The London School of Economics and Political Sciences

Dimensionality Reduction in Non-Parametric Conditional Density Estimation with Applications to Nonlinear Time Series

A thesis submitted for the degree of

Doctor of Philosophy

by

Roy Rosemarin

Department of Statistics

The London School of Economics

This work was carried out under the supervision of

Professor Qiwei Yao

London

June 2012

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of approximately 30,000 words.

ABSTRACT

Nonparametric methods of estimation of conditional density functions when the dimension of the explanatory variable is large are known to suffer from slow convergence rates due to the 'curse of dimensionality'. When estimating the conditional density of a random variable Y given random d-vector X, a significant reduction in dimensionality can be achieved, for example, by approximating the conditional density by that of a Y given $\theta^T X$, where the unit-vector θ is chosen to optimise the approximation under the Kullback-Leibler criterion. As a first step, this thesis pursues this 'single-index' approximation by standard kernel methods. Under strong-mixing conditions, we derive a general asymptotic representation for the orientation estimator, and as a result, the approximated conditional density is shown to enjoy the same first-order asymptotic properties as it would have if the optimal θ was known. We then proceed and generalise this result to a 'multi-index' approximation using a Projection Pursuit (PP) type approximation. We propose a multiplicative PP approximation of the conditional density that has the form $f\left(y|x\right) = f_0\left(y\right) \prod_{m=1}^{M} h_m\left(y, \theta_m^T x\right)$, where the projection directions θ_m and the multiplicative elements, h_m, $m = 1, ..., M$, are chosen to minimise a weighted version of the Kullback-Leibler relative entropy between the true and the estimated conditional densities. We first establish the validity of the approximation by proving some probabilistic properties, and in particular we show that the PP approximation converges weakly to the true conditional density as M approaches infinity. An iterative procedure for estimation is outlined, and in order to terminate the iterative estimation procedure, a variant of the bootstrap information criterion is suggested. Finally, the theory established for the single-index model serve as a building block in deriving the asymptotic properties of the PP estimator under strong-mixing conditions. All methods are illustrated in simulations with nonlinear time-series models, and some applications to prediction of daily exchange-rate data are demonstrated.

ACKNOWLEDGMENTS

I would like to take this opportunity to thank some of the people who have helped me bring this thesis to its completion.

I am deeply indebted to Professor Qiwei Yao for introducing me to the topic of dimensionality reduction and directing me throughout this research. Thanks to his willingness to share from his own insightful perspective, but also his readiness to allow me to develop my research directions independently, this research thesis took shape. I am sincerely grateful to him for his invaluable and constructive criticism during my thesis writing and for his encouragement at all stages of my work.

I am most grateful to Professor Yingcun Xia and Professor Piotr Fryzlewicz for willingly accepting to be part of the examination committee and to evaluate this thesis. Their comments made this thesis much more interesting in terms of theoretical and numerical results, and hopefully more reader-friendly.

A very special thanks goes to my beloved partner, Georgia Ladbury, for her ongoing personal support and encouragement. I very much appreciate her effort and kind assistance in proofreading this entire thesis and correcting my grammatical errors. All remaining errors are my own. But more than anything else, I am thankful for her belief in me, which spurred me on towards the completion of this thesis.

This thesis would not have been possible without the financial support provided to me by more than one party. I am very grateful to the London School of Economics and the Department of Statistics, who provided me with generous financial support and facilities to carry my research throughout my Ph.D. studies. I very much appreciate all other friends and family, who were willing to support me financially from time to time. These obviously include my parents, Yehuda and Margalit Rosemarin, my partner, Georgia Ladbury, and also my close friend, Michael Yosipov.

I am very thankful to Ian Marshall, the Research Administrator of the Department of Statistics at the London School of Economics, for the ongoing support and kind assistance that goes beyond simply “doing his job”.

Finally, I would like to dedicate this thesis to my parents for their unconditional love and belief in me.

Roy Rosemarin

February 2012

Contents

Chapter 1

Introduction

1.1 Motivation and Objectives

Conditional probability density functions (c.p.d.f.) provide complete information on the relationship between independent and dependent random variables. As such, they play a pivotal role in applied statistical analysis. Applications include regression analysis (Yin and Cook 2002), interval predictions (Hyndman 1995, Fan and Yao 2003), sensitivity to initial conditions in nonlinear stochastic dynamic systems (Yao and Tong 1994, Fan, Yao and Tong 1996), quantiles estimation and measuring Value-at-Risk (Engle and Manganelli 2004, Wu, Yu and Mitra 2008), and asset pricing (Aït-Sahalia 1999, Engle 2001), among others.

If the conditional density has a known parametric form, then the estimation of the c.p.d.f. reduces to estimation of a finite number of parameters. In particular, if the c.p.d.f. is assumed to be Gaussian then it can be fully characterised by a model for the conditional mean and the variance, e.g. ARMA and GARCH time-series models. However, it is often the case that probability densities are characterised by asymmetry, heavy-tails, multimodality, and possibly other a priorily unknown features. Furthermore, even for known parametric models, c.p.d.f. of nonlinear systems may be hard to derive analytically (see Fan and Yao 2003). In such cases when the form of the c.p.d.f. is unknown or hard to derive, adopting a nonparametric approach can be beneficial.

In this thesis we consider a nonparametric estimation of the c.p.d.f. $f_{Y|X}\left(y|x\right)$ of a random scalar Y given a random d-vector $X = x$. Even for a small dimension of X, $d \geq 2$, a purely nonparametric approach may suffer from poor performance due to the 'curse of dimensionality' and the 'empty space phenomenon' (see Silverman 1986, Section 4.5).

In order to overcome this 'curse', a vast number of techniques have emerged in the literature for reducing the dimensionality of the problem, without losing too many of the main characteristics of the data. These include Principal Component Analysis (see Jolliffe 2002), Factor Analysis (see Gorsuch 1983), Independent Component Analysis (Comon 1994) additive and generalised-additive models (Hastie and Tibshirani 1990, Linton and Nielsen 1995, Horowitz and Mammen 2007), single index models (Powell, stock and Stoker 1989, Härdle and Stoker1989, Ichimura 1993, Delecroix, Härdle and Hristache 2003), inverse regression estimation methods (Li 1991, Cook and Weisberg 1991), MAVE and OPG methods (Xia et al 2002, see also Xia 2007, 2008), and successive direction estimation (Yin and Cook 2005, Yin, Li and Cook 2008), among many others. Dimension reduction techniques aimed directly at estimation of conditional densities were studied by Hall, Racine and Li (2004) and Efromovich (2010), where dimensionality reduction is achieved by attenuation of irrelevant covariates. Hall and Yao (2005) and Fan et al (2009) offered a single-index approximation.

This aim of this thesis is to contribute to this line of research by suggesting two related approximation techniques of the c.p.d.f., based on the information gained by univariate projections of the X-data. In addition, by allowing the data to be stationary strong-mixing, the suggested approximations are shown to be applicable for dependent data, and in particular to the estimation of predictive densities in time-series.

Definition: A stationary process $\{Z_t; t = 0, \pm 1, \pm 2, ...\}$ is said to be strong-mixing or alpha-mixing if

$$\alpha_k = \sup_{A \in F^0_{-\infty},\ B \in F^\infty_k} |P(A)P(B) - P(AB)| \to 0 \quad \text{as} \quad k \to \infty,$$

where F_i^j denotes the σ-algebra generated by $\{Z_t; i \leq t \leq j\}$. We call $\{\alpha_k\}_{k \in \mathbb{N}}$ the mixing coefficients.

As an example, ARMA, GARCH and stochastic volatility processes were proved to be strong-mixing under some mild conditions (cf. Pham and Tran 1985, Carrasco and Chen 2002, Davis and Mikosch 2009), and our method can be applied to these series when the assumption of Gaussianity is not applicable. For a general univariate strong-mixing series $\{z_t\}_{t=1}^{n+d+k-1}$, let

$$y_t = Z_{t+d+k-1}, \quad x_t = (Z_{t+d-1}, ..., Z_t)^T, \quad t = 1, ..., n.$$

Then $f_{Y|X_t}(y_t|x_t)$ provides a k-steps ahead conditional density based on the d-lagged vector x_t, which allows generalising standard time-series models to possibly nonlinear or non-gaussian processes.

1.2 Thesis Outline and Research Contributions

In the second chapter of the thesis, we suggest approximating the conditional density $f(y|x)$ by $f\left(y|\theta^T x\right)$, the conditional density of Y given $\theta^T X = \theta^T x$, where the orientation θ is a scalar-valued d-vector that minimises the Kullback-Leibler (K-L) relative entropy,

$$E \log f(y|x) - E \log f\left(y|\theta^T x\right).$$

The approximated conditional density $f\left(y|\theta^T x\right)$ is estimated nonparametrically by a kernel estimator. In doing so, our approach provides a low dimensional approximation of the conditional density which is optimal under the Kullback-Leibler criterion.

The approach of using the K-L relative entropy for estimation of orientation has been utilised by Delecroix, Härdle and Hristache (2003) in single-index regression, Yin and Cook (2005) for dimension reduction subspace estimation, and by Fan et al (2009), who similar to us, dealt with conditional densities. Yin and Cook (2005) discuss several equivalent presentations of the K-L relative entropy and they show relations to inverse regression, maximum likelihood and other ideas from information theory.

Our work extends the approaches taken by the above papers in two main aspects; First, by allowing the data to be stationary strong-mixing, as discussed in the previous section. As a second contribution, we derive a general asymptotic representation for the difference between the orientation estimator $\widehat{\theta}$ and the unknown optimal orientation θ_0 that is equal to a sum of zero-mean asymptotic Gaussian components with $\sqrt{n}$-rate of convergence and two other, stochastic and deterministic, components. The representation holds for kernels of any order, while the asymptotically dominant terms are determined by the order of kernels in use and the choice of kernel bandwidths.

Kernels of high-order benefit from reduced asymptotic bias in the estimation, yet they take negative values and thus often produce negative density estimates. An investigation by Marron and Wand (1992) of higher order kernels for density estimation concluded that the practical gain from higher order kernels is often absent or insignificant for realistic sample sizes (see also Marron 1992 for graphical insight into the effectiveness of high-order kernels). Our proposed procedure allows estimating θ_0 with high-order kernels, while then estimating the conditional density with non-negative second-order kernels.

The method is illustrated in simulations with nonlinear time-series models, and an application to prediction of daily exchange-rate volatility is demonstrated.

In Chapter 3 of the thesis, we proceed and generalise the result of Chapter 2 to a ‘multi-index’ approximation using a Projection Pursuit type approximation. More precisely, motivated by the Projection Pursuit Density Estimation (PPDE) of Friedman, Stuetzle and Schroeder (1984), we propose a multiplicative projection pursuit approximation of the conditional density that has the form $f\left(y|x\right) = f_0\left(y\right)\prod_{m=1}^{M} h_m\left(y, \theta_m^T x\right)$, where the projection directions θ_m and the multiplicative elements, h_m, $m = 1, ..., M$, are chosen to minimise a weighted version of the Kullback-Leibler relative entropy between the true and the estimated conditional densities. In particular, the single-index approximation of Chapter 2 can be seen as a private case of the projection pursuit approximation when $M = 1$. Indeed, in Chapter 3, the single-index approximation serves as a theoretical building block for the projection pursuit approximation, which allows us to derive the asymptotic properties of the projection pursuit estimator under similar settings.

Other ‘multi-index’ extensions of the single-index c.p.d.f. approximation have been proposed in the literature by Xia (2007) and by Yin, Li and Cook (2008). Both these papers aim to estimate the central dimension reduction subspace spanned by the column of $d \times q$ orthogonal matrix B, $q \leq d$, such that $f\left(y|x\right) = f\left(y|B^T x\right)$ (see Cook 1998). However, while these papers offer a method to estimate the central dimension reduction subspace, estimation of the c.p.d.f. can still be cumbersome to implement, even in the reduced subspace, which may still be of high-dimension. The projection pursuit method offers a different generalisation of the single-index c.p.d.f. approximation, in that it attempts to approximate the c.p.d.f. directly by a multi-index approximation, while it does not necessarily produce an effective estimate of the dimension reduction subspace. Unfortunately, the flexibility of the Projection Pursuit approximation comes at the cost of interpretability, as the obtained estimates for M, θ_m's and h_m's can be hard to interpret in practice.

In the third chapter, we first establish the validity of the projection pursuit approximation by proving some probabilistic properties, and in particular we show that the projection pursuit approximation converges weakly to the true conditional density as M approaches infinity. Similar properties have been proved to hold for the PPDE by Friedman, Stuetzle and Schroeder (1984) and Huber (1985). However, some adaptations of their arguments are required to account for the different nature of the problem discussed in this thesis and the modified Kullback-Leibler criterion for c.p.d.f's, which is in use.

After establishing the theoretical approximation, an iterative procedure for estimation is outlined, based on similar principles as for the projection pursuit density estimation. However, due to the nature of the problem, there is no need to incorporate cumbersome Monte Carlo samplings as in the projection pursuit density estimation, rendering our method simple and computationally undemanding even for very large datasets. In order to terminate the iterative estimation procedure, a variant of the bootstrap information criterion is suggested that has the advantage of avoiding the need to solve an optimisation problem for each bootstrap sample. The asymptotic results derived in Chapter 2 are used to derive the asymptotic properties of the proposed projection pursuit estimator under strong similar mixing conditions.

Finally, the projection pursuit method is illustrated in simulations with nonlinear time-series models, and an application to prediction of daily exchange-rate data is demonstrated.

Chapter 4 briefly concludes and summarises the achieved results and possible directions for future research which arise directly out of the thesis.

Chapter 2

Semiparametric Estimation of Single-Index Conditional Densities for Dependent Data

2.1 Introduction

In this chapter, we consider an approximation of the conditional density $f_{Y|X}(y|x)$ by $f_{Y|\theta^T X}\left(y|\theta^T x\right)$, the conditional density of Y given $\theta^T X = \theta^T x$, where the orientation θ is a scalar-valued d-vector that minimises the Kullback-Leibler (K-L) relative entropy,

$$E \log f_{Y|X}(y|x) - E \log f_{Y|\theta^T X}\left(y|\theta^T x\right). \tag{2.1}$$

The approximated conditional density $f_{Y|\theta^T X}\left(y|\theta^T x\right)$ is estimated nonparametrically by a kernel estimator. In doing so, our approach provides a low dimensional 'single-index' approximation of the conditional density which is optimal under the Kullback-Leibler criterion.

