psi}_t^2(\tilde{e}_t)-\Psi^2(\tilde{e}_t)=\hat{f}_t^{-2}(\tilde{e}_t)(\hat{f}_t'^2(\tilde{e}_t)-f'^2(\tilde{e}_t))+\hat{f}_t^{-2}(\tilde{e}_t)\Psi^2(\tilde{e}_t)(f^2(\tilde{e}_t)-\hat{f}_t^2(\tilde{e}_t))$ and $\hat{\Psi}_t(\tilde{e}_t)-\Psi(\tilde{e}_t)=\hat{f}_t^{-1}(\tilde{e}_t)(\hat{f}_t'(\tilde{e}_t)-f'(\tilde{e}_t))+\hat{f}_t^{-1}(\tilde{e}_t)\Psi(\tilde{e}_t)(f(\tilde{e}_t)-\hat{f}_t(\tilde{e}_t))$ with $\inf_t\{|\hat{f}_t(\tilde{e}_t)|\}>b\sim n^{-a}$ and $|G_b(\hat{f}_t)|<C$ we conclude that

$$I_{11}\leq Cn^{2a}\Big\{\frac{1}{n}\sum_{t=1}^{n}|\hat{f}_t'^2(\tilde{e}_t)-f'^2(\tilde{e}_t)|\tilde{e}_t^2H^{-1}\mathbf{Z}_t\mathbf{Z}_t^TH^{-1}K_h(\mathbf{X}_t-\mathbf{x})$$

$$+\frac{1}{n}\sum_{t=1}^{n}|f^2(\tilde{e}_t)-\hat{f}_t^2(\tilde{e}_t)|\Psi^2(\tilde{e}_t)\tilde{e}_t^2H^{-1}\mathbf{Z}_t\mathbf{Z}_t^TH^{-1}K_h(\mathbf{X}_t-\mathbf{x})\Big\}$$

$$+Cn^a\Big\{\frac{1}{n}\sum_{t=1}^{n}|\hat{f}_t'(\tilde{e}_t)-f'(\tilde{e}_t)|\tilde{e}_tH^{-1}\mathbf{Z}_t\mathbf{Z}_t^TH^{-1}K_h(\mathbf{X}_t-\mathbf{x})$$

$$+\frac{1}{n}\sum_{t=1}^{n}|f(\tilde{e}_t)-\hat{f}_t(\tilde{e}_t)|\Psi(\tilde{e}_t)\tilde{e}_tH^{-1}\mathbf{Z}_t\mathbf{Z}_t^TH^{-1}K_h(\mathbf{X}_t-\mathbf{x})\Big\}.$$

From A3 and the fact that $||\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^0||\leq C$ since $\boldsymbol{\theta}^0\in\Theta$ a compact set, it follows that $|\tilde{e}_t|\leq C$. Assumptions A2-A3 yield

$$\frac{1}{n}\sum_{t=1}^{n}H^{-1}\mathbf{Z}_t\mathbf{Z}_t^TH^{-1}K_h(\mathbf{X}_t-\mathbf{x})=O_p(1) \tag{3.12}$$

and next we prove that

$$\frac{1}{n}\sum_{t=1}^{n}\Psi^k(\tilde{e}_t)\tilde{e}_t^kH^{-1}\mathbf{Z}_t\mathbf{Z}_t^TH^{-1}K_h(\mathbf{X}_t-\mathbf{x})=O_p(1)\ \text{ for }\ k=1,2 \tag{3.13}$$

then combining (3.6),(3.12) and (3.13) it follows that $I_{11}=o_p(n^{2a-2/(d+4)+\beta})$ which is of order $o_p(1)$ for the trimming parameter satisfying $0<a<1/(d+4)-\beta/2$, see A5. It remains to prove (3.13). We present the case of $k=2$. Note that from A1 it holds $\mathrm{E}|\Psi(\epsilon)\epsilon+1|^2<\infty$, then from A2-A3, the ergodic theorem yields:

$$\frac{1}{n}\sum_{t=1}^{n}\Psi^2(\epsilon_t)\epsilon_t^2H^{-1}\mathbf{Z}_t\mathbf{Z}_t^TH^{-1}K_h(\mathbf{X}_t-\mathbf{x})=O_p(1). \tag{3.14}$$

Henceforth, it is sufficient to show that

$$\mathcal{J} = \frac{1}{n}\sum_{t=1}^{n}\Big(\Psi^2(\tilde{e}_t)\tilde{e}_t^2 - \Psi^2(\epsilon_t)\epsilon_t^2\Big)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x}) = o_p(1). \tag{3.15}$$

Substitute the Taylor expansion $\Psi^2(\tilde{e}_t)\tilde{e}_t^2 - \Psi^2(\epsilon_t)\epsilon_t^2 = 2\Omega(\epsilon_t)\Psi(\epsilon_t)\epsilon_t(\tilde{e}_t - \epsilon_t) + o_p(\tilde{e}_t - \epsilon_t)$ to obtain $\mathcal{J} = 2/n\sum_{t=1}^{n}\Omega(\epsilon_t)\Psi(\epsilon_t)\epsilon_t(\tilde{e}_t - \epsilon_t)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x}) + o_p(\tilde{e}_t - \epsilon_t)$. Note that

$$\tilde{e}_t - \epsilon_t = Y_t(e^{-\mathbf{Z}_t^T\tilde{\boldsymbol{\theta}}} - e^{-s(\mathbf{X}_t)}) = \epsilon_t(e^{-\mathbf{Z}_t^T\tilde{\boldsymbol{\theta}} + s(\mathbf{X}_t)} - 1) \tag{3.16}$$

where

$$s(\mathbf{X}_t) - \mathbf{Z}_t^T\tilde{\boldsymbol{\theta}} = s(\mathbf{X}_t) - \mathbf{Z}_t^T\boldsymbol{\theta}^0 + \mathbf{Z}_t^T\boldsymbol{\theta}^0 - \mathbf{Z}_t^T\tilde{\boldsymbol{\theta}}. \tag{3.17}$$

Substitution of the second order Taylor expansion of $s(\mathbf{X}_t)$ around $\mathbf{x}$ in (3.17) yields

$$s(\mathbf{X}_t) - \mathbf{Z}_t^T\tilde{\boldsymbol{\theta}} = \frac{1}{2}\sum_{i,j=1}^{d}\ddot{s}_{ij}(\mathbf{x}')(X_{t,i} - x_i)(X_{t,j} - x_j) + s(\mathbf{x}) - \tilde{s}(\mathbf{x}) + \sum_{j=1}^{d}(\theta_j^0 - \tilde{\theta}_j)(X_{t,j} - x_j).$$

Therefore,

$$\begin{aligned}\mathcal{J} = {} & \frac{2}{n}\sum_{t=1}^{n}\Omega(\epsilon_t)\Psi(\epsilon_t)\epsilon_t^2\frac{1}{2}\sum_{i,j=1}^{d}\ddot{s}_{ij}(\mathbf{x}')(X_{t,i} - x_i)(X_{t,j} - x_j)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x}) \\ & + \frac{2}{n}\sum_{t=1}^{n}\Omega(\epsilon_t)\Psi(\epsilon_t)\epsilon_t^2\sum_{j=1}^{d}(\theta_j^0 - \tilde{\theta}_j)(X_{t,j} - x_j)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x}) \\ & - (\tilde{s}(\mathbf{x}) - s(\mathbf{x}))\frac{2}{n}\sum_{t=1}^{n}\Omega(\epsilon_t)\Psi(\epsilon_t)\epsilon_t^2 H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x}) = \sum_{i=1}^{3}\mathcal{J}_i.\end{aligned}$$

It is easy to see that under A1: $\mathrm{E}|\Omega(\epsilon)\Psi(\epsilon)\epsilon^2| < \infty$. Then, ergodic theorem implies that $\mathcal{J}_1$ is of order $O_p(h^2)$ while similar arguments along with the fact that $|\theta_j^0 - \tilde{\theta}_j| \leq C$, $j = 1, \ldots, d$ ensure that the second term $\mathcal{J}_2$ is of order $O_p(h)$ while Fan and Yao (1998) have shown that the local linear LS-estimator converges in distribution and therefore it holds that $\tilde{\sigma}(\mathbf{x}) - \sigma(\mathbf{x}) = O_p(n^{-1/2}h^{-d/2})$ i.e. $\mathcal{J}_3 = O_p(n^{-1/2}h^{-d/2})$. Under

A5, summing up the above results we conclude (3.15). Note here that for the case of $k = 1$ in (3.13) we prove

$$\frac{1}{n}\sum_{t=1}^{n}\Big(\Psi(\tilde{e}_t)\tilde{e}_t - \Psi(\epsilon_t)\epsilon_t\Big)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x}) = o_p(1) \tag{3.18}$$

in the same way as (3.14) but substituting $\Psi^2(y)y^2$ with $\Psi(y)y$. Details are omitted. Then, combining (3.14) and (3.18) we conclude

$$\frac{1}{n}\sum_{t=1}^{n}\Big\{\Big(\Psi(\tilde{e}_t)\tilde{e}_t + 1\Big)^2 - \Big(\Psi_t(\epsilon_t)\epsilon_t + 1\Big)^2\Big\}H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x}) = o_p(1) \tag{3.19}$$

which completes the proof for $I_{11} = o_p(1)$. We focus on I_{12}:

$$\frac{1}{n}\sum_{t=1}^{n}\Big\{\Big(\Psi(\tilde{e}_t)\tilde{e}_t + 1\Big)^2 - \Big(\Psi(\epsilon_t)\epsilon_t + 1\Big)^2\Big\}H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x})(G_b(\hat{f}_t(\tilde{e}_t)) - 1)$$

$$+\frac{1}{n}\sum_{t=1}^{n}\{\Psi(\epsilon_t)\epsilon_t + 1\}^2 H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x})(G_b(\hat{f}_t(\tilde{e}_t)) - 1) = I_{121} + I_{122}.$$

Based on (3.19) it is easy to see that $I_{121} = o_p(1)$. Further, I_{122} is decomposed to

$$I_{1221} = \frac{1}{n}\sum_{t=1}^{n}\{\Psi(\epsilon_t)\epsilon_t + 1\}^2 H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x})(G_b(\hat{f}_t(\tilde{e}_t)) - G_b(f(\tilde{e}_t)))$$

$$I_{1222} = \frac{1}{n}\sum_{t=1}^{n}\{\Psi(\epsilon_t)\epsilon_t + 1\}^2 H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x})(G_b(f(\tilde{e}_t)) - G_b(f(\epsilon_t)))$$

$$I_{1223} = \frac{1}{n}\sum_{t=1}^{n}\{\Psi(\epsilon_t)\epsilon_t + 1\}^2 H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x})(G_b(f(\epsilon_t)) - 1).$$

Similar to the proof of Lemma 2.3 that involves the information matrix, it holds that

$$\frac{1}{n}\sum_{t=1}^{n}\{\Psi(\epsilon_t)\epsilon_t + 1\}^2 H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x}) = O_p(1). \tag{3.20}$$

Taylor expansion of the smooth trimming function $G_b(.)$ yields

$$\sup_t |G_b(\hat{f}_t(\tilde{e}_t)) - G_b(f(\tilde{e}_t))| \leq \sup_{f^*:|f^*-f|\to 0} |g_b(f^*)| \sup_t |\hat{f}_t(\tilde{e}_t) - f(\tilde{e}_t)|$$

where f^* lies between $\hat{f}_t$ and f hence $|f^* - f| \to 0$. Since $\sup_{f^*:|f^*-f|\to 0} |g_b(f^*)| = O(n^{-a})$ then from (3.6) it follows that

$$\sup_t |G_b(\hat{f}_t(\tilde{e}_t)) - G_b(f(\tilde{e}_t))| = o_p(n^{a-2/(d+4)+\beta}) \tag{3.21}$$

which along with (3.20) and A5 yields $I_{1221} = o_p(1)$. Similarly, Taylor expansion of $G_b(f(.))$ yields $G_b(f(\tilde{e}_t)) - G_b(f(\epsilon_t)) = g_b(f(\epsilon_t))f'(\epsilon_t)(\tilde{e}_t - \epsilon_t) + o_p(1)$. Therefore,

$$I_{1222} = \frac{1}{n}\sum_{t=1}^{n}\{\Psi(\epsilon_t)\epsilon_t + 1\}^2 g_b(f(\epsilon_t))f'(\epsilon_t)(\tilde{e}_t - \epsilon_t)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x}) + o_p(1)$$

while substitution of (3.16) and (3.17) yields

$$I_{1222} = \frac{1}{n}\sum_{t=1}^{n}\{\Psi(\epsilon_t)\epsilon_t + 1\}^2\epsilon_t g_b(f(\epsilon_t))f'(\epsilon_t)\sum_{j=1}^{d}\ddot{s}_{jj}(\mathbf{x}')H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x})$$

$$+\frac{1}{n}\sum_{t=1}^{n}\{\Psi(\epsilon_t)\epsilon_t + 1\}^2\epsilon_t g_b(f(\epsilon_t))f'(\epsilon_t)\sum_{j=1}^{d}(\theta_j^0 - \tilde{\theta}_j)(X_{t,j} - x_j)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x})$$

$$-(\tilde{s}(\mathbf{x}) - s(\mathbf{x}))\frac{1}{n}\sum_{t=1}^{n}\{\Psi(\epsilon_t)\epsilon_t + 1\}^2\epsilon_t g_b(f(\epsilon_t))f'(\epsilon_t)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h((\mathbf{X}_t - \mathbf{x})) + o_p(1).$$

From Cauchy-Schwartz inequality: $(\mathrm{E}|\{\Psi(\epsilon)\epsilon + 1\}^2\epsilon g_b(f(\epsilon))f'(\epsilon)|)^2 \leq$

$$\leq C\int\{\Psi(\epsilon)\epsilon + 1\}^4\epsilon^2 f(\epsilon)d\epsilon\int g^2(\frac{f(\epsilon)}{b} - 1)\frac{f(\epsilon)}{b}d\frac{f(\epsilon)}{b} < \infty \tag{3.22}$$

bounded from A1 and the definition of $g(.)$ as a pdf defined in $[0, 1]$. Consequently, the first and the second term of the decomposition of I_{1222} are of order $O_p(h)$ and $O_p(h^2)$ respectively as a result of the ergodic theorem while for the last term, note that $\tilde{s}(\mathbf{x}) - s(\mathbf{x}) = O_p(n^{-1/2}h^{-d/2})$ and again application of the ergodic theorem yields that the last term is $o_p(1)$ from A5. For I_{1223} note that

$$E\Big(\{\Psi(\epsilon_t)\epsilon_t + 1\}^2(1 - G_b(f(\epsilon_t)))\Big) = \int_{0\leq f(\epsilon)\leq b}\{\Psi(\epsilon)\epsilon + 1\}^2 f(\epsilon)d\epsilon$$

$$+\int_{b\leq f(\epsilon)\leq 2b}\{\Psi(\epsilon)\epsilon + 1\}^2\Big(\int_{f(\epsilon)}^{\infty} g_b(z)dz\Big)f(\epsilon)d\epsilon \to 0, \quad \text{as } n \to \infty$$

from A1,A5 ($b \to 0$) and the dominated convergence theorem. Therefore $I_{1223} = o_p(1)$. Consequently, we proved that $I_{12} = o_p(1)$ which along with $I_{11} = o_p(1)$ entail the result for I_1. We continue with I_2:

$$I_2 = \frac{1}{n}\sum_{t=1}^{n}\Big(\{\Psi(\tilde{e}_t)\tilde{e}_t + 1\}^2 - \{\Psi(\epsilon_t)\epsilon_t + 1\}^2\Big)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x})$$

$$+\frac{1}{n}\sum_{t=1}^{n}\Big(\{\Psi(\epsilon_t)\epsilon_t + 1\}^2 - \{\Psi(\bar{e}_t)\bar{e}_t + 1\}^2\Big)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x}) = I_{21} + I_{22}$$

where $I_{21} = o_p(1)$ directly from (3.19). For the term I_{22}, using the Taylor expansion $(\Psi(\bar{e}_t)\bar{e}_t + 1)^2 - (\Psi(\epsilon_t)\epsilon_t + 1)^2 = 2(\Psi(\epsilon_t)\epsilon_t + 1)\Omega(\epsilon_t)(\bar{e}_t - \epsilon_t) + O((\bar{e}_t - \epsilon_t)^2)$ yields

$$I_{22} = -\frac{1}{n}\sum_{t=1}^{n}2(\Psi(\epsilon_t)\epsilon_t + 1)\Omega(\epsilon_t)(\bar{e}_t - \epsilon_t)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t - \mathbf{x}) + o_p(1).$$

Substitute the expansion:

$$(\bar{e}_t - \epsilon_t) = \epsilon_t(e^{s(\mathbf{X}_t) - \mathbf{Z}^T\theta^0} - 1) = \epsilon_t\frac{1}{2}\sum_{i,j=1}^{d}\ddot{s}_{ij}(\mathbf{x}')(X_{t,j} - x_j)(X_{t,i} - x_i) \qquad (3.23)$$

in I_{22} to get that $-I_{22} =$

$$\frac{1}{n}\sum_{t=1}^{n}(\Psi(\epsilon_t)\epsilon_t+1)\Omega(\epsilon_t)\epsilon_t\sum_{i,j=1}^{d}\ddot{s}_{ij}(\mathbf{x}')(X_{t,j}-x_j)(X_{t,i}-x_i)H^{-1}\mathbf{Z}_t\mathbf{Z}_t^T H^{-1}K_h(\mathbf{X}_t-\mathbf{x})+o_p(1).$$

From A2-A5 and $\mathrm{E}|(\Psi(\epsilon)\epsilon + 1)\Omega(\epsilon)\epsilon| < \infty$, ergodic theorem yields $I_{22} = O_p(h^2) = o_p(1)$ and that completes the proof of $I_2 = o_p(1)$ which combined with $I_1 = o_p(1)$ entail the result in (i). It remains to show (ii). It holds that

$$\frac{1}{n}H^{-1}\hat{S}_n(\theta^0) = \frac{1}{n}\sum_{t=1}^{n}\{\hat{\Psi}_t(\bar{e}_t)\bar{e}_t - \Psi(\bar{e}_t)\bar{e}_t\}H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x})G_b(\hat{f}_t(\bar{e}_t))$$

$$+\frac{1}{n}\sum_{t=1}^{n}\{\Psi(\bar{e}_t)\bar{e}_t + 1\}H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x})G_b(\hat{f}_t(\bar{e}_t)) = J_1 + J_2.$$

Note that

$$J_1 \leq Cn^a \sup_t|\hat{f}_t'(\bar{e}_t) - f'(\bar{e}_t)|\ \frac{1}{n}\sum_{t=1}^{n}Y_t H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x})$$

$$+Cn^a \sup_t |\hat{f}_t(\bar{e}_t) - f(\bar{e}_t)| \frac{1}{n}\sum_{t=1}^{n} \Psi(\bar{e}_t)\bar{e}_t H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}) = J_{11} + J_{12}.$$

The process $(Y_t, \mathbf{X}_t)$ is a strictly stationary, strongly mixing process with $\mathrm{E}|Y_t|^6 < \infty$ and $\sum_{k=1}^{\infty} \alpha^{2/3}(k) < \sum_{k=1}^{\infty} ß(k) < \infty$ from A3. Theorems 1.5 and 1.7 (Bosq 1998) yield $1/n \sum_{t=1}^{n} Y_t H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}) = O_p(n^{-1/2}h^{-d/2})$. Consequently from (3.6), $J_{11} = o_p(n^{-1/2}h^{-d/2})$ for the trimming parameter satisfying $0 < a < 1/(d+4) - \beta/2$ from A5. For J_{12}, we argue that it is sufficient to prove that

$$\mathcal{T} = \frac{1}{n}\sum_{t=1}^{n} \{\Psi(\bar{e}_t)\bar{e}_t - \Psi(\epsilon_t)\epsilon_t\} H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}) = o_p(n^{-1/2}h^{-d/2}). \tag{3.24}$$

Indeed if (3.24) holds, then from A1-A3 and the CLT we have that

$$\frac{1}{n}\sum_{t=1}^{n} \{\Psi(\epsilon_t)\epsilon_t + 1\} H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}) = O_p(n^{-1/2}h^{-d/2}) \tag{3.25}$$

which along with (3.6) and (3.24) ensure that $J_{12} = o_p(n^{-1/2}h^{-d/2})$ for the same choice of the trimming parameter a as above. So focus on (3.24). Substitute

$$\Psi(\bar{e}_t)\bar{e}_t - \Psi(\epsilon_t)\epsilon_t = \Omega(\epsilon_t)(\bar{e}_t - \epsilon_t) + O((\bar{e}_t - \epsilon_t)^2) \tag{3.26}$$

in $\mathcal{T}$ to obtain $\mathcal{T} = n^{-1}\sum_{t=1}^{n} \Omega(\epsilon_t)(\bar{e}_t - \epsilon_t) H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}) + o_p(1)$ and using (3.23), it follows that

$$\mathcal{T} = \frac{1}{n}\sum_{t=1}^{n} \Omega(\epsilon_t)\epsilon_t \frac{1}{2}\sum_{i,j=1}^{d} \ddot{s}_{ij}(\mathbf{x}^*)(X_{t,i} - x_i)(X_{t,j} - x_j) H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x}) + o_p(1).$$

Call $Q_t = h^{-2}\Omega(\epsilon_t)\epsilon_t 2^{-1}\sum_{i,j=1}^{d} \ddot{s}_{ij}(\mathbf{x})(X_{t,i} - x_i)(X_{t,j} - x_j) H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x})$ and note that $h^{-2}\mathcal{T} = n^{-1}\sum_{t=1}^{n}(Q_t - \mathrm{E}(Q_t)) + O(1)$, from $\mathrm{E}(Q_t) = O(1)$. The process $Q_t - \mathrm{E}(Q_t)$ is a zero-mean, strictly stationary, strongly mixing process with mixing coefficients satisfying $\sum_{k=1}^{n} \alpha^{1/3}(k) < \infty$ and $\mathrm{E}||Q_t - \mathrm{E}(Q_t)||^3 < \infty$. The latter holds from $\mathrm{E}|\Omega(\epsilon_t)\epsilon_t|^3 < \infty$ (implied from A1) along with A2-A4. Then, direct application of the Central Limit Theorem 2.21 (Fan and Yao 2003), yields that $\sqrt{nh^d}h^{-2}\mathcal{T} \xrightarrow{d} N$

consequently $\mathcal{T} = O_p(n^{-1}h^{-d/2+2}) = o_p(n^{-1}h^{-d/2})$ which along with the result for J_{11} yields $J_1 = o_p(n^{-1}h^{-d/2})$. For the second term J_2:

$$J_2 = \frac{1}{n}\sum_{t=1}^{n}\{\Psi(\bar{e}_t)\bar{e}_t + 1\}H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x})\Big(G_b(\hat{f}_t(\bar{e}_t)) - G_b(f(\bar{e}_t))\Big)$$

$$+\frac{1}{n}\sum_{t=1}^{n}\{\Psi(\bar{e}_t)\bar{e}_t + 1\}H^{-1}\mathbf{Z}_t K_h(\mathbf{X}_t - \mathbf{x})G_b(f(\bar{e}_t)) = J_{21} + J_{22}.$$

Note that (3.21) holds for $\tilde{e}_t$ substituted by $\bar{e}_t$ and combined with (3.25) and (3.8) implies that $J_{21} = o_p(n^{a-2/(d+4)+\beta})O_p(n^{-1/2}h^{-d/2}) = o_p(n^{-1/2}h^{-d/2})$ under A5. Further, recall the process V_t defined in Lemma 2.5 of Chapter 2. Then we can write $J_{22} = n^{-1}\sum_{t=1}^{n}\Big(V_tG_b(f(\bar{e}_t)) - \mathrm{E}(V_tG_b(f(\bar{e}_t)))\Big) + \mathrm{E}(V_tG_b(f(\bar{e}_t)))$ and note that from expansions (3.23) and (3.26) $\mathrm{E}(V_tG_b(f(\bar{e}_t))) =$

$$= p(\mathbf{x})\int\int \Omega(\epsilon_t)\epsilon_t G_b(f(\bar{e}_t))f(\epsilon_t)d\epsilon_t\frac{h^2}{2}\sum_{i,j=1}\ddot{s}_{ij}(\mathbf{x})u_iu_j(1,\mathbf{u}^T)K(\mathbf{u})d\mathbf{u} + o(h^2)$$

with $\int \Omega(\epsilon_t)\epsilon_t G_b(f(\bar{e}_t))f(\epsilon_t)d\epsilon_t =$

$$\int \mathbb{I}(b \le f(\bar{e}_t) \le 2b)\Omega(\epsilon_t)\epsilon_t f(\epsilon_t)\Big(\int_{-\infty}^{f(\bar{e}_t)} g_b(z)dz\Big)d\epsilon_t + \int \mathbb{I}(f(\bar{e}_t) > 2b)\Omega(\epsilon_t)\epsilon_t f(\epsilon_t)d\epsilon_t$$

where $\mathbb{I}(x)$ the indicator function. Note that $b \to 0$ as $n \to \infty$ and by dominated convergence theorem $\lim_{b\to 0}\int \mathbb{I}(b \le f(\bar{e}_t) \le 2b)\Omega(\epsilon_t)\epsilon_t f(\epsilon_t)\Big(\int_{-\infty}^{f(\bar{e}_t)} g_b(z)dz\Big)d\epsilon_t = 0$ and $\lim_{b\to 0}\int \mathbb{I}(f(\bar{e}_t) > 2b)\Omega(\epsilon_t)\epsilon_t f(\epsilon_t)d\epsilon_t = \int \Omega(\epsilon_t)\epsilon_t f(\epsilon_t)d\epsilon_t$. Hence, $\mathrm{E}(V_tG_b(f(\bar{e}_t))) = (2^{-1}h^2\mu_2 p(\mathbf{x})\int \Omega(\epsilon_t)\epsilon_t f(\epsilon_t)d\epsilon_t\sum_{j=1}^{d}\ddot{s}_{jj}(\mathbf{x}), o(h^2), \ldots, o(h^2))^T$ and similar to Lemma 2.4, third order Taylor expansion of $s(\mathbf{x})$ will yield the bias term for the derivatives estimator so we conclude that $\mathrm{E}(V_tG_b(f(\bar{e}_t))) = -h^2p(\mathbf{x})I(f)H\mathcal{M}_{K,1}\mathbf{H}_s = H\mathcal{I}\mathbf{b}$. Denote with $U_t = V_tG_b(f(\bar{e}_t)) - \mathrm{E}(V_tG_b(f(\bar{e}_t)))$ then it is easy to see that

$$\sqrt{nh^d}(J_{22} + H\mathcal{I}\mathbf{b}) = \frac{1}{\sqrt{nh^d}}\sum_{t=1}^{n}h^dU_t \tag{3.27}$$

with $\mathrm{Var}(U_t) = \mathrm{E}((\int_{-\infty}^{f(\bar{e}_t)} g_b(z)dz)^2\mathbb{I}(b \le f(\bar{e}_t) \le 2b)V_tV_t^T) + \mathrm{E}(\mathbb{I}(f(\bar{e}_t) \ge 2b)V_tV_t^T)$. The dominated convergence theorem yields

$$\lim_{b\to 0}\mathrm{E}\Big((\int_{-\infty}^{f(\bar{e}_t)} g_b(z)dz)^2\mathbb{I}(b \le f(\bar{e}_t) \le 2b)V_tV_t^T\Big) = 0$$

and $\lim_{b\to 0} \mathrm{E}(\mathbb{I}(f(\bar{e}_t) \geq 2b)V_t V_t^T) = \mathrm{E}(V_t V_t^T)$. Hence, based on the results for the process V_t, derived in Lemma 2.5, we conclude that

$$nh^d \mathrm{Var}(n^{-1}\sum_{t=1}^{n} U_t) = p(\mathbf{x})I(f)\mathbf{S}_K^*. \tag{3.28}$$

Further, it holds that $\mathrm{E}||U_t||^3 \leq C\mathrm{E}||G_b(f)V_t||^3$ where

$$\mathrm{E}||G_b(f)V_t||^3 = \mathrm{E}((\int_{-\infty}^{f(\bar{e}_t)} g_b(z)dz)^3 \mathbb{I}(b \leq f(\bar{e}_t) \leq 2b)||V_t||^3) + \mathrm{E}(\mathbb{I}(f(\bar{e}_t) \geq 2b)||V_t||^3)$$

and again, dominated convergence theorem yields that the first term of the right hand side is zero and the second is $\mathrm{E}||V_t||^3$. Therefore $\mathrm{E}||U_t||^3 \leq C\mathrm{E}||V_t||^3 \leq Ch^{-2d}$ the latter as a direct application of the result in (2.30) in Lemma 2.5 for $\delta = 3$. Consequently, $\mathrm{E}||h^d U_t|| = O(h^d)$ while the α-mixing coefficients for $(Y_t, \mathbf{X}_t)$ satisfy $\sum_{k\geq 1} \alpha^{1/3}(k) < \infty$ from A3. Application of the Central Limit Theorem 2.21 (Fan and Yao 2003), yields

$$\frac{1}{\sqrt{nh^d}} \sum_{t=1}^{n} h^d U_t \xrightarrow{d} N(0, \mathbf{\Sigma}_n)$$

with

$$\mathbf{\Sigma}_n = \mathrm{Var}(\frac{1}{\sqrt{nh^d}} \sum_{t=1}^{n} h^d U_t) = nh^d \mathrm{Var}(\frac{1}{n} \sum_{t=1}^{n} U_t) = p(\mathbf{x})I(f)\mathbf{S}_K^*$$

from (3.28). Based on (3.27) we conclude $\sqrt{nh^d}(J_{22} + H\mathcal{I}\mathbf{b}) \xrightarrow{d} N(0, \mathbf{\Sigma})$ which combined with the results for the remaining terms completes the proof of (ii).