In the single-index regression model (see Ichimura 1993) it is typically assumed that $Y = g\left(\theta^T X\right) + \varepsilon$, where g is some link function and ε is a noise term such that $E(\varepsilon|X) = 0$. Our methodology differs from this regression model by aiming for the most informative projection $\theta^T X$ of X to explain the conditional density of Y given X, rather than just the conditional mean. However, that is not to say that the true conditional distribution of $Y|X$ is assumed to be the same as that of $Y|\theta^T X$. The method aims to provide the optimal single-index conditional density approximation possible for a general $f_{Y|X}(y|x)$.

The asymptotic theory developed throughout the chapter is justified by appealing to a result by Gao and King (2004), who established a moment inequality for degenerate U-statistics of strongly dependent processes, given in Lemma 2.6.4.

Definition 2.1.1 *A U-statistic of general order $m \geq 2$ is a random variable of the form*

$$U_n = \sum_{1 \leq i_1 < ... < i_m \leq n} H\left(Z_{i_1}, ..., Z_{i_m}\right), \tag{2.2}$$

where H is a real-valued function, symmetric in its m arguments, and $X_1, ..., X_n$ are stationary random variables (or vectors). If for any fixed $z_{i_2}, ..., z_{i_m}$ we have

$$E\left(H\left(Z_{i_1}, z_{i_2}, ..., z_{i_m}\right)\right) = 0,$$

then the U-statistic is said to be degenerate.

U-statistics play a key role in the literature in deriving the asymptotic properties of semiparametric index-models for independent observations (e.g. Powell, Stock, and Stoker 1989, Delecroix, Härdle and Hristache 2003 and Delecroix, Hristache and Patilea 2006), and in order to extend this theory to dependent observations we rely heavily on Gao and King's (2004) result.

The outline for the rest of the Chapter is as follows. Section 2.2 states the model's general setting and estimation methodology; Section 2.3 contains the assumptions and main theoretical results; and Section 2.4 presents a numerical study with three simulated time-series examples and exchange-rate volatility series. The proofs of the main theorems are given in Section 2.5, while some other technical lemmas are outlined in Section 2.6.

2.2 Model and Estimation

Let $\{y_j, x_j\}_{j=1}^n$ be strictly stationary strong-mixing observations with the same distribution as (Y, X), where Y is a random scalar and X is a random d-vector. Our aim is to estimate the conditional density $f_{Y|\theta^T X}\left(y|\theta^T x\right)$ of Y given a random d-vector $\theta^T X{=}\theta^T x$, where θ is a vector in $\mathbb{R}^d$ that minimises the K-L relative entropy (2.1). Since the first term of the K-L relative entropy does not depend on θ, minimising K-L relative entropy is equivalent to maximising the expected log-likelihood $E \log f_{Y|\theta^T X}\left(y|\theta^T x\right)$. Clearly, the orientation θ is identifiable only with regards to its direction and sign inversion, and we therefore consider unit-vectors that belong to the compact parameter space

$$\Theta = \left\{\theta \in \mathbb{R}^d : \theta^T\theta = 1,\ \theta_1 \geq c > 0\right\},$$

where θ_1 is the first element of the orientation and $c > 0$ is arbitrarily small. For example, if Y_t is the k-step ahead observation of a time-series and X_t consists of d lagged values of the series, then the constraint that $\theta_1 \neq 0$ represents the belief that the k-step ahead observation depends on the most recent observed value.

In order to ensure the uniform convergence of our estimator, we need to restrict ourselves to a compact subset of the support of $Z = (Y, X)$ such that for any $\theta \in \Theta$ the probability density $f_{Y|\theta^T X}\left(y|\theta^T x\right)$ is well defined and bounded away from 0. Denote such a subspace by $\mathbb{S}$, and let also $\mathbb{S}_X = \left\{x \in \mathbb{R}^d : \exists y \text{ s.t. } (y, x) \in \mathbb{S}\right\}$. Let θ_0 be the maximiser of expected log-likelihood conditional on $Z \in \mathbb{S}$, that is,

$$\theta_0 = \arg\max_{\theta\in\Theta} E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\left(Y|\theta^T X\right)\right), \tag{2.3}$$

where $E_{\mathbb{S}}$ is the conditional expectation given $Z \in \mathbb{S}$. Note that the condition $Z \in \mathbb{S}$ should not have any significant effect on θ_0 if the subset $\mathbb{S}$ is large enough. For ease of presentation, we shall assume that all observations $\{y_j, x_j\}_{j=1}^n$ belong to $\mathbb{S}$.

To estimate θ_0 one can maximise a sample version of (2.3). Define the orientation estimator by $\widehat{\theta} = \arg\max_{\theta\in\Theta} \mathcal{L}(\theta)$, where $\mathcal{L}(\theta)$ is the likelihood function

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \widehat{f}^{-i}_{Y|\theta^T X}\left(y_i|\theta^T x_i\right)\widehat{\rho}_i^{\theta}. \tag{2.4}$$

Here, $\widehat{\rho}_i^{\theta}$ is a trimming term, which is discussed below, and with probability 1 it is eventually equals to 1 for large enough n. The unknown conditional density is estimated by a nonparametric kernel estimator

$$\widehat{f}^{-i}_{Y|\theta^T X}\left(y_i|\theta^T x_i\right) = \frac{\widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right)}{\widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right)}.$$

where $\widehat{f}_{Y,\theta^T X}\left(y, \theta^T x\right)$ and $\widehat{f}_{\theta^T X}\left(\theta^T x\right)$ denote the standard kernel probability density estimates, whereas the superscript '$-i$' indicates exclusion of the i'th observation from the calculation, that is,

$$\begin{aligned} \widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right) &= \left\{(n-1) h_y h_x\right\}^{-1} \sum_{j \neq i} K\left(\frac{y_j - y_i}{h_y}\right) K\left(\frac{\theta^T\left(x_j - x_i\right)}{h_x}\right), \\ \widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right) &= \left\{(n-1) h_x\right\}^{-1} \sum_{j \neq i} K\left(\frac{\theta^T\left(x_j - x_i\right)}{h_x}\right), \end{aligned}$$

where $h_y,\ h_x$ are bandwidths and K is a fixed, bounded-support, kernel function.

The exclusion of a single observation from the calculation should not have asymptotic effect, but is mainly used for theoretical convenience, as it removes any extraneous bias terms that may arise from reusability of data. While this 'leave-one-out' formulation becomes necessary when smoothing parameters are estimated along with the orientation parameter (see Härdle, Hall, and Ichimura 1993), it has been widely used in the literature for single-index modelling even when no estimation of smoothing parameter is involved (cf. Powell, Stok, Stoker 1993, Ichimura 1993, Hall 1989). Note that adding the i'th observation, i.e. the $j = i$ case, contributes a deterministic term of the form $nh^{-1}K(0)$ to each density estimate, and therefore it may stabilise the finite-sample performances as it ensures that all density estimates are positive. On the other hand, adding the term $nh^{-1}K(0)$ to each density estimate seems heuristic, and in general one may consider adding any non-stochastic decaying term $\varepsilon_n \to 0$ to the density estimates with the purpose of ensuring positivity. However, as this method can be very sensitive to the choice of ε_n, here we applied a trimming operator for this purpose.

The trimming term $\widehat{\rho}^{\theta}_i$ that appears in (2.4) is introduced to stabilise the finite-sample performances of the algorithm. To appreciate the role of this term, observe that even if observation (y_i, x_i) belongs to $\mathbb{S}$ it may still be the case that the kernel estimates $\widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right)$, $\widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right)$ rely on very few neighbouring observations, or even none, and as a result these estimates may be close to zero and even non-positive when high-order kernels are used. Including $\log \widehat{f}^{-i}_{Y|\theta^T X}\left(y_i|\theta^T x_i\right)$ in the computation of the likelihood function in such cases may have a drastic adverse effect on the accuracy of the likelihood surface estimates, and it is therefore preferable to trim such terms. Here, we adopted the following simple data-driven trimming scheme, which works very well in practice (For alternative trimming schemes, cf. Härdle and Stoker 1989, Ichimura 1993, Delecroix, Hristache and Patilea 2006, Ichimura and Todd 2006 and Xia Härdle and Linton 2012). For a given observation (y_i, x_i) and $\theta \in \Theta$, let

$$I^i_{n,\theta} = \begin{cases} \mathbf{1}, & \text{if } \min\left\{\widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right), \widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right)\right\} > a_0 n^{-c}, \\ 0, & \text{otherwise,} \end{cases}$$

for some small constants $a_0, c > 0$. As $I^i_{n,\theta}$ depends on θ it needs to be normalised to account for the actual number of observations considered in the computation of $\mathcal{L}(\theta)$, and hence we take

$$\widehat{\rho}^{\theta}_i = I^i_{n,\theta} \Big/ \frac{1}{n} \textstyle\sum_{i=1}^n I^i_{n,\theta}. \tag{2.5}$$

Thus, the trimming term $\widehat{\rho}^{\theta}_i$ is completely data-driven in the sense that it depends on the set of observations, and in addition, it also depends on the value of the parameter θ, evaluated by the likelihood. However, it does not assume any prior knowledge or applying a pilot estimation of θ_0. We show in appendix B that if c is sufficiently small, then $\widehat{\rho}^{\theta}_i$ eventually equals to 1 for any large enough n with probability 1. Therefore, $\widehat{\rho}^{\theta}_i$ has no asymptotic effect on the method performance.

It is common in single-index regression models, that the optimal kernel's bandwidths for orientation estimation undersmooth the nonparametric estimator of the link function, in the sense that these bandwidths have a faster rate of decay than the optimal rate for purely nonparametric estimation (cf. Hall 1989, p. 583). Our theory, presented in the next section, indicates that a similar property arises in single-index conditional density estimation. It is therefore the reason that a second stage of estimation is utilised when $f_{Y|\theta^T X}\left(y|\theta^T x\right)$ is estimated with the orientation estimate $\widehat{\theta}$ and with optimal-rate bandwidths H_y and H_x for nonparametric estimation. Moreover, the 'two stages' procedure allows estimating θ_0 with high-order kernels, while then estimating the conditional density with non-negative second-order kernels.

Notice that although θ_0 is defined with respect to the true conditional density, it is possible that the orientation estimator, $\widehat{\theta}$, will be more suitable to use at the second-stage with the same bandwidths and kernel functions as in the first-stage. However, this is likely to happen only in very small sample sizes as the increase in accuracy achieved by using optimal-rate bandwidths and non-negative kernels is likely to take effect very quickly. For further discussion, see also Hall (1989, pp. 583-4)

Our final c.p.d.f. approximation is obtained by using all observations in the calculation, non-negative symmetric kernels $\widetilde{K}(\cdot)$, and with bandwidths H_y and H_x in place of h_y and h_x, that is

$$\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right) = \frac{\frac{1}{nH_yH_x}\sum_{j=1}^{n}\widetilde{K}\left(\frac{y_j-y}{H_y}\right)\widetilde{K}\left(\frac{\widehat{\theta}^T(x_j-x)}{H_x}\right)}{\frac{1}{nH_x}\sum_{j=1}^{n}\widetilde{K}\left(\frac{\widehat{\theta}^T(x_j-x)}{H_x}\right)}.$$

The following section presents the asymptotic properties of $\widehat{\theta}$ and $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)$.

2.3 Asymptotic Results

We introduce some new notations that will be used throughout the section and in the proofs. For a function $g(\theta)$ of $\theta \in \Theta$ and possibly of other variables, let $\nabla g(\theta)$ and $\nabla^2 g(\theta)$ be the vector and matrix of partial derivatives of $g(\theta)$ with respect to θ, i.e.,

$$\{\nabla g(\theta)\}_k = \frac{\partial g(\theta)}{\partial \theta_k} \quad \text{and} \quad \left\{\nabla^2 g(\theta)\right\}_{k,l} = \frac{\partial^2 g(\theta)}{\partial\theta_k \partial\theta_l}, \quad k,l \in \{1,...,d\}.$$

Denote now $Z = (X, Y)$ and

$$\begin{aligned}\Psi(\theta) &= E_{\mathbb{S}}\left[\nabla \log f_{Y|\theta^T X}\left(Y|\theta^T X\right) \nabla \log f_{Y|\theta^T X}\left(Y|\theta^T X\right)^T\right], \\ \Omega(\theta) &= E_{\mathbb{S}}\left[-\nabla^2 \log f_{Y|\theta^T X}\left(Y|\theta^T X\right)\right].\end{aligned}$$

where $E_{\mathbb{S}}$ is the conditional expectation given $Z \in \mathbb{S}$. For some small $\delta > 0$, define also the set $\mathbb{S}_\delta$ distant no further than $\delta > 0$ from some $\left(y, \theta^T x\right)$ such that $(y, x) \in \mathbb{S}$ and $\theta \in \Theta$.

The following assumptions are required to obtain the asymptotic results for the orientation estimator $\widehat{\theta}$.

(A1) The sequence $\{y_j, x_j\}_{j=1}^n$ is strictly stationary strong-mixing series with mixing coefficients that satisfy $\alpha_t \leq A\alpha^t$ with $0 < A < \infty$ and $0 < \alpha < 1$.

(A2) $K(\cdot)$ is a symmetric, compactly supported, boundedly differentiable kernel.

(A3) The bandwidths satisfy $h_y, h_x = o(1)$ and $n^{1-\delta} h_y h_x \to \infty$ for some $\delta > 0$.

(A4) For all $\theta \in \Theta$, $\left(Y, \theta^T X\right)$ has probability density $f_{Y,\theta^T X}(y, t)$ with respect to Lebesgue measure on $\mathbb{S}_\delta$ and $\inf_{(y,t)\in\mathbb{S}_\delta} f_{Y,\theta^T X}(y, t) > 0$. $f_{Y,\theta^T X}(y, t)$, $E\left(X|Y = y, \theta^T X = t\right)$ and $E\left(XX^T|Y = y, \theta^T X = t\right)$ are twice continuously differentiable with respect to $(y, t) \in \mathbb{S}_\delta$. Moreover, there is some j^* such that for all $j > j^*$ and $\left(y_1, \theta^T x_1\right), \left(y_j, \theta^T x_j\right) \in \mathbb{S}_\delta$ the joint probability density of $\left(y_1, \theta^T x_1, y_j, \theta^T x_j\right)$ is bounded.

(A5) For the trimming operator, we require that $a_0, c > 0$ and $n^c\left(h_y^2 + h_x^2\right) = o(1)$ and $n^{1-2c-\delta} h_y h_x \to \infty$ for some $\delta > 0$.

(A6) For all $\theta \in \Theta$, $E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\right)$ is finite and it has a unique global maximum θ_0 that lies in the interior of Θ.