Theorem 3.3 reveals that the proposed ML-estimator shares the same asymptotic properties with the infeasible ML-estimator in Chapter 2. Equivalently, it is adaptive in respect to the error density function and therefore, knowing f or using an estimator of f does not affect asymptotically the MSE of the estimator. Due to this property, from now on we refer to the proposed estimator as the adaptive ML-estimator. It is understood that the adaptive ML-estimator requires a number of initial estimators and hence the number of computations involved has increased. However, the numerical examples below show that the improvement of the estimator is sufficient and compensates for the increase in computations.

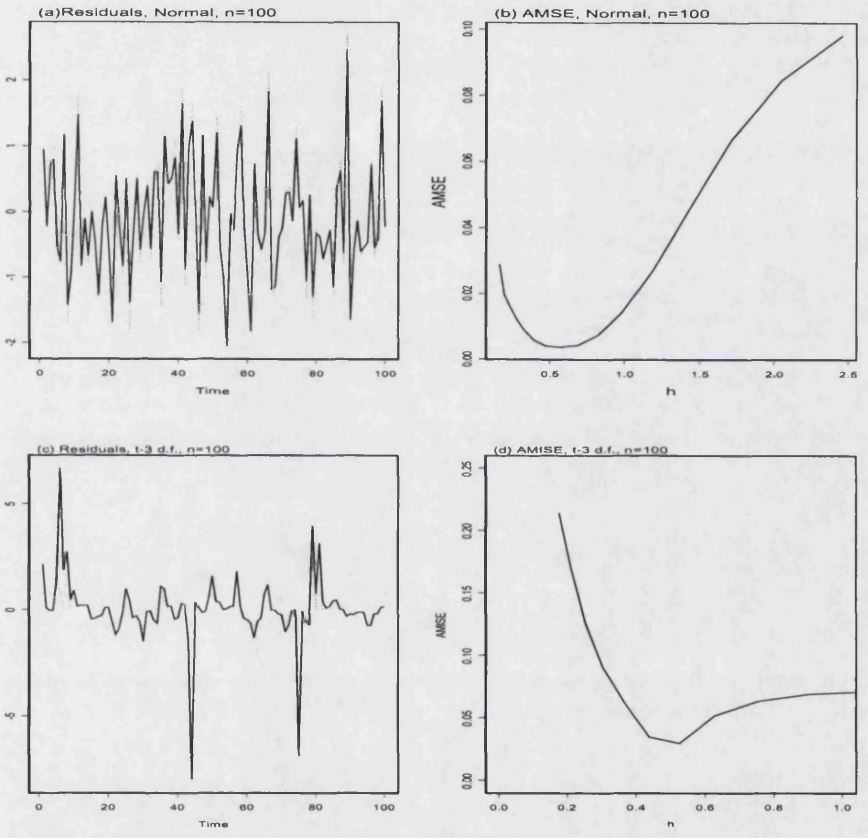
3.3 Numerical applications

Does likelihood estimator remain more efficient in practice than LS-estimator when the error density is estimated? Do adaptive and infeasible estimators perform similarly? Although we have answered these questions theoretically for large n, it remains to see if the same conclusions are reached numerically for small sample size n. Next, we consider two simulation examples with real data analysis postponed to Chapter 5. For the simulated data, the true values are known and hence, we can evaluate directly the performance of the estimators. It is understood that some of the assumptions of the model may not be met by the simulated data. For instance, error distribution is assumed to have compact support in A1. This precludes most of the interesting distributions. Nevertheless, we choose to investigate numerically the behavior of the adaptive ML-estimator even for non regular cases. The results of the simulation will show whether the departure from a particular assumption is influential or not. Particularly, in the first example, we give a thorough report about the initial estimators, the density and derivative bandwidths and the performance of the adaptive estimator while in the second case, we visit the example 2.1 from Chapter 2 and present the deviation errors in order to evaluate the performance of the estimators.

3.3.1 Numerical example 3.1

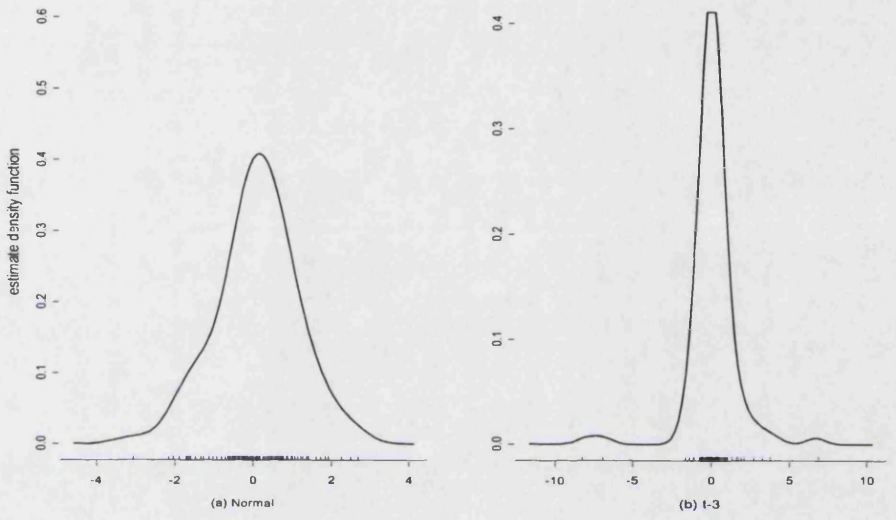
We generate 100 random samples of size (a) $n = 100$ and (b) $n = 500$ from the model $Y_t = \sigma(Y_{t-1})\varepsilon_t$, with $\sigma(x) = e^{-x^2/2}$ where the error distribution is assumed to be (i) standard normal, (ii) t-distribution with 3 d.f. and (iii) cauchy distribution. We continue using the Mean Absolute Deviation Error as the measure to evaluate the performance of the estimators. We estimate the function on 20 equally distanced grid points in $[-2, 2]$. The Epanechnikov kernel is preferred due to its bounded support. First, we calculate the Nadaraya-Watson initial estimates $\tilde{\sigma}_t^2$, equivalently $\tilde{s}_t = 2^{-1} \log \tilde{\sigma}_t^2$ and hence the error estimates $e_t = Y_t e^{-\tilde{s}_t}$.

Figure 3.1: Plot of true residuals (solid line) and estimated residuals based on the initial NW-variance estimates (dotted line) for (a) normal-(c) t_3-dist. Plot of estimated AMSE vs bandwidth for (b) normal (d) t_3-dist.



In Figure 3.1 we plot the true residuals (solid line) and the estimated residuals (dotted line) for (a) normal and (c) t_3 distributed errors. It appears that the initial estimators perform rather well capturing most of the dynamics. Further, in (b) and (d) the estimated AMSE for the initial estimates, calculated in the grid points, is plotted against the bandwidth. We identify the global optimal bandwidth being around $\hat{h}_{opt} = 0.5$ independent of the error distribution while from Corollary 3.1, the

Figure 3.2: Density estimators based on the estimated errors e_t for (a) gaussian and (b) t_3 error distribution.



theoretical bandwidth that yields the optimal uniform rates for the initial estimators is $h \simeq (\log(n)/n)^{1/5} \simeq 0.457$ for $n = 100$. Therefore, the required bandwidth for the optimal rates is achieved by minimizing the estimated AMSE. Using the estimated errors e_t, we calculate the density and the derivative estimator, Figure 3.2 (a) and (b). Apparently, the density estimator of the t-distribution in (b) is leptokurtic, equivalently, the mass of the observation is narrowly concentrated in the center with fatter and longer tails than the estimator of gaussian density in (a). The latter is a well known result of the kurtosis. Moreover, the bandwidths for the density and the derivative estimators are selected as the minimizers of the estimated AMSE. Table 3.1 contains the result for the average of the bandwidths on the grid points. The first column includes the selected density's bandwidth and it is compared to the third column that is the required bandwidth rate to achieve the uniform optimal rates of

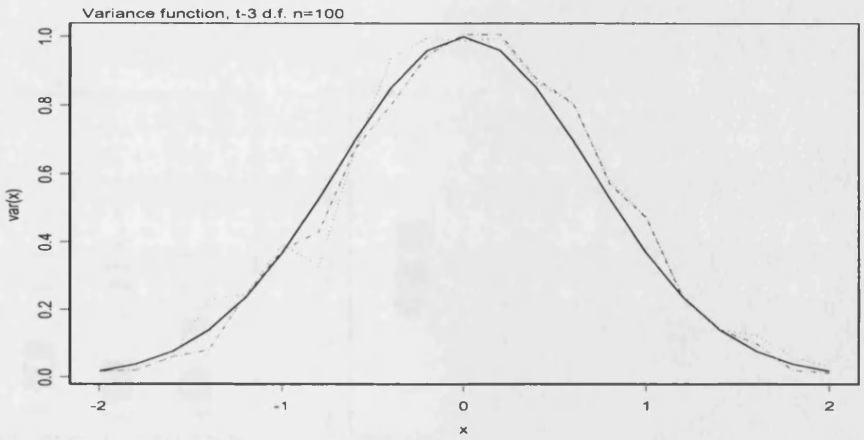
Table 3.1: Bandwidth for error density and derivative estimator.

Distribution	size	$h(f)$	$h(df)$	$n^{-1/5}$	$n^{-1/7}$
Normal	$n = 100$	0.3826	0.8683	0.3981	0.5179
Normal	$n = 500$	0.3482	0.8152	0.2885	0.4116
t_3	$n = 100$	0.5561	0.9872	0.3981	0.5179
t_3	$n = 500$	0.5127	0.9354	0.2885	0.4116

Theorem 3.1. For gaussian distribution these are relatively close. For instance, the selected density bandwidth from minimizing AMSE is $\hat{h}_{opt} = 0.3826$ compared to the rate $n^{-1/5} = 0.3981$ for $n = 100$. For the t-distribution we found $\hat{h}_{opt} = 0.5561$ which is not close to 0.3981 but we expected this discrepancy due to the existence of extreme values. Further, the same pattern appears if we increase the sample size to $n = 500$. As far as density's derivative bandwidth is concerned, there is more significant departure from the theoretical value. Note that the calculated bandwidth is $\hat{h}_{opt} = 0.8683$ while Theorem 3.1 requires rate around $n^{-1/7} = 0.5179$. At this point, we argue that apart from the rate, the constant terms abbreviated as C, affect significantly the bandwidth and should also be taken into account when we compare the optimal values. In any case, the final result will reveal whether there is departure from the bandwidth assumption with significant effect on the performance of the adaptive ML-estimator. Further in Figure 3.3 we plot the variance function (solid line) along with the LS-estimator (dotted line) and the adaptive ML-estimator (dashed-dotted line) for t_3 error distribution. It appears that the ML-estimator is a smoother and improved estimator of the variance function.

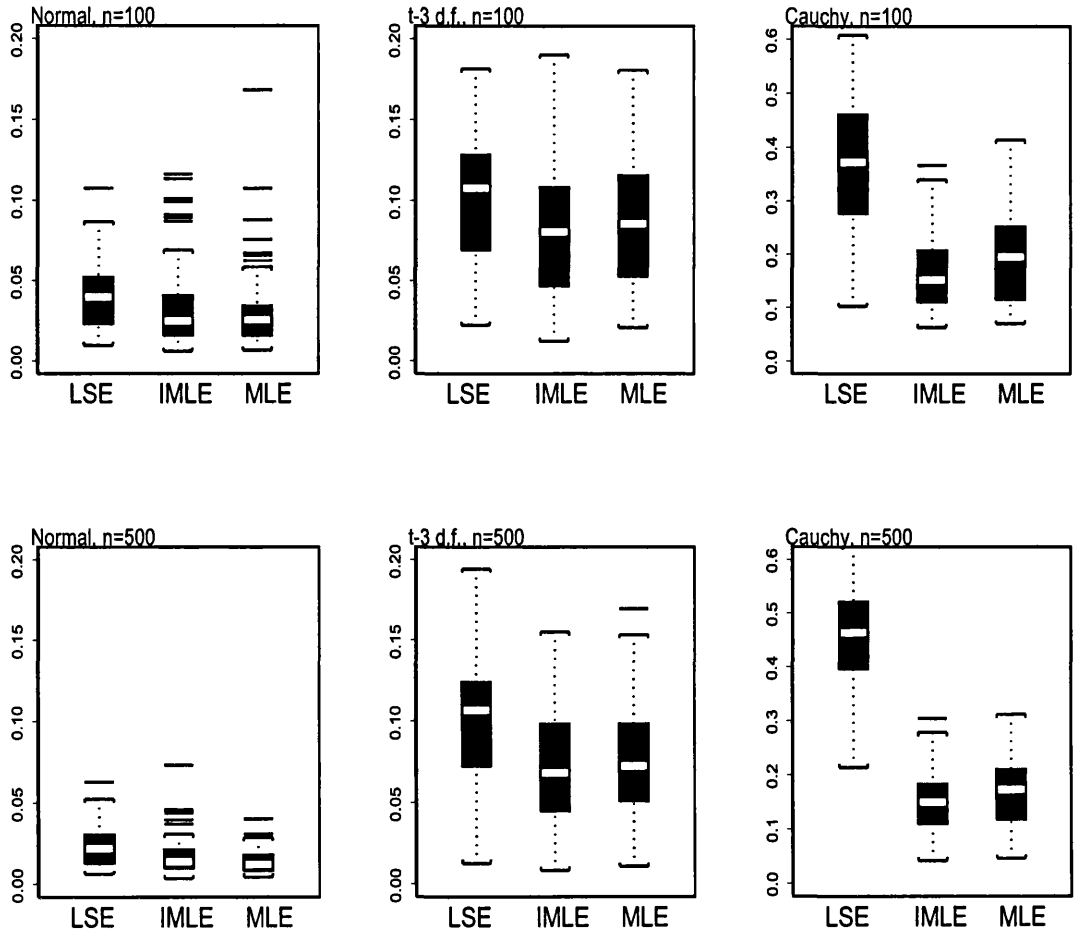
Figure 3.4 contains the box-plots of the Mean Absolute Deviation Error for LS-estimator (LSE), the infeasible ML-estimator (IMLE) from Chapter 2 and the adaptive ML-estimator (MLE). For gaussian errors, there is no significant difference between the three estimators, especially for large n. Indeed, for $n = 100$, IMLE and MLE seem to perform slightly better than LSE, though the differences are smoothed

Figure 3.3: Plot of the true variance function (solid line), the LSE (dotted line) and adaptive MLE (dashed-dotted line).



out when we increase the sample size to $n = 500$. However, if the errors follow a t_3 distribution, LSE is outperformed by IMLE and MLE. Further, note that the infeasible ML-estimator of Chapter 2 and the adaptive estimator perform equally well especially for sample size $n = 500$. The latter agrees with our findings in Theorem 3.3 about adaptiveness in respect to the error density function. Equivalently, for large n, using the density and derivative estimators $\hat{f}$ and $\hat{f}'$ instead of the true functions does not have any significant effect on the performance of the adaptive ML-estimator. The same conclusions, but now more emphatic, are drawn when looking at the results for the Cauchy error distribution. Likelihood-based estimators totally outperformed the LS-estimator which fails to capture the dynamics of a fat tailed distribution like the Cauchy distribution. Further, even in the extreme case of an error distribution with infinite first moment, the adaptive MLE performs rather similar to the infeasible-MLE confirming once more that the proposed estimator is adaptive to the error distribution.

Figure 3.4: Box-Plot of the MADE of the LSE, the Infeasible-MLE and the adaptive MLE for gaussian, t_3 and Cauchy errors for $n = 100,\ 500$.

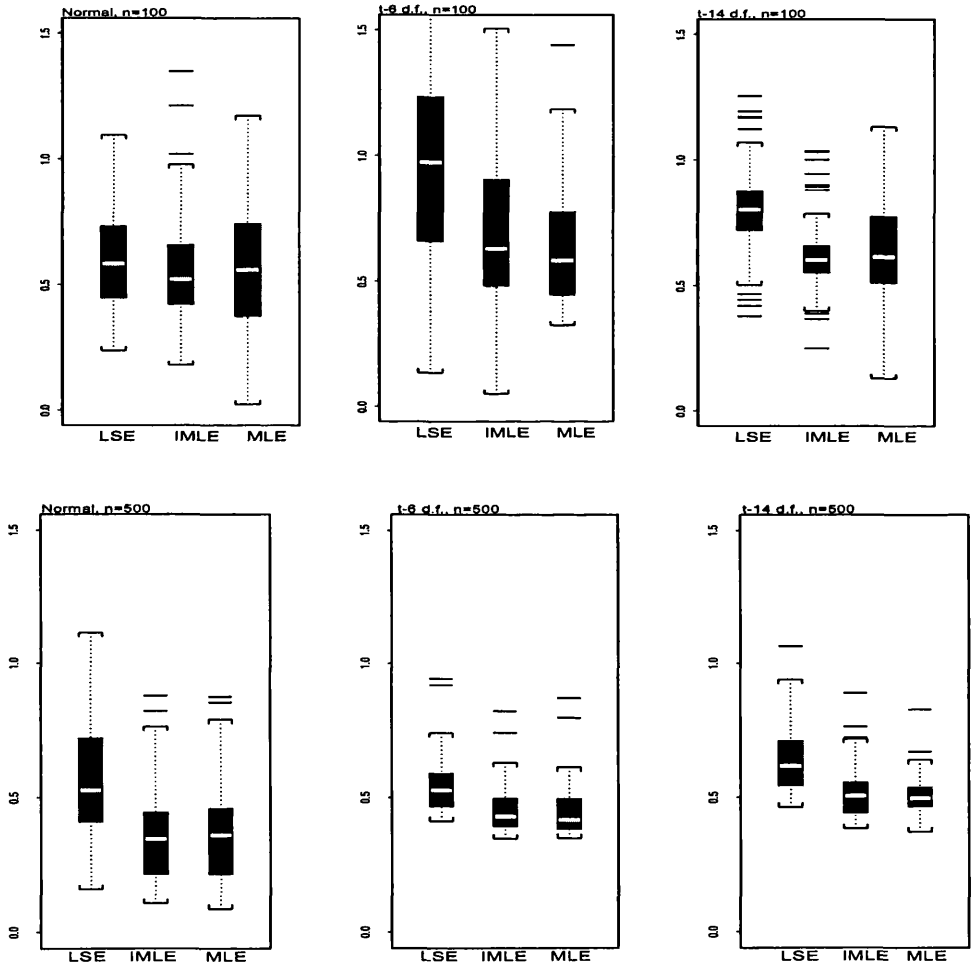


3.3.2 Numerical example 3.2

Recall the conditional heteroscedastic model considered in example 2.1, Chapter 2. Particularly, we generate 100 random samples of size $n = 100$ and 500 from the model $Y_t = \sigma(Y_{t-1}, Y_{t-2})\varepsilon_t$ with $\sigma(x_1, x_2) = 0.3\sqrt{1 + x_1^2} + \log(1 + x_2^2)$ where the error distribution is assumed to be (i) standard normal and standardized t-distribution with (ii) 6 and (iii) 14 degrees of freedom. The function is estimated in $9 \times 9 = 81$ grid points equally distanced in the interval $[-2, 2] \times [-2, 2]$. In Chapter 2, the direct comparison between the LSE and the infeasible ML-estimator revealed that there was significant improvement in the performance, due to the use of information from the error distribution. It is interesting to see if this improvement is maintained when the error distribution is replaced by the nonparametric density estimator.

The results for the Mean Absolute Deviation Error are presented in box plots in Figure 3.5. Clearly, the LS-estimator appears to have the poorest performance for small sample size especially for t-distributed errors. When the sample size is increased the difference in the performance is maintained though it is not as evident as in the case of small sample size. Note here that the latter observation is not true for the case of gaussian errors, where the difference is more emphatic for $n = 500$. On the other hand, the infeasible and the adaptive ML-estimators perform similarly. Their only difference is that for normally and t_{14}-distributed errors with $n = 100$, the interquartile range for the adaptive estimator is wider than of the infeasible implying that it is less robust and has larger variation. However, this difference is smoothed out when the sample size is increased to $n = 500$. But the main point here is that on average their deviation from the true values does not vary significantly and hence there is little evidence for significant difference on the performance between the infeasible and the adaptive likelihood-based estimators. Consequently, overall, the numerical results support the theoretical findings, summarized in Theorem 3.3, on the adaptiveness for the proposed estimator.

Figure 3.5: Box-Plot of the MADE of the LSE, the Infeasible-MLE and the adaptive MLE for gaussian, t_6 and t_{14} distributed errors with $n = 100,\ 500$.



Chapter 4

A Two-Step Cross-Validation Selection Method For Partially Linear Models

4.1 Existence of a partially linear regression model

The estimation of the variance function has been so far the focus of discussion. Note that the nonparametric model investigated in earlier sections required a fixed d-dimensional set of regressors. However, it is not clear in advance which variables should be used as regressors. As the rate of convergence slows down when the number of regressors is increased, it is important that only variables with significant effect on the dependent variable are included in the model. Furthermore, if the true model includes a linear term then the nonparametric approach is less efficient compared to a partially linear model. Hence, in this chapter, we address the issue of model selection for the special class of partially linear models that contain a linear as well as a nonparametric component. The advantage of these models is that the parametric part is estimated efficiently while the flexibility of the nonparametric model is retained.

We begin with the introduction of the partially linear mean regression model before we extend the results to the variance function in later section. Let $(Y_t, \mathbf{X}_t, \mathbf{Z}_t)$ be a strictly stationary process with a scalar Y_t, vectors $\mathbf{X}_t = (X_{t,1}, \ldots, X_{t,P})^T$ and $\mathbf{Z}_t = (Z_{t,1}, \ldots, Z_{t,Q})^T$. In the context of time series analysis, $\mathbf{X}_t$ and $\mathbf{Z}_t$ may contain lagged variables of Y_t. Consider the regression model

$$Y_t = \mathrm{E}(Y_t|\mathbf{X}_t, \mathbf{Z}_t) + \epsilon_t = \mathbf{X}_t^T\boldsymbol{\theta} + g(\mathbf{Z}_t) + \epsilon_t \tag{4.1}$$

where $g : \mathbb{R}^Q \to \mathbb{R}$ is an unknown function, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)^T$ is the vector of the parameters and $\epsilon_t = Y_t - \mathrm{E}(Y_t|\mathbf{X}_t, \mathbf{Z}_t)$ is the error term. It is easy to see that $\mathrm{E}(\epsilon_t|\mathbf{X}_t, \mathbf{Z}_t) = 0$. Härdle, Liang, and Gao (2000) propose an estimation procedure for both the parametric and nonparametric component and establish the asymptotic properties. More recently Gao and Tong (2002) proposed a combination of the leave-one-out cross validation criterion for selecting the regressors of the nonparametric component and the leave-n_v-out cross validation criterion for selecting the regressors of the linear component. However, the computations involved are quite intensive. For example, if we have P linear and Q nonparametric candidate regressors then the number of variable subsets to be investigated, is $(2^P - 1) \times (2^Q - 1)$. We propose an alternative procedure which is computationally more efficient reducing this number to $2^Q + 2^P - 2$ possibilities. However, before we go into details about the proposed two-step cross-validation selection procedure, let us introduce some conditions to ensure the existence of a true model as a reduced form of model (4.1). Denote with $U_t = Y_t - \mathbf{X}_t^T\boldsymbol{\theta}$ the residuals after removing the linear effect, then from (4.1) it holds that $\mathrm{E}(U_t|\mathbf{Z}_t) = g(\mathbf{Z}_t)$. Define the variance function $\sigma^2 : \mathbb{R}^k \to \mathbb{R}$

$$\sigma^2(A) = \mathrm{E}[U_t - \mathrm{E}(U_t|\mathbf{Z}_t^A)]^2 \tag{4.2}$$

with $\mathbf{Z}_t^A = (Z_{t,i} : i \in A)^T$, $A = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, Q\}$ and $|A| = k$ with $1 \leq k \leq Q$. Then, the optimal nonparametric regressors subset is the subset with the minimum dimension k and equal variability with the full set of nonparametric regressors denoted as $\mathbf{Z}_t$. Equivalently,

Definition 4.1 *Assume that there is a subset* $A_0 \equiv \{1, \ldots, q\}$ *of* $\{1, \ldots, Q\}$ *with* $q \leq Q$ *for which*

(a) $\sigma^2(A_0) = \sigma^2(1, \ldots, Q)$ *and*

(b) for any $A = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, Q\}$ *with* $k \leq q$ *and* $A \neq A_0$ *it holds that* $\sigma^2(A) > \sigma^2(A_0)$

then the set $\mathbf{Z}_t^{A_0} = \{Z_{t,1}, \ldots, Z_{t,q}\}$ *is called the optimal regression subset of the nonparametric component.*

Given the optimal nonparametric regression subset $\mathbf{Z}_t^{A_0}$, let $V_t = Y_t - g(\mathbf{Z}_t^{A_0})$ and define the function $\bar{\sigma}^2 : \mathbb{R}^k \to \mathbb{R}$,

$$\bar{\sigma}^2(M) = E[V_t - \mathrm{E}(V_t|\mathbf{X}_t^M)]^2 \tag{4.3}$$

for any $M = \{j_1, \ldots, j_k\} \subseteq \{1, \ldots, P\}$ and $1 \leq k \leq P$. Similar to Definition 4.1,

Definition 4.2 *Assume that there is a subset* $M_0 \equiv \{1, \ldots, p\}$ *of* $\{1, \ldots, P\}$ *with* $p \leq P$ *for which*

(a) $\bar{\sigma}^2(M_0) = \bar{\sigma}^2(1, \ldots, P)$ *and*

(b) for any $M = \{j_1, \ldots, j_k\} \subseteq \{1, \ldots, P\}$ *with* $k \leq p$ *and* $M \neq M_0$ *it holds that* $\bar{\sigma}^2(M) > \bar{\sigma}^2(M_0)$

then the set $\mathbf{X}_t^{M_0} = \{X_{t,1}, \ldots, X_{t,p}\}$ *is called the optimal regression subset of the parametric component.*

At this point, we are ready to impose the necessary condition that will ensure identifiability of the true model. In particular,

A1 The true model is the model with the optimal nonparametric $\mathbf{Z}_t^{A_0} = \{Z_{t,1}, \ldots, Z_{t,q}\}$ and parametric $\mathbf{X}_t^{M_0} = \{X_{t,1}, \ldots, X_{t,p}\}$ components. Further, we assume that $\min_{j \in \{1,\ldots,Q\} - A_0} \inf_{\alpha,\beta} E(E(g(\mathbf{Z}_t^{A_0})|Z_{t,j}) - \alpha - \beta Z_{t,j})^2 > C$, with $C > 0$.

It is easy to see that if A1 holds, Definition 4.1 yields that $\mathrm{E}(U_t|\mathbf{Z}_t) = \mathrm{E}(U_t|\mathbf{Z}_t^{A_0})$ almost surely. In other words, the optimal subset contains almost all the information on U_t available from $\{Z_{t,1}, \ldots, Z_{t,Q}\}$. Further, from Definition 4.2 $\mathrm{E}(V_t|\mathbf{X}_t) = \mathrm{E}(V_t|\mathbf{X}_t^{M_0})$ almost surely, so we conclude that some of the linear predictors are insignificant and should be omitted. Note also that in A1, the contribution of the nonparametric component cannot be explained by any linear term combination. In a similar situation, Chen and Chen (1991) impose a condition to ensure that the nonparametric component is not a linear function of the regressors.

Given the existence of the true model as a reduced form of model (4.1), we try to identify the optimal predictors for both linear and nonlinear components of the regression function. We propose a two-step selection procedure. The first step is the selection of the nonparametric component. We include all the candidate regressors in the parametric component and use the leave-one-out cross validation procedure to select the optimal subset $\mathbf{Z}_t^{A_0}$. We do not repeat this procedure for any other subset of linear regressors, reducing our computations to $2^Q - 1$. Having selected the optimal nonparametric subset, at the second step we employ the leave-n_v-out cross validation criterion to reduce, if necessary, the number of parametric regressors. This will require the investigation of $2^P - 1$ cases which implies that the overall calculations are reduced from $(2^P - 1) \times (2^Q - 1)$ to $2^Q + 2^P - 2$ cases.

4.2 Selection of the nonparametric component

We begin with the selection of the nonparametric component. In practice $\boldsymbol{\theta}$ is an unknown parameter and therefore is replaced by $\hat{\boldsymbol{\theta}}$ the LS-estimator calculated by regressing Y_t over $X_{t,1}, \ldots, X_{t,P}$ and $\hat{U}_t = Y_t - \mathbf{X}_t^T\hat{\boldsymbol{\theta}}$ the estimated residuals. For any $A \subseteq \{1, \ldots, Q\}$, define the Nadaraya-Watson estimation for $g(.)$,

$$g_n(\mathbf{z}) = \sum_{t=1}^{n} w_{t,A}(\mathbf{z})(Y_t - \mathbf{X}_t^T\boldsymbol{\theta}) \tag{4.4}$$

with $w_{t,A} : \mathbb{R}^k \to \mathbb{R}$, $w_{t,A}(\mathbf{z}) = K_h(\mathbf{Z}_t^A - \mathbf{z}) / \sum_{r=1}^n K_h(\mathbf{Z}_r^A - \mathbf{z})$ a weighting function and $K : \mathbb{R}^k \to \mathbb{R}$, $K_h(.) = 1/h^k K(./h)$, a k-dimensional kernel function. Similarly, we define

$$\hat{g}_n(\mathbf{z}) = \sum_{t=1}^n w_{t,A}(\mathbf{z})(Y_t - \mathbf{X}_t^T \hat{\boldsymbol{\theta}}) \tag{4.5}$$

which is the estimator defined in (4.4) with $\boldsymbol{\theta}$ replaced by $\hat{\boldsymbol{\theta}}$. The leave-one-out estimators are $g_n^{(-s)}(\mathbf{z}) = \sum_{t=1,t\neq s}^n w_{t,A}^{(-s)}(\mathbf{z})(Y_t - \mathbf{X}_t^T \boldsymbol{\theta})$ and $\hat{g}_n^{(-s)}(\mathbf{z}) = \sum_{t=1,t\neq s}^n w_{t,A}^{(-s)}(\mathbf{z})(Y_t - \mathbf{X}_t^T \hat{\boldsymbol{\theta}})$ respectively, where $w_{t,A}^{(-s)}(\mathbf{z}) = K_h(\mathbf{Z}_t^A - \mathbf{z}) / \sum_{r=1,r\neq s}^n K_h(\mathbf{Z}_r^A - \mathbf{z})$. Then, the cross validation function is defined:

$$\mathrm{CV}(A) = \frac{1}{n} \sum_{s=1}^n \{\hat{U}_s - \hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\}^2 \text{ for all } A \subset \{1, \ldots, Q\}. \tag{4.6}$$

Definition 4.3 *The estimator for the optimal regression subset of the nonparametric component is defined as*

$$\hat{A} = \arg \min_{A \subseteq \{1,\ldots,Q\}} \mathrm{CV}(A). \tag{4.7}$$

Next we state the assumptions and introduce the notation. Let $C > 0$ be a constant that can take different values in different places.

A2 For the least squares estimator $\hat{\boldsymbol{\theta}}$: $\mathrm{E} \parallel \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \parallel^2 = O(n^{-1})$.

A3 The density functions $f(.)$ and $p(.)$ of the random processes $\mathbf{Z}_t$ and $\mathbf{X}_t$, satisfy the Lipschitz condition and the sets $B_1 = \{\mathbf{z} : f(\mathbf{z}) > 0\}$, $B_2 = \{\mathbf{z} : p(\mathbf{z}) > 0\}$ are compact subsets of $\mathbb{R}^Q$ and $\mathbb{R}^P$ respectively.

A4 For the strictly stationary process $\{(Y_t, \mathbf{X}_t, \mathbf{Z}_t) : t = 1, 2, \ldots\}$ define $\beta(n) = \sup_{k\geq 1} \mathrm{E}\{\sup_{B \in \Im_{k+n}^\infty} |P(B|\Im_1^k) - P(B)|\}$ where $\Im_k^n$ the sigma-field generated by $\{(Y_t, \mathbf{X}_t, \mathbf{Z}_t) : k \leq t \leq n\}$. Then $\beta(n) = O(n^{-(2+\delta)/\delta})$ where $0 \leq \delta \leq 2/5$. In addition, there are positive integers m_n and $l_n = [n/(2m_n)]$ such that $\limsup_{n\to\infty}(1 + 6\sqrt{e}\beta(m_n)^{1/(1+l_n)})^{l_n} < \infty$.

A5 The kernel function is $K_h(.) = 1/h^k K(./h)$ where $K : \mathbb{R}^k \to \mathbb{R}$ is a symmetric density function with bounded support satisfying the Lipschitz condition. Further, for the bandwidth $h = n^{-\lambda(k)}$ it holds that $0 < k\lambda(k) < 1/2$ for $1 \leq k \leq Q$.

A6 For m_n defined in A4, $\limsup_{n\to\infty} l_n n^{-\lambda(k)} < \infty$ for all $1 \leq k \leq Q$.

A7 $\mathrm{E}|Y_t|^6 < \infty$, $\mathrm{E}||\mathbf{X}_t||^6 < \infty$ and $\mathrm{E}(Y_t|\mathbf{X}_t, \ldots, \mathbf{X}_1, \mathbf{Z}_t, \ldots, \mathbf{Z}_1) = \mathrm{E}(Y_t|\mathbf{X}_t, \mathbf{Z}_t)$ for $\mathbf{X}_t = (X_{t,1}, \ldots, X_{t,P})^T$ and $\mathbf{Z}_t = (Z_{t,1}, \ldots, Z_{t,Q})^T$.

A8 It holds that $|g(\mathbf{z}_1) - g(\mathbf{z}_2)| \leq C \parallel \mathbf{z}_1 - \mathbf{z}_2 \parallel^\gamma$ for $\gamma > 0$.

A9 $(k + \gamma)\lambda(k) > 1/2$ for γ in A8, and $1 \leq k \leq Q$.

A10 $k\lambda(k)$ is a strictly increasing function of k.

The above assumptions are not the weakest possible and may be altered at the cost of a lengthier proof. Assumption A1, section 4.1, of existence of the true model is a common assumption in the context of regressor's selection while A2 is a standard result of the linear regression theory and requires no further explanation. Further, in A3 we follow Yao and Tong (1994) who point out that in practice any reasonably stationary real data could be considered as bounded set and thus we assume density with bounded support. This is a technical assumption which facilitates the proof and can be relaxed by introducing a weight function in the definition of the cross-validation function. Assumption A4 implies that we are dealing with absolutely regular processes while the assumption on the rate of $\beta(n)$ and A6 allow us to use the results of Yoshihara (1976) and Roussas (1988). Assumptions A5, A7-A8 are self-explanatory, A9 is standard in nonparametric order determination while A10 is essential in the proof of convergence in probability of the CV-estimator. Next, we present some preliminary results that are used in the proof of the main theorem of consistency of the regressors subset estimator.

Lemma 4.1 *Under assumptions A2-A9 it holds that*

(a) For any $A = \{i_1, \ldots, i_k\}$, $1 \leq k \leq q$: $\mathrm{CV}(A) \xrightarrow{P} \sigma^2(A)$.

(b) If for some A *it holds that* $\mathrm{E}(U_t|\mathbf{Z}_t^A) = \mathrm{E}(U_t|\mathbf{Z}_t^{A_0})$ *a.s. then*

$$\mathrm{CV}(A) = \frac{1}{n}\sum_{s=1}^{n}\epsilon_s^2 + \frac{1}{nh^k}\mathrm{E}(\epsilon_t^2/f(\mathbf{Z}_t^A))\int K^2(u)du + o_p(n^{-1}h^{-k}).$$

Proof of Lemma 4.1 Note that: $\mathrm{CV}(A) = n^{-1}\sum_{s=1}^{n}\{\hat{U}_s - U_s\}^2 + n^{-1}\sum_{s=1}^{n}\{U_s - \hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\}^2 + 2n^{-1}\sum_{s=1}^{n}\{\hat{U}_s - U_s\}\{U_s - \hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\} = I_1 + I_2 + I_3$. Assumptions A2-A4 along with Slutsky's theorem yield

$$I_1 = \frac{1}{n}\sum_{s=1}^{n}\{\mathbf{X}_t^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\}^2 = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\frac{1}{n}\sum_{s=1}^{n}\mathbf{X}_t^T\mathbf{X}_t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{P} 0.$$

Call $\epsilon_s^A = U_s - g(\mathbf{Z}_s^A)$, then $I_2 = \sum_{j=1}^{3} I_{2,j}$ with $I_{2,1} = 1/n\sum_{s=1}^{n}(\epsilon_s^A)^2$, $I_{2,2} = 1/n\sum_{s=1}^{n}\{g(\mathbf{Z}_s^A) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\}^2$ and $I_{2,3} = 2/n\sum_{s=1}^{n}\epsilon_s^A\{g(\mathbf{Z}_s^A) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\}$. Using the ergodic theorem (Fan and Yao 2003), we have that

$$I_{2,1} \xrightarrow{P} E[U_t - \mathrm{E}(U_t|\mathbf{Z}_t^A)]^2 = \sigma^2(A). \tag{4.8}$$

For the remaining terms, Lemma 4.2 below shows that they converge in probability to zero and therefore $I_2 \xrightarrow{P} \sigma^2(A)$. Cauchy-Schwartz inequality and the results for I_1, I_2 ensure $I_3 \xrightarrow{P} 0$. Hence, conclude (a). Further, note that if $\mathrm{E}(U_t|\mathbf{Z}_t^A) = \mathrm{E}(U_t|\mathbf{Z}_t^{A_0})$ then $\epsilon_s^A = \epsilon_s$, so $I_{2,1} = 1/n\sum_{s=1}^{n}\epsilon_s^2$. This along with the results in Lemma 4.3 yield (b).

Lemma 4.2 *Suppose A2-A8 hold. Then it follows that,*

(a) $1/n\sum_{s=1}^{n}\epsilon_s^A\{g(\mathbf{Z}_s^A) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\} \xrightarrow{P} 0.$

(b) $1/n\sum_{s=1}^{n}\{g(\mathbf{Z}_s^A) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\}^2 \xrightarrow{P} 0.$

Proof of Lemma 4.2 (a): It holds that

$$n^{-1}\sum_{s}^{n}\epsilon_s^A\{g(\mathbf{Z}_s^A)-\hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\}=n^{-1}(\sum_{s}^{n}\epsilon_s^A\{g(\mathbf{Z}_s^A)-g_n^{(-s)}(\mathbf{Z}_s^A)\}$$

$$+\sum_{s}^{n}\epsilon_s^A\{g_n^{(-s)}(\mathbf{Z}_s^A)-\hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\})=J_1+J_2.$$

Note that $n^{-1}|\sum_{r=1,r\neq s}^{n}K_h(\mathbf{Z}_s^A-\mathbf{Z}_r^A)-\sum_{r=1}^{n}K_h(\mathbf{Z}_s^A-\mathbf{Z}_r^A)|=o_p(n^{-1+k\lambda(k)})=o_p(1)$ and $\sup_{\mathbf{z}:f_k(\mathbf{z})>0}|n^{-1}\sum_{r=1}^{n}K_h(\mathbf{z}-\mathbf{Z}_r^A)-f_k(\mathbf{z})|\to 0$. The latter is implied from assumption A3-A6 and Theorem 3.1 of Roussas (1988). See also Yao and Tong (1994) or Cheng and Tong (1992). Consequently,

$$g(\mathbf{Z}_s^A)-g_n^{(-s)}(\mathbf{Z}_s^A)=\Big(g(\mathbf{Z}_s^A)-g_n^{(-s)}(\mathbf{Z}_s^A)\Big)\Big(\frac{1}{n}\sum_{r=1,r\neq s}^{n}K_h(\mathbf{Z}_s^A-\mathbf{Z}_r^A)\Big)(f_k(\mathbf{Z}_s^A))^{-1}(1+o_p(1)).$$

Further, it holds that $(g(\mathbf{Z}_s^A)-g_n^{(-s)}(\mathbf{Z}_s^A))n^{-1}\sum_{r=1,r\neq s}^{n}K_h(\mathbf{Z}_s^A-\mathbf{Z}_r^A)=$

$$\frac{1}{n}\sum_{t=1,t\neq s}^{n}K_h(\mathbf{Z}_s^A-\mathbf{Z}_t^A)(g(\mathbf{Z}_s^A)-U_t)=\frac{1}{n}\sum_{t=1}^{n}K_h(\mathbf{Z}_s^A-\mathbf{Z}_t^A)(g(\mathbf{Z}_s^A)-g(\mathbf{Z}_t^A))+\frac{1}{n}K_h(0)g(\mathbf{Z}_s^A)$$

$$-\frac{1}{n}\sum_{t=1,t\neq s}^{n}K_h(\mathbf{Z}_s^A-\mathbf{Z}_t^A)(U_t-g(\mathbf{Z}_t^A))=\frac{1}{nh^k}\Big(\sum_{t=1}^{n}C_{s,t}-\sum_{t=1,t\neq s}^{n}\epsilon_t^A d_{s,t}+K(0)g(\mathbf{Z}_s^A)\Big)$$

where $d_{s,t}=K((\mathbf{Z}_s^A-\mathbf{Z}_t^A)/h)$ and $C_{s,t}=d_{s,t}\{g(\mathbf{Z}_s^A)-g(\mathbf{Z}_t^A)\}$. Hence, we write $g(\mathbf{Z}_s^A)-g_n^{(-s)}(\mathbf{Z}_s^A)=(f_k(\mathbf{Z}_s^A)nh^k)^{-1}(\sum_{t=1}^{n}C_{s,t}-\sum_{t=1,t\neq s}^{n}\epsilon_t^A d_{s,t}+K(0)g(\mathbf{Z}_s^A))(1+o_p(1))$ and therefore

$$J_1=n^{-2}h^{-k}\Big(\sum_{s=1}^{n}\sum_{t=1,t\neq s}^{n}\epsilon_s^A(C_{s,t}-\epsilon_t^A d_{s,t})f_k^{-1}(\mathbf{Z}_s^A)$$

$$+K(0)\sum_{s=1}^{n}\epsilon_s^A f_k^{-1}(\mathbf{Z}_s^A)g(\mathbf{Z}_s^A)\Big)(1+o_p(1))=J_{1,1}+J_{1,2}+o_p(J_{1,1}+J_{1,2}).$$

From the ergodic theorem $nh^kJ_{1,2}\xrightarrow{P}\mathrm{E}(\epsilon_s^A f_k^{-1}(\mathbf{Z}_s^A)g(\mathbf{Z}_s^A))$ and using conditional argument and the fact that $nh^k\to\infty$, it follows that $J_{1,2}\xrightarrow{P}0$. For $J_{1,1}$ we follow Yao and Tong (1994) and apply the decomposition of U-statistic as proposed by Yoshihara

(1976). The latter has already been employed in Chapter 3. Under certain conditions, Yoshihara showed that a U-statistics $U_n = 2!(n-2)!/n! \sum_i^n g(\xi_{i_1}, \xi_{i_2})$ which is an estimator of a functional form $\vartheta(F) = \int_{R^{2p}} g(x_1, x_2) dF(x_1) dF(x_2)$, can be decomposed as

$$\vartheta(F) + \sum_{1 \le i_1 < i_2 \le n} \{g(\xi_{i_1}, \xi_{i_2}) - \int g(\xi_{i_1}, \xi_{i_2}) dF(\xi_{i_2}) - \int g(\xi_{i_1}, \xi_{i_2}) dF(\xi_{i_1}) + \vartheta(F)\}.$$

Call $\eta_t = (\mathbf{Z}_t^A, \epsilon_s^A)^T$ and define $H(\eta_t, \eta_s) = \epsilon_s^A \Big(C_{s,t} - \epsilon_t^A d_{s,t}\Big) f_k^{-1}(\mathbf{Z}_s^A) + \epsilon_t^A \Big(C_{t,s} - \epsilon_s^A d_{t,s}\Big) f_k^{-1}(\mathbf{Z}_t^A)$ and $H(\eta_t) = \int H(\eta_t, \eta_s) dP(\eta_s) = \epsilon_t^A f_k^{-1}(\mathbf{Z}_t^A) \int C_{t,s} dP(\mathbf{Z}_s^A)$ then,

$$J_{1,1} = \frac{1}{2n^2 h^k} \sum_{t \neq s}^{n} \{H(\eta_t, \eta_s) - H(\eta_s) - H(\eta_t)\} + \frac{n-1}{n^2 h^k} \sum_{t=1}^{n} H(\eta_t).$$

$H(\eta_s, \eta_t)$ is symmetric and η_t is a strictly stationary absolutely regular process. Hence, applying the above results for $g(\xi_{i_1}, \xi_{i_2}) = H(\eta_s, \eta_t)$ and noticing that $\int H(\eta_t) dP(\eta_t) = 0 \Rightarrow \vartheta(F) = 0$ we have that the first term of $J_{1,1}$ is equal to the remainder in Hoeffding's projection decomposition of the U-statistics. Further, A3 yields $|H(\eta_t, \eta_s)| \le C\Big(|\epsilon_s^A| + |\epsilon_t^A|\Big) + C|\epsilon_s^A||\epsilon_t^A|$ and from A7 $\int |H(\eta_t, \eta_s)|^3 dP(\eta_t) dP(\eta_s) < \infty$. Therefore, Lemma 1 of Yoshihara (1976) yields $\sup_{s<t} \mathrm{E}|H(\eta_t, \eta_s)|^3 < \infty$. The latter ensures that Lemma 2 (Yoshihara 1976) holds, implying that $\mathrm{E}(n^{-2} \sum_{t \neq s}^{n} \{H(\eta_t, \eta_s) - H(\eta_s) - H(\eta_t)\})^2 = o(n^{-2})$ thus $n^{-2} h^{-k} \sum_{t \neq s}^{n} \{H(\eta_t, \eta_s) - H(\eta_s) - H(\eta_t)\} = o_p(n^{-1} h^{-k})$.

For the second term note that A3, A8 yield $|H(\eta_t)| \le C h^{k+\gamma} |\epsilon_t^A|$ almost surely, so $(n-1)/(n^2 h^k) |\sum_{t=1}^{n} H(\eta_t)| \le C h^{\gamma} n^{-1} \sum_{t=1}^{n} |\epsilon_t^A|$ which converges to zero by the ergodic theorem. Summing up, we conclude $J_{1,1} = o_p(1)$ which completes the proof of $J_1 = o_p(1)$. It remains to prove $J_2 \xrightarrow{P} 0$. From assumption A3 it follows that $\mathbf{X}_t$ is bounded process i.e. $\| \mathbf{X}_t \| \le M < \infty$. Consequently, from

$$\frac{1}{n} \Big| \sum_{s=1}^{n} \epsilon_s^A \Big(\sum_{t=1, t \neq s}^{n} w_{t,A}^{(-s)}(\mathbf{Z}_s^A) \mathbf{X}_t^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \Big) \Big| \le M \| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \| \frac{1}{n} \sum_{s=1}^{n} |\epsilon_s^A|$$

ergodic theorem and A2 yield $J_2 \xrightarrow{P} 0$. The proof of (b) contains similar technical details with the proof of (a) so is omitted.

Lemma 4.3 *Suppose A2-A9 hold and let for some* $\mathbf{Z}_t^A$,

$$\mathrm{E}(Y_t|\mathbf{X}_t, \mathbf{Z}_t^A) = \mathrm{E}(Y_t|\mathbf{X}_t, \mathbf{Z}_t^{A_0}) \quad a.s. \tag{4.9}$$

(a) $1/n\sum_{s=1}^n \epsilon_s^A\{g(\mathbf{Z}_s^A) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\} = o_p(n^{-1}h^{-k})$.

(b) $1/n\sum_{s=1}^n\{g(\mathbf{Z}_s^A) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\}^2 = n^{-1}h^{-k}\mathrm{E}(\epsilon_t^2/f(\mathbf{Z}_t^A))\int K^2(u)du + o_p(n^{-1}h^{-k})$.