We further require that $K(\cdot)$ is a p'th-order kernel function, such that

$$\int u^j K(u)\, du = 0 \quad \text{for} \quad j = 1, ..., p-1, \quad \text{and} \quad \int u^p K(u)\, du \neq 0,$$

for $p \geq 2$. We then make the following assumptions.

(A7) $K(\cdot)$ is p'th-order kernel with $p \geq 2$, and it is three times boundedly differentiable.

(A8) The bandwidths $h_y,\ h_x$ satisfy $n^{2-\delta} h_y h_x^5 \to \infty$ for some $\delta > 0$.

(A9) $f_{Y,\theta^T X}(y,t)$ and $E\left(X|Y=y,\theta^T X=t\right)$ and $E\left(XX^T|Y=y,\theta^T X=t\right)$ are $(2+p)$-times continuously differentiable with respect to $(y,t)\in\mathbb{S}_\delta$.

Conditions (A1)-(A6) are needed for uniform consistency of the log-likelihood function on $\Theta\times\mathbb{S}$, and therefore for consistency of $\widehat{\theta}$. In particular, condition (A1) allows the data to come from a strong-mixing process. As an example, ARMA, GARCH and stochastic volatility processes satisfy condition (A1) (cf. Pham and Tran 1985, Carrasco and Chen 2002, Davis and Mikosch 2009). Condition (A2) requires that $K(\cdot)$ is symmetric and therefore it is of second-order at the least. Condition (A3) on the bandwidths is needed to obtain uniform convergence of the kernel density estimators. In condition (A4), the bound on the joint probability density of $\left(\theta^T X_1, Y_1, \theta^T X_j, Y_j\right)$ may not hold for $j\leq j^*$, which allows components of X_1 and X_j to overlap for some small j's, as in the case where X_t consists of multiple lags of Y_t. For example, if $x_t=\left(y_{t-1},...,y_{t-d}\right)^T$ for $d\geq 2$ then the joint probability density of $\left(Y_0,\theta^T X_0, Y_j,\theta^T X_j\right)$ is unbounded for $j<d$ because the components of X_0 and X_j overlap. Condition (A5) for the trimming operator terms is derived from Lemma 2.6.5 in the appendix. (A6) is an identifiability requirement for θ_0. In the case where $E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\right)$ has more than a single global maximum, θ_0 can be any one within the set of maxima points, and our asymptotic results will still apply as long as this θ_0 is a local maximum in a small neighbourhood. Note that in that case, the choice of θ_0 within the set of optimum points is not crucial as long as the approximation of $f_{Y|X}(Y|X)$ is concerned. Conditions (A7)-(A9) are stronger versions of (A2)-(A4) and are needed for the derivation of the rate of consistency of $\widehat{\theta}$. In these conditions, the order of the kernel function $K(\cdot)$ is set to $p\geq 2$, which has to be an even number if $K(\cdot)$ is a symmetric by (A2). Kernels of high-order may take negative values, and thus they often produce negative density estimates. However, the trimming scheme, introduced in Section 2.2, is designed to trim non-positive estimates from the log-likelihood calculation, and therefore any potential problems caused by non-positive density estimates are avoided. Condition (A8) discusses rate of decay for the bandwidths. For example, if both bandwidths h_y, h_x are taken to be proportional to $n^{-\gamma}$, $\gamma>0$, then γ must satisfy $0<\gamma<\frac{1}{3}$.

The following theorem, proved in Section 2.5. shows the consistency of $\widehat{\theta}$.

Theorem 2.3.1 *Let (A1)-(A6) hold. Then as $n \to \infty$*

$$\widehat{\theta} \to_p \theta_0.$$

As an implication of Theorem 2.3.1 and the fact that both θ_0 and $\widehat{\theta}$ are unit-vectors, it follows from a simple geometric argument that the difference $\widehat{\theta} - \theta_0$ can be approximated up to first-order asymptotics by $\widehat{\theta}^{\perp}$, the projection of $\widehat{\theta}$ into the plane orthogonal to θ_0, i.e.,

$$\widehat{\theta} - \theta_0 = \widehat{\theta}^{\perp} + o_p\left(\left\|\widehat{\theta} - \theta_0\right\|\right).$$

Since $f_{Y|\theta^T X}\left(Y|\theta^T X\right)$ depends only on the direction of θ, then for any vector $\theta \in \mathbb{R}^d$ we get that both vector $\nabla \log f_{Y|\theta^T X}\left(Y|\theta^T X\right)$ and the column (row) space spanned by matrix $\nabla^2 \log f_{Y|\theta^T X}\left(y|\theta^T x\right)$ are perpendicular to θ. Indeed, this can also be seen directly from Lemma 2.6.1 in appendix B. Note, however, that by conditions (A6) and (A9) there is a generalised inverse of $\Omega\left(\theta_0\right)$, denoted $\Omega\left(\theta_0\right)^{-}$, that is well defined in the perpendicular space to θ_0 (cf. Theorem 3.1 of White 1982). Let now

$$V\left(\theta_0\right) = \Omega\left(\theta_0\right)^{-} \Psi\left(\theta_0\right) \Omega\left(\theta_0\right)^{-}.$$

The next theorem gives a general second-order asymptotic representation for $\widehat{\theta} - \theta_0$.

Theorem 2.3.2 *Let (A1)-(A10) hold. Then*

$$\widehat{\theta} - \theta_0 = n^{-1/2} V\left(\theta_0\right)^{1/2} Z + O_p\left(n^{2-\delta} h_y h_x^3\right)^{-1/2} + O(h_y^p + h_x^p),$$

where Z is asymptotically normal $N\left(0, I\right)$ random d-vector and $\delta > 0$ arbitrarily small.

Similar results to Theorems 2.3.1 and 2.3.2 were derived by Delecroix, Härdle and Hristache (2003) in the context of single-index regression. However, they assumed independent observations and fourth-order kernels, and they obtained

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right)\rightarrow N(0,V\left(\theta_0\right)).$$

Yin and Cook (2005) also dealt with a similar model for the purpose dimension reduction subspace estimation, and they derived consistency of $\widehat{\theta}$ under independence.

Fan et al (2009) were the first to suggest applying the Kullback-Leibler criterion to a single-index approximation of the conditional density. They also assumed independent observations and obtained

$$\widehat{\theta}-\theta_0=O_p\left(n^{2-\delta}h_yh_x^3\right)^{-1/2}.$$

Thus, Theorems 2.3.1 and 2.3.2 extends these papers by allowing the data to be stationary strong-mixing, and by offering a general second-order asymptotic representation for $\widehat{\theta}$ that is holds for kernels of any order.

The proofs of Theorems 2.3.1 and 2.3.2 are given in Appendix A of the Chapter. The idea of the proofs of is to look first at

$$L\left(\theta\right)=n^{-1}\sum_{i=1}^{n}\log f_{Y|\theta^TX}\left(y_i|\theta^Tx_i\right),$$

which is a version of the likelihood $\mathcal{L}\left(\theta\right)$ when conditional density estimates are replaced by the true conditional densities. Deriving the asymptotic properties of $L\left(\theta\right)$ is straightforward by the ergodic theorem and CLT for strong-mixing processes (see Fan and Yao 2003, Proposition 2.8 and Theorem 2.21).

At a second step in the proofs, we look at the difference $\mathcal{L}(\theta)-L\left(\theta\right)$. By Lemma 2.6.5, all trimming-terms, $\widehat{\rho}_i^{\theta}$, $i=1,...,n$, defined in (2.5), are eventually equal to 1 for any large enough n with probability 1. Therefore, we can consider n to be large enough so $\widehat{\rho}_i^{\theta}\equiv 1$ and

$$\mathcal{L}(\theta)-L\left(\theta\right)=n^{-1}\sum_{i=1}^{n}\log\left(\widehat{f}_{Y|\theta^TX}^{-i}\left(y_i|\theta^Tx_i\right)\Big/f_{Y|\theta^TX}\left(y_i|\theta^Tx_i\right)\right).$$

For Theorem 2.3.1, it is then sufficient to prove that

$$\sup_{\theta\in\Theta}\left|\mathcal{L}\left(\theta\right)-L\left(\theta\right)\right|=o_p\left(1\right).$$

In the proof of Theorem 2.3.1, the main effort is in the establishment of an asymptotic bound for the difference

$$\nabla\mathcal{L}\left(\widehat{\theta}\right)-\nabla L\left(\widehat{\theta}\right). \tag{2.6}$$

We write this difference as a sum of few U-statistic terms. The main idea in the derivation is to perform the Hoeffding's decomposition on the U-statistic terms (Lemma A, pp. 178 in Serfling 1980), and then apply the result for degenerate U-statistics of strong-mixing processes by Gao and King (2004, Lemma C.2). We then manage to establish a uniform bound for (2.6) over a shrinking neighbourhood of θ_0, in the sense that $\theta\to_p\theta_0$ implies that

$$\nabla\mathcal{L}\left(\theta\right)-\nabla L\left(\theta\right)=O_p\left(n^{2-\delta}h_yh_x^3\right)^{-1/2}+O(h_y^p+h_x^p)+o_p\left(\theta-\theta_0\right). \tag{2.7}$$

Nolan and Pollard (1987) and Sherman (1994) developed a general uniform convergence theory for U-statistics, and some applications include Ichimura (1995), Zheng (1998) and Wang (2006). However, these results were obtained under assumption of independence, while as far as we are aware, there is no general theory for uniform convergence of U-statistics under mixing conditions. In the proof of Theorem 2.3.1, the property (2.7) is achieved using a Lipschitz continuity property of the kernel functions in a similar manner to Theorem 2 of Hansen (2008), see also the proof of Lemma 2.6.3 in Appendix B.

Theorem 2.3.2 implies that the choice of bandwidth h_x has a greater impact on the rate of convergence of $\widehat{\theta}$ than that of the bandwidth h_y. This is due to the fact the orientation vector, which appears within the kernel density estimates, is related to the X variable through the function $K\left(\frac{\theta^T(x_j-x_i)}{h_x}\right)$. In particular, the symmetry between h_x and h_y breaks when considering the score function $\nabla\mathcal{L}(\theta)$, as one gets an additional factor of h_x from the inner derivative w.r.t. θ.

It is clear from this theorem that for $\widehat{\theta}$ to be $\sqrt{n}$-consistent estimator of θ_0, one needs

$$\left(n^{2-\delta}h_y h_x^3\right)^{-1/2} \leq n^{-1/2} \quad \text{and} \quad h_y^p + h_x^p \leq n^{-1/2}. \tag{2.8}$$

However, it is easy to see that both conditions cannot be satisfied if $p = 2$, and hence the $\sqrt{n}$-convergence rate is not achieved in that case (cf. Remark 2 of Fan et al 2009), although the convergence rate can still become arbitrarily close to $\sqrt{n}$. By increasing the order of the kernel to $p = 4$, the condition (2.8) can be fulfilled under $h_y, h_x \leq n^{-1/2p}$ and $h_y h_x^3 \geq n^{\delta-1}$, and if the two last inequalities are strict, then the Theorem implies asymptotic normality of the estimate.

The asymptotic expression given by Theorem 2.3.2 at the limit $\delta \to 0$ suggests that the optimal bandwidths h_y and h_x have both the asymptotic rate $n^{-1/(p+2)}$, where p is the kernel's order. Taking $p = 2$, for example, we have that the optimal bandwidths are of asymptotic order $n^{-1/4}$. This optimal rate reflects undersmoothing of the kernel estimator, which is a typical requirement in many single-index models.

As an immediate consequence of Theorems 2.3.1 and 2.3.2, we get that under appropriate choice of bandwidths, $\widehat{\theta}$ can converge fast enough to θ_0 so that $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)$ estimates $f_{Y|\theta_0^T X}\left(y|\theta_0^T x\right)$ with the same first-order asymptotic properties as if θ_0 was known. The Theorem below formalises this idea.

Theorem 2.3.3 *Let (A1)-(A10) hold and* $H_y H_x / h_y h_x^3 = o\left(n^{1-\delta}\right)$ *for some* $\delta > 0$ *and* $H_y H_x \left(h_y^{2p} + h_x^{2p}\right) = o\left(n^{-1}\right)$. *In addition let* $\widetilde{K}$ *be a symmetric, compactly supported, boundedly differentiable kernel, and* $H_y, H_x = O\left(n^{-1/6}\right)$ *and* $\frac{\ln n}{nH_y H_x} = o(1)$. *Then for any* $\delta > 0$,

$$\sup_{(y,x)\in\mathbb{S}} \left|\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right) - f_{Y|\theta_0^T X}\left(y|\theta_0^T x\right)\right| = O_p\left(\left(\frac{\ln n}{nH_y H_x}\right)^{1/2}\right). \tag{2.9}$$

Notice that the exact rate of consistency for conditional density kernel estimator is $(nH_y H_x)^{-1/2}$ (Robinson 1983). The $\ln n$ term in the RHS of (2.9) is needed to get a uniform rate of convergence by Lemma 2.6.2. This upper bound for the uniform rate of convergence was proved to be accurate for i.i.d. case by Bickel and Rosenblatt (1973, see also Fan and Yao 2003, Theorem 5.4). As far as we are aware, Bickel and Rosenblatt's result was not generalised to the dependent case. Nevertheless, we would expect this to be the case for general stationary process under certain mixing conditions.

Theorem 2.3.3 is proved in Appendix A of the Chapter.

2.4 Implementation and Simulations

In this section, we discuss implementation of the proposed method and we examine its finite-sample properties over few simulated time-series models.

In all of the simulations we used the three-time differentiable and IMSE optimal kernels with support $(-1,1)$, derived by Müller (1984) and specified below. The second-order Müller's kernel, also known as the Triweight kernel, is given for $u \in (-1,1)$ by

$$K(u) = 35/32 \cdot \left(1 - 3u^2 + 3u^4 - u^6\right), \tag{2.10}$$

and the fourth-order Müller's kernel is given for $u \in (-1,1)$ by

$$K(u) = 315/512 \cdot \left(3 - 20u^2 + 42u^4 - 36u^6 + 11u^8\right).$$

For the estimation of the conditional density $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)$ we use only the non-negative Triweight kernel.

In order to facilitate the implementation, we standardised $x_j = (x_{j1}, ..., x_{jd})$ by setting $x_j \leftarrow S_x^{-1}(x_j - \overline{x})$ and we standardised y_j by setting $y_j \leftarrow (y_j - \overline{y})/s_y$, where $\overline{x}$ and $\overline{y}$ are the vector and scalar sample means of $\{x_j\}_{j=1}^n$ and $\{y_j\}_{j=1}^n$, and S_x^2 and s_y^2 are the $d \times d$-matrix and the scalar sample variances. This procedure allows us to disregard scaling parameters and to apply the same smoothing parameter for each direction of $\theta \in \Theta$, in accordance with Scott's (1992) normal reference rule. Once the two-stage estimation procedure was complete, the estimates of the orientation and the conditional density were transformed back to the original coordinates by setting $\widehat{\theta} \leftarrow S_x^{-1}\widehat{\theta}/\left\|S_x^{-1}\widehat{\theta}\right\|$

and $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right) \leftarrow \widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)/s_y$.