Proof of Lemma 4.3 (a): For $J_{1,1}$ we have that $n^{-2}h^{-k}\sum_{t\neq s}^n\{H(\eta_t,\eta_s) - H(\eta_s) - H(\eta_t)\} = o_p(n^{-1}h^{-k})$. Under condition (4.9) we have that $\mathrm{E}(H(\eta_t)H(\eta_s)) = 0$ for $s < t$, the latter from $\mathrm{E}(H(\eta_t)|\mathbf{X}_t,\ldots,\mathbf{X}_1,\mathbf{Z}_t,\ldots,\mathbf{Z}_1) = C(\mathbf{Z}_t)\mathrm{E}(\epsilon_t|\mathbf{X}_t,\mathbf{Z}_t) = 0$. Hence, $\mathrm{E}(n^{-1}\sum_{t=1}^n H(\eta_t))^2 = n^{-1}\mathrm{E}(H^2(\eta_t)) \leq Ch^{2k+2\gamma}n^{-1}$ which along with A9 yields

$$\frac{n-1}{n^2h^k}\sum_{t=1}^n H(\eta_t) = O_p\Big(\frac{n-1}{nh^k}(\frac{h^{2k+2\gamma}}{n})^{1/2}\Big) = O_p(h^\gamma n^{-1/2}) = o_p(n^{-1}h^{-k})$$

and combined with $J_{1,2} = o_p(n^{-1}h^{-k})$ from Lemma 4.2, implies that $J_1 = o_p(n^{-1}h^{-k})$. Similarly, from assumption A7, $\mathrm{E}(\epsilon_t\epsilon_s) = 0$ so $\mathrm{E}|J_2|^2 \leq n^{-1}M^2\mathrm{E}\parallel\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\parallel^2\mathrm{E}(\epsilon_t^2) \Rightarrow \mathrm{E}|J_2|^2 = O(n^{-2})$ from A2. Therefore, $J_2 = O_p(\mathrm{E}|J_2|^2)^{1/2} = O_p(n^{-1}) = o_p(h^{-k}n^{-1})$ so conclude. For (b):

$$\frac{1}{n}\sum_{s=1}^n\{g(\mathbf{Z}_s^A)-\hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\}^2 = \frac{1}{n}\sum_{s=1}^n\{g(\mathbf{Z}_s^A)-g_n^{(-s)}(\mathbf{Z}_s^A)\}^2+\frac{1}{n}\sum_{s=1}^n\{g_n^{(-s)}(\mathbf{Z}_s^A)-\hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\}^2$$

$$+\frac{2}{n}\sum_{s=1}^n\{g(\mathbf{Z}_s^A) - g_n^{(-s)}(\mathbf{Z}_s^A)\}\{g_n^{(-s)}(\mathbf{Z}_s^A) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\} = \sum_{j=1}^3 R_j.$$

Concentrate on R_1. Recall the notation from Lemma 4.2, then it holds that

$$R_1 = \frac{1}{n^3h^{2k}}\sum_{s=1}^n\sum_{t=1,t\neq s}^n (C_{s,t} - \epsilon_t d_{s,t})^2 f^{-2}(\mathbf{Z}_s^A)+$$

$$+\frac{1}{n^3h^{2k}}\sum_{s=1}^n\sum_{t_1,t_2=1,t_1\neq t_2\neq s}^n (C_{s,t_1} - \epsilon_{t_1}d_{s,t_1})(C_{s,t_2} - \epsilon_{t_2}d_{s,t_2})f^{-2}(\mathbf{Z}_s^A)$$

$$+\frac{2}{n^3h^{2k}}K^2(0)\sum_{s=1}^n\sum_{t=1,t\neq s}^n g(\mathbf{Z}_s^A)(C_{s,t} - \epsilon_t d_{s,t})f^{-2}(\mathbf{Z}_s^A)$$

$$+\frac{1}{n^3h^{2k}}K^2(0)\sum_{s=1}^{n} g^2(\mathbf{Z}_s^A)f^{-2}(\mathbf{Z}_s^A) + o_p(R_1) = \sum_{j=1}^{4} R_{1,j}(1+o_p(1)).$$

Under (4.9), the process η_t defined in Lemma 4.2 is written as $\eta_t = (\mathbf{Z}_t^A, \epsilon_t)^T$ and let

$$H_1(\eta_t,\eta_s) = (C_{s,t}^2 - 2C_{s,t}\epsilon_t d_{s,t} + \epsilon_t^2 d_{s,t}^2)f^{-2}(\mathbf{Z}_s^A) + (C_{t,s}^2 - 2C_{t,s}\epsilon_s d_{t,s} + \epsilon_s^2 d_{t,s}^2)f^{-2}(\mathbf{Z}_t^A),$$

$$H_1(\eta_t) = \int C_{s,t}^2 f^{-2}(\mathbf{Z}_s^A)dP(\mathbf{Z}_s^A) - 2\epsilon_t \int C_{s,t}d_{s,t}f^{-2}(\mathbf{Z}_s^A)dP(\mathbf{Z}_s^A)$$

$$+\epsilon_t^2 \int d_{s,t}^2 f^{-2}(\mathbf{Z}_s^A)dP(\mathbf{Z}_s^A) + f^{-2}(\mathbf{Z}_t^A)\int C_{t,s}^2 dP(\mathbf{Z}_s^A) + f^{-2}(\mathbf{Z}_t^A)\int \epsilon_s^2 d_{t,s}^2 dP(\eta_s).$$

Thus, it holds that,

$$R_{1,1} = \frac{1}{2n^3h^{2k}}\sum_{t\neq s}^{n}\{H_1(\eta_t,\eta_s) - H_1(\eta_t) - H_1(\eta_s) - \int H_1(\eta_t)dP(\eta_t)\}$$

$$+\frac{1}{2nh^{2k}}\int H_1(\eta_t)dP(\eta_t) + \frac{n-1}{n^3h^{2k}}\sum_{t=1}^{n} H_1(\eta_t).$$

Like in Lemma 4.2 the first term can be interpreted as the remainder in Hoeffding's projection decomposition of the U-statistics generated by $H_1(\eta_t,\eta_s)$ and using similar arguments we show that

$$\frac{1}{2n^3h^{2k}}\sum_{t\neq s}^{n}\{H_1(\eta_t,\eta_s) - H_1(\eta_t) - H_1(\eta_s) - \int H_1(\eta_t)dP(\eta_t)\} = o_p(n^{-1}h^{-k}). \quad (4.10)$$

For the second term note that $2^{-1}n^{-1}h^{-2k}\int H_1(\eta_t)dP(\eta_t) =$

$$\frac{1}{nh^{2k}}\int C_{s,t}^2 f^{-2}(\mathbf{Z}_t^A)dP(\mathbf{Z}_s^A)dP(\mathbf{Z}_t^A) + \frac{1}{nh^{2k}}\int \epsilon_t^2 d_{s,t}^2 f^{-2}(\mathbf{Z}_s^A)dP(\mathbf{Z}_s^A)dP(\eta_t)$$

but from A3 and A8-A9 the first part is $O(h^{2\gamma+2k}n^{-1}h^{-2k}) = o(n^{-1}h^{-k})$ while the second part is

$$\frac{1}{nh^{2k}}\int \epsilon_t^2 d_{s,t}^2 f^{-2}(\mathbf{Z}_s^A)dP(\mathbf{Z}_s^A)dP(\eta_t) \sim n^{-1}h^{-k}\mathrm{E}(\epsilon_t^2/f(\mathbf{Z}_t^A))\int K^2(u)du$$

where with $\sim$ we denote the fact that they are asymptotically equivalent. Thus

$$\frac{1}{2nh^{2k}}\int H_1(\eta_t)dP(\eta_t) = n^{-1}h^{-k}\mathrm{E}(\epsilon_t^2/f(\mathbf{Z}_t^A))\int K^2(u)du + o_p(n^{-1}h^{-k}). \quad (4.11)$$

For the third term of $R_{1,1}$ note that

$$E\Big(\frac{1}{n}\sum_{t=1}^{n} H_1(\eta_t)\Big)^2 = E\Big(\frac{1}{n^2}\sum_{t=1}^{n} H_1^2(\eta_t)\Big) + E\Big(\frac{2}{n}\sum_{t=1}^{n-1}(1-\frac{t}{n})H_1(\eta_1)H_1(\eta_{t+1})\Big)$$

$$= \frac{1}{n}\mathrm{E}(H_1^2(\eta_t)) + \mathrm{E}(\frac{2}{n}\sum_{t=1}^{n-1}(1-\frac{t}{n})H_1(\eta_1)H_1(\eta_{t+1})) = O(n^{-1}h^{4k+2\gamma})$$

the latter from A8. Hence

$$\frac{n-1}{n^3h^{2k}}\sum_{t=1}^{n} H_1(\eta_t) = O_p(n^{-3/2}h^{2\gamma-k}) = o_p(n^{-1}h^{-k}). \tag{4.12}$$

Combine (4.10),(4.11) and (4.12) to get $R_{1,1} = n^{-1}h^{-k}\mathrm{E}(\epsilon_t^2/f(\mathbf{Z}_t^A))\int K^2(u)du + o_p(n^{-1}h^{-k})$. Using similar arguments of the decomposition of U-statistics we can show that $R_{1,j} = o_p(n^{-1}h^{-k})$ for $j = 2,3$ while the ergodic theorem yields $R_{1,4} = o_p(n^{-1}h^{-k})$. Summing up,

$$R_1 = n^{-1}h^{-k}\mathrm{E}(\epsilon_t^2/f(\mathbf{Z}_t^A))\int K^2(u)du + o_p(n^{-1}h^{-k}). \tag{4.13}$$

Note that $R_2 = n^{-1}\sum_{s=1}^{n}(\sum_{t=1,t\neq s}^{n} w_{t,A}^{(-s)}(\mathbf{Z}_s^A)\mathbf{X}_t^T(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}))^2$ thus $|R_2| \leq C \parallel \hat{\boldsymbol{\theta}}-\boldsymbol{\theta} \parallel^2 n^{-1}\sum_{s=1}^{n}|\sum_{t=1,t\neq s}^{n} w_{t,A}^{(-s)}(\mathbf{Z}_s^A)|^2$ and since $\sum_{t=1,t\neq s}^{n} w_{t,A}^{(-s)}(\mathbf{Z}_s^A) = O_p(1)$ assumption A2 yields $\mathrm{E}|R_2| = O(n^{-1}) \Rightarrow$

$$R_2 = O_p(n^{-1}) = o_p(n^{-1}h^{-k}). \tag{4.14}$$

Cauchy-Schwartz inequality yields $\mathrm{E}|R_3| \leq (\mathrm{E}(R_1))^{1/2}(\mathrm{E}(R_2))^{1/2}$ and therefore $\mathrm{E}|R_3| = O(h^{-k/2}n^{-1/2})o(h^{-k/2}n^{-1/2}) = o(n^{-1}h^{-k})$, so

$$R_3 = O_p(\mathrm{E}|R_3|) = o_p(n^{-1}h^{-k}) \tag{4.15}$$

and the proof of (b) concludes from (4.13),(4.14) and (4.15).

We now state the main theorem that proves the consistency of the proposed leave-one-out cross validation criterion.

Theorem 4.1 *Under assumptions A1-A10, it holds that*

$$\lim_{n\to\infty} P(\hat{A} = A_0) = 1.$$

Proof of Theorem 4.1 For any $A \subset \{1,\ldots,Q\}$, if $\sigma^2(A) > \sigma^2(1,\ldots,Q) = \sigma^2(A_0)$ then from Lemma 4.1 it follows that $P\Big(\mathrm{CV}(A_0) < \mathrm{CV}(A)\Big) \to 1$. Alternatively, if $\sigma^2(A) = \sigma^2(1,\ldots,Q) = \sigma^2(A_0)$ then condition (4.9) in Lemma 4.3 holds. Note also that $k > q$ by Definition 4.1. Hence, from A10,

$$h^q/h^k = n^{k\lambda(k)-q\lambda(q)} \to \infty \text{ as } n \to \infty \tag{4.16}$$

thus Lemma 4.1 (b) along with (4.16) yield:

$$P\Big(nh^q(\mathrm{CV}(A) - \mathrm{CV}(A_0)) > 0\Big)$$

$$= P\Big(\int K^2(u)du\{\frac{h^q}{h^k}\mathrm{E}(\epsilon_t^2/f(\mathbf{Z}_t^A)) - \mathrm{E}(\epsilon_t^2/f(\mathbf{Z}_t^{A_0}))\} + o_p(\frac{h^q}{h^k}) > 0\Big) \to 1$$

$\Rightarrow P(\hat{A} = A_0) \to 1$ as $n \to \infty$ that completes the proof**.**

The calculation of the residuals requires regressing over the full linear regressors set, even if some of the linear regressors are insignificant, in order to make sure that there is no linear contribution left on the residuals. As it appears from the result of Theorem 4.1, the inclusion of insignificant linear regressors in the calculation of the residuals U_t does not affect the asymptotic property of consistency for the cross validation criterion. In fact, the main terms of the decomposition of the CV-function are the same as those derived in Yao and Tong (1994) for a fully nonparametric model. It is therefore understood that, having removed the parametric component's contribution, the selection of the nonparametric component is performed on the residuals as a standard nonparametric variable selection, assuming that A1-A10 hold.

4.3 Selection of parametric component

The second step of the proposed procedure is the selection of the parametric regressors. For any $M \subseteq \{1, \ldots, P\}$, we write

$$Y_t = (\mathbf{X}_t^M)^T \boldsymbol{\theta}_M + g(\mathbf{Z}_t^{A_0}) + \epsilon_{t,M} \tag{4.17}$$

where $\mathbf{X}_t^M = (X_{t,i} : i \in M)^T$ and $\epsilon_{t,M} = Y_t - E(Y_t|\mathbf{X}_t^M, \mathbf{Z}_t^{A_0})$. We denote this model as $\mathcal{M}_M$. To this end, we classify the models $\mathcal{M}_M$ into two groups

Category I: at least one nonzero component of $\boldsymbol{\theta}$ is not in $\boldsymbol{\theta}_M$,

Category II: $\boldsymbol{\theta}_M$ contains all the nonzero components of $\boldsymbol{\theta}$.

In the first category, we have models $\mathcal{M}_M$ that are incorrect in the sense that they do not include all the significant regressors. Models in the second category include all the significant regressors but they may include regressors unrelated to the response variable. Substituting $g(.)$ with the nonparametric estimator $g_n(\mathbf{Z}_t^{A_0}) = \sum_{s=1}^n w_{s,A_0}(\mathbf{Z}_t^{A_0})(Y_s - \mathbf{X}_s^T\boldsymbol{\theta})$ yields that the least squares estimator of $\boldsymbol{\theta}_M$ is

$$\hat{\boldsymbol{\theta}}_M = (\tilde{\mathbf{X}}_M^T \tilde{\mathbf{X}}_M)^{-1} \tilde{\mathbf{X}}_M^T \tilde{\mathbf{Y}} \tag{4.18}$$

where $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^T$, $\tilde{Y}_t = Y_t - \sum_{s=1}^n w_{s,A_0}(\mathbf{Z}_t^{A_0}) Y_s$ and $\tilde{\mathbf{X}}_M = (\tilde{\mathbf{X}}_{1,M}, \ldots, \tilde{\mathbf{X}}_{n,M})^T$ with $\tilde{\mathbf{X}}_t^M = \mathbf{X}_t^M - \sum_{s=1}^n w_{s,A_0}(\mathbf{Z}_t^{A_0}) \mathbf{X}_s^M$. Note that for model $\mathcal{M}_M$ the mean square prediction error is given by

$$\mathrm{MSE}_n(M) = \frac{1}{n}\tilde{\epsilon}^T\tilde{\epsilon} - \frac{1}{n}\tilde{\epsilon}^T \mathbf{P}_M \tilde{\epsilon} + \frac{1}{n}\boldsymbol{\theta}^T \tilde{\mathbf{X}}^T \mathbf{H}_M \tilde{\mathbf{X}} \boldsymbol{\theta} + \frac{2}{n}\tilde{\epsilon}^T \mathbf{H}_M \tilde{\mathbf{X}} \boldsymbol{\theta} \tag{4.19}$$

where $\tilde{\epsilon} = (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n)$, $\tilde{\epsilon}_t = \tilde{Y}_t - \tilde{\mathbf{X}}_t^T\boldsymbol{\theta}$, $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_n)^T$, $\mathbf{P}_M = \tilde{\mathbf{X}}_M(\tilde{\mathbf{X}}_M^T\tilde{\mathbf{X}}_M)^{-1}\tilde{\mathbf{X}}_M^T$ and $\mathbf{H}_M = I_n - \mathbf{P}_M$. By definition, $\tilde{\epsilon}_t = \tilde{Y}_t - \tilde{\mathbf{X}}_t^T\boldsymbol{\theta} = \epsilon_t - \sum_{s=1}^n w_{s,A_0}(\mathbf{Z}^{A_0})\epsilon_s$ and using Lemma 4.4 below, with $p = 1 - k\lambda(k)$ and $r = 6$ (see A5, A7) we conclude that $\max_{1\le t\le n} |\sum_{s=1}^n w_{s,A_0}(\mathbf{Z}^{A_0})\epsilon_s| = o(n^{k\lambda(k)-1/2}) = o(1)$. Consequently, the nonparametric term does not affect the rate of convergence of the mean square prediction

error under the assumptions imposed. Surprising as it may look, similar conclusion has been observed by Speckman (1988) who noticed that for a certain choice of bandwidth, the parametric estimator $\hat{\boldsymbol{\theta}}$ remains a $\sqrt{n}$-consistent estimator, see also Härdle, Liang, and Gao (2000). It follows from (4.19) and assumption A4 that the conditional expected mean square error is given by

$$\frac{1}{n}\mathrm{E}(\tilde{\epsilon}^T\tilde{\epsilon}|\tilde{\mathbf{X}}) - \frac{1}{n}\mathrm{E}(\tilde{\epsilon}^T P_M\tilde{\epsilon}|\tilde{\mathbf{X}}) + \frac{1}{n}\mathrm{E}(\boldsymbol{\theta}^T\tilde{\mathbf{X}}^T\mathbf{H}_M\tilde{\mathbf{X}}\boldsymbol{\theta}|\tilde{\mathbf{X}}) + \frac{2}{n}\mathrm{E}(\tilde{\epsilon}^T\mathbf{H}_M\tilde{\mathbf{X}}\boldsymbol{\theta}|\tilde{X}) \Rightarrow$$

$$\mathrm{EMSE}_n(M) = \sigma^2_{\tilde{\epsilon}} - \frac{m}{n}\sigma^2_{\tilde{\epsilon}} + \Omega_{n,M} \quad \text{a.s.} \tag{4.20}$$

where $\sigma^2_{\tilde{\epsilon}} = n^{-1}E(\tilde{\epsilon}^T\tilde{\epsilon})$ and $\Omega_{n,M} = n^{-1}\boldsymbol{\theta}^T\tilde{\mathbf{X}}^T\mathbf{H}_M\tilde{\mathbf{X}}\boldsymbol{\theta}$. Note that for every $M \subseteq \{1,\ldots,P\}$ with $\mathcal{M}_M$ from category II, it follows that

$$\mathrm{MSE}_n(M) = \frac{1}{n}\tilde{\epsilon}^T\tilde{\epsilon} - \frac{1}{n}\tilde{\epsilon}^T\mathbf{P}_M\tilde{\epsilon} + \frac{2}{n}\tilde{\epsilon}^T\mathbf{H}_M\tilde{\mathbf{X}}\boldsymbol{\theta} \quad \text{and}$$

$$\mathrm{EMSE}_n(M) = \frac{1}{n}(n-m)\sigma^2_{\tilde{\epsilon}}$$

the latter from the fact that in category II $\tilde{\mathbf{X}}\boldsymbol{\theta} = \tilde{\mathbf{X}}_M\boldsymbol{\theta}_M$. In addition to assumptions A1-A10, we require that

B1 For every $M \subseteq \{1,\ldots,P\}$

(i) $\mathrm{E}(\tilde{\mathbf{X}}_M^T\tilde{\mathbf{X}}_M)$ is a positive definite matrix with order $m \times m$.

(ii) If $\mathcal{M}_M$ in category I it holds: $\liminf_{n\to\infty}\Omega_{n,M} > 0$ in probability.

Assumption (i) is necessary for the consistency of $\hat{\boldsymbol{\theta}}_M$, (see Härdle, Liang, and Gao 2000). Assumption (ii) is an identifiability condition which is a very minimal argument for asymptotic analysis. Further let $d_j(\mathbf{z}) = \mathrm{E}(X_{t,j}|\mathbf{Z}_t^{A_0} = \mathbf{z})$ and define $u_{t,j} = X_{t,j} - d_j(\mathbf{Z}_t^{A_0})$ for $j = 1,\ldots,P$ and $\mathbf{u}_t = (u_{t,1},\ldots,u_{t,P})^T$. Gao and Tong (2002) show that B1(ii) can be replaced by $\liminf_{n\to\infty} n^{-1}(\mathbf{u}\boldsymbol{\theta})^T(I_n - \mathbf{u}_M(\mathbf{u}_M^T\mathbf{u}_M)^{-1}\mathbf{u}_M^T)(\mathbf{u}\boldsymbol{\theta}) > 0$ where $\mathbf{u} = (\mathbf{u}_1,\ldots,\mathbf{u}_n)^T$, $\mathbf{u}_M = (\mathbf{u}_{1,M},\ldots,\mathbf{u}_{n,M})^T$, $\mathbf{u}_t = \mathbf{X}_t - \mathrm{E}(\mathbf{X}_t|\mathbf{Z}_t^{A_0})$ and $\mathbf{u}_{t,M} = \mathbf{X}_{t,M} - \mathrm{E}(\mathbf{X}_{t,M}|\mathbf{Z}_t^{A_0})$ an extension of the identifiability condition 2.5 in Shao (1993) to the partial linear context.

We split the data into two parts: $\{(\tilde{Y}_t, \tilde{\mathbf{X}}_t) : t \in N\}$ and $\{(\tilde{Y}_t, \tilde{\mathbf{X}}_t) : t \in N^c\}$ where $N \subseteq \{1, \ldots, n\}$ and N^c is its compliment. Hence, if we call n_v and n_c the size of N and N^c respectively then $n_v + n_c = n$. Model $\mathcal{M}_M$ is fitted using the sub-sample N^c which is called the construction data while the prediction error is calculated using N called the validation data. Then, the leave-n_v-out cross validation function is defined by $\text{CV}(M, n_v) = n_v^{-1} \parallel \tilde{Y}_N - \tilde{\mathbf{X}}_{N,M}\hat{\boldsymbol{\theta}}_{N^c,M} \parallel^2$. The simplest case would have been the leave-one-out cross validation. However, it has been shown that the leave-one-out cross validation criterion yields an asymptotically inconsistent estimator, see Shao (1993). On the other hand, for n large there are too many possible sub-samples which is computationally inconvenient. A good compromise is to use the Monte Carlo-CV(n_v). We randomly draw a collection $\mathcal{B}$ of b subsets of $\{1, \ldots, n\}$ each one with size n_v and we choose the model that minimizes

$$\text{MCCV}(M, n_v) = \frac{1}{b}\sum_{N \in \mathcal{B}} \text{CV}(M, n_v) = \frac{1}{bn_v}\sum_{N \in \mathcal{B}} \parallel \tilde{Y}_N - \tilde{\mathbf{X}}_{N,M}\hat{\boldsymbol{\theta}}_{N^c,M} \parallel^2 . \qquad (4.21)$$

Definition 4.4 *The estimator for the optimal regression subset of the linear component is defined as*

$$\hat{\mathcal{M}}_M = \arg \min_{M \subseteq \{1,\ldots,P\}} \text{MCCV}(M, n_v). \qquad (4.22)$$

The consistency of the estimator of the optimal parametric regressors subset is entailed in the next theorem. But first we need to impose an additional assumption:

B2 When $n \to \infty$ it holds that $n_v/n \to 1$, $n_c = n - n_v \to \infty$ and $n^2/(n_c^2 b) \to 0$.

Theorem 4.2 *Assume that A1-A10 and B1-B2 hold, then*

(1) If $\mathcal{M}_M$ is in category I, then there exists $R_n \geq 0$ such that

$$\text{MCCV}(M, n_v) = \frac{1}{b}\sum_{N \in \mathcal{B}} \tilde{\epsilon}_N^T \tilde{\epsilon}_N + \Omega_{n,M} + R_n + o_p(1)$$

where $\tilde{\epsilon}_N = \tilde{Y}_N - \tilde{\mathbf{X}}_N \boldsymbol{\theta}$.

(2) If $\mathcal{M}_M$ in category II, then

$$\text{MCCV}(M, n_v) = \frac{1}{b}\sum_{N\in\mathcal{B}} \tilde{\epsilon}_N^T \tilde{\epsilon}_N + \frac{m}{n_c}\sigma_{\tilde{\epsilon}}^2 + o_p(n_c^{-1}).$$

(3) Combining (1) and (2) we conclude

$$\lim_{n\to\infty} P(\hat{\mathcal{M}}_M = \mathcal{M}_{M_0}) = 1.$$

Proof of Theorem 4.2 The proof is based on Theorem 2 in Shao (1993). Similar results for the partial linear model can be found in Theorem 2.2 (Gao and Tong 2002). Hence we only present an outline of the proof and particularly we show that conditions in Theorem 2 (Shao 1993) hold. Indeed condition 2.5, 3.12 and 3.22 have been introduced in B1 and B2. Therefore, it remains to show

$$\max_{N\in\mathcal{B}} \| \frac{1}{n_v}\sum_{t\in N} \tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T - \frac{1}{n_c}\sum_{t\in N^c} \tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T \| = o_p(1) \tag{4.23}$$

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = O_p(n),\ (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1} = O_p(n^{-1}) \text{ and} \tag{4.24}$$

$$\lim_{n\to\infty} \max p_{t,M} = 0 \text{ for all } M \tag{4.25}$$

where $p_{t,M}$ is the t-th diagonal element of $\mathbf{P}_M$. Lemma 4.4 and 4.5 below establish the above conditions that entail the proof of the theorem.

The following lemma is an extension of Lemma A.3 in Härdle, Liang, and Gao (2000) for α-mixing processes. It is a novel result on the uniform rates of the weighted sum of α-mixing random variables and can be used independently of the context of variable selection.

Lemma 4.4 *Let X_i, $i = 1,\ldots,n$ be zero mean, strictly stationary, α-mixing, real valued random variables. Let the mixing coefficients follow $t^6\alpha(t) \to 0$. Suppose that $\sup_{1\le i\le n} \mathrm{E}|X_i|^r < C < \infty$ for $r > 2$ and denote with $\alpha_{i,j}$ with $i,j = 1,\ldots,n$ a sequence of positive numbers such that $\sup_{1\le i,j\le n} |\alpha_{i,j}| \le n^{-p}$ for some $0 < p < 1$. Then it holds that*

$$\max_{1\le j\le n} |\sum_{i=1}^{n} \alpha_{i,j}X_i| = O_p(n^{-p+1/3+1/r}\log n).$$

Proof of Lemma 4.4 Define $X_i' = X_i I(|X_i| \leq n^{1/r})$ and $X_i'' = X_i - X_i'$. Note that $\sup_{1\leq i\leq n} |\alpha_{i,j} X_i'| < Cn^{-p} n^{1/r} \equiv M$. The exponential-type inequality in Theorem 1.3 (Bosq 1998) with $\varepsilon = Cn^{-p-2/3+1/r} \log n$ and $q = n^{2/3}$ yields:

$$P\Big(\max_{1\leq j\leq n} |\sum_{i=1}^{n} \alpha_{i,j}(X_i' - EX_i')| > n\varepsilon \Big) \leq \sum_{j=1}^{n} P\Big(|\sum_{i=1}^{n} \alpha_{i,j}(X_i' - \mathrm{E}(X_i'))| > n\varepsilon \Big)$$

$$\leq 4n \exp\left(-\frac{\varepsilon^2 q}{8v^2(q)} \right) + 22n^{1+2/3}(1 + \frac{4M}{\varepsilon})^{1/2} \alpha([n^{1/3}/2])$$

where with $\alpha(k)$ we denote the mixing coefficient and

$$v^2(q) \leq \frac{8}{n^{2/3}} \{ \max_{0\leq t\leq n} \mathrm{E}(\alpha_{i,j}(X_i' - EX_i'))^2 + 8M^2 \sum_{k=1}^{[n^{2/3}]+1} \alpha(k) \} + \frac{M\varepsilon}{2}.$$

It holds that $v^2(q) \leq CM^2 n^{-2/3} + 2^{-1} M\varepsilon \leq Cn^{-2p+2/r-2/3}$ thus,

$$P\Big(\max_{1\leq j\leq n} |\sum_{i=1}^{n} \alpha_{i,j}(X_i' - EX_i')| > n\varepsilon \Big) \leq 4n \exp\Big(- \frac{Cn^{-2p-2+4/3+2/r} \log^2 n}{n^{-2p-2/3+2/r}} \Big)$$

$$+22n^{5/3}(1 + \frac{4Cn^{-p+1/r}}{n^{-p-2/3+1/r} \log n})^{1/2} \alpha([n^{1/3}/2]) \leq 4n \exp\left(-C \log^2 n\right)$$

$$+22n^2(n^{-2/3} + 4C \log^{-1} n)^{1/2} \alpha([n^{1/3}/2]) \leq n^{1-C\log n} + C_2 n^2 \alpha(n^{1/3}) \to 0$$

given that $t^6 \alpha(t) \to 0$ when $t \to \infty$. Hence, it follows that

$$\max_{1\leq j\leq n} |\sum_{i=1}^{n} \alpha_{i,j}(X_i' - EX_i')| = O_p(n^{-p+1/r+1/3} \log n). \tag{4.26}$$

For the second process X_i'', ergodic theorem yields

$$\frac{1}{n} \sum_{i=1}^{n} \Big(|X_i'' - EX_i''|^l - \mathrm{E}|X_i'' - EX_i''|^l \Big) \xrightarrow{a.s.} 0. \tag{4.27}$$

Note that $X_i'' = X_i - X_i' = X_i - X_i I(|X_i| \leq n^{1/r}) = X_i I(|X_i| \geq n^{1/r})$ and $\mathrm{E}|X_i''|^l = \mathrm{E}(|X_i|^l I(|X_i| \geq n^{1/r})) \leq \Big(\mathrm{E}|X_i|^r \Big)^{l/r} \Big(\mathrm{E}(I(|X_i| \geq n^{1/r})) \Big)^{1-l/r}$

$$= \Big(\mathrm{E}|X_i|^r \Big)^{l/r} \Big(P(|X_i| \geq n^{1/r}) \Big)^{1-l/r} \leq \Big(\mathrm{E}|X_i|^r \Big)^{l/r} \Big(\frac{\mathrm{E}|X_i|^r}{n} \Big)^{1-l/r}$$

the latter from Markov inequality. Consequently, $\mathrm{E}|X_i''|^l \leq \mathrm{E}|X_i|^r n^{l/r-1}$. Therefore, $\mathrm{E}|X_i'' - EX_i''|^l \leq C\mathrm{E}|X_i''|^l \leq C\mathrm{E}|X_i|^r n^{l/r-1} \leq Cn^{l/r-1}$ which along with (4.27) yields $\sum_{i=1}^n |X_i'' - EX_i''|^l \leq Cn^{l/r}$. But Hölder's inequality (with m,l such that $1/m \leq 1/3$ and $1/m + 1/l = 1$) implies that

$$\max_{1\leq j\leq n} |\sum_{i=1}^n \alpha_{i,j}(X_i'' - EX_i'')| \leq \max_{1\leq j\leq n} \Big(\sum_{i=1}^n |\alpha_{i,j}|^m\Big)^{\frac{1}{m}} \Big(\sum_{i=1}^n |X_i'' - EX_i''|^l\Big)^{\frac{1}{l}}$$

$$\leq Cn^{-p+1/m}\Big(\sum_{i=1}^n |X_i'' - EX_i''|^l\Big)^{\frac{1}{l}} \Rightarrow$$

$$\max_{1\leq j\leq n} |\sum_{i=1}^n \alpha_{i,j}(X_i'' - EX_i'')| = O_p(n^{-p+1/3+1/r}\log n) \tag{4.28}$$

and the final result is entailed in (4.26) and (4.28).