We now provide a brief discussion on the topic of bandwidths selection. Typical bandwidths selection methods proposed in the literature of single-index models usually suffer from heavy computational burden. Härdle, Hall, and Ichimura (1993) proposed to optimise the least squares criterion function over the orientation coefficients as well as the bandwidth. A related iterative procedure of alternately estimating the orientation and the bandwidth was suggested by Xia, Tong, Li (1999). Fan and Gijbels (1995a,b) combined goodness of fit and plug-in steps to achieve variable bandwidth. Fan and Yim (2004) and Hall, Racine and Li (2004) discussed cross-validation techniques for bandwidth selection. Hall and Yao (2005) utilised a bootstrap approach based on the linear model to choose a bandwidth. Moreover, most of the mentioned methods require applying the algorithm to any $\theta \in \Theta$, or at least to some pilot estimator of θ_0. The computational burden in such methods may be particularly noticeable in models like ours, where the estimation requires solving a numerical multivariate optimisation problem. In practice, however, various prior numerical studies that we carried out with different selection rules for h_y and h_x demonstrated that the orientation estimator is very robust to the choice of bandwidths as long as the bandwidths are not too small. Motivated by the single-index regression algorithm of Xia, Härdle, Linton (2012), we propose the following iterative procedure that successfully reconciles effective bandwidth selection with fast and robust numerical optimisation.

Step 0. Let $\widehat{\theta}^0 \in \Theta$ be any initial guess for θ_0, for example $\widehat{\theta}^0 = (1, 0, ..., 0)$. Set also a finite sequences of decreasing bandwidths $h_y^i = h_x^i = a^i n^{-1/(p+2)}$, $i = 1, ..., I$, where p is the kernel-order and $\{a^i\} > 0$ is a decreasing sequence such that the first bandwidths notably oversmooth the unconditional density and the last one is chosen, e.g., by Scott's (1992) normal reference rule. In our simulations, we used $\left(a^1, a^2, ..., a^I\right) = (9, 8, ..., 3)$, which yields good results. Set the iteration number $i = 1$.

Step 1. Apply a multivariate variant of the Newton-Raphson method with starting point $\widehat{\theta}^{i-1}$ to find a maximum log-likelihood estimate $\widehat{\theta}^i$ numerically based on bandwidths h_y^i and h_x^i. In our simulations, we use the Broyden-Fletcher-Goldfarb-Shanno BFGS method* (see Nocedal and Wright 2006, Chapter 6).

Step 2. Stop the procedure and use the estimate $\widehat{\theta} = \widehat{\theta}^i$ either if $i = I$ or if a certain convergence criterion is met, i.e. if $\left(\widehat{\theta}^i\right)^T \widehat{\theta}^{i-1} > 1 - \varepsilon$ for some small $\varepsilon > 0$. Otherwise, set $i \leftarrow i + 1$ and $h_y^i = h_x^i = a^i n^{-1/(p+2)}$, and return to Step 1.

Note that since $h_y^1 = h_x^1 = 9n^{-1/(p+2)}$ are chosen to oversmooth the conditional density in the first iteration of estimation, the corresponding likelihood surface is thus oversmoothed as well, and the optimisation algorithm is insensitive to the choice of $\widehat{\theta}^0$. On the other hand, if we simply use one step of maximization with only $h_y = h_x = 3n^{-1/(p+2)}$, then the algorithm is very likely to converge to some local maximum, depending on the starting point $\widehat{\theta}^0$ provided. Having said that, one needs to be aware that if the expected likelihood surface is truly multimodal, oversmoothing the likelihood surface may lead to convergence of the procedure to a locally optimal parameter, rather than to a globally optimal one.

For the second stage estimator of the conditional density, $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)$, Theorem 2.3.3 assumes the bandwidths H_y and H_x do not change when the fit is carried out on the data. However, in practice, if the variables Y and $\widehat{\theta}^T X$ have a nonuniform distribution, a constant bandwidths may lead to problems caused by undersmoothing in some sparse neighborhoods. In order to overcome this problem, the nearest neighbour bandwidth estimator chooses the bandwidths to ensure that sufficient amount data is contained in the calculation. More specifically, the nearest neighbor bandwidth for a density estimate at point z, given stationary observations $z_1, ..., z_n$, is computed as the distance from z to its k'th nearest neighbour among $z_1, ..., z_n$ (cf. Silverman 1986, Section 2.5, Loader 1999, Section 2.2.1, Scot 1999, Section 6.6). Thus, for example, in the tails of the distribution, the nearest neighbour bandwidth will be larger than in the centre of the distribution, and so the sparsity problems in the tails are reduced. However, nearest neighbour estimates are generally not probability densities as they do not necessarily integrate to one as the tails of the estimate may approach zero very slowly (see Silverman 1986, Section 2.5).

*The code for the BFGS algorithm was published by Daniel F. Heitjan

Other alternative methods for adaptive bandwidth are typically more complicated, such as the locally adaptive two-stage kernel estimator of Breiman, Meisel and Purcell (1977), and the supersmoother algorithm of Friedman and Stuetzle (1982), which chooses bandwidths based on a local cross validation method. Few other local goodness of fit approaches for bandwidth selection are discussed in Chapter 11 of Loader (1999). However, according to Loader (1999), these locally adaptive procedures work very well on problems with plenty of data, obvious structure and low noise, while when the data structure is not obvious, simpler methods of bandwidth selection are generally preferable. Moreover, locally adaptive procedures are charachterised by tuning parameters, or penalties, and the estimates can be quite sensitive to the choice of these parameters.

In our simulation study, we have considered at first an application of the nearest neighbour procedure for the second stage of estimation by using the R statistical package 'locfit' of Clive Loader. This package allows combining a fixed bandwidth with a nearest neighbor bandwidth, such that the final bandwidth is determined as the maximum amongst both components. An asymptotic analysis by Moore and Yackel (1977) suggest that the parameter k in the nearest neighbour estimator is linked to the bandwidth h in a nonparametric kernel estimation by $k = nh^d$. Hence, for our study, Scott's (1992) normal reference rule for bandwidth selection suggests using bandwidths that are taken as the maximum between $an^{-1/6}$ and the k'th nearest neighbour with parameter $k = an^{5/6}$, where for Triweight kernel (2.10) $a \approx 3$. Nevertheless, after gaining some experience with the method, we came to conclude that while the nearest neighbour method is more demanding in terms of computational time, it offered no significant benefit relatively to the fixed bandwidth in terms of minimising $RMSPE$ in our Monte Carlo study or prediction of out-of-sample conditional tail quantiles in a real-data example. In all of our simulations reported below, we have therefore used fixed bandwidths according to Scott's (1992) normal reference rule, which suggests using bandwidths given by $H_y = H_x = 3n^{-1/6}$.

As with most other trimming schemes proposed in the literature (cf. Härdle and Stoker 1989, Ichimura 1993, Delecroix, Hristache and Patilea 2006, Ichimura and Todd 2006 and

Xia, Härdle and Linton 2012), our trimming method requires setting values for the trimming parameters; these are a_0 and c. Since the trimming factor (2.5) serves also as a normalising factor, our method is expected to be relatively robust to the trimming parameters as long as $a_0 n^{-c}$ is not too small such that any single observation may have a strong effect on the likelihood function value. At the same time, a good choice of values for a_0 and c should aim to trim only as few observations as possible, in the sense that the probability that $\min\left\{f_{Y,\theta^T X}\left(y_i, \theta^T x_i\right), f_{\theta^T X}\left(\theta^T x_i\right)\right\}$ is smaller than $a_0 n^{-c}$ is very low. In practice, in all of our simulations, we used all observations in both stages of the estimation, and we set $\widehat{\rho}_i^{\theta}$ to trim down only observations whose density estimates were lower than 0.001.

The complexity of the proposed algorithm is calculated as follows. Computing the likelihood has a computational complexity of $O\left(n^2 d\right)$, since it is calculated as a sum of $n\left(n-1\right)$ terms, $j \neq i$, where each term requires calculating the inner product of θ with $\left(x_j - x_i\right)$. Note that each term also requires applying the kernel function which may take some expensive computational time, and therefore it is generally recommended to choose kernels of simple form (i.e., polynomials). The BFGS optimisation method is a hill-climbing optimisation technique, and it requires $O\left(d^2 F\left(n,d\right)\right)$ time per step of the optimisation algorithm to optimise a system with d parameters, where $F\left(n,d\right)$ is the cost of calculating the objective function (Nocedal and Wright 2006, Chapter 6). Thus, in our case, assuming the number of steps of the BFGS optimisation algorithm is limited to a finite number, the computational complexity of finding the MLE, $\widehat{\theta}$, using the the BFGS method is of order $O\left(n^2 d^3\right)$. Similar considerations show that the computational complexity of the second step of the estimation, i.e., calculating $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)$, is of order $O\left(nd\right)$, which is relatively insignificant.

In the simulations we used R 2.14.1 programme on a computer with 3.4ghz intel core i7-2600 processor. In all of the examples considered below, the dimensionality of the problem was $d = 4$. The average computational times of the method (for a single estimation based on Example 1 below) with dimension $d = 4$ and sample sizes $n = 100, 200, 400$ and 800 were $1.5\,\text{sec}, 3.5\,\text{sec}, 11\,\text{sec}$ and $31\,\text{sec}$, respectively.

An R code PPCDE.txt for the calculations below is available at

http://personal.lse.ac.uk/rosemari/

For comparison, we also tested the *dOPG* method of Xia (2007)[†]. The *dOPG* method performs estimation the central dimension reduction subspace, and when it is exercised with a one-dimensional central dimension reduction subspace, it can be applied for the the first-stage estimation of the orientation vector. For the *dOPG* method the average computational times of the method (for a single estimation based on Example 1 below) with dimension $d = 4$ and sample sizes $n = 100, 200, 400$ and 800 were 1 sec, 1 sec, 2 sec and 5 sec, respectively.

The performances of the proposed methods are demonstrated in the following three examples of simulated time-series models.

Example 1. As a first example, we consider the linear AR(4) model

$$y_t = 0.5 \cdot \sum\nolimits_{j=1}^{4} \theta_{0,j} y_{t-j} + 0.5 \cdot \varepsilon_t,$$

where $\theta_0^T \equiv (\theta_{0,1}, .., \theta_{0,4}) = (3, 2, 0, -1) / \sqrt{14}$ and ε_t are i.i.d. $N(0, 1)$.

Example 2. In the next example we consider the nonlinear AR(4) model

$$y_t = g\left(\sum\nolimits_{j=1}^{4} \theta_{0,j} y_{t-j}\right) + 0.5 \cdot \varepsilon_t,$$

where $g(u) = \exp\left(\left(0.4 - 2u^2\right) u\right)$, $\theta_0^T \equiv (\theta_{0,1}, .., \theta_{0,4}) = (1, 2, -1, 0) / \sqrt{6}$, and the ε_t are as in Example 1.

Example 3. Finally we would like to examine how the method works where the optimal projection $\theta_0^T X$ is related to higher moments of X. For the third example, we consider the nonlinear ARCH(4) model

$$y_t = g\left(\sum\nolimits_{j=1}^{4} \theta_{0,j} y_{t-j}\right) \cdot \varepsilon_t,$$

[†] I thank Professor Yingcun Xia for providing a code for *dOPG* at http://www.stat.nus.edu.sg/~staxyc/

TABLE 2.1: Mean and Standard error (in brackets) of the inner product $\widehat{\theta}^T \theta_0$.

	$n = 100$	$n = 200$	$n = 400$	$n = 800$
Example 1				
$p = 2$	0.9241 (0.0776)	0.9630 (0.0312)	0.9758 (0.0258)	0.9864 (0.0182)
$p = 4$	0.8670 (0.1344)	0.8828 (0.1277)	0.9034 (0.0615)	0.9113 (0.0529)
$dOPG$	0.8943 (0.0793)	0.9458 (0.0309)	0.9617 (0.0267)	0.9749 (0.0155)
Example 2				
$p = 2$	0.8867 (0.2193)	0.9632 (0.1325)	0.9809 (0.1043)	0.9936 (0.0654)
$p = 4$	0.6865 (0.3063)	0.6903 (0.2917)	0.7074 (0.3164)	0.7439 (0.3154)
$dOPG$	0.8900 (0.0770)	0.9336 (0.0575)	0.9612 (0.0292)	0.9757 (0.0177)
Example 3				
$p = 2$	0.6412 (0.2864)	0.7374 (0.2636)	0.8703 (0.1689)	0.9301 (0.0961)
$p = 4$	0.6858 (0.2757)	0.8131 (0.2201)	0.8914 (0.1534)	0.9195 (0.0975)
$dOPG$	0.4881 (0.2613)	0.4792 (0.2892)	0.5085 (0.2755)	0.4734 (0.2770)

where $g(u) = \frac{1}{2}\sqrt{1+u^2}$. Here, $\theta_{0,j} = \exp(-j)/\sqrt{\sum_{k=1}^{4} \exp(-2k)}$, $j = 1, ..., 4$, and the ε_t are as in the previous examples.All the three models can easily be verified to be geometrically ergodic by either Theorem 3.1 or Theorem 3.2 of An and Huang (1996), and hence they are strictly stationary and strong-mixing with exponential decaying rates (see Fan and Yao 2003, p. 70). In all examples, our goal was to estimate the optimal orientation θ_0 and the single-index predictive density $f_{Y|\theta^T x}(y_t|\theta^T x_t)$ of y_t given the lagged observations $x_t = (y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4})$. For each model 200 replications were generated with sample sizes $n = 100, 200, 400$ and 800, and we implemented the method to produce the corresponding estimates $\widehat{\theta}$ and $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)$.

In practice, of-course, one does not know a priorily the optimal number of lagged observations to be considered in the model, and the lag should be chosen according to some preliminary analysis or model selection criterion. Cross-validatory techniques were shown to have successful applications to model selection in semiparametric settings (Gao and Tong 2004, Kong and Xia 2007), and they can be used to produce a stopping rule to the single-index c.p.d.f. model. However, these computationally intensive techniques are less desirable as $\widehat{\theta}$ has to found by numerical optimisation. The topic of model selection is thus left open for some further research, and a relevant discussion is given in the concluding Chapter 4 of the thesis.