Lemma 4.5 *Under assumptions A4-A5 and A7 it holds that*

(a) $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = O_p(n)$, $(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1} = O_p(n^{-1})$.

(b) $\lim_{n\to\infty} \max p_{t,M} = 0$ for all $M \subseteq \{1,\ldots,P\}$.

(c) $\max_{N\in\mathcal{B}} \| n_v^{-1}\sum_{t\in N} \tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T - n_c^{-1}\sum_{t\in N^c} \tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T \| = o_p(1)$.

Proof of Lemma 4.5 (a) Note that $u_{t,j} = X_{t,j} - d_j(\mathbf{Z}_t^{A_0})$ for $j = 1,\ldots,P$, where $d_j(\mathbf{z}) = \mathrm{E}(X_{t,j}|\mathbf{Z}_t^{A_0} = \mathbf{z})$, is a strictly stationary α-mixing process. Call $A_n = n^{-1}\sum_{t=1}^n \tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T$ then we prove that $A_{n,i,j} = 1/n\sum_{t=1}^n \tilde{X}_{t,i}\tilde{X}_{t,j} \to A_{i,j}$ for $i,j = 1,\ldots,P$, where $A = [A_{i,j}]$ a positive definite matrix, see B1(i). Indeed, from $\tilde{X}_{t,j} = u_{t,j} - \sum_{s=1}^n w_s(\mathbf{Z}_t^{A_0})u_{s,j} + D_{t,j}$ with $D_{t,j} = d_j(\mathbf{Z}_t^{A_0}) - \sum_{s=1}^n w_s(\mathbf{Z}_t^{A_0})d_j(\mathbf{Z}_s^{A_0})$ we obtain,

$$A_{n,i,j} = \frac{1}{n}\sum_{t=1}^n u_{t,i}u_{t,j} + \frac{1}{n}\sum_{t=1}^n D_{t,j}D_{t,i} + \frac{1}{n}\sum_{t=1}^n \Big(\sum_{s=1}^n w_s(\mathbf{Z}_t^{A_0})u_{s,j}\sum_{k=1}^n w_k(\mathbf{Z}_t^{A_0})u_{k,i}\Big)$$

$$-\frac{1}{n}\sum_{t=1}^n D_{t,j}\sum_{k=1}^n w_k(\mathbf{Z}_t^{A_0})u_{k,i} - \frac{1}{n}\sum_{t=1}^n D_{t,i}\sum_{s=1}^n w_k(\mathbf{Z}_t^{A_0})u_{s,j} - \frac{1}{n}\sum_{t=1}^n u_{t,j}\sum_{k=1}^n w_k(\mathbf{Z}_t^{A_0})u_{k,i}$$

$$-\frac{1}{n}\sum_{t=1}^{n} u_{t,i}\sum_{s=1}^{n} w_k(\mathbf{Z}_t^{A_0})u_{s,j} + \frac{1}{n}\sum_{t=1}^{n} D_{t,j}u_{t,i} + \frac{1}{n}\sum_{t=1}^{n} D_{t,i}u_{t,j} = \sum_{m=0}^{8} J_m.$$

From A4, A7 and ergodic theorem $J_0 = n^{-1}\sum_{t=1}^{n} u_{t,j}u_{t,i} \xrightarrow{P} \mathrm{E}(u_{1,j}u_{1,i})$. Kernel's Lipschitz property and $\max_{1\le t\le n}\sum_{s=1}^{n} w_s(\mathbf{Z}_t^{A_0})I(\parallel \mathbf{Z}_t^{A_0} - \mathbf{Z}_s^{A_0} \parallel > n^{-1/2}) = O_p(n^{-1/2})$ yield

$$\max_{1\le t\le n} |d_j(\mathbf{Z}_t^{A_0}) - \sum_{s=1}^{n} w_s(\mathbf{Z}_t^{A_0})d_j(\mathbf{Z}_s^{A_0})| = O_p(n^{-1/2}). \tag{4.29}$$

Thus, using Abel's inequality along with (4.29), we get $J_1 = O_p(n^{-1}) = o(1)$. Further, Lemma 4.4 for $p = 1 - k\lambda(k)$, $r = 6$ and B3 yields

$$|\max_{1\le t\le n}\sum_{s=1}^{n} w_s(\mathbf{Z}_t^{A_0})u_{s,j}| = O_p(n^{k\lambda(k)-1/2}\log n) \tag{4.30}$$

which combined with Abel's inequality yields $J_2 = O_p(n^{2k\lambda(k)-1}\log n) = o_p(1)$ from $k\lambda(k) < 1/2$. Similarly for $m = 3, 4$ equations (4.29) and (4.30) imply

$$J_m \le C\max_{1\le t\le n}|D_{t,j}|\max_{1\le t\le n}|\sum_{s=1}^{n} w_s(\mathbf{Z}_t^{A_0})u_{s,i}| = O_p(n^{k\lambda(k)-1}\log n) = o_p(1)$$

while Cauchy-Schwartz inequality for $m = 5, 6$ along with the ergodic theorem and equation (4.30) yield

$$J_m \le Cn^{-1/2}\Big(\sum_{t=1}^{n} u_{t,i}^2\Big)^{1/2}\max_{1\le t\le n}|\sum_{s=1}^{n} w_k(\mathbf{Z}_t^{A_0})u_{s,j}| = O_p(n^{k\lambda(k)-1/2}\log n) = o_p(1).$$

Finally from Cauchy-Schwartz inequality, ergodic theorem and (4.29) it follows that $J_m \le Cn^{-1/2}(\sum_{t=1}^{n} u_{t,i}^2)^{1/2}\max_{1\le t\le n}|D_{t,j}| = O_p(n^{-1/2}) = o_p(1)$ for $m = 7, 8$ and conclude (a). The t-th diagonal element of $\mathbf{P}_M$ is $p_{t,M} = \sum_{r,j=1}^{m}\tilde{\mathbf{X}}_{t,i_r}^2 C_{t,j}$ where $C_{t,j} = O_p(n^{-1})$ from (a). Under assumption A4, A5 and A7 conclude (b). For (c), we have that

$\max_{N\in\mathcal{B}} \parallel \frac{1}{n_u}\sum_{t\in N}\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T - \frac{1}{n_c}\sum_{t\in N^c}\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T \parallel \le$

$$\max_{N\in\mathcal{B}} \parallel \frac{1}{n_u}\sum_{t\in N}\{\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T - \mathrm{E}(\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T)\} \parallel + \max_{N\in\mathcal{B}} \parallel \frac{1}{n_c}\sum_{t\in N^c}\{\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T - \mathrm{E}(\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T)\} \parallel .$$

Call $\tau_t = \tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T - \mathrm{E}(\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T)$ then using similar arguments as in (a) it can be shown that $\max_{N\in\mathcal{B}} \parallel \sum_{t\in N}\tau_t \parallel = o_p(n_u)$, $\max_{N\in\mathcal{B}} \parallel \sum_{t\in N^c}\tau_t \parallel = o_p(n_c)$. Details are omitted.

4.4 Bandwidth selection

The cross validation function defined in (4.6) depends directly on the bandwidth. Using standard asymptotic results from the partial linear theory, it is easy to see that for the minimizer h_0, of the Asymptotic Mean Square Error of the model with the true nonparametric component:

$$\text{AMSE}(h) = \frac{1}{n}\sum_{t=1}^{n} \text{E}\Big(\mathbf{X}_t^T\hat{\boldsymbol{\theta}} - \mathbf{X}_t^T\boldsymbol{\theta} + \hat{g}(\mathbf{Z}_t^{A_0}) - g(\mathbf{Z}_t^{A_0})\Big)^2$$

holds that $h_0 = Cn^{-1/(4+q)}$. Further, define $(\hat{A}, \hat{h})$ the simultaneous estimator of the regressors optimal subset and the bandwidth that minimizes the CV-function, i.e.

$$(\hat{A}, \hat{h}) = \arg \min_{A\subseteq\{1,\ldots,Q\}, 1\leq k\leq Q} \min_{h\in H_n(k)} \text{CV}(A; h)$$

then it follows that $\hat{h}/h_0 \xrightarrow{P} 1$ as $n \to \infty$ with $H_n(k) = [A_k n^{-1/(4+k)-c_k}, B_k n^{-1/(4+k)+c_k}]$ and $A_k, B_k > 0$, $0 < c_k < 1/2(4+k)$ constants, see Gao and Tong (2002) for details of the proof. The result implies that the CV-criterion not only identifies the correct dimensionality of the nonparametric function but it automatically adjusts the bandwidth to have the same rate with the optimal bandwidth h_0 that minimizes the AMSE, equivalently $h \sim n^{-1/(4+q)}$.

Consistency of the estimator for the optimal linear regressors subset was proven under the assumption that the bandwidth is of rate $n^{-\lambda(k)}$ with $k\lambda(k) < 1/2$. In other words, for a bandwidth of this order, the nonparametric component does not affect the rate of convergence of the parametric estimator. The above condition may look strong but in practice the rate of the bandwidth is as important as its constant, especially for large n. For a similar case, Yao and Tong (1994) suggested that the data-driven bandwidths do not depart in principle from the bandwidth's assumptions. They also allowed some minor modification to ensure for instance the monotonicity of $k\lambda(k)$ when necessary. The numerical examples, presented below, with the bandwidth chosen as the minimizer of the CV-function support this remark.

4.5 Extension to the variance function

The model considered in (4.1) is a mean regression model. However, with some modifications in the assumptions, the results can be extended to include modelling of variance functions. In particular, a second order partial linear model is defined as

$$Y_t = \sigma_t \epsilon_t, \;\; \sigma_t^2 = \mathbf{X}_t^T \boldsymbol{\theta} + g(\mathbf{Z}_t) \tag{4.31}$$

where Y_t scalar, $\mathbf{X}_t = (X_{t,1}, \ldots, X_{t,P})^T$, $X_{t,j} \geq 0$, $\mathbf{Z}_t = (Z_{t,1}, \ldots, Z_{t,Q})^T$. Further, $g(.) \geq 0$ is a Q-dimensional function and ϵ_t a zero mean, error process independent of $\{\mathbf{X}_s, \mathbf{Z}_s, s \leq t\}$ satisfying $\mathrm{E}(\epsilon_t^2) = 1$. Note that if $X_{t,i} = Y_{t-i}^2$ then model (4.31) is a partial linear in respect to squared-Y_t, ARCH model. Re-arranging (4.31) yields $Y_t^2 = \sigma_t^2 \epsilon_t^2 = \sigma_t^2 + \sigma_t^2(\epsilon_t^2 - 1) \equiv \sigma_t^2 + \sigma_t^2 \xi_t$ with $\xi_t = \epsilon_t^2 - 1$. Obviously, $\mathrm{E}(\xi_t) = 0$ so $\mathrm{E}(\sigma_t^2 \xi_t | \mathbf{X}_t, \mathbf{Z}_t) = \sigma_t^2 \mathrm{E}(\xi_t) = 0$. Thus, (4.31) is written as

$$\mathrm{E}(Y_t^2 | \mathbf{X}_t, \mathbf{Z}_t) = \sigma_t^2 = \mathbf{X}_t^T \boldsymbol{\theta} + g(\mathbf{Z}_t) \tag{4.32}$$

which is in the form of model (4.1). The main concern here is that the error term is heteroscedastic. However, Härdle, Liang, and Gao (2000) have already shown that under some assumptions on initial estimates of σ_t^2, the weighted-LS estimator of $\boldsymbol{\theta}$ is $\sqrt{n}$-consistent and asymptotically normally distributed. The generalization of the previous results to the weighted cross-validation function is straightforward. Consequently, by introducing weights to the proposed selection procedure we ensure that the asymptotic results for the estimator of the optimal subset are still applicable. In practice, we first regress Y_t^2 on all the candidate parametric regressors $X_{t,j}$ and use the weighted leave-one-out CV criterion to find the optimal nonparametric regressors set. Then, we apply the weighted leave-n_v-out cross validation to exclude the insignificant parametric regressors. The weights are based on initial estimates of the variance σ_t^2 using a fully nonparametric method, e.g. the Nadaraya Watson estimates, including all the candidate variables. Numerical evaluation of the method is presented in the following section.

4.6 Numerical examples

Two simulated examples for mean regression models and one for variance modelling are considered. We calculate the probabilities of selection for the predictor subsets as the minimizers of the CV-criteria for the nonparametric and parametric components. Knowledge of the true model helps us to evaluate the procedure. In addition, we calculate the probabilities of selection using a fully nonparametric method in order to emphasize the advantage of the proposed selection procedure against a less flexible selection method that does not take into consideration the partially linear form of the model. We use the multivariate kernel $K(\mathbf{x}) = \prod_{i=1}^{k} K(x_i)$ where $K(.)$ is the Epanechnikov kernel. Note that the selected kernel satisfies assumption A5 on the kernel function. Bandwidth is selected by minimizing the cross validation criterion in grid points $h = 0.2\ 1.2^a\ \sigma$ for $a = 1, \ldots, 15$, where σ is the sample standard deviation, (see discussion in section 2.8 about the choice of the grid points). Further, B2 is met by choosing $b = n$ and, $n_v = n - n_c$ with $n_c = [n^{3/4}]$ the largest integer part of $n^{3/4}$.

4.6.1 Mean regressors selection

We generate a time series data set from the model

$$Y_t = 0.5Y_{t-1} - 0.35Y_{t-2} - 0.75\exp(-Y_{t-3}^2) + \frac{0.85}{1 + Y_{t-4}^2} + \epsilon_t$$

with ϵ_t following a uniform distribution in $[-1, 1]$. Note that the error distribution has bounded support, hence there is no need for introducing a weighting function in the CV-function. Having identified the nonparametric components, the candidate linear components are $M_1 = \{1\}$, $M_2 = \{2\}$ and $M_0 = \{1, 2\}$ the true one. Let $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$, $u_{t,1} = Y_{t-1} - E(Y_{t-1}|Y_{t-3}, Y_{t-4})$ and $u_{t,2} = Y_{t-2} - E(Y_{t-2}|Y_{t-3}, Y_{t-4})$, $u_t = (u_{t,1}, u_{t,2})^T$, $u_{t,M_1} = u_{t,1}$, $u_{t,M_2} = u_{t,2}$, $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{u}_{M_i} = (u_{1,M_i}, \ldots, u_{n,M_i})$, for $i = 1, 2$. Further, B1(ii) holds from $\liminf_{n\to\infty} \frac{1}{n}(\mathbf{u}\boldsymbol{\theta})^T(I_n - \mathbf{u}_{M_i}(\mathbf{u}_{M_i}^T\mathbf{u}_{M_i})^{-1}\mathbf{u}_{M_i}^T)(\mathbf{u}\boldsymbol{\theta}) = (\theta_{3-i}^2 \sum_{t=3}^n u_{t,1}^2 \sum_{t=3}^n u_{t,2}^2 - (\sum_{t=3}^n u_{t,1}u_{t,2})^2)/\sum_{t=3}^n u_{t,i}^2 > 0$ with probability one, since $P(u_{t,1} = u_{t,2}) = 0$.

Table 4.1: Probabilities of nonparametric regressors selection: the leave-one-out CV.

	Two-Step CV			Fully Nonparametric		
subset	$n = 50$	$n = 100$	$n = 300$	$n = 50$	$n = 100$	$n = 300$
$\{Y_{t-1}\}$	0.1875	0.125	0.100	0.075	0.0375	0.0125
$\{Y_{t-2}\}$	0.025	0.0125	0.0125	0.0625	0.025	0.00
$\{Y_{t-1}, Y_{t-2}\}$	0.225	0.175	0.1125	0.2125	0.2625	0.2375
$\{Y_{t-1}, Y_{t-2}, Y_{t-3}\}$	0.075	0.075	0.0375	0.2375	0.25	0.2625
$\{Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}\}$	0.0875	0.05	0.05	0.3	0.375	0.475
$\{Y_{t-2}, Y_{t-3}\}$	0.0125	0.0	0.0	0.0125	0.0	0.0
$\{Y_{t-1}, Y_{t-3}\}$	0.025	0.025	0.0125	0.025	0.0125	0.0125
$\{Y_{t-3}, Y_{t-4}\}$	0.3875	0.55	0.6875	0.075	0.0375	0.0

Table 4.2: Probabilities of selection based on the MCCV with Y_{t-3}, Y_{t-4} the nonparametric regressors.

subset	$n = 50$	$n = 100$	$n = 300$
$\{Y_{t-1}\}$	0.2875	0.2375	0.025
$\{Y_{t-2}\}$	0.1375	0.05	0.0
$\{Y_{t-1}, Y_{t-2}\}$	0.575	0.7125	0.975

It is easy to see that A3 is met while the generated process satisfies A7. Note that A8 holds for $g(y, x) = -0.75e^{-y^2} + 0.85/(1 + x^2)$. We first regress Y_t against all Y_{t-j}, for $j = 1, 2, 3, 4$. Then using the residuals $\hat{U}_t$ we calculate the leave-one-out cross validation. The first three columns of Table 4.1 contain the probabilities, calculated from 80 iterations, of selection for each candidate subset. Combinations with calculated zero probability have been omitted from the table. Apparently, $\{Y_{t-3}, Y_{t-4}\}$ has the highest probability of selection even in a small sample size of $n = 50$ observations that is 0.3875. Moreover, when the sample size increases, the probability of selection increases, reaching up to 0.6875 for a sample of size $n = 300$ which implies consistency of the estimator. Then, using $\{Y_{t-3}, Y_{t-4}\}$ as the nonparametric component, we

calculate the leave-n_v-out CV. The results are presented in Table 4.2. The parametric component is identified successfully from the MCCV, even for small samples, while the probability of selecting the true regressors increases up to 0.975 for $n = 300$. Furthermore, in the last three columns of Table 4.1 we present the results using a fully nonparametric cross validation selection procedure. It is understood that the fully parametric selection method fails to distinguish the linear term from the nonparametric component while the convergence rate appears significantly slower. It seems that the linear term dominates the nonparametric component and this is the reason why, even when the sample size is large, $\{Y_{t-1}, Y_{t-2}\}$ and $\{Y_{t-1}, Y_{t-2}, Y_{t-3}\}$ have high selection probabilities, 0.2375 and 0.2625 respectively. This is an example of the importance of employing a combined selection method instead of a fully nonparametric one when working with a semiparametric model. Note here that the proposed procedure reduces the total number of investigated models from $(2^4 - 1)\text{x}(2^2 - 1) = 45$ into $2^4 + 2^2 - 2 = 18$. The reduction is significant and it consists the main contribution of the proposed selection procedure.

4.6.2 Mean regressors selection with two processes

We generate data from the model

$$Y_t = 0.35Y_{t-1} - 0.15Y_{t-2} + \frac{0.5X_t}{1 + X_t^2} + e_t, \quad X_t = 0.3X_{t-1} + 0.2X_{t-2} + \epsilon_t$$

with $e_t \sim U[-0.25, 0.25]$ and $\epsilon_t \sim U[-0.5, 0.5]$. The example is also examined by Gao and Tong (2002). They showed that assumptions A4, A7-A8 and B1 hold. We proceed by regressing Y_t against the candidate linear regressors $Y_{t-1}, Y_{t-2}, Y_{t-3}$ to calculate the residuals U_t. This set should always be the largest possible even if it includes insignificant linear regressors. Note here that Y_{t-3} is not a linear regressor. This is to show that the procedure works even when insignificant regressors are used in the calculation of the residuals U_t. The results of the leave-one-out CV are summarized

Table 4.3: Probabilities of nonparametric regressors selection: the leave-one-out CV.

Regressors subset	$n = 50$	$n = 150$	$n = 300$
$\{X_t\}$	0.4625	0.575	0.7875
$\{X_{t-1}\}$	0.2875	0.225	0.125
$\{X_t, X_{t-1}\}$	0.25	0.2	0.0875

Table 4.4: Probabilities of selection based on the MCCV with X_t the nonparametric regressor.

Parametric Regressors subset	$n = 50$	$n = 150$	$n = 300$
$\{Y_{t-1}\}$	0.25	0.225	0.125
$\{Y_{t-2}\}$	0.0625	0.025	0.0
$\{Y_{t-3}\}$	0.025	0.0	0.0
$\{Y_{t-1}, Y_{t-2}\}$	0.4375	0.575	0.8375
$\{Y_{t-1}, Y_{t-3}\}$	0.15	0.1125	0.025
$\{Y_{t-2}, Y_{t-3}\}$	0.0	0.0	0.0
$\{Y_{t-1}, Y_{t-2}, Y_{t-3}\}$	0.075	0.0625	0.0125

in Table 4.3 while Table 4.4 contains the results for the MCCV, using X_t as the nonparametric regressor. Apparently, the true nonparametric component is identified with a probability 0.4625 and 0.575 for sample sizes $n = 50$ and 150 while for $n = 300$ the probability is 0.7875. Therefore, there is strong evidence that the single predictor X_t should be selected. The increase in the probability due to the increase of the sample size is in line with the property of convergence shown earlier. On the other hand, the MCCV distinguishes the insignificant linear regressors and successfully identifies the linear regressors of the underlying model with probability as large as 0.8375 for $n = 300$. The required computations have been reduced from $(2^3 - 1) \times (2^2 - 1) = 21$ that would include all the possible combinations, to $2^3 + 2^2 - 2 = 10$ cases under investigation.

4.6.3 Variance regressors selection

The third example is an application in variance modelling. The data is a time series generated from the conditional heteroscedastic model

$$Y_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = (0.45 Y_{t-1}^2 + 1.1 \sin(Y_{t-2}^2) \exp(-0.85 Y_{t-2}))_+$$

where ϵ_t is the sum of 35 independent random variables each uniformly distributed on $[-0.05, 0.05]$. According to the Central Limit Theorem, the resulting process is a good approximation of the standard normal random variable but in fact the support of the error density is bounded, given by $[-1.75, 1.75]$. Note also that variance of the error term is equal to one a necessary assumption for the identifiability of the variance component. Since we deal with heteroscedastic data, it is suggested that we use the weighted least squares to calculate the residuals U_t. The latter requires some initial estimates of the variance function to be used as weights when fitting the regression line. Consequently, we calculate the fully nonparametric Nadaraya-Watson estimates, denoted by $\tilde{\sigma}_t^2$. Then, we regress Y_t^2 on Y_{t-j}^2 for $j = 1, 2, 3$ using the weighted least squares with $\tilde{\sigma}_t^{-2}$ as the weights. Then U_t are the standardized residuals. This is equivalent to introducing weights in the leave-one-out cross validation function as required from earlier discussion in order to account for the heteroscedasticity of the model. The probabilities of selection for the nonparametric regressors calculated after 80 iterations are presented in Table 4.5. The nonparametric optimal subset Y_{t-2} is identified with probability equal to 0.3375 while the probability of $\{Y_{t-1}, Y_{t-2}\}$ is 0.225 for sample size $n = 50$. Hence, though successful, the CV-criterion is not very decisive. The evidence is more significant when the sample size is increased. Note that the probability is 0.6375 for size $n = 300$, indicating that the single variable Y_{t-2} is the nonparametric regressor. The MCCV criterion for the selection of the linear regressors is calculated using the residuals $\hat{V}_t = Y_t^2 - \sum_{s=1}^{n} w_s(Y_{t-2}) Y_s$. Table 4.6 contains the probabilities for the linear regressors. The true linear regressor is identified with probabilities 0.5375 and 0.7875 for sample size $n = 50$ and $n = 300$.

Table 4.5: Probabilities of the nonparametric regressors: the leave-one-out CV

	Two-step CV			Fully Nonparametric		
subset	$n = 50$	$n = 100$	$n = 300$	$n = 50$	$n = 100$	$n = 300$
$\{Y_{t-1}\}$	0.125	0.1375	0.075	0.1375	0.125	0.1375
$\{Y_{t-2}\}$	0.3375	0.4375	0.6375	0.1875	0.2	0.1375
$\{Y_{t-3}\}$	0.05	0.0375	0.0375	0.0625	0.0375	0.0
$\{Y_{t-1}, Y_{t-2}\}$	0.225	0.15	0.1125	0.2625	0.325	0.4125
$\{Y_{t-1}, Y_{t-3}\}$	0.0875	0.0875	0.0625	0.1375	0.0875	0.0875
$\{Y_{t-2}, Y_{t-3}\}$	0.0625	0.0625	0.0125	0.0375	0.0375	0.0125
$\{Y_{t-1}, Y_{t-2}, Y_{t-3}\}$	0.1125	0.0875	0.0625	0.175	0.1875	0.2125

Table 4.6: Probabilities of selection based on the MCCV with Y_{t-2} for nonparametric regressor

Parametric Regressors subset	$n = 50$	$n = 100$	$n = 300$
$\{Y_{t-1}^2\}$	0.5375	0.7125	0.7875
$\{Y_{t-3}^2\}$	0.3	0.1875	0.1375
$\{Y_{t-1}^2, Y_{t-3}^2\}$	0.1625	0.1	0.075

Overall, the selection procedure is successful but the probabilities are relatively smaller, equivalently, the rate of convergence is slower especially for small sample size. This feature is related to the convergence rate of the initial estimators that were used as weights in the cross validation function. At the same time, the fully nonparametric selection procedure identifies Y_{t-1}, Y_{t-2} but with a slower rate since the corresponding selection probability was found to be 0.4125 for $n = 300$. The latter observation emphasizes the need of a more flexible and efficient selection method than a fully nonparametric method when the underlying model includes a linear term.

Chapter 5

Applications of the adaptive ML-estimator to Value at Risk

5.1 Introduction to VaR theory

Due to the increase of traded assets and the complexity of the market, risk measurement has been the favorite topic of recent discussion and research. Newly imposed regulations have now made necessary that financial institutions and banks should hold a certain amount of capital as a cushion against adverse market movements. Many of the proposed theories are highly dependent on the accurate modelling, estimation and prediction of the market volatility.

A commonly used quantile-based risk measure is the Value at Risk (VaR). In Value at Risk theory, we are interested in the quantile of the *Profit-Loss* distribution[1] over a defined period of time. Hence, from a statistical point of view, VaR is a simple quantile calculation of the returns distribution. Equivalently, VaR is defined as $\mathrm{F}^{-1}(\alpha) = \mathrm{VaR}$ where F is the probability distribution of the returns over a defined period of time and α is the given probability losses. The existing estimation methods of the returns

[1]Most often the Profit-Loss distribution corresponds to the distribution of the log-returns of a single asset (univariate) or a portfolio (multivariate).

distribution are mainly divided in three groups (McNeil and Frey 2000): (1) the nonparametric historical simulation (2) the parametric methods based on modelling the conditional variance (Riskmetrics and ARCH/GARCH) and (3) Extreme Value Theory (EVT). See also Duffie and Pan (1997) for an overview of VaR theory.

Historical simulation is easy to implement but it has been proven to perform poorly while parametric conditional variance models have a major drawback that assume conditional normality which contradicts with the observed heavy tailed sample distribution of the most financial data sets. On the other hand, in Extreme Value Theory the unconditional distribution has a parametric form at the tails known as tail index, (Embrechts, Resnick, and Samorodnitsky 1998, 1999 and Danielsson and Vries 1997) or comes from a particular family of distributions like hyperbolic distribution (Eberlein, Kallsen, and Kristen 2001). McNeil and Frey (2000) quote that "*none of the previous EVT-based methods for quantile estimation yields* VaR *estimates which reflects the current volatility background. Given the conditional heteroscedasticity of most financial data, which is well documented by the considerable success of the models from the ARCH/GARCH family, we believe this is to be a major drawback of any kind of* VaR-*estimator.*" In response, they propose a combination of historical simulation (for the central part) along with threshold methods from EVT (for the tails) of the error distribution. Then, the conditional return distribution is constructed from the estimated error distribution and the conditional variance estimators calculated by fitting a GARCH model. Similarly, Danielsson and Vries (2000) propose a semiparametric method as a mixture of the two approaches. A common point shared by both approaches, is that estimating the returns conditional distribution requires the calculation of the conditional variance. Consequently, accurate estimators for the conditional variance will eventually result in an improvement in the conditional distribution of the returns. But as noted above, the parametric conditional variance models employed to produce these estimates assume Gaussian error distribution and often fail to capture the dynamics of the heavy tailed data.

The proposed likelihood-based estimator of the conditional variance function does not require such an assumption. Consequently, we apply the proposed nonparametric likelihood estimator in the estimation of the VaR. Details about the implementation of the adaptive ML-estimator in the calculation of VaR are described in the following sections. There, we introduce a number of combinations between quantile estimators of error distribution and conditional variance estimators. In fact, as we shall see later on, the adaptive ML-estimator performs relatively better compared to existing nonparametric estimators especially when the error distribution is heavy tailed. Of course, this does not guarantee that it outperforms the parametric estimators like GARCH-estimators. However, it provides a nonparametric alternative to the parametric models that have dominated the VaR-theory.

5.2 Real data applications

Within the nonparametric context, simulated examples showed that for heavy tailed data, the adaptive ML-estimator is an improvement to the LS-estimator. It remains to see if this improvement is maintained when dealing with real data. In this chapter, we analyze a number of time series data sets and study the performance of the two nonparametric estimators. Furthermore, we attempt a comparison between the proposed nonparametric fittings and parametric fittings. Parametric models are less complex yielding parameter estimates that have faster convergence rates. However, it is often observed that they lack flexibility and fail to capture the underlying dynamics especially if a non-linear structure is present. On the other hand, nonparametric estimators converge more slowly and are computationally demanding. But they reduce the model bias when the true model contains non-linear terms.

The examples considered in this chapter involve exclusively financial data. In particular, we study three different types: stock indices, stocks and exchange rates. The first two data sets were found in the web site: http://finance.yahoo.com/. The ex-

change rates data set was retrieved from the web site: http://fx.sauder.ubc.ca/data.html. The reason for choosing to work with financial data is due to the degree of heteroscedasticity and the existence of many extreme observations, and therefore, heavy tailed underlying distributions. Furthermore, risk is a concept closely related to financial data. Like all the risk measures, Value at Risk (VaR) has been developed within the context of finance. However, we would like to point out that our methodology is independent of the nature of the data set used, and can be employed to analyze any time series data which exhibit conditional heteroscedasticity.

5.2.1 Stock indices

We begin with the analysis of four stock indices and particularly we look at the daily log-returns of Standard and Poor 500 (SP), Dow-Jones (DJ), FTSE 100 (FTSE) and DAX 100 (DAX) indices for the period of September 1997 to December 2003. The sample size is 1500 observations. We split the sample into two parts, the "*pre-sample*" period of size $n_f = 1000$ which we use to fit the models and the "*post-sample*" period with $n_e = 500$ used for the evaluation of the models. We proceed by plotting the data. Figure 5.1 indicates that volatility of the returns changes over time. Note that clusters of high volatile peaks are followed by clusters of low volatile peaks indicating non constant variation. In Figure 5.2 we calculate the autocorrelation function (ACF) for the original returns (first column) along with the ACF for the squared returns (second column). Looking at the ACF functions for the returns, we see that there is little evidence for correlation within the original series which was expected given that the trend has been removed by taking the difference. However, ACF for the squared returns reveals that there is strong evidence for second order correlation implying that the volatility of the series at time t depends significantly on the past variables.

We consider three different conditional heteroscedastic models, two parametric and one nonparametric. The first parametric model is GARCH(1,1) a popular model in analyzing financial data mainly due to its practicality as well as good performance.

Figure 5.1: Time series plot for the returns of the stock indices.

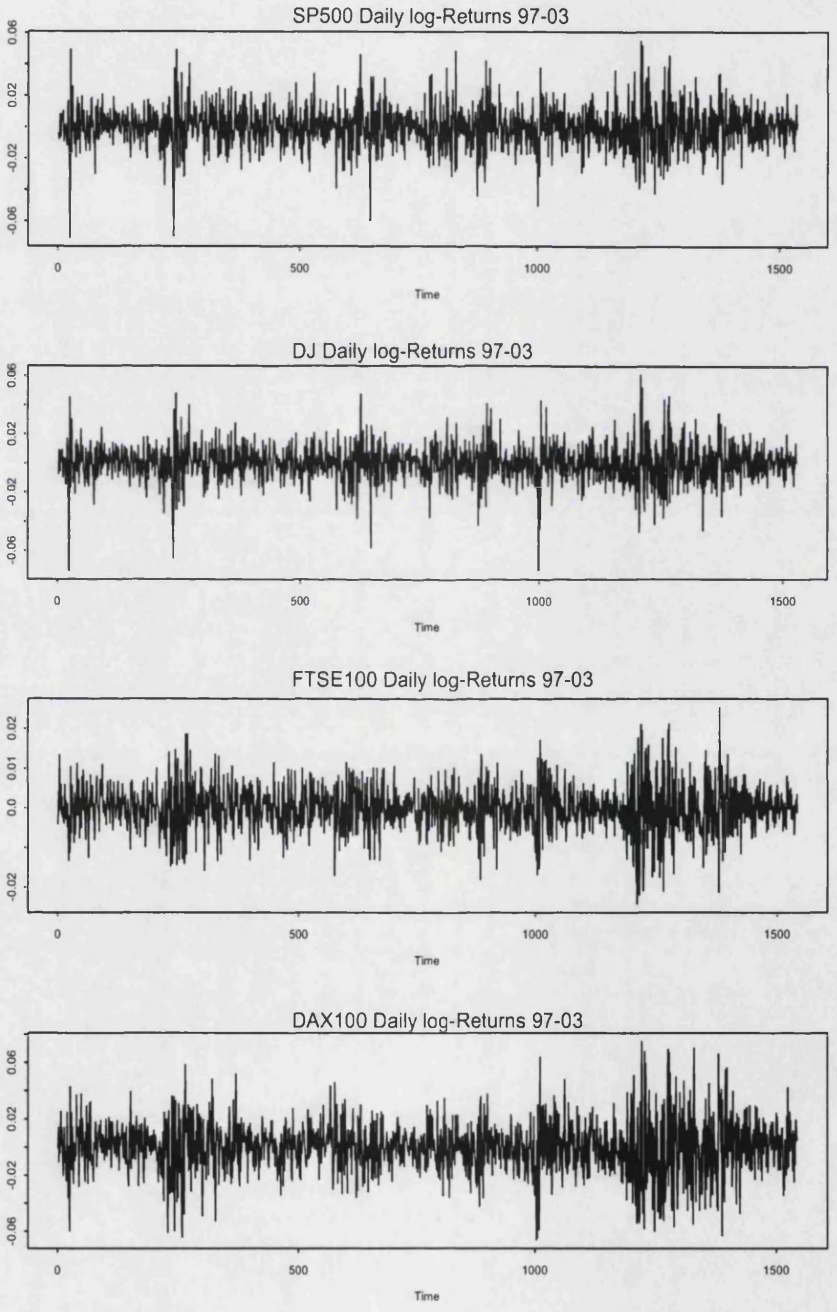
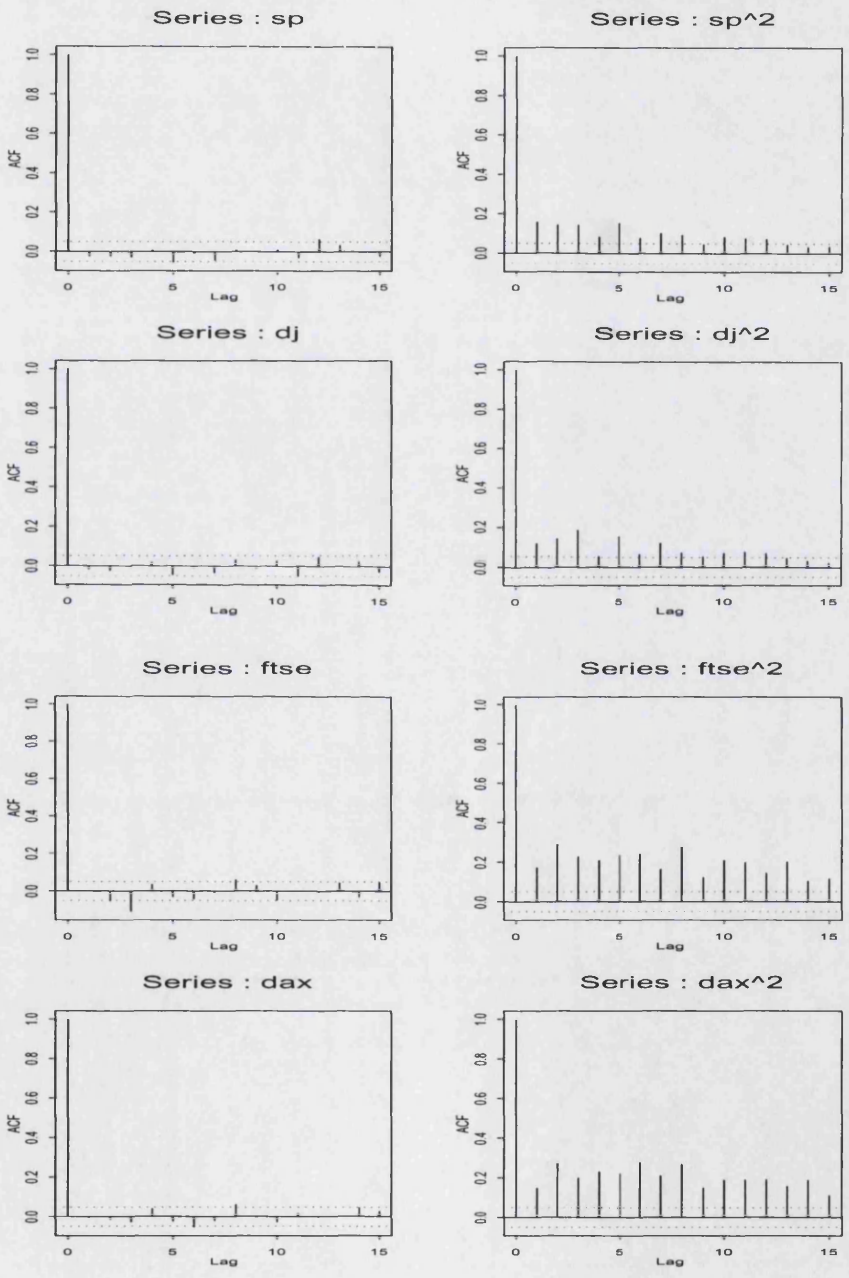


Figure 5.2: Returns and squared returns autocorrelation function.



Nevertheless, the basic GARCH model fails to take into account the asymmetry of the news impact where negative news has larger impact on volatility than good news, also known as "*leverage effect*". With the introduction of the absolute value of the error term, exponential-GARCH is able to model this asymmetry. Hence, the second parametric model under investigation, is the exponential-GARCH, or EGARCH, see Nelson (1991). On the other hand, the nonparametric model is in the form of model (2.2) with the mean function $m(.)$ set equal to zero, equivalently, assuming no trend (see discussion on ACF of the returns). The selection of the predictors is made using the nonparametric CV-criterion. Among all the combinations of Y_{t-i}, $i = 1, 2, 3$, the regressor set $\{Y_{t-1}, Y_{t-2}\}$ yields the lowest CV value of 1.186×10^{-7} with $\{Y_{t-1}\}$ coming second with a value of 1.264×10^{-7} for the SP500 data set. Similar conclusions are drawn for the remaining data sets all in favor of $\{Y_{t-1}, Y_{t-2}\}$. Note here that the combination of $\{Y_{t-1}, Y_{t-2}, Y_{t-3}\}$, yields a slightly lower CV value for FTSE and DJ data set. Nevertheless, we argue that the difference is probably not significant enough to compensate for the increase in the computations. Hence, we end up with two predictors for the variance function, namely $\{Y_{t-1}, Y_{t-2}\}$, in nonparametric fitting. Equivalently, we write model (2.2) as $Y_t = \sigma(Y_{t-1}, Y_{t-2})\epsilon_t$. For the nonparametric model we calculate (i) the LS-estimator, (ii) the adaptive ML-estimator introduced in Chapter 3. Overall, we end up with four fitted models for the conditional variance: two parametric, GARCH and EGARCH and two nonparametric, LSE and MLE.

Figures 5.3-5.6 contain the calculated conditional standard deviations for the pre-sample period, along with the original series. Note that the conditional standard deviation calculated from the parametric models is smoother than that from the nonparametric models. In other words the fluctuation is higher for the nonparametric standard deviation. The latter could be attributed to the fact that the nonparametric weights are calculated using a proportion of the total observations and give more importance to the values that lie close enough contrary to the parametric models where the equivalent weights take into account the full set of observations.

Figure 5.3: SP Returns: Series and conditional standard deviation, LSE and MLE.

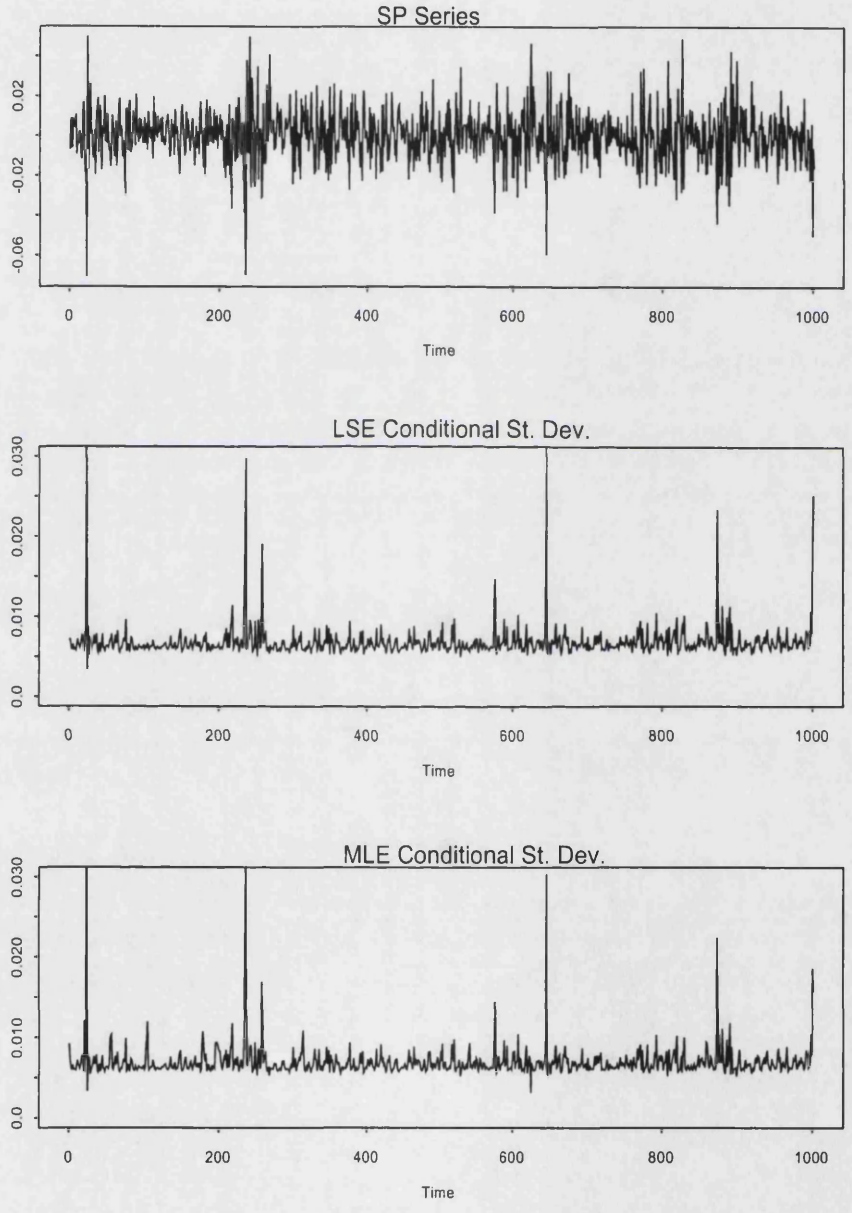


Figure 5.4: DJ Returns: Series and conditional standard deviation, GARCH and EGARCH.

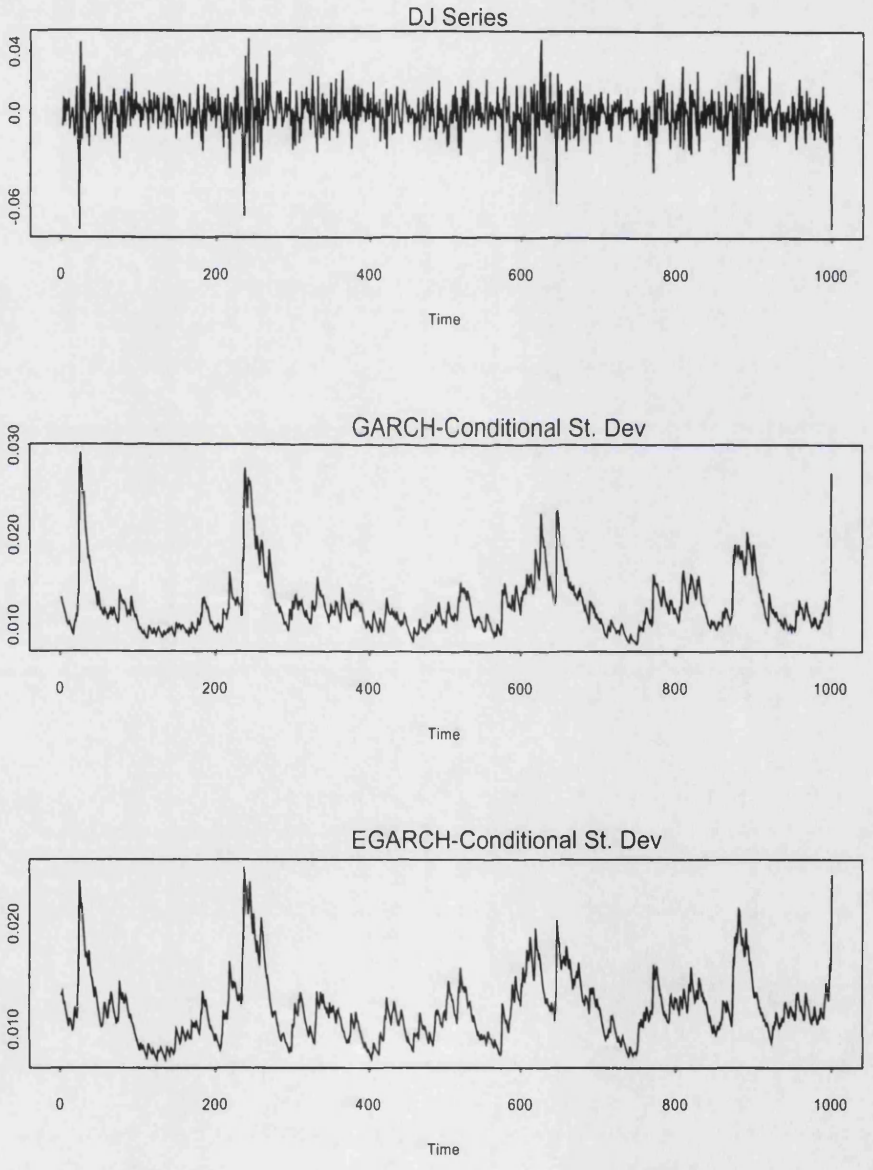


Figure 5.5: FTSE Returns: Series and conditional standard deviation, EGARCH and MLE.

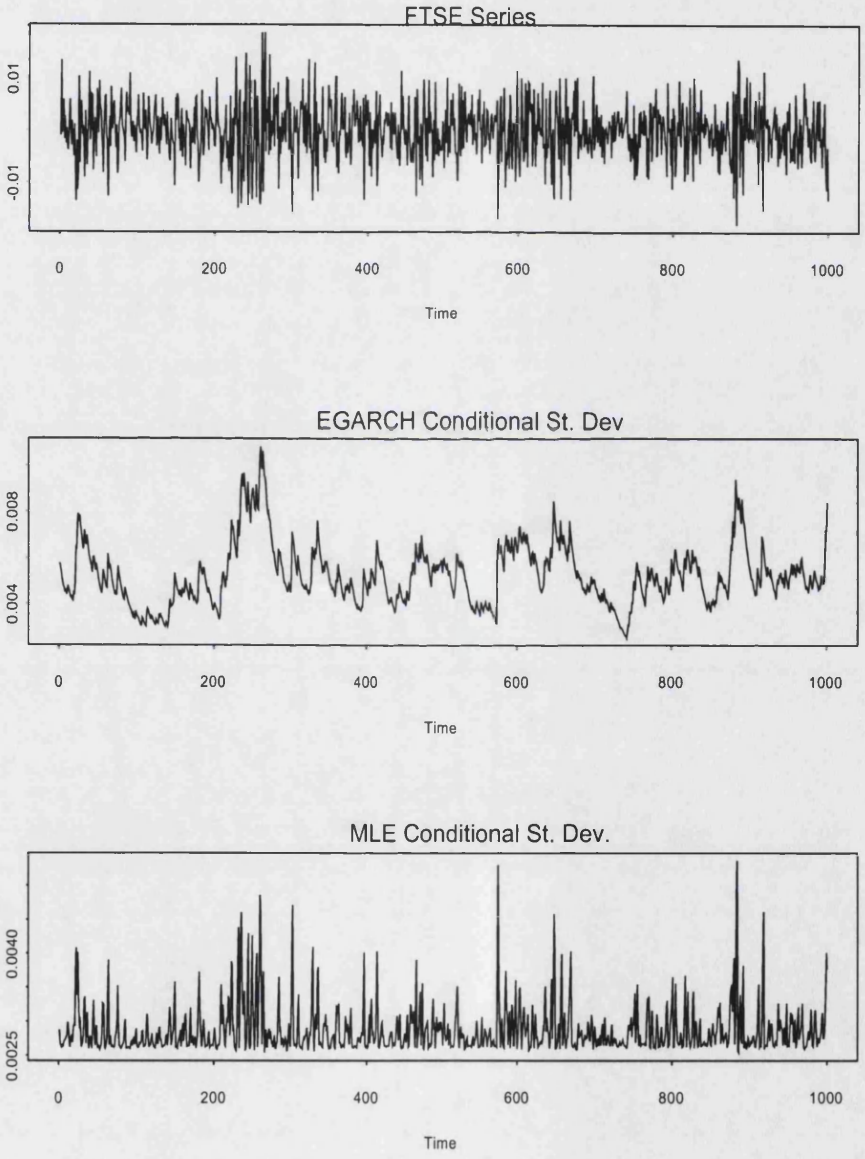


Figure 5.6: DAX Returns: Series and conditional standard deviation, GARCH and LSE.

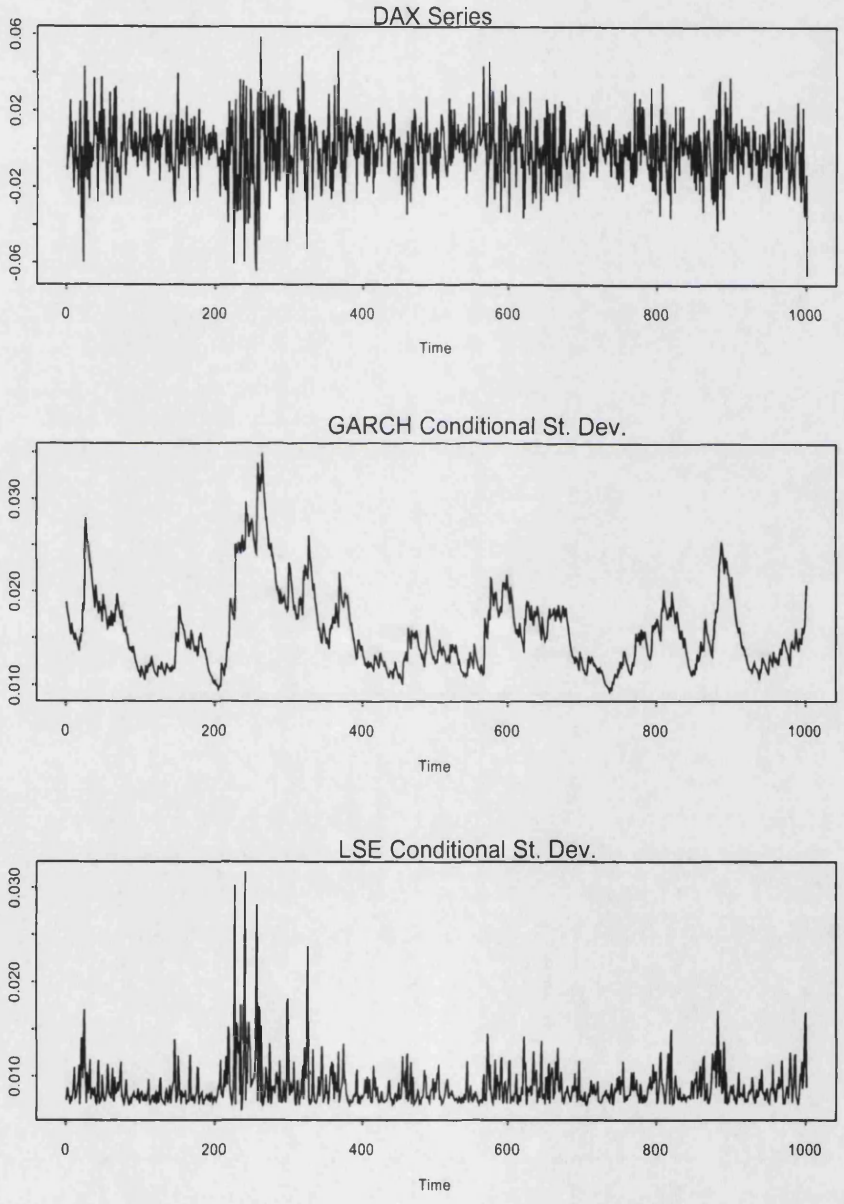


Figure 5.7: Normal probability plot for the SP500 and DAX returns.