Table 2.1 presents the average and standard error (over 200 replications) of the inner products $\left|\widehat{\theta}^T\theta_0\right|$ obtained for the three models with different sample sizes, and where the estimation was performed using a kernels of order $p = 2$, $p = 4$ or by the *dOPG* method of Xia (2007). Since $\widehat{\theta}$ and θ_0 are both unit vectors, $\left|\widehat{\theta}^T\theta_0\right|$ is simply $|\cos\alpha|$, where α is the angle between $\widehat{\theta}$ and θ_0, and it is 1 if and only if $\widehat{\theta} = \theta_0$. Note also that this inner product is directly related to the sum of square error measure

$$\left\|\widehat{\theta} - \theta_0\right\|^2 = \left\|\widehat{\theta}\right\|^2 + \|\theta_0\|^2 - 2\left|\widehat{\theta}^T\theta_0\right| = 2\left(1 - \left|\widehat{\theta}^T\theta_0\right|\right).$$

As a general conclusion from Table 2.1, we can see that the orientation estimates become more accurate as the sample size increases, although the rate of improvement is not as fast as suggested by the theoretical asymptotic results. Comparing between the accuracy of the orientation estimates across the three different models, one can see that the method seems to be less accurate for the nonlinear models, and in particular for the nonlinear ARCH model with relatively small sample sizes ($n = 100$ or 200). However, when the number of observations is increased to 800, the average of the inner product $\left|\widehat{\theta}^T\theta_0\right|$ is consistently higher than 0.9 for all of the models with second-order kernels, and two out of the three models with fourth-order kernels.

The generally better performances of the second-order kernels compared with the fourth-order kernels in terms of the accuracy of the corresponding orientation estimates are particularly striking in Examples 1 and 2. In Example 3, on the other hand, the fourth-order kernel yields some more accurate estimates for θ_0 with sample sizes $n = 100,\ 200$ or 400. However, when the sample size is increased to $n = 800$ the accuracy of the second-order kernels 'catches up' with that of the fourth-order kernels. An extensive investigation performed by Marron and Wand (1992) of the effectiveness of high-order kernels in nonparametric density estimation provides an explanation for this discrepancy between theory and practice as it shows that it may take extremely large sample sizes (with a typical order of magnitude of few thousands and up to hundreds of thousands) for the asymptotic domi-

TABLE 2.2: Mean and standard error (in brackets) of the sample RMSPE.

	$n = 100$	$n = 200$	$n = 400$	$n = 800$
Example 1				
$p = 2$	0.0460 (0.0201)	0.0327 (0.0117)	0.0245 (0.0097)	0.0167 (0.0055)
$p = 4$	0.0524 (0.0218)	0.0420 (0.0156)	0.0343 (0.0130)	0.0289 (0.0102)
$dOPG$	0.0491 (0.0280)	0.0279 (0.0131)	0.0115 (0.0070)	0.0154 (0.0050)
Example 2				
$p = 2$	0.0756 (0.0333)	0.0511 (0.0205)	0.0370 (0.0167)	0.0272 (0.0086)
$p = 4$	0.1080 (0.0348)	0.0952 (0.0381)	0.0845 (0.0413)	0.0722 (0.0453)
$dOPG$	0.0963 (0.0275)	0.0706 (0.0230)	0.0499 (0.0154)	0.0382 (0.0108)
Example 3				
$p = 2$	0.0712 (0.0271)	0.0500 (0.0172)	0.0374 (0.0148)	0.0276 (0.0100)
$p = 4$	0.0626 (0.0241)	0.0455 (0.0165)	0.0347 (0.0135)	0.0256 (0.0093)
$dOPG$	0.0483 (0.0284)	0.0387 (0.0155)	0.0371 (0.0109)	0.0302 (0.0083)

nant effect to begin to be realised, and for the high-order kernels to produce more accurate estimates. In particular, Marron and Wand (1992) conclude that high-order kernels are not recommended in practice for kernels density estimation with realistic sample sizes.

The $dOPG$ method seems to perform very well in Examples 1 and 2, although its performance is inferior to that achieved with the second-order kernels. However, the $dOPG$ performs very poorly in Example 3, even for sample size $n = 800$, which suggests that $dOPG$ have difficulties in estimation of the optimal orientation where it is related to higher moments of X.

Notice that at the second-stage of estimation, $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)$ is estimated using the same second-order kernel. Thus, for the purpose of comparison between the approach obtained with different implementation of the first-stage of estimation of the orientation estimator, $\widehat{\theta}$, it is sufficient to examine directly the performance of $\widehat{\theta}$. Nevertheless, for the sake of completeness and to illustrate the finite-sample properties of the procedure, we now continue and assess the accuracy of the conditional density estimator, $\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right)$. To this end, we used the sample Root Mean Square Percentage Error (RMSPE),

$$RMSPE = \sum_{i=1}^{n} \left[\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y_i|\widehat{\theta}^T x_i\right) - f_{Y|X}\left(y_i|x_i\right)\right]^2 \Bigg/ \sum_{i=1}^{n} f_{Y|X}\left(y_i|x_i\right)^2,$$

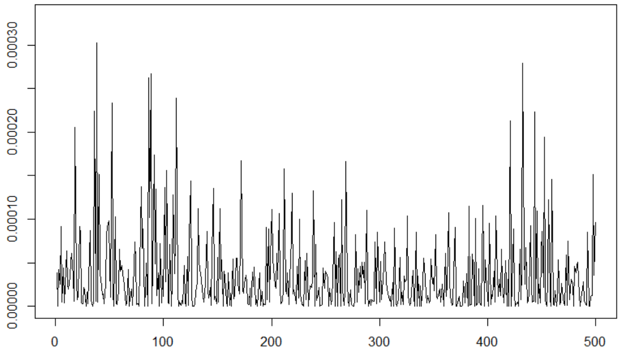


FIGURE 2.1. *Example 4:* Daily exchange rate squared returns of the USD-GDP between 04/01/2010 and 30/12/2011.

where $f_{Y|X}\left(y_i|x_i\right)$ is the real conditional density. The average and standard error (over 200 replications) of the sample RMSPE are given in Table 2.2 for orientation estimates that were obtained at the first stage of estimation using a kernels of order $p = 2$, $p = 4$ or by the *dOPG* method of Xia (2007).

Here, we see that the estimation error given by the sample RMSPE consistently decreases as the sample size increases for all the simulation settings. Observe that although the average accuracy of the orientation estimates did not improve in Examples 1 and 2 between $n = 200$ and $n = 400$, the approximated conditional density obtained at the second stage is more accurate on average for the larger sample size $n = 400$. Finally, as a consequence of the orientation estimation performances, we see that in Examples 1 and 2 the conditional density estimates obtained by using second-order kernels (at the first-stage of the estimation) outperforms the ones obtained with fourth-order kernels or with *dOPG*. In Example 3, however, the estimates corresponding to fourth-order kernels are slightly more accurate on average.

Example 4: Finally, we demonstrate a real-data application of the proposed method. In the standard ARCH(p) model, it is assumed that

$$y_t = \sigma_t \cdot \varepsilon_t,$$

where

$$\sigma_t = \alpha_0 + \sum\nolimits_{j=1}^{p} \alpha_j y_{t-j}^2.$$

Here, ε_t is a white noise process, $\alpha_0 > 0$ and $\alpha_j \geq 0,\ j = 1, ..., p$. The ARCH(p) model can be written as an AR(p) model in y_t^2, by the relation

$$y_t^2 = \sigma_t^2 + \sigma_t^2 \left(\varepsilon_t^2 - 1\right) = \alpha_0 + \sum\nolimits_{j=1}^{p} \alpha_j y_{t-j}^2 + \varepsilon_t^*,$$

where $\varepsilon_t^* = \sigma_t^2 \left(\varepsilon_t^2 - 1\right)$ is an heteroscedastic white noise process. This last formulation motivates us to consider an application of the proposed method to prediction of the volatility process in terms of the squared returns (cf. Andersen and Bollerslev 1988). We use a time-series of the daily exchange-rates' squared returns between the US Dollar (USD) and the British Pound (GBP) between 4 January 2010 and 30 December 2011. The data consists of 501 data points, out of which we allocate the last 100 points for prediction. Figure 2.1 presents the time-series data over the full period. We implement the approximation to estimate the predictive density $f_{Y^2|\theta^T x}\left(y_t^2|x_t\right)$ of y_t given the 4-lagged data $x_t = \left(y_{t-1}^2, y_{t-2}^2, y_{t-3}^2, y_{t-4}^2\right)$. Using only the first 401 data points, we estimate first the orientation vector, and the obtained estimate is $\widehat{\theta} = (0.921, 0.082, 0.250, -0.288)$. This estimate suggests that the most recent lag has the strongest effect on the predictive density of y_t^2, although the third and fourth lag also have some significant effect. Next, for any observation y_t that belongs to the last 100 observations, we iteratively construct a predictive density model using the estimated orientation, $\widehat{\theta}$, where all nonparametric functional estimators rely on past information $y_1^2, ..., y_{t-1}^2$ (that may include some past observations from the last 100 data points). In order to examine the predictive capability of the models, we construct the corresponding one-sided $(1-\alpha)$ −prediction confidence intervals, based on the right tail of the density function, for the squared returns in the last 100 observations. The reason we considered one-sided prediction intervals, rather than standard two-sided ones, is that the density function of the squared returns is supported on $[0, \infty)$, while $f_{Y^2}\left(y_t^2 = 0\right)$ seems to be non-zero and perhaps infinite, while the kernel estimates cannot

Model	$\alpha = 1\%$		$\alpha = 5\%$		$\alpha = 10\%$		$\alpha = 25\%$	
	Cover.	Length	Cover.	Length	Cover.	Length	Cover.	Length
Uncond.	0.99	24.54	0.94	13.50	0.91	9.98	0.78	5.64
One lag	0.99	22.99	0.94	12.96	0.90	9.50	0.79	5.23
Sing. Ind.	0.99	22.66	0.94	12.85	0.89	9.46	0.77	5.24
Kernel	0.96	18.23	0.93	12.33	0.90	9.67	0.80	5.95

TABLE 2.3: *Results for Example 4*: Prediction coverage (%) and avg. length ($\cdot 10^5$) of $(1-\alpha)$ −prediction intervals.

completely capture the probability mass in the left tail of the volatility density. On the other hand, the right tail of the volatility distribution is very heavy, and for practical purposes, such as for risk management, the right tail of the volatility distribution seems to be of much more interest than the left tail (see, for example, Windsor and Thyagaraja 2001). Table 2.3 gives the prediction coverage (% of observations y_t^2 that fall inside the prediction interval) and the average length of the prediction interval over the last 100 observations for all obtained models with $\alpha = 1\%,\ 5\%,\ 10\%$ and 25%. For comparison, this table presents the result obtained with the unconditional kernel density of y_t^2 (Uncond.), the conditional density based only on the most recent lag (One lag), the single-index approximation (Sing. Ind.) and the standard multivariate kernel estimator (Kernel). Also, for visual illustration, Figure 2.1 shows plots of the last 100 observations and the corresponding right tail 90%−prediction interval obtained by each model.

For all of the confidence level values examined, the unconditional density estimator produced the widest prediction-intervals on average, while the standard unconditional density estimator produced relatively narrow confidence intervals. In terms of prediction coverage, both the unconditional density estimator and the PPCDE provide relatively accurate estimates, while the standard conditional density kernel estimator has much less similar to reality. We thus see that the PPCDE manages to provide increased accuracy and predictive power relative to standard kernel methods.

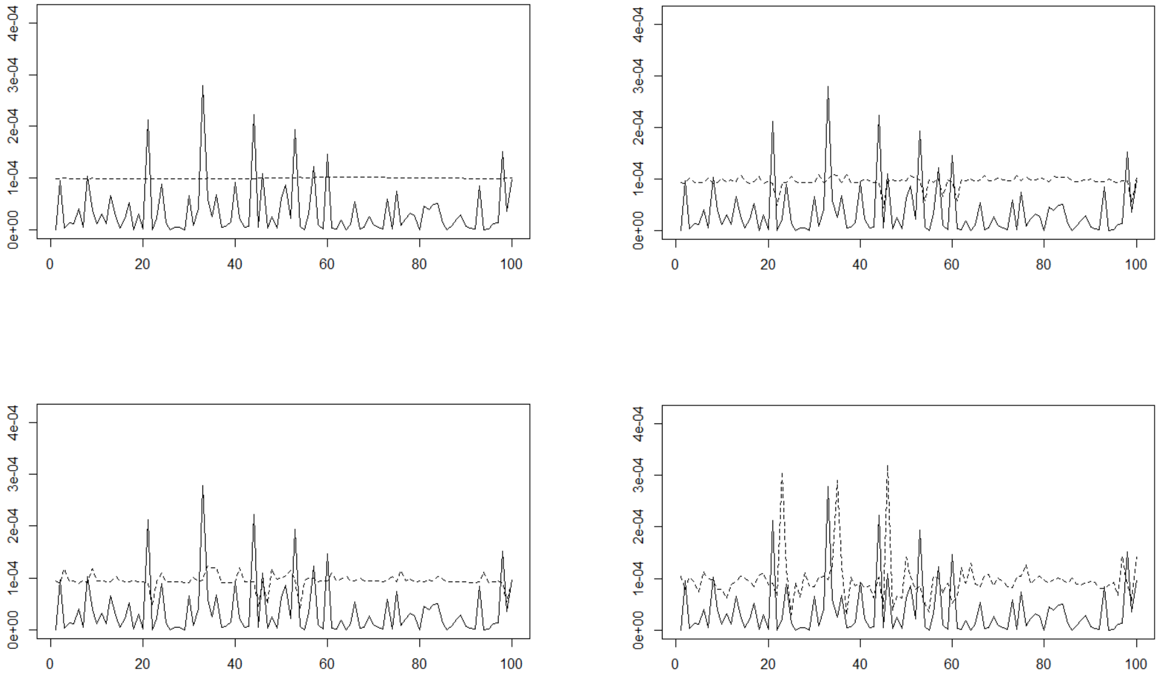


Figure 2.1. *Example 4*: 90%−prediction intervals for the daily USD-GDP exchange-rate squared returns between 19/10/2010 and 30/12/2011 based on (a) Unconditional kernel density estimator; (b) Conditional kernel density estimator based on the most recent lag; (c) The single-index approximation; (d) Multivariate conditional density kernel estimator.

2.5 Appendix A - Proofs of the Theorems

Proof of Theorem 2.3.1. By assumptions (A4), (A6) it is sufficient (Amemiya 1985, Theorem 4.1.1) to prove that

$$\sup_{\theta\in\Theta}\left|\left(\mathcal{L}\left(\theta\right)-E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\left(y|\theta^T x\right)\right)\right)\right|=o_p\left(1\right). \tag{2.11}$$

Denote $L\left(\theta\right)$ as the version of $\mathcal{L}\left(\theta\right)$ when conditional density estimates are replaced by the true conditional densities, that is,

$$L\left(\theta\right)=n^{-1}\sum_{i=1}^{n}\log f_{Y|\theta^T X}\left(y_i|\theta^T x_i\right).$$

By smoothness condition (A4) we have that for any $\varepsilon>0$ there exists a positive constant $\delta>0$ such that for any $(\theta_1,y,x)\in\Theta\times\,\mathbb{S}$ and $\theta\in U_\delta\left(\theta_1\right)$, a δ-ball with centre at θ_1,

$$\left|\log f_{Y|\theta^T X}\left(y|\theta^T x\right)-\log f_{Y|\theta^T X}\left(y|\theta_1^T x\right)\right|<\varepsilon.$$

As a result, we have

$$\begin{aligned}&\sup_{\theta\in U_\delta(\theta_1)}\left|\left(L\left(\theta\right)-E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\left(y|\theta^T x\right)\right)\right)\right|\\ =\;& 2\varepsilon+\left|\left(L\left(\theta_1\right)-E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\left(y|\theta_1^T x\right)\right)\right)\right|.\end{aligned} \tag{2.12}$$

Note also that since is Θ compact, it is possible to construct a finite open covering of Θ by δ-balls, $U_\delta\left(\theta_k\right)$, $k=1,...,K$. Thus, using (2.12) we have that for any $\varepsilon>0$

$$\begin{aligned}&P\left(\sup_{\theta\in\Theta}\left|L\left(\theta\right)-E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\left(y|\theta^T x\right)\right)\right|>3\varepsilon\right)\\ \le\;& K\max_{k=1,\ldots,K}P\left(\sup_{\theta\in U_\delta(\theta_k)}\left|L\left(\theta\right)-E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\left(y|\theta^T x\right)\right)\right|>3\varepsilon\right)\\ \le\;& K\max_{k=1,\ldots,K}P\left(\left|L\left(\theta_k\right)-E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\left(y|\theta_k^T x\right)\right)\right|>\varepsilon\right)\end{aligned}$$

The series $\log f_{Y|\theta^T X}\left(y_i|\theta^T x_i\right)$ is itself strong-mixing (see, for instance, White 1984), and by the ergodic theorem for strong-mixing processes (see Fan and Yao 2003, Proposition 2.8) we get for any $\theta \in \Theta$

$$\left|\left(L\left(\theta\right) - E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\left(y|\theta^T x\right)\right)\right)\right| \to 0 \ \ a.s.$$

We thus established that

$$\sup_{\theta\in\Theta}\left|L\left(\theta\right) - E_{\mathbb{S}}\left(\log f_{Y|\theta^T X}\left(y|\theta^T x\right)\right)\right| = o_p\left(1\right). \tag{2.13}$$

Next, by Lemma 2.6.5, all trimming-terms, $\widehat{\rho}_i^{\theta}$, $i = 1, ..., n$, defined in (2.5), are eventually equal to 1 for any large enough n with probability 1. Therefore, we can consider n to be large enough so we can ignore the trimming-terms, i.e. set $\widehat{\rho}_i^{\theta} \equiv 1$. Since $f_{Y,\theta^T X}\left(y, \theta^T x\right)$, $f_{\theta^T X}\left(\theta^T x\right)$ are bounded from below by $\varepsilon > 0$ on $\Theta \times \mathbb{S}$ and $\Theta \times \mathbb{S}_X$, by the uniform consistency result of Lemma 2.6.2 and the continuous mapping theorem (Amemiya 1985, Theorem 3.2.6) applied to the log-function we get with $z_j = (y_j, x_j)$,

$$\begin{aligned}
\sup_{\theta\in\Theta, z\in\mathbb{S}}\left|\log \widehat{f}_{Y,\theta^T X}\left(y, \theta^T x\right) - \log f_{Y,\theta^T X}\left(y, \theta^T x\right)\right| &= o_p\left(1\right), \\
\sup_{\theta\in\Theta, x\in\mathbb{S}_X}\left|\log \widehat{f}_{\theta^T X}\left(\theta^T x\right) - \log f_{\theta^T X}\left(\theta^T x\right)\right| &= o_p\left(1\right).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
& \sup_{\theta\in\Theta}\left|\mathcal{L}\left(\theta\right) - L\left(\theta\right)\right| \\
\leq\ & \max_{1\leq i\leq n}\sup_{\theta\in\Theta}\left|\log \widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right) - \log f_{Y,\theta^T X}\left(y_i, \theta^T x_i\right)\right| \\
& + \max_{1\leq i\leq n}\sup_{\theta\in\Theta}\left|\log \widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right) - \log f_{\theta^T X}\left(\theta^T x_i\right)\right| \\
\leq\ & \sup_{\theta\in\Theta, z\in\mathbb{S}}\left|\log \widehat{f}_{Y,\theta^T X}\left(y, \theta^T x\right) - \log f_{Y,\theta^T X}\left(y, \theta^T x\right)\right| \\
& + \sup_{\theta\in\Theta, x\in\mathbb{S}_X}\left|\log \widehat{f}_{\theta^T X}\left(\theta^T x\right) - \log f_{\theta^T X}\left(\theta^T x\right)\right| + o\left(1\right) \\
=\ & o_p\left(1\right).
\end{aligned} \tag{2.14}$$

Results (2.13) and (2.14) imply (2.11), and therefore the Theorem is proved. ■

Proof of Theorem 2.3.2. As in the proof of Theorem 2.3.1, let $L(\theta)$ be a version of the likelihood $\mathcal{L}(\theta)$ when conditional density estimates are replaced by the true conditional densities,

$$L(\theta) = n^{-1} \sum_{i=1}^{n} \log f_{Y|\theta^T X}\left(y_i|\theta^T x_i\right).$$

Furthermore, let $\widetilde{\theta} = \arg\max_{\theta\in\Theta} L(\theta)$. Note that by (2.13) in the proof of Theorem 2.3.1 and Theorem 4.1.1 of Amemiya (1985) $\widetilde{\theta}$ is a consistent estimator to θ_0.

Due to smoothness condition (A9), the mean value theorem, applied to the function $\nabla L(\theta)$ with mean-value $\overline{\theta}$ such that $\left|\overline{\theta} - \theta_0\right| \leq \left|\widetilde{\theta} - \theta_0\right|$, yields

$$\nabla L\left(\widetilde{\theta}\right) - \nabla L(\theta_0) = \nabla^2 L\left(\overline{\theta}\right)\left(\widetilde{\theta} - \theta_0\right). \tag{2.15}$$

Since θ_0 lies in the interior of Θ, the consistency of $\widetilde{\theta}$ implies that for all $\varepsilon > 0$,

$$\left(\sqrt{n}\nabla L\left(\widetilde{\theta}\right) > \varepsilon\right) \to_p 0. \tag{2.16}$$

Moreover, By the central limit theorem (CLT) for α-mixing processes (cf. Fan and Yao 2003, Theorem 2.21),

$$n^{1/2}\nabla L(\theta_0) \to_d N(0, \Psi(\theta_0)). \tag{2.17}$$

By smoothness condition (A9) and the ergodic theorem for strong-mixing processes (see Fan and Yao 2003, Proposition 2.8), one can show by a similar way to (2.13), that

$$\sup_{\theta\in\Theta}\left|\nabla^2 L(\theta) + \Omega(\theta)\right| = o_p(1). \tag{2.18}$$

Note that by conditions (A6) and (A9) there is a generalised inverse of $\Omega(\theta_0)$, denoted $\Omega(\theta_0)^-$, that is well defined in the perpendicular space to θ_0 (cf. Theorem 3.1 of White

1982). Results (2.15)-(2.18) imply

$$\widetilde{\theta} - \theta_0 = n^{-1/2} V\left(\theta_0\right)^{1/2} Z, \tag{2.19}$$

where Z is asymptotically normal $N\left(0, I\right)$ random d-vector.

Next, by the mean value theorem, applied to the function $\nabla L\left(\theta\right)$ again, with mean-value $\overline{\theta}$ such that $\left|\overline{\theta} - \widehat{\theta}\right| \le \left|\widetilde{\theta} - \widehat{\theta}\right|$,

$$\nabla L\left(\widetilde{\theta}\right) - \nabla L\left(\widehat{\theta}\right) = \nabla^2 L\left(\overline{\theta}\right)\left(\widetilde{\theta} - \widehat{\theta}\right). \tag{2.20}$$

By using results (2.16), (2.18) again, the consistency of $\widetilde{\theta}$ and $\widehat{\theta}$, and

$$\left(\sqrt{n}\nabla \mathcal{L}\left(\widehat{\theta}\right) > \varepsilon\right) \to_p 0,$$

we can write (2.20) as

$$-\Omega\left(\theta_0\right)\left(\widetilde{\theta} - \widehat{\theta}\right) = \nabla \mathcal{L}\left(\widehat{\theta}\right) - \nabla L\left(\widehat{\theta}\right) + o_p\left(n^{-1/2}\right). \tag{2.21}$$

Thus, by (2.19) and (2.21) and the triangle inequality, Theorem 2.3.2 will be established if we show that for some $\delta > 0$,

$$\nabla \mathcal{L}\left(\widehat{\theta}\right) - \nabla L\left(\widehat{\theta}\right) = O_p\left(n^{2-\delta} h_y h_x^3\right)^{-1/2} + O(h_y^p + h_x^p) + o_p\left(\widehat{\theta} - \theta_0\right). \tag{2.22}$$

Since by Lemma 2.6.5, all trimming-terms, $\widehat{\rho}_i^{\theta}$ are eventually equal to 1 for any large enough n with probability 1, we may set $\widehat{\rho}_i^{\theta} \equiv 1$ for large enough n. We then have

$$\nabla \mathcal{L}\left(\theta\right) = n^{-1} \sum_{i=1}^{n} \left(\frac{\nabla \widehat{f}_{Y,\theta^T X}^{-i}\left(y_i, \theta^T x_i\right)}{\widehat{f}_{Y,\theta^T X}^{-i}\left(y_i, \theta^T x_i\right)} - \frac{\nabla \widehat{f}_{\theta^T X}^{-i}\left(\theta^T x_i\right)}{\widehat{f}_{\theta^T X}^{-i}\left(\theta^T x_i\right)} \right).$$

Assertion (2.22) will follow if we prove that the following two assertions hold. For some arbitrarily small $\delta > 0$, $\theta \rightarrow_p \theta_0$ implies that

$$n^{-1} \sum_{i=1}^{n} \left(\frac{\nabla \widehat{f}^{-i}_{\theta^T X} \left(\theta^T x_i\right)}{\widehat{f}^{-i}_{\theta^T X} \left(\theta^T x_i\right)} - \frac{\nabla f_{\theta^T X} \left(\theta^T x_i\right)}{f_{\theta^T X} \left(\theta^T x_i\right)} \right) = O_p \left(n^{2-\delta} h_x^3\right)^{-1/2} + O(h_x^p) + o_p \left(\theta - \theta_0\right), \tag{2.23}$$

and

$$n^{-1} \sum_{i=1}^{n} \left(\frac{\nabla \widehat{f}^{-i}_{Y,\theta^T X} \left(y_i, \theta^T x_i\right)}{\widehat{f}^{-i}_{Y,\theta^T X} \left(y_i, \theta^T x_i\right)} - \frac{\nabla f_{Y,\theta^T X} \left(y_i, \theta^T x_i\right)}{f_{Y,\theta^T X} \left(y_i, \theta^T x_i\right)} \right) = O_p \left(n^{2-\delta} h_y h_x^3\right)^{-1/2} + O(h_y^p + h_x^p) + o_p \left(\theta - \theta_0\right). \tag{2.24}$$

The proof of (2.23)-(2.24) is long and tedious. However, both assertions are established similarly with Gao and King's (2004) result for degenerate U-statistics of strongly dependent processes (see Definition 2.1.1) given in Lemma 2.6.4. For the sake of brevity, we shall focus here on proving (2.23), while (2.24) follows similarly.

In the following, whenever confusion does not occur we denote $f_{\theta^T X} \equiv f_{\theta^T X} \left(\theta^T x_i\right)$ and $\widehat{f}^{-i}_{\theta^T X} \equiv \widehat{f}^{-i}_{Y,\theta^T X} \left(\theta^T x_i\right)$ for some $\theta \in \Theta$ and $x_i \in \mathbb{S}_X$. We now have by the Taylor's theorem applied to the function $\xi\left(x\right) = \frac{1}{x}$ for $x \geq \varepsilon > 0$, with mean value $\overline{f}_{\theta^T x}$ such that $\left|\overline{f}_{\theta^T x} - f_{\theta^T x}\right| < \left|\widehat{f}^{-i}_{\theta^T X} - f_{\theta^T x}\right|$,

$$\frac{1}{\widehat{f}^{-i}_{\theta^T X}} - \frac{1}{f_{\theta^T X}} = -\frac{1}{\left(f_{\theta^T X}\right)^2} \left(\widehat{f}^{-i}_{\theta^T X} - f_{\theta^T X}\right) + \frac{2}{\left(\overline{f}_{\theta^T X}\right)^3} \left(\widehat{f}^{-i}_{\theta^T X} - f_{\theta^T X}\right)^2. \tag{2.25}$$