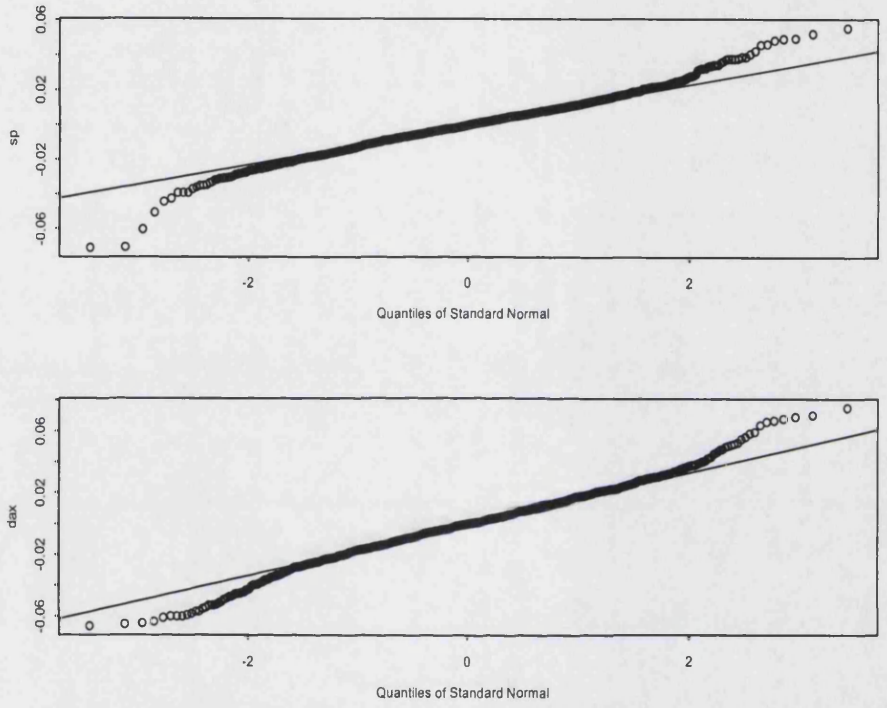
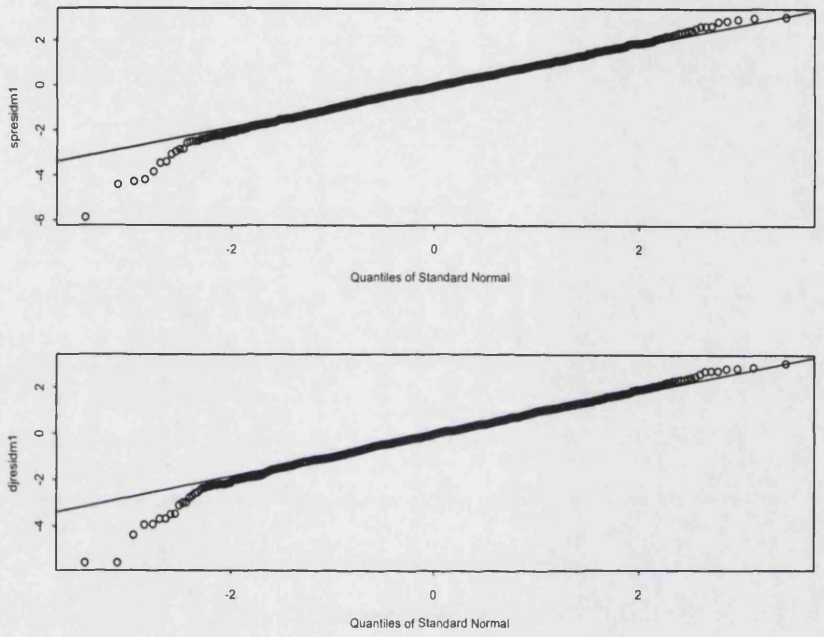


Figure 5.7 is the normal probability plot for the SP500 and DAX returns. There is strong evidence that empirical distribution of the returns of the stock indices exhibit fatter tails than Gaussian distribution while the same applies for the the normal probability plot of the GARCH-residuals for SP500 and DJ as revealed by Figure 5.8. This characteristic is called "*leptokurtosis*" and it is a well known property exhibited by many financial data sets, see Engle and Gonzales-Rivera (1991) and Danielsson and Vries (2000) for more details. It is understood that leptokurtosis of the error distribution raises serious concerns about the use of Gaussian distribution when fitting the GARCH model. In response, Bollerslev (1987) concluded that SP500

Figure 5.8: Normal probability plot for SP500 and DJ residuals from GARCH model.



monthly returns fit better to a GARCH model under the assumption of a t-distribution. Similar conclusions were observed for the FTSE returns while the probability plot of residuals for the DAX showed a rather satisfactory fit with the straight line suggesting normality for the error term.

In order to investigate if there is any evidence for remaining second order correlation we calculate the autocorrelation function for the squared residuals. It appears that the underlying correlation was explained well by the fitted models with the exception again of the DAX series, where there is unexplained autocorrelation present at the residuals of GARCH model. The latter is suggested by the Ljung-Box autocorrelation test for the squared residuals of DAX with p-value 0.0327 with the null hypothesis of no autocorrelation.

Our primary aim is to compare the volatility models. This requires the introduction of some functions that will be used as the means of comparison. The first measure is the Mean Absolute Deviations Error, MADE $= n_e^{-1} \sum_{t=1}^{n_e} |\hat{\sigma}_t^2 - \sigma_t^2|$ already introduced in earlier chapter. Alternatively Fan and Gu (2003) propose the square-Root Absolute Deviation Error, defined as RADE $= n_e^{-1} \sum_{t=1}^{n_e} |\sqrt{2/\pi}\hat{\sigma}_t - |\sigma_t||$. Of course, the true values of the volatility denoted by σ_t^2, are not directly observed and can only be approximated. It is understood that the latter raises some issues regarding to the validity of the selected approximation. Nevertheless, the choice of the proxies for the volatilities is beyond of the scope of this work. Therefore, here we use the conventional square of the observed return. Apart from these deviation measures, Fan and Gu (2003) suggest two tests that can be used for evaluation of volatility models. The independence test is based on the idea that the sequence of events exceeding a given quantile should behave like an i.i.d. Bernoulli distribution. Let $\Phi(.)$ be the normal distribution function and call $I_t = \mathbb{I}(Y_t < \Phi^{-1}(\alpha)\hat{\sigma}_t)$ the indicator of an exceeding event at time t, then I_t takes values zero and one. Define n_{ij} the observed numbers of events from state $i \in \{0,1\}$ to $j \in \{0,1\}$ and let $\hat{\pi}_{i,j} = n_{ij}/(n_{i0} + n_{i1})$, $n_j = n_{0j} + n_{1j}$ and $\hat{\pi} = n_0/(n_0 + n_1)$ then, under the null hypothesis of i.i.d. Bernoulli, the likelihood ratio

$$\text{LR1} = 2\log\left(\frac{\hat{\pi}_{0,0}^{n_{00}}\hat{\pi}_{0,1}^{n_{01}}\hat{\pi}_{1,0}^{n_{10}}\hat{\pi}_{1,1}^{n_{11}}}{\hat{\pi}^{n_0}(1-\hat{\pi}^{n_1})}\right)$$

follows a χ_1^2-distribution. The second test is a test against a given confidence level with $H_0 : P(I_t = 1) = \alpha$ vs $H_1 : P(I_t = 1) \neq \alpha$ and the likelihood ratio

$$\text{LR2} = 2\log\left(\frac{\hat{\pi}^{n_0}(1-\hat{\pi}^{n_1})}{\alpha^{n_0}(1-\alpha^{n_1})}\right)$$

follows a χ_1^2 under H_0. For both tests see Christoffersen (1998) for more details.

Table 5.1 contains the results for the four stock indices. With the exception of Dow-Jones index, the nonparametric ML-estimator has the smallest mean absolute deviation but EGARCH model yields the smallest square root absolute deviation with ML-estimator coming second. Further, all p-values for the independence test LR1 are

Table 5.1: Stock indices: deviation measures and hypothesis tests.

Index	Method	MADE ($\times 10^{-4}$)	RADE ($\times 10^{-3}$)	LR1 p-value	LR2 p-value	LR1* p-value	LR2* p-value
SP	GARCH	1.836	6.380	0.951	0.714	0.827	0.898
	EGARCH	1.701	**6.110**	0.218	0.140	0.872	0.999
	LSE	1.577	6.268	0.842	0.001*	0.790	0.690
	MLE	**1.525**	6.201	0.830	0.003*	0.593	0.892
DJ	GARCH	1.679	6.046	0.234	0.736	0.827	0.898
	EGARCH	1.595	**5.853**	0.882	0.957	0.918	0.511
	LSE	1.430	6.062	0.673	0.000*	0.188	0.893
	MLE	**1.419**	6.020	0.612	0.000*	0.472	0.690
FTSE	GARCH	**0.228**	2.623	0.309	0.445	0.665	0.998
	EGARCH	0.281	**2.578**	0.487	0.165	0.736	0.690
	LSE	0.251	2.689	0.539	0.001*	0.716	0.893
	MLE	**0.228**	2.668	0.337	0.001*	0.873	0.999
DAX	GARCH	2.522	7.533	0.698	0.691	0.816	0.690
	EGARCH	2.493	**7.429**	0.658	0.610	0.827	0.898
	LSE	**2.170**	7.759	0.591	0.000*	0.388	0.893
	MLE	2.202	7.764	0.686	0.000*	0.430	0.789

not significant and hence the null hypothesis that the sequence of extreme events follows an i.i.d. Bernoulli cannot be rejected. Finally the p-values for the confidence level test LR2, are significant for the two nonparametric estimators for all indices hence we reject the hypothesis of $P(I_t = 1) = \alpha$. Note here that the indicator I_t was defined under the assumption of normally distributed errors. Hence, one probability would be that the rejection of the null hypothesis is a result of the departure from

normality for the nonparametric model. In response, we recalculate the tests, namely LR1* and LR2*, but now using the nonparametric α-quantile estimator found directly from the sample distribution of the estimated errors. Clearly, all p-values are now greater than 5%. Hence we conclude that (1) the hypothesis of independent identically Bernoulli distributed extreme events and (2) the hypothesis that the probability of an extreme event being greater than the 95%-quantile is 0.05, cannot be rejected. These conclusions hold for all four volatility models.

We continue with the calculation of the VaR using conditional arguments. The VaR involves the extreme values of the series over a predetermined period, called τ, given a confidence level $1-\alpha$. Recall that the log-return at time $t+1$ is denoted by Y_{t+1} while the conditional variance is σ_{t+1}^2. Let $Y_{t+1,\tau}$ be the aggregate return at time $t+1$ over a period of τ and $\sigma_{t+1,\tau}^2 = \mathrm{Var}(\mathrm{Y_{t+1,\tau}}|\Im_t)$, the corresponding conditional variance where $\Im_t$ is the information up to time t. Let $V_{t+1,\tau}$ be the α-quantile of the conditional distribution of $Y_{t+1,\tau}$. We write $P(|Y_{t+1,\tau}| > V_{t+1,\tau}|\Im_t) = 1-\alpha$. Then $\mathrm{VaR}_{t+1,\tau}$ is the quantile $V_{t+1,\tau}$ of the conditional distribution that yields probability equal to $1-\alpha$ for the given level of losses α. Consequently, if $q(\alpha,\tau)$ is the α-quantile of the error distribution then it holds that $\mathrm{VaR}_{t+1,\tau} = q(\alpha,\tau)\sigma_{t+1,\tau}$. In other words, calculation of the VaR is reduced to the calculation of the conditional standard deviation $\sigma_{t+1,\tau}$ and the α-quantile $q(\alpha,\tau)$ of the error distribution.

Note that, if $\hat{\sigma}_{t+1}$ is the predicted conditional standard deviation for one holding period, i.e. $\tau = 1$, then we use the square-root rule for the τ-period volatility prediction, that is $\hat{\sigma}_{t+1,\tau} = \sqrt{\tau}\hat{\sigma}_{t+1}$, see Fan and Gu (2003) among others, for further details about the $\sqrt{\tau}$-rule. On the other hand, we have already argued that the distribution of the errors $\epsilon_{t,\tau} = Y_{t,\tau}/\sigma_{t,\tau}$ is likely to depart from the normal distribution. Therefore, the α-quantile of the error distribution $q(\alpha,\tau)$, should not be calculated by the normal tables. Instead, we introduce a nonparametric estimator. The natural nonparametric α-quantile estimate $\hat{q}(\alpha,\tau)$ for the error distribution is found as the sample α-quantile of the estimated errors $\hat{\epsilon}_{t,\tau} = Y_{t,\tau}/\hat{\sigma}_{t,\tau}$. At this point we follow Fan and Gu

(2003) who showed that the choice of this natural quantile estimator is not as efficient as the parametric quantile estimator. In response, they propose the nonparametric quantile, defined as $\hat{q}_1(\alpha,\tau) = 2^{-1}(\hat{q}(\alpha,\tau) - \hat{q}(1-\alpha,\tau))$, that improves the symmetry of the sample error distribution. Hence we denote with $\text{VaR}^{(1)}_{t+1,\tau} = \hat{q}_1(\alpha,\tau)\hat{\sigma}_{t+1,\tau}$ the VaR estimator based on the improved nonparametric α-quantile estimator of the error distribution $\hat{q}_1(\alpha,\tau)$. Alternatively, we define a parametric estimator $\hat{q}_2(\alpha,\tau)$, for the α-quantile of the error distribution. It is calculated assuming that the estimated errors $\hat{\epsilon}_{t,\tau}$ follow a scaled t-distribution. The degrees of freedom ν and the scaling factor λ are found using the nonparametric estimator $\hat{q}_1(\alpha,\tau)$ and particularly by solving the equations:

$$\frac{t(\alpha_1,\hat{\nu})}{t(\alpha_2,\hat{\nu})} = \frac{\hat{q}_1(\alpha_1,\tau)}{\hat{q}_1(\alpha_2,\tau)}, \quad \hat{\lambda} = \frac{\hat{q}_1(\alpha_1,\tau)}{t(\alpha_1,\hat{\nu})}. \tag{5.1}$$

In the theoretical comparisons of the quantile estimators Fan and Gu (2003) found that the choice of $\alpha_1 = 0.15$ and $\alpha_2 = 0.3$ is near optimal in terms of efficiency for all degrees of freedom ν. Note here that the assumption of error variance equal to one is violated. For instance, looking at SP returns, the estimated error variance $\hat{\lambda}^2\hat{\nu}/(\hat{\nu}-2)$ is 0.679 and 0.723 for GARCH and EGARCH, 1.273 for LSE and 1.332 for MLE while similar results are found for the remaining series. The latter suggests that the departure from the assumption is not significant, though we need to be cautious when drawing any inference based on this quantile estimator. We denote the VaR estimator based on the parametric quantile with $\text{VaR}^{(2)}_{t+1,\tau}(\alpha) = \hat{q}_2(\alpha,\tau)\hat{\sigma}_{t+1,\tau}$. Hence, define $\Phi^{(i)}(\alpha,\tau) = \{t : |Y_{t,\tau}| > \text{VaR}^{(i)}_{t+1,\tau}(\alpha), t \in [1,n_e]\}$ for $i = 1,2$ then the Exceedence Ratio (ER) in the post sample period for given level α is the percentage of exceeding observations calculated from $\text{ER}^{(i)}(\alpha,\tau) = \#(\Phi^{(i)}(\alpha,\tau))/n_e$.

Overall, for the calculation of the VaR, we consider the combinations of the four conditional variance models with the two quantile estimators introduced above. We perform our calculations for three different holding periods $\tau = 1$ (daily), 10 (fortnight) and 25 (monthly).

Table 5.2: Stock indices: ratio of exceeding observations ($\times 10^{-2}$) for α=5%.

Index	Holding	GARCH		EGARCH		LSE		MLE	
	Period (τ)	$ER^{(1)}$	$ER^{(2)}$	$ER^{(1)}$	$ER^{(2)}$	$ER^{(1)}$	$ER^{(2)}$	$ER^{(1)}$	$ER^{(2)}$
SP	1	5.8	5.4	6.6	5.6	6.4	5.6	6.6	5.6
	10	3.8	5.2	3.0	4.2	4.8	4.8	5.4	5.0
	25	0.4	1.2	0.0	0.8	1.0	1.4	1.2	1.6
DJ	1	5.4	5.4	4.8	4.8	6.8	5.8	6.2	5.4
	10	2.8	4.0	2.4	2.8	5.6	4.8	6.0	5.0
	25	0.2	1.2	0.0	0.6	1.0	0.8	1.8	2.4
FTSE	1	3.8	3.8	2.8	3.0	7.4	6.2	6.6	5.8
	10	1.0	1.8	0.6	1.4	4.2	4.2	4.2	4.2
	25	0.6	0.8	0.0	0.4	1.2	1.6	1.0	2.0
DAX	1	5.8	5.8	6.0	5.2	8.2	7.0	8.2	6.0
	10	2.2	2.2	1.2	1.0	6.2	4.2	7.4	5.4
	25	0.2	0.4	0.0	0.0	2.2	3.0	3.2	3.4

The results for α =5% level, summarized in Table 5.2, reveal that for the holding period $\tau = 1$, the parametric models outperform the nonparametric models with the exception of FTSE and this holds for both quantile estimators. The optimal combination is the EGARCH with the parametric quantile $q_2(\alpha, \tau)$ with the combination of GARCH and $q_2(\alpha, \tau)$ coming second. Note here that, though the nonparametric models performed poorly, the proposed MLE performs relatively better than the LSE and this holds independent of the quantile estimator. When the holding period is increased the variability increases since the prediction involves longer time horizon. In that case, the nonparametric fittings deal in a better way with this increase in variability than the parametric models. Moreover, the proposed MLE along with $q_2(\alpha, \tau)$ is ranked first for $\tau = 10$ and $\tau = 25$ with the combination of LSE and $q_2(\alpha, \tau)$ coming second. Note here that in all cases the parametric quantile $q_2(\alpha, \tau)$ is preferred.

Table 5.3: Stock indices: ratio of exceeding observations ($\times 10^{-2}$) for α=1%.

Index	Holding	GARCH		EGARCH		LSE		MLE	
	Period (τ)	$ER^{(1)}$	$ER^{(2)}$	$ER^{(1)}$	$ER^{(2)}$	$ER^{(1)}$	$ER^{(2)}$	$ER^{(1)}$	$ER^{(2)}$
SP	1	0.6	0.0	1.0	0.4	1.8	0.8	1.6	0.8
	10	0.0	0.4	0.0	0.2	0.2	0.6	0.4	0.8
	25	0.0	0.0	0.0	0.0	0.2	0.2	0.2	0.4
DJ	1	1.0	0.8	0.0	0.0	2.2	1.8	1.8	1.4
	10	0.0	0.2	0.0	0.0	0.4	0.0	0.6	0.4
	25	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.4
FTSE	1	0.6	0.2	0.4	0.4	2.8	1.2	2.6	1.0
	10	0.0	0.4	0.0	0.0	1.4	1.0	1.2	1.0
	25	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.4
DAX	1	1.0	0.2	1.2	0.6	2.4	1.2	2.2	1.4
	10	0.0	0.0	0.0	0.0	1.2	0.2	1.6	1.2
	25	0.0	0.0	0.0	0.0	0.0	0.2	1.2	1.2

Similar conclusions hold for the results of the $\alpha = 1\%$-quantile in Table 5.3. The fitted EGARCH model combined with the parametric quantile yields the closest to 1% ratio for holding period $\tau = 1$. However, VaR estimates using the MLE are more consistent as the holding period increases for all four indices.

Summing up, the analysis of the four stock indices is rather promising for the adaptive ML-estimator. The measures of comparison reveal that between the two nonparametric estimators, the MLE manages to capture the dynamics of the data better than the LSE. Concurrently, the MLE has the smallest Mean Absolute Deviation Error. It also provides more accurate calculation of VaR between the two nonparametric estimators. But even when compared to the parametric models, with the exemption of the holding period of $\tau = 1$ where GARCH provides better VaR estimates, the proposed ML-estimator should be preferred.

5.2.2 Stocks

The second data set consists of the daily log-returns of the following stocks: CitiBank (CB), Coca-Cola (CC), Hewlett-Packard (HP), IBM, JP-Morgan (JPM), Microsoft (MCS), Xerox (XRX), McDonald's (MCD) and Intel (INL) for the period of January 1994 to December 2003. We use the first $n_f = 1500$ observations to fit the models and the last $n_e = 1000$ observations for evaluation of the models. We investigate the behavior of each stock individually. In addition, we study the average performance of the ten different stocks in an attempt to approximate the performance of a portfolio of the stocks by using a univariate approach. It is understood that a multivariate approach that would take into account the interactions is likely to yield safer conclusions and should be preferred when the primary aim is the portfolio analysis. We fit two parametric models: the GARCH(1,1) model with (i) Gaussian and (ii) t-distributed errors. The latter was selected as a result of the earlier discussion concerning the departure of the sample from the Gaussian distribution and the existence of heavy tailed error distribution. Further, we fit the nonparametric model in (2.2) using least-squares (LSE) and the adaptive likelihood estimation (MLE). The selection of the nonparametric regressors is based on the minimization of the Cross-Validation function. In particular, we find that for seven out of ten stocks, the first two lagged variables $\{Y_{t-1}, Y_{t-2}\}$, are significant. However, for the stocks HP, IBM and XRX we conclude that only the first lagged variable $\{Y_{t-1}\}$, is significant. A summary of the results for the CV-function is given in Table 5.4.

We continue using the deviation errors MADE and RADE along with the nonparametric Exceedence Ratio $\mathrm{ER}^{(1)}$ as the performance measures. However, the Exceedence Ratio as Davé and Stahl (1997) quoted "*is only sensitive to the frequency and not the degree with which the loss exceeds the predicted-VaR*". In response, they propose a measure that takes into account the degree of exceedence. This measure depends on the logarithm of the probability of the realized event y_t, in terms of the predicted conditional distribution $\hat{p}_t$, denoted as $\mathcal{L}_t = \log \hat{p}_t(y_t)$.

Table 5.4: Stocks: nonparametric Cross-Validation function ($\times 10^{-6}$).

Stock	$i=1$	$i=2$	$i=3$	$i=1,2$	$i=1,2,3$
CB	6.239	6.156	6.938	6.125	6.131
CC	1.681	1.681	1.697	1.675	1.689
GE	0.237	0.236	0.301	0.231	0.233
HP	5.400	5.481	5.488	5.433	5.467
IBM	0.639	0.643	0.677	0.642	0.641
INL	2.037	2.025	2.121	2.020	2.034
JPM	0.775	0.764	0.808	0.753	0.771
MCD	0.261	0.266	0.289	0.259	0.267
MCS	0.716	0.713	0.759	0.700	0.720
XRX	8.510	8.683	8.831	8.879	8.667

More precisely, if $T(y) = \{t \mid |Y_t| > y,\ t \in [1, n_e]\}$ then the Mean log-Likelihood is defined as

$$l(\alpha) = \frac{\sum_{t \in T(y_\alpha)} \mathcal{L}_t}{\#(T(y_\alpha))}$$

with the exceedence level y_α calculated from $(\#(T(y_\alpha)) - 0.5)/n_e = 0.5(1+\alpha)$ for different values of α. It is understood that the higher the likelihood value the better the performance of the predicted volatility. Indeed, high Mean log-Likelihood values, at large percentile levels α, indicate that the fitted volatility model has captured with success the dynamics of large financial movements.

The results for the RADE and MADE measures are presented in Table 5.5. It appears that for seven out of ten stocks MLE produces the smallest Mean Absolute Deviation Error. At the same time, for 8 out of 10 the square-Root Absolute Deviation Error is in favor of the MLE. Naturally, both deviations measures are in favor of MLE when looking at the average of the ten stocks.

Table 5.5: Stocks: Mean Absolute Deviation Error ($\times 10^{-4}$) and square-Root-Absolute Deviation Error ($\times 10^{-3}$).

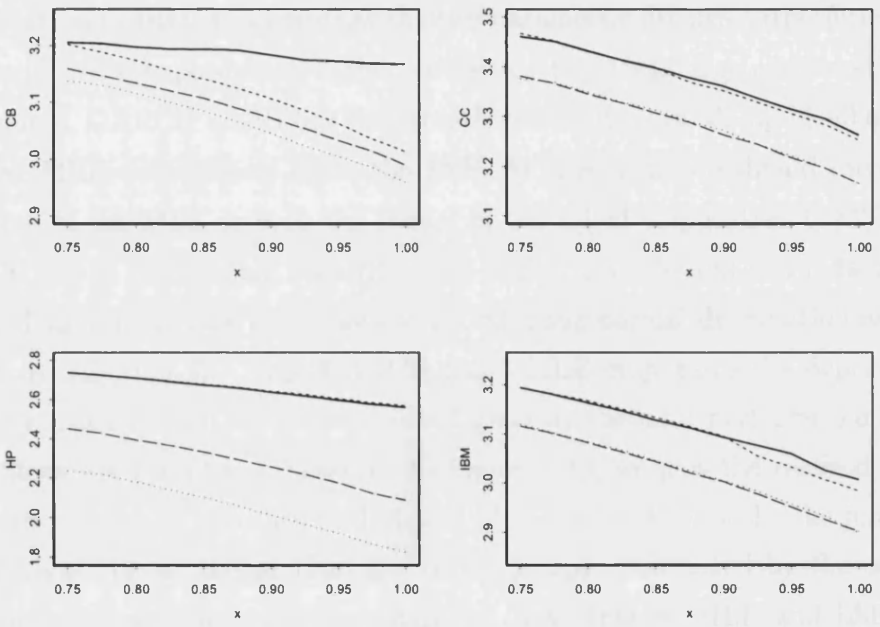
Stock	Measure	GARCH	GARCH-t	LSE	MLE
CB	MADE	1.141	**1.130**	1.154	1.221
	RADE	4.916	**4.878**	5.096	5.028
CC	MADE	0.794	0.700	0.630	**0.629**
	RADE	3.948	3.931	3.695	**3.078**
GE	MADE	0.984	0.989	0.856	**0.848**
	RADE	4.801	4.807	4.526	**4.477**
HP	MADE	2.919	3.000	2.425	**2.242**
	RADE	7.913	8.031	7.340	**7.295**
IBM	MADE	1.164	1.147	1.087	**1.064**
	RADE	5.034	5.029	4.908	**4.813**
INL	MADE	2.46	2.641	2.313	**2.311**
	RADE	7.322	7.463	7.091	**7.007**
JPM	MADE	1.591	1.597	1.350	**1.312**
	RADE	5.911	5.915	5.500	**5.453**
MCD	MADE	0.951	0.954	0.806	**0.778**
	RADE	4.688	4.696	4.323	**4.226**
MCS	MADE	**1.432**	1.448	4.689	2.325
	RADE	5.586	5.585	5.732	**5.502**
XRX	MADE	4.675	4.268	**3.370**	3.418
	RADE	9.748	9.116	**8.055**	8.141
AVER	MADE	1.802	1.789	1.867	**1.635**
	RADE	5.987	5.945	5.627	**5.576**

Table 5.6: Stocks: exceedence ratio ($\times 10^{-2}$) for $\alpha = 5\%$.