We then obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\frac{\nabla\widehat{f}^{-i}_{\theta^T X}}{\widehat{f}^{-i}_{\theta^T X}} \\
= \quad & \frac{1}{n}\sum_{i=1}^{n}\frac{\nabla\widehat{f}^{-i}_{\theta^T X}}{f_{\theta^T X}} + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{\widehat{f}^{-i}_{\theta^T X}} - \frac{1}{f_{\theta^T X}}\right]\left[\nabla f_{Y,\theta^T X} + \left(\nabla\widehat{f}^{-i}_{\theta^T X} - \nabla f_{Y,\theta^T X}\right)\right] \\
= \quad & \frac{1}{n}\sum_{i=1}^{n}\frac{\nabla f_{\theta^T X}}{f_{\theta^T X}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\nabla\widehat{f}^{-i}_{\theta^T X}}{f_{\theta^T X}} - \frac{\widehat{f}^{-i}_{\theta^T X}\nabla f_{Y,\theta^T X}}{\left(f_{\theta^T X}\right)^2}\right)}_{U_\theta^{(A)}} \\
& \underbrace{-\frac{1}{n}\sum_{i=1}^{n}\frac{\left(\widehat{f}^{-i}_{\theta^T X} - f_{\theta^T X}\right)\left(\nabla\widehat{f}^{-i}_{\theta^T X} - \nabla f_{\theta^T X}\right)}{\left(f_{Y,\theta^T X}\right)^2}}_{U_\theta^{(B)} + R_{\theta,1}} \\
& \underbrace{+\frac{1}{n}\sum_{i=1}^{n}\frac{2\nabla f_{\theta^T X}\left(\widehat{f}^{-i}_{\theta^T X} - f_{\theta^T X}\right)^2}{\left(\overline{f}_{\theta^T X}\right)^3}}_{R_{\theta,2}} \\
& \underbrace{+\frac{1}{n}\sum_{i=1}^{n}\frac{2\left(\nabla\widehat{f}^{-i}_{\theta^T X} - \nabla f_{\theta^T X}\right)\left(\widehat{f}^{-i}_{\theta^T X} - f_{\theta^T X}\right)^2}{\left(\overline{f}_{\theta^T X}\right)^3}.}_{R_{\theta,3}}
\end{aligned}$$

Thus,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\nabla\widehat{f}^{-i}_{\theta^T X}}{\widehat{f}^{-i}_{\theta^T X}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\nabla f_{\theta^T X}}{f_{\theta^T X}} + U_\theta^{(A)} - U_\theta^{(B)} - R_{\theta,1} + R_{\theta,2} + R_{\theta,3}, \tag{2.26}$$

where

$$U_\theta^{(A)} = \frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j\neq i}\zeta_\theta^{(A)}\left(x_i, x_j\right), \quad U_\theta^{(B)} = \frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j\neq i}\varsigma_\theta^{(B)}\left(x_i, x_j\right),$$

are second order $\mathbb{R}^d$-vector U-statistics with arguments

$$\begin{aligned}
& \varsigma_\theta^{(A)}\left(x_i, x_j\right) \qquad\qquad (2.27) \\
= \quad & \frac{1}{h_x^2}\frac{1}{f_{\theta^T X}\left(\theta^T x_i\right)}\left(x_j - x_i\right)K'\left(\frac{\theta^T\left(x_j - x_i\right)}{h_x}\right) - \frac{1}{h_x}\frac{\nabla f_{\theta^T X}\left(\theta^T x_i\right)}{f^2_{\theta^T X}\left(\theta^T x_i\right)}K\left(\frac{\theta^T\left(x_j - x_i\right)}{h_x}\right),
\end{aligned}$$

and

$$\begin{aligned}
& \varsigma_{\theta}^{(B)}\left(x_i, x_j\right) \qquad (2.28)\\
= & \frac{1}{h_x^3} \int \left(x_j - E\left(X|\theta^T X = t\right)\right) K\left(\frac{\theta^T x_i - t}{h_x}\right) K'\left(\frac{\theta^T x_j - t}{h_x}\right) f\left(t\right)^{-1} dt \\
& -E\left(\nabla \log f_{\theta^T X}\left(\theta^T X\right)|\theta^T X = \theta^T x_i\right) - E\left(\nabla \log f_{\theta^T X}\left(\theta^T X\right)|\theta^T X = \theta^T x_j\right) \\
& +E\left(\nabla \log f_{\theta^T X}\left(\theta^T X\right)\right).
\end{aligned}$$

Note that the term $U_{\theta}^{(B)}$ was added to (2.26) simply to make $R_{\theta,1}$ a degenerate U-statistic (see Definition 2.1.1), which will be proved later in the proof. Now, $R_{\theta,1}, R_{\theta,2}, R_{\theta,3}$ are the high-order remainder terms,

$$\begin{aligned}
R_{\theta,1} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\left(\widehat{f}_{\theta^T X}^{-i} - f_{\theta^T X}\right)\left(\nabla \widehat{f}_{\theta^T X}^{-i} - \nabla f_{\theta^T X}\right)}{\left(f_{\theta^T X}\right)^2} - U_{\theta}^{(B)}, \\
R_{\theta,2} &= \frac{1}{n}\sum_{i=1}^{n} \frac{2\nabla f_{\theta^T X}\left(\widehat{f}_{\theta^T X}^{-i} - f_{\theta^T X}\right)^2}{\left(\overline{f}_{\theta^T X}\right)^3}, \\
R_{\theta,3} &= \frac{1}{n}\sum_{i=1}^{n} \frac{2\left(\nabla \widehat{f}_{\theta^T X}^{-i} - \nabla f_{\theta^T X}\right)\left(\widehat{f}_{\theta^T X}^{-i} - f_{\theta^T X}\right)^2}{\left(\overline{f}_{\theta^T X}\right)^3}.
\end{aligned}$$

We proceed to handle the terms in the expansion (2.26) and we prove for an arbitrarily small $\delta > 0$ that $\theta \to_p \theta_0$ implies that

$$U_{\theta}^{(A)}, U_{\theta}^{(B)} = O_p\left(n^{2-\delta} h_x^3\right)^{-1/2} + O\left(h_x^p\right) + o_p\left(\theta - \theta_0\right), \qquad (2.29)$$

and

$$R_{\theta,1}, R_{\theta,2}, R_{\theta,3} = o_p\left(\left(n^{2-\delta} h_x^3\right)^{-1/2}\right) + O\left(h_x^p\right) + o_p\left(\theta - \theta_0\right). \qquad (2.30)$$

Write $U_\theta^{(A)}$ as a symmetrical function by

$$\begin{aligned} U_\theta^{(A)} &= \frac{2}{n\left(n-1\right)} \sum_{1\le i<j\le n} \frac{1}{2}\left(\varsigma_\theta^{(A)}\left(x_i,x_j\right)+\varsigma_\theta^{(A)}\left(x_j,x_i\right)\right) \\ &\equiv \frac{2}{n\left(n-1\right)} \sum_{1\le i<j\le n} \phi_\theta^{(A)}\left(x_i,x_j\right). \end{aligned} \tag{2.31}$$

We show now that $U_\theta^{(A)}$ is a degenerate U-statistic up to a $O(h_x^p)$ term that does not depend on $\theta\in\Theta$. Denote

$$\eta_\theta^{(A)}\left(\cdot\right)\equiv E\left(\phi_\theta^{(A)}\left(X,\cdot\right)\right) \text{ and } \mu_\theta^{(A)}=E\left(\eta_\theta^{(A)}\left(X\right)\right), \tag{2.32}$$

and by the Hoeffding's decomposition of U-processes (Lemma A, pp. 178 in Serfling 1980),

$$U_\theta^{(A)}=U_\theta^{*(A)}+\frac{1}{n}\sum_{1\le i\le n}\left(\eta_\theta^{(A)}\left(x_i\right)-\mu_\theta^{(A)}\right)+\mu_\theta^{(A)}, \tag{2.33}$$

where $U_\theta^{*(A)}$ is the degenerate U-statistic,

$$U_\theta^{*(A)}=\frac{2}{n\left(n-1\right)}\sum_{1\le i<j\le n}\phi_\theta^{*(A)}\left(x_i,x_j\right), \tag{2.34}$$

with elements

$$\phi_\theta^{*(A)}\left(x_i,x_j\right)=\phi_\theta^{(A)}\left(x_i,x_j\right)-\eta_\theta^{(A)}\left(x_i\right)-\eta_\theta^{(A)}\left(x_j\right)+\mu_\theta^{(A)}. \tag{2.35}$$

Let

$$\begin{aligned} Z_\theta^{(1)}\left(x_i,x_j\right) &= \frac{1}{h_x} f_{\theta^T X}\left(\theta^T x_i\right)^{-1} K\left(\frac{\theta^T\left(x_j-x_i\right)}{h_x}\right), \\ Z_\theta^{(2)}\left(x_i,x_j\right) &= \frac{1}{h_x^2} f_{\theta^T X}\left(\theta^T x_i\right)^{-1}\left(x_j-x_i\right) K'\left(\frac{\theta^T\left(x_j-x_i\right)}{h_x}\right), \end{aligned} \tag{2.36}$$

so that $\phi_\theta^{(A)}(x_i,x_j)=\frac{1}{2}\left(\varsigma_\theta^{(A)}(x_i,x_j)+\varsigma_\theta^{(A)}(x_j,x_i)\right)$ and

$$\varsigma_\theta^{(A)}(x_i,x_j)=Z_\theta^{(2)}(x_i,x_j)-\frac{\nabla f_{\theta^TX}\left(\theta^Tx_i\right)}{f_{\theta^TX}^2\left(\theta^Tx_i\right)}Z_\theta^{(1)}(x_i,x_j). \tag{2.37}$$

For a fixed $x_i\in\mathbb{S}_X$ we obtain with a change of variables that

$$\begin{aligned}&E\left(\frac{\nabla f_{\theta^TX}\left(\theta^Tx_i\right)}{f_{\theta^TX}^2\left(\theta^Tx_i\right)}Z_\theta^{(1)}(x_i,X)\right)\\ =\;&\frac{\nabla f_{\theta^TX}\left(\theta^Tx_i\right)}{f_{\theta^TX}^2\left(\theta^Tx_i\right)}\frac{1}{h_x}\int K\left(\frac{t-\theta^Tx_i}{h_x}\right)dt\\ =\;&\frac{\nabla f_{\theta^TX}\left(\theta^Tx_i\right)}{f_{\theta^TX}^2\left(\theta^Tx_i\right)}\int K(u)\,du\\ =\;&\frac{\nabla f_{\theta^TX}\left(\theta^Tx_i\right)}{f_{\theta^TX}\left(\theta^Tx_i\right)}.\end{aligned}$$

Since by Lemma 2.6.1

$$\nabla f_{\theta^TX}\left(\theta^Tx\right)=\left.\frac{d}{dt}\right|_{t=\theta^Tx}\left\{E\left(x-X|\theta^TX=t\right)f_{Y,\theta^TX}(t)\right\},$$

we also get with integration by parts, a change of variables, Taylor expansion around $u=0$, and utilising the order of the kernel,

$$\begin{aligned}&E\left(Z_\theta^{(2)}(x_i,X)\right)\\ =\;&\frac{1}{h_x^2}f_{\theta^TX}\left(\theta^Tx_i\right)^{-1}\int E\left[X-x_i|\theta^TX=t\right]K'\left(\frac{t-\theta^Tx_i}{h_x}\right)f(t)dt\\ =\;&f_{\theta^TX}\left(\theta^Tx_i\right)^{-1}\int\nabla f_{\theta^TX}\left(\theta^Tx-h_xu\right)K(u)\,dt\\ =\;&f_{\theta^TX}\left(\theta^Tx_i\right)^{-1}\int\left\{\sum_{j=1}^{p-1}\left[\left.\frac{d^j}{dt^j}\right|_{t=\theta^Tx}\nabla f_{\theta^TX}\left(\theta^Tx\right)(-h_xu)^j\right]+O\left(h_x^p\right)\right\}K(u)\,dt\\ =\;&\frac{\nabla f_{\theta^TX}\left(\theta^Tx_i\right)}{f_{\theta^TX}\left(\theta^Tx_i\right)}+O\left(h_x^p\right).\end{aligned}$$

Thus, we established that both $E\left(\frac{\nabla f_{\theta^TX}\left(\theta^TX\right)}{f_{\theta^TX}\left(\theta^TX\right)}Z^{(1)}(x_i,X)\right)$ and $E\left(Z^{(2)}(x_i,X)\right)$ are equal to

$$\frac{\nabla f_{\theta^T X}\left(\theta^T x_i\right)}{f_{\theta^T X}\left(\theta^T x_i\right)} + O(h_x^p), \tag{2.38}$$

where the constants in the $O\left(\cdot\right)$ terms are independent of $\theta \in \Theta$. Similarly, for a fixed $x_j \in \mathbb{S}_X$, using the same line of arguments yields that both $E\left(\frac{\nabla f_{\theta^T X}\left(\theta^T x_j\right)}{f_{\theta^T X}\left(\theta^T x_j\right)} Z_\theta^{(1)}\left(X, x_j\right)\right)$ and $E\left(Z_\theta^{(2)}\left(X, x_j\right)\right)$ are equal to

$$\begin{aligned} & -\left.\frac{d}{dt}\right|_{t=\theta^T x} E\left(X|\theta^T X = t\right) + O(h_x^p) \\ = \;& E\left(\nabla \log f_{\theta^T X}\left(\theta^T X\right)|\theta^T X = \theta^T x\right) + O(h_x^p). \end{aligned} \tag{2.39}$$

Thus, it follows from definitions (2.32), (2.37), and results(2.38), (2.39), that

$$\eta_\theta^{(A)}\left(\cdot\right) = O(h_x^p) \text{ and } \mu_\theta^{(A)} = O(h_x^p),$$

where the constants in the $O\left(\cdot\right)$ terms are independent of $\theta \in \Theta$. Hence, by (2.33)

$$U_\theta^{(A)} = U_\theta^{*(A)} + O(h_x^p), \tag{2.40}$$

uniformly on Θ.

Next, we proceed to handle the degenerate U-statistic $U_\theta^{*(A)}$. We first apply a uniformity argument based on a stochastic equicontinuity property (see definition in Andrews 1992). If $\theta \rightarrow_p \theta_0$, then assumption (A7) and compactness of Θ and $\mathbb{S}_X$ imply that

$$\left| K\left(\frac{\theta^T\left(x_j - x_i\right)}{h_x}\right) - K\left(\frac{\theta_0^T\left(x_j - x_i\right)}{h_x}\right)\right| \le \frac{\|\theta - \theta_0\|}{h_x}\ \widetilde{K}_1\left(\frac{\theta_0^T\left(x_j - x_i\right)}{h_x}\right),$$

and

$$\left| K'\left(\frac{\theta^T\left(x_j - x_i\right)}{h_x}\right) - K'\left(\frac{\theta_0^T\left(x_j - x_i\right)}{h_x}\right)\right| \le \frac{\|\theta - \theta_0\|}{h_x}\widetilde{K}_2\left(\frac{\theta_0^T\left(x_j - x_i\right)}{h_x}\right)$$

for some non-negative, compactly supported and bounded functions $\widetilde{K}_1$ and $\widetilde{K}_2$.

Correspondingly to $Z_{\theta}^{(1)}(x_i,x_j)$ and $Z_{\theta}^{(2)}(x_i,x_j)$, define

$$\begin{aligned}\widetilde{Z}_{\theta}^{(1)}(x_i,x_j) &= \frac{1}{h_x} f_{\theta^T X}\left(\theta^T x_i\right)^{-1} \widetilde{K}_1\left(\frac{\theta^T(x_j-x_i)}{h_x}\right),\\ \widetilde{Z}_{\theta}^{(2)}(x_i,x_j) &= \frac{1}{h_x^2} f_{\theta^T X}\left(\theta^T x_i\right)^{-1}(x_j-x_i)\,\widetilde{K}_2\left(\frac{\theta^T(x_j-x_i)}{h_x}\right).\end{aligned}$$

Hence, by smoothness condition (A9),

$$\begin{aligned}&\left|\frac{\nabla f_{\theta^T X}\left(\theta^T x_i\right)}{f_{\theta^T X}^2\left(\theta^T x_i\right)} Z_{\theta}^{(1)}(x_i,x_j) - \frac{\nabla f_{\theta_0^T X}\left(\theta_0^T x_i\right)}{f_{\theta_0^T X}^2\left(\theta_0^T x_i\right)} Z_{\theta_0}^{(1)}(x_i,x_j)\right|\\ \leq\ & \frac{\|\theta-\theta_0\|}{h_x}\left[\frac{\nabla f_{\theta_0^T X}\left(\theta_0^T x_i\right)}{f_{\theta_0^T X}^2\left(\theta_0^T x_i\right)} + o(1)\right]\widetilde{Z}_{\theta_0}^{(1)}(x_i,x_j),\end{aligned}$$

and

$$\left|Z_{\theta}^{(2)}(x_i,x_j) - Z_{\theta_0}^{(2)}(x_i,x_j)\right| \leq \frac{\|\theta-\theta_0\|}{h_x}\widetilde{Z}_{\theta}^{(2)}(x_i,x_j)$$

According to definitions (2.27), (2.31), (2.34), (2.35) and (2.37), the last inequalities yield

$$\begin{aligned}\left|U_{\theta}^{*(A)}\right| &\leq \left|U_{\theta_0}^{*(A)}\right| + \left|U_{\theta}^{*(A)} - U_{\theta_0}^{*(A)}\right|\\ &\leq \left|U_{\theta_0}^{*(A)}\right| + \frac{\|\theta-\theta_0\|}{h_x}\left|\widetilde{U}_{\theta_0}^{*(A)}\right|, \qquad (2.41)\end{aligned}$$

where $\widetilde{U}_{\theta_0}^{*(A)}$ is a version of $U_{\theta_0}^{*(A)}$ at $\theta=\theta_0$ and with $K(\cdot)$ and $K'(\cdot)$ replaced by $\widetilde{K}_1(\cdot)$ and $\widetilde{K}_2(\cdot)$ respectively. The right term in the RHS of (2.41) represents a stochastic equicontinuity term for $U_{\theta}^{*(A)}$, since showing that $\frac{1}{h_x}\widetilde{U}_{\theta_0}^{*(A)} = O_p(1)$ implies that $\widetilde{U}_{\theta}^{*(A)}$ is stochastic equicontinuous at θ_0 by a a Lipschitz condition (see Andrews 1992, Lemma 1(a)). We now bound in probability the terms in the RHS of (2.41) using the same argument. We therefore prove the bound only for the term $\left|U_{\theta_0}^{*(A)}\right|$.

Since by an applications of Chebyshev's inequality (Gut 2005, Chapter 3, Theorem

1.4), $X = O_p\left([E(X^2)]^{1/2}\right)$, for any random variable X, we then have by Lemma 2.6.4,

$$\begin{aligned} U_{\theta_0}^{*(A)} &= O_p\left(\frac{2}{n(n-1)}[E(\sum_{1\le i<j\le n}\phi_{\theta_0}^{*(A)}(x_i,x_j))^2]^{1/2}\right) \\ &= O_p\left(n^{-1}\left(M_{\theta_0}^{(A)}\right)^{1/(2+\delta)}\right), \end{aligned} \tag{2.42}$$

where

$$\begin{aligned} M_{\theta_0}^{(A)} &= \max_{1\le i<j\le T}\max\left\{E\left|\phi_{\theta_0}^{*(A)}(x_i,x_j)\right|^{2+\delta}, \int\left|\phi_{\theta_0}^{*(A)}(x_i,x_j)\right|^{2+\delta}dP(x_i)\,dP(x_j)\right\} \\ &\le C\max_{1\le i<j\le T}\max\left\{E\left|\phi_{\theta_0}^{(A)}(x_i,x_j)\right|^{2+\delta}, \int\left|\phi_{\theta_0}^{(A)}(x_i,x_j)\right|^{2+\delta}dP(x_i)\,dP(x_j)\right\} \end{aligned}$$

Here, $P(X)$ denotes the probability measure of r.v. X and $0<\delta<1$, and $C>0$ is a constant obtained by the C_r inequality (Gut 2005, Chapter 3, Theorem 2.2).

Using the bound of the kernels function and the probability density functions, and the compactness of $\mathbb{S}$, an integration with a change of variables leads to

$$M_{\theta_0}^{(A)} = O\left(h_x^{-2(2+\delta)+1}\right) = O\left(h_x^{-3-2\delta}\right). \tag{2.43}$$

Hence, we obtain with results (2.42) and (2.43) that $U_{\theta_0}^{(A)} = O_p\left(n^{2-\delta}h_x^3\right)^{-1/2} + O\left(h_x^p\right)$, and by (2.41) we have that $\theta \to_p \theta_0$ implies

$$\begin{aligned} U_{\theta}^{*(A)} &= O_p\left(n^{2-\delta}h_x^3\right)^{-1/2} + O\left(h_x^p\right) + O_p\left(\frac{\|\theta-\theta_0\|}{h_x}\right)\left[O_p\left(n^{2-\delta}h_x^3\right)^{-1/2} + O\left(h_x^p\right)\right] \\ &= O_p\left(n^{2-\delta}h_x^3\right)^{-1/2} + O\left(h_x^p\right) + O_p\left(\theta-\theta_0\right)\left[O_p\left(n^{2-\delta}h_x^5\right)^{-1/2} + O\left(h_x^{p-1}\right)\right] \\ &= O_p\left(n^{2-\delta}h_x^3\right)^{-1/2} + O\left(h_x^p\right) + o_p\left(\theta-\theta_0\right). \end{aligned}$$

where the last step results from assumption (A8). Finally, by (2.40),

$$U_{\theta}^{(A)} = O_p\left(n^{2-\delta}h_x^3\right)^{-1/2} + O\left(h_x^p\right) + o_p\left(\theta-\theta_0\right).$$

We now turn to deal with $U_\theta^{(B)}$ in a similar way. Note that in the definition of $\varsigma_\theta^{(B)}(x_i, x_j)$ (see (2.28)), the first term, say $\varsigma_\theta^{(B,1)}(x_i, x_j)$, is

$$\varsigma_\theta^{(B,1)}(x_i, x_j) \equiv \frac{1}{h_x^3} \int \left(x_j - E\left(X|\theta^T X = t\right)\right) K\left(\frac{\theta^T x_i - t}{h_x}\right) K'\left(\frac{\theta^T x_j - t}{h_x}\right) f(t)^{-1}\, dt. \tag{2.44}$$

Since by assumptions (A2) and (A4) $\varsigma_\theta^{(B,1)}(x_i, x_j)$ is bounded, an application of Fubini's theorem (Gut 2005, Chapter 2,Theorem 9.1) and similar arguments as the ones that led to (2.38) and (2.39), yield that for a fixed $x \in \mathbb{S}_X$ both $E\left(\varsigma_\theta^{(B,1)}(X, x)\right)$ and $E\left(\varsigma_\theta^{(B,1)}(x, X)\right)$ are equal to

$$\begin{aligned} & -\left.\frac{d}{dt}\right|_{t=\theta^T x} E\left(X|\theta^T X = t\right) + O(h_x^p) \qquad (2.45) \\ = \; & E\left(\nabla \log f_{\theta^T X}\left(\theta^T X\right)|\theta^T X = \theta^T x\right) + O(h_x^p), \end{aligned}$$

Thus,

$$E\left(\varsigma_\theta^{(B)}(X, x)\right) = O(h_x^p) \text{ and } E\left(\varsigma_\theta^{(B)}(x, X)\right) = O(h_x^p),$$

and similarly to $U_\theta^{(A)}$, we have that $U_\theta^{(B)}$ is a degenerate U-statistic up to a $O(h_x^p)$ term, where the constants in the $O(\cdot)$ terms are independent of $\theta \in \Theta$. Applying Chebyshev's inequality (Gut 2005, Chapter 3, Theorem 1.4) and Lemma 2.6.4 to the U-statistic $U_{\theta_0}^{(B)}$ in a similar manner as to $U_{\theta_0}^{(A)}$ (see (2.42)) we get that

$$U_{\theta_0}^{(B)} = O_p\left(n^{2-\delta} h_x^3\right)^{-1/2} + O\left(h_x^p\right).$$

A similar uniformity argument as the one that led to (2.41) finally completes the proof of (2.29).

We continue to prove the asymptotic bounds in probability for the remainder terms $R_{\theta,1}, R_{\theta,2}$ and $R_{\theta,3}$.

We start with

$$R_{\theta,1} = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(\widehat{f}_{\theta^T X}^{-i} - f_{\theta^T X}\right)\left(\nabla \widehat{f}_{\theta^T X}^{-i} - \nabla f_{\theta^T X}\right)}{\left(f_{Y,\theta^T X}\right)^2} - U_\theta^{(B)}.$$

Put

$$\begin{aligned} \rho_{1,\theta}\left(x_i, x_j, x_k\right) &\equiv Z_\theta^{(1)}\left(x_i, x_j\right) Z_\theta^{(2)}\left(x_i, x_k\right) - Z_\theta^{(1)}\left(x_i, x_j\right) \frac{\nabla f_{\theta^T X}\left(\theta^T x_i\right)}{f_{\theta^T X}\left(\theta^T x_i\right)} \\ &\quad - Z_\theta^{(2)}\left(x_i, x_k\right) + \frac{\nabla f_{\theta^T X}\left(\theta^T x_i\right)}{f_{\theta^T X}\left(\theta^T x_i\right)}, \end{aligned} \tag{2.46}$$

where $Z^{(1)}(\cdot,\cdot)$ and $Z^{(2)}(\cdot,\cdot)$ are defined in (2.36). Note that $R_{\theta,1}$ can be redefined as a sum of third and second order $\mathbb{R}^d$-vector U-statistics in the following way

$$\begin{aligned} R_{\theta,1} &= \frac{1}{n(n-1)^2} \sum_i \sum_{j \neq i} \sum_{k \neq i} \rho_\theta^{(1)}\left(x_i, x_j, x_k\right) - U_\theta^{(B)} \\ &= \frac{n-2}{n-1} \cdot \underbrace{\frac{1}{n(n-1)(n-2)} \sum_{1 \leq i \neq j \neq k \leq n} \left(\rho_{1,\theta}\left(x_i, x_j, x_k\right) - \varsigma_\theta^{(B)}\left(x_j, x_k\right)\right)}_{U^{(1,A)}} \\ &\quad + \frac{1}{n-1} \cdot \underbrace{\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \rho_{1,\theta}\left(x_i, x_j, x_j\right).}_{U^{(1,B)}} \end{aligned} \tag{2.47}$$

We have by construction for fixed $x_j, x_k \in \mathbb{S}_X$,

$$E\left(Z_\theta^{(1)}\left(X, x_j\right) Z_\theta^{(2)}\left(X, x_k\right)\right) = \varsigma_\theta^{(B,1)}\left(x_j, x_k\right), \tag{2.48}$$

with $\varsigma_\theta^{(B,1)}\left(x_i, x_j\right)$ as in (2.44), and a tedious but straightforward calculation with results (2.38), (2.39), (2.45) and (2.48) implies that $U^{(1,A)}$ is a degenerate U-statistic up to a $O(h_x^p)$ term, in the sense that for any fixed $x_i, x_j, x_k \in \mathbb{S}_X$,

$$\begin{aligned} E\left(\rho_{1,\theta}\left(X, x_j, x_k\right) - \varsigma_\theta^{(B)}\left(x_j, x_k\right)\right) &= O(h_x^p), \\ E\left(\rho_{1,\theta}\left(x_i, X, x_k\right) - \varsigma_\theta^{(B)}\left(X, x_k\right)\right) &= O(h_x^p), \end{aligned}$$

and

$$E\left(\rho_{\theta}^{(1)}\left(x_i,x_j,X\right)-\varsigma_{\theta}^{(B)}\left(x_j,X\right)\right)=O(h_x^p).$$

As an Applications of Chebyshev's inequality (Gut 2005, Chapter 3, Theorem 1.4) and Lemma 2.6.4 (see (2.42)) we now obtain

$$U^{(1,A)}=O_p\left(n^{3-\delta}h_x^5\right)^{-1/2}+O\left(h_x^p\right). \tag{2.49}$$

For the term $U^{(1,B)}$, defined by (2.47), it is enough to note that by Lemma 2.6.2, uniformly on $x_i\in\mathbb{S}_X$,

$$\begin{aligned}\frac{1}{n-1}\sum_{j\neq i}Z_{\theta}^{(1)}\left(x_i,x_j\right) &= 1+O_p\left(\left(\frac{\ln n}{nh_x}\right)^{1/2}\right)+O\left(h_x^p\right),\\ \frac{1}{n-1}\sum_{j\neq i}Z_{\theta}^{(2)}\left(x_i,x_j\right) &= \frac{\nabla f_{\theta^TX}\left(\theta^Tx_i\right)}{f_{\theta^TX}\left(\theta^Tx_i\right)}+O_p\left(\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right)+O\left(h_x^p\right),\end{aligned}$$

and thus, by the definition (2.46) and the continuous mapping theorem (Amemiya 1985, Theorem 3.2.6) applied to a product function, we have

$$\frac{1}{n-1}\sum_{j\neq i}\rho_{1,\theta}\left(x_i,x_j,x_j\right)=O_p\left(\frac{\ln n}{nh_x^2}\right)+O\left(h_x^{2p}\right).$$

Hence,

$$U^{(1,B)}=O_p\left(\frac{\ln n}{nh_x^2}\right)+O\left(h_x^{2p}\right). \tag{2.50}$$

Thus, we get from (2.47), (2.49) and (2.50) that

$$R_{\theta,1}=o_p\left(\left(n^{2-\delta}h_x^3\right)^{-1/2}\right)+O\left(h_x^p\right).$$

A similar uniformity argument as the one that led to (2.41) and assumption (A8) then shows that $\theta\to_p\theta_0$ implies

$$\begin{aligned} |R_{\theta,1}| &\leq |R_{\theta_0,1}| + \frac{\|\theta-\theta_0\|}{h_x} |R_{\theta