Stock	GARCH	GARCH-t	LSE	MLE
CB	4.93	4.93	6.71	5.49
CC	5.98	5.44	5.68	5.66
GE	8.09	7.54	8.08	7.87
HP	4.51	3.88	9.32	9.36
IBM	6.71	7.44	4.19	4.83
INL	9.01	5.87	9.64	9.32
JPM	6.19	6.08	8.60	7.45
MCD	6.65	6.04	9.82	4.85
MCS	7.56	7.23	6.93	6.09
XRX	5.56	7.44	9.52	8.50
AVER	6.52	6.19	7.85	6.94

Table 5.6 entails the results for the exceedence ratio $\mathrm{ER}^{(1)}$ based on the nonparametric quantile estimator $\hat{q}^{(1)}(\alpha, \tau)$ with holding period $\tau = 1$. For five stocks, the GARCH model fitted using t-distribution yields the closest to 5% values with the nonparametric MLE and the GARCH model based on Gaussian conditional error distribution coming second. In addition, a direct comparison between the nonparametric estimators indicates that the adaptive ML-estimator yields better results than the LS-estimator. For instance, by looking at the exceedence ratio for the average of the ten stocks, it is revealed that MLE managed to reduce the deviation from the true percentile to 1.94% from 2.85% for LSE. This corresponds to a relative reduction of almost 32% in the deviation from the true value.

Figure 5.9: Mean log-Likelihood vs percentile for GARCH (solid), GARCH-t (small dashed), MLE (large dashed), LSE (dotted).



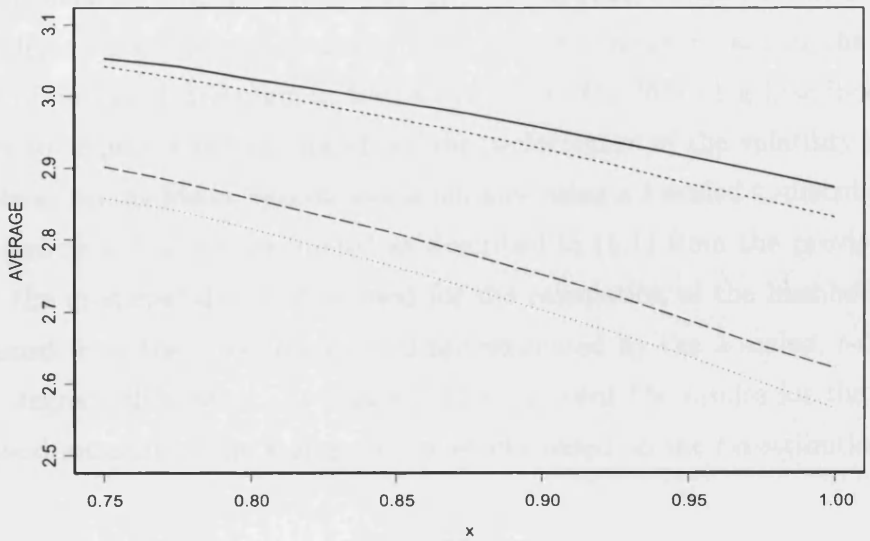


Figure 5.9 summarizes the results for the Mean log-Likelihood measure. It has already been pointed out that higher likelihood values for increasing percentile indicate better model fit that captures the dynamics of the series especially for the extreme events. The first four graphs display the performance of the models for the four stocks CB, CC, HP and IBM. It seems that the two parametric fittings outperform the non-parametric fittings, producing higher values for the Mean log-Likelihood measure. Furthermore, GARCH model has consistently the higher overall log-likelihood values while the MLE outperforms again the LSE. At this point we should mention that the failure of the MLE or even the failure of the tailed emphasized GARCH model (GARCH model fitted using t-distribution) could be attributed to the fact that the likelihood values $\mathcal{L}_t = \log \hat{p}_t(y_t)$ are calculated using normal distribution as the conditional distribution $\hat{p}_t$. Since GARCH-t and MLE emphasize the departure from normality, using normal distribution could generate the poor performance in respect to the Mean log-Likelihood measure. In Figure 5.10, we plot the realized volatility, against the predicted volatility for INL and JPM stocks. Obviously, the jumps of the realized volatility are higher than any of the jumps constructed by the estimators, indicating the need for stochastic volatility. Nevertheless, MLE and LSE seem to capture the dynamics of the data better than the two parametric fittings.

This conclusion, though not mathematically rigorous, contradicts with the results of the Mean log-Likelihood measure, enforcing the concern regarding the validity of the use of normal distribution in the calculation of the Mean log-Likelihood. Hence, in order to acquire a better idea about the performance of the volatility models, we calculate again the Mean log-Likelihood but now using a λ-scaled t_ν-distribution. The two parameters λ, ν, are estimated as described in (5.1) from the previous section. Hence, the predicted distribution used for the calculation of the likelihood measure is obtained from the error distribution approximated by the $\hat{\lambda}$-scaled, t-distribution with $\hat{\nu}$ degrees of freedom. In Figure 5.11 we present the results for the Mean log-Likelihood measure of the average of the stocks based on the t-distribution.

Figure 5.10: Realized and predicted volatility for INL and JPM: LSE (solid top), GARCH (dashed top), MLE (solid bottom), GARCH-t (dashed bottom).

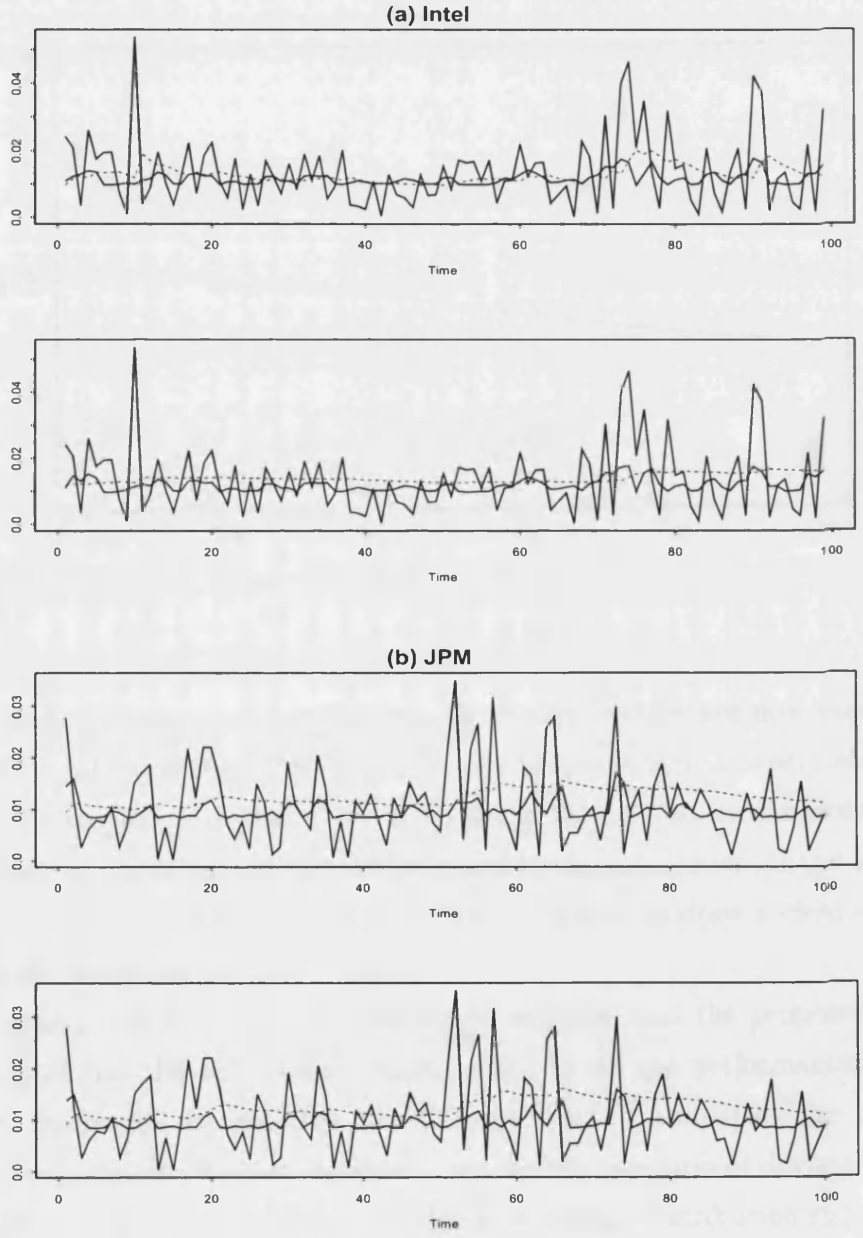
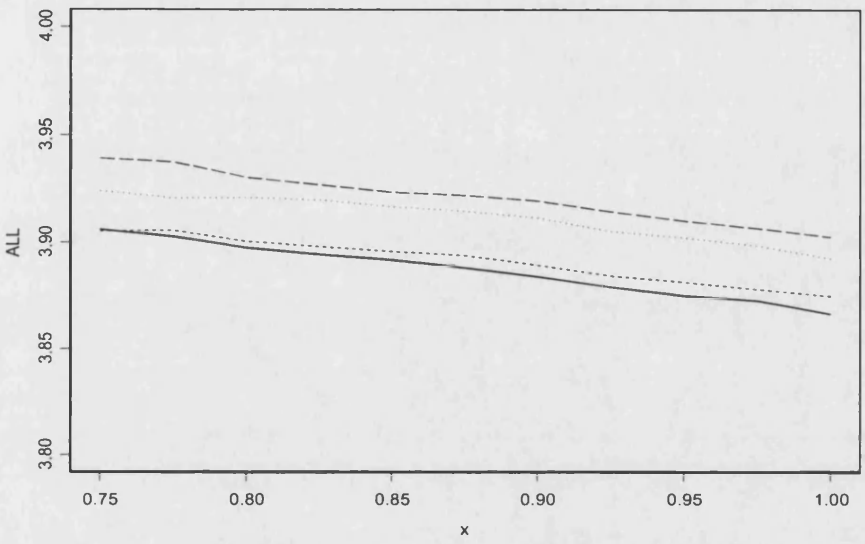


Figure 5.11: Mean log-Likelihood using t-dist. vs percentile for the stock average: LSE (dotted), GARCH (solid), MLE (large dashed), GARCH-t (small dashed).



The result changes dramatically since the highest values are now coming from the adaptive ML-estimator with the alternative nonparametric LS-estimator coming second. It becomes clear that there is no single fitting that outperforms the rest independent of the predicted distribution used in the calculation of the likelihood-based measure. Given this contradiction, we are unable to draw a clear conclusion based on the Mean log-Likelihood measure.

In summary, the analysis of the ten stocks suggests that the proposed adaptive MLE outperforms the LSE, a conclusion backed by all the performance measures. However, the results are not that straightforward when comparing the parametric and the nonparametric fittings together, since the two measures of deviation (MADE and RADE) along with the Mean log-Likelihood using t-distribution suggest a bet-

ter fit for the MLE but the Exceedence Ratio and the Mean log-Likelihood using Gaussian distribution indicate that the parametric models follow more successfully the dynamics of the data.

5.2.3 Exchange rates

The final data set involves the daily log-returns of exchange rates. We define American dollar AD($) as the reference currency and we examine the behavior of the exchange rates with British pound GBP(£), Japanese yen JPY(¥) and Swiss franc SFr. The models are fitted using the data points from the period between Jan 2000 to Dec 2001 i.e. $n_f = 500$ while the observations from period Jan 2002 to Dec 2003, $n_e = 500$, are used for the evaluation of the fitted models. Note here that in our analysis the Euro currency has been omitted due to lack of sufficient observations.

Diebold and Nason (1989) showed that, for the exchange rates, the use of a nonparametric model does not improve the parametric prediction. They studied the exchange rates of American dollar against British pound, Japanese yen, Canadian dollar, along with most of the main European currencies that have now been replaced by euro, for the period of Jan 1973 to Sep 1986. Nevertheless, in their comparison, they only considered the Least Squares estimator of a local polynomial approximation for different polynomial orders, the performance of which has been disappointing so far. Here, similar to the previous sections, we employ two parametric models namely the GARCH using normal and t-distributed errors and the nonparametric model in (2.2) estimated using least squares (LSE) and the proposed likelihood procedure (MLE). The nonparametric regressors are selected by the nonparametric CV-criterion. For the GBP and SFr the model with the first three lagged variables, $\{Y_{t-1}, Y_{t-2}, Y_{t-3}\}$, yields the smallest CV-value of 3.01 and 3.31 ($\times 10^{-7}$) respectively while the smallest CV-value for JPY, 2.14 ($\times 10^{-7}$) corresponds to the model with the two regressors $\{Y_{t-1}, Y_{t-2}\}$. In Figure 5.12 we plot the realized volatility, the predicted volatility from the parametric models as well as the nonparametric MLE and LSE.

Figure 5.12: Realized and predicted volatility for exchange rates, GARCH (dashed-dotted), GARCH-t (dotted), LSE (dashed) and MLE (solid).

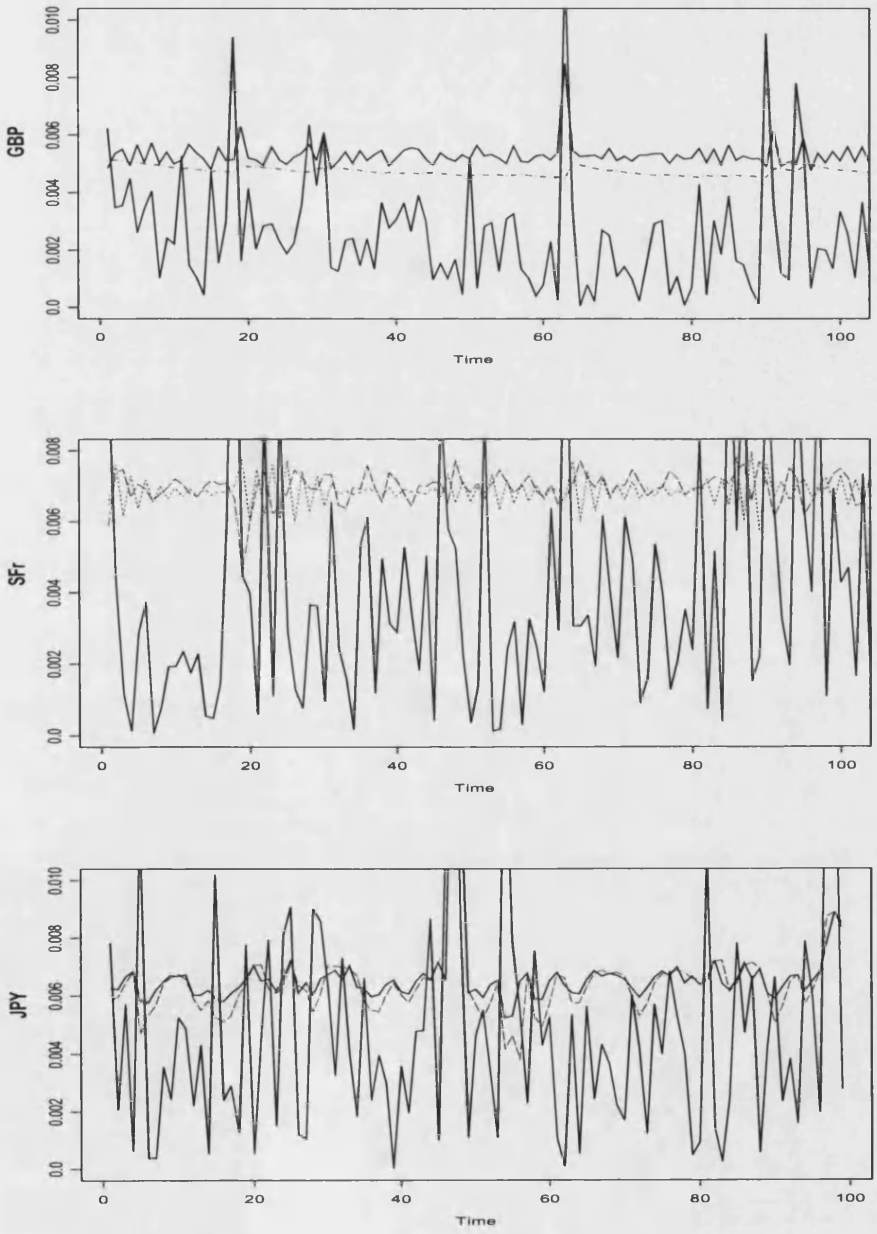


Table 5.7: Exchange rates: deviation measures and hypothesis tests.

Currency	Method	MADE ($\times 10^{-4}$)	RADE ($\times 10^{-3}$)	LR1* p-value	LR2* p-value
GBP	GARCH	0.249	2.492	0.348	0.999
	GARCH-t	0.251	2.411	0.387	0.893
	LSE	0.252	2.420	0.188	0.893
	MLE	0.253	2.431	0.243	0.679
JPY	GARCH	0.389	3.068	0.872	0.999
	GARCH-t	0.387	3.055	0.920	0.891
	LSE	0.408	3.157	0.932	0.578
	MLE	0.403	3.120	0.348	0.999
SFr	GARCH	0.487	3.553	0.867	0.597
	GARCH-t	0.487	3.550	0.816	0.690
	LSE	0.533	3.558	0.920	0.891
	MLE	0.529	3.556	0.928	0.898

We continue with the calculation of the volatility measures of deviation MADE and RADE and the two tests LR1* and LR2* introduced in section 5.2.1. Recall that the hypothesis of the first test was that the α%-exceeding events occur independently according to a bernoulli distribution. The second hypothesis entailed the probability of occurrence of such an event, namely $P(I_t = 1)$. In our case, since $\alpha = 5\%$, the hypothesis is $H_0 : P(I_t = 1) = 0.05$. Note here that in the calculation of the two tests we use the nonparametric quantile that assumes no particular distribution and is calculated from the ordered estimated residuals. Table 5.7 contains the results for MADE and RADE along with the p-values for the two tests. Both the deviation measures are in favor of the parametric estimates and particularly of the GARCH with t-distributed errors. In addition, there is no p-value small enough to reject the

Table 5.8: Exchange rates: exceedence ratio ($\times 10^{-2}$) for $\alpha = 5\%$.

Currency	Method	Normal	t-dist.	Nonpar.
GBP	GARCH	4.4	4.6	5.2
	GARCH-t	4.2	5.2	<u>5.0</u>
	LSE	4.4	6.0	5.8
	MLE	4.2	<u>5.0</u>	4.2
JPY	GARCH	4.8	4.6	5.2
	GARCH-t	<u>5.0</u>	5.8	4.8
	LSE	<u>5.0</u>	6.0	4.2
	MLE	4.0	5.8	4.8
SFr	GARCH	<u>5.0</u>	5.6	5.8
	GARCH-t	5.2	5.2	5.6
	LSE	4.8	5.4	4.8
	MLE	4.8	<u>5.0</u>	4.8

null hypothesis and therefore all models have captured rather satisfactorily the dynamics of the exceeding events for the three exchange rates. The evaluation continues with the calculation of the ratio of exceedence. Apart from the nonparametric $ER^{(1)}$ and the parametric $ER^{(2)}$ ratios of exceedence, introduced earlier, we calculate a third ratio that uses a parametric quantile estimator based on the normal distribution. The results are summarized in Table 5.8. For the exchange rate of British pound, the combination of the nonparametric quantile with the GARCH-t conditional volatility and the combination of the parametric quantile based on t-distribution with the MLE yield the optimal ratio. On the other hand, for the Japanese yen the optimal exceedence ratio is generated by the combination of the parametric quantile based on normal distribution with GARCH-t volatilities and LSE. For the Swiss franc, the re-

sults are in favor of the parametric quantile based on normal distribution combined with the MLE and the parametric quantile based on t-distribution along with the conditional volatilities predicted by GARCH model. Apparently, there is no distinct conclusion drawn from these results though it should be noted that the GARCH-t model and the nonparametric MLE appear more frequently than the other two models as the optimal in respect to the exceedence ratio. At the same time, deviation errors suggest GARCH-t yields the better fit, while between the two nonparametric estimators, ML-estimator once more outperforms the LS-estimator.

5.3 Conclusion

Summing up the results from the analysis of the stock indices, stocks and exchange rates, we conclude the following. Direct comparison of the two nonparametric volatility estimators indicates that the proposed likelihood-based estimator performs better than the estimator calculated using least squares. This conclusion seems to hold for all the financial data sets considered above and is confirmed by the deviation measures, the performance of the VaR-estimators in respect to the hypothesis tests and the exceedence ratio. As far as the parametric and the nonparametric fittings concerns, mixed messages are coming from the different measures. It is understood that all these inferences are drawn from the results derived for the particular data sets considered above. Another important observation is that the restrictive conditions imposed at the introduction of the likelihood estimator do not seem to be an obstacle in the use of the estimator. On the contrary, the performance of the adaptive ML-estimator smooths out any suspicion about its practical implementation.

Undoubtedly, likelihood estimation is computationally demanding. On the other hand, it is an alternative method to the dominant in nonparametric theory least squares estimation that, at least in the above studied cases, improves the accuracy of the prediction. This characteristic encapsulates the main contribution of this work.

Bibliography

Akaike, H. (1974). A new look at statistical model indentification. *IEEE Transactions on Automatic Control 19*, 716–723.

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

Bollerslev, T. (1987). A conditional heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics 69*, 542–547.

Bollerslev, T., R. Y. Chu, and K. F. Kroner (1994). Arch modeling in finance: a selective review of the theory and empirical evidence. *Journal of Econometrics 52*, 5–59.

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction.* New York: Springer.

Bradley, R. C. (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.

Chen, H. and K. Chen (1991). Selection of the splined variables and convergence rates in a partial spline model. *Canadian Journal of Statistics 19*, 323–339.

Cheng, B. and H. Tong (1992). On consistent nonparametric order determination and chaos. *Journal of the Royal Statistical Society, Ser.B 54*, 427–474. (with discussion).

Cheng, B. and H. Tong (1993). On residual sums of squares in non-parametric autoregression. *Stochastic Process and their Applications 48*, 154–174.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review 39*, 841–862.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions. *Numerische Mathematik 31*, 377–403.

Danielsson, J. and C. G. Vries (1997). Tail index and quantile estimation with very high frequency data. *Journal of Empirical Finance 4*, 241–257.

Danielsson, J. and C. G. Vries (2000). Value-at-risk and extreme returns. *Annales d'Economie et de Statistique 60*, 236–269.

Davé, R. and G. Stahl (1997). On the accuracy of var estimates based on the variance-covariance approach. www.citeseer.ist.psu.edu/dave97accuracy.html.

Diebold, F. X. and J. A. Nason (1989). Nonparametric exchange rate prediction? *Journal of International Economics 28*, 315–332.

Duffie, D. and J. Pan (1997). An overview of value at risk. *The journal of Derivatives 4*, 7–49. Spring 1997.

Eberlein, E., J. Kallsen, and J. Kristen (2001). Risk management based on stochastic volatility. FDM-Preprint, University of Freiburg, http://citeseer.ist.psu.edu/eberlein01risk.html.

Embrechts, P., S. Resnick, and G. Samorodnitsky (1998). Living on the edge. *Risk Magazine 11*, 96–100.

Embrechts, P., S. Resnick, and G. Samorodnitsky (1999). Extreme value theory as a risk management tool. *North American Actuarial Journal 3*, 30–41.

Engle, R. and G. Gonzales-Rivera (1991). Semiparametric arch models. *Journal of Bussiness and Economic Statistics 9*, 345–359.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of variance of uk inflation. *Econometrica 50*, 987–1008.

Engle, R. F., D. M. Lilien, and R. P. Robins (1987). Estimating time varying risk premia in the term-structure: the arch-m model. *Econometrica 55*, 391–407.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association 87*, 998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *Annals of Statistics 21*, 196–216.

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

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and its Applications.* London: Chapman and Hall.

Fan, J. and J. Gu (2003). Semiparametric estimation of value at risk. *Journal of Econometrics 6*, 261–290.

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

Fan, J. and Q. Yao (2003). *Nonlinear Time Series.* New York: Springer.

Fan, J., Q. Yao, and Z. Cai (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society, Ser.B 65*, 57–80.

Gao, J. and H. Tong (2002). Model selection and inference in semiparametric regression. Research report, http://www.lse.ac.uk/collections/statistics/research.

Gasser, T. and H. G. Müller (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics 11*, 171–185.

Gouriéroux, C. (1997). *ARCH models and Financial Application.* New York: Springer.

Hall, P. and T. Tao (2002). Relative efficiencies of kernel and local likelihood density estimators. *Journal of the Royal Statistical Society, Ser.B 64*, 537–547.

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

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press.

Härdle, W., P. Hall, and J. S. Marron (1988). How far are automatically chosen regression smoothing parameters from their optimum. *Journal of the American Statistical Association 83*(401), 86–95.

Härdle, W., H. Liang, and J. Gao (2000). *Partially Linear Models.* New York: Springer Series in Contributions to Statistics Physica-Verlag.

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

Hjort, N. L. and M. C. Jones (1996). Locally parametric nonparametric density estimation. *Annals of Statistics 24*(4), 1619–1647.

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

McNeil, A. and R. Frey (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. Preprint, Dept. of Mathematics, ETH Zürich, Switzerland.

Müller, H. G. and U. Stadtmüller (1987). Estimation of heteroscedasticity in regression analysis. *Annals of Statistics 15*, 610–625.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications 9*, 141–142.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica 59*, 347–370.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Roussas, G. G. (1988). Non-parametric estimation in mixing sequencies of random variables. *Journal of Statistical Planning and Inference 18*, 135–149.

Ruppert, D., S. J. Sheather, and M. P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association 90*(432), 1257–1270.

Ruppert, D. and M. P. Wand (1994). Multivariate weighted least squares regression. *Annals of Statistics 22*, 1346–1370.

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

Severini, T. A. (2000). *Likelihood Methods in Statistics*, Volume 22. Oxford Statistical Science Series.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association 88*, 486–494.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika 68*, 45–54.

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

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Ser.B 50*, 413–436.

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

Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics 5*, 595–645.

Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction. *Journal of the Royal Statistical Society, Ser.B 36*, 111–147.

Tjøstheim, D. and B. H. Auestad (1994). Nonparametric identification of nonlinear time series: selecting significant lags. *Journal of the American Statistical Association 89*, 1410–1419.

Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing.* London: Chapman and Hall.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A 26*, 101–116.

Wei, C. Z. (1992). On predictive least squares principles. *Annals of Statistics 20*, 1–42.

Yao, Q. and H. Tong (1994). On subset selection in non-parametric stochastic regression. *Statistica Sinica 4*, 51–70.

Yoshihara, K. (1976). Limiting behavior of u-statistics for stationary absolutely regular processes. *Z. Wahr. verw. Geb. 35*, 237–252.

Yu, K. and M. C. Jones (2004). Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association 99*, 139–144.

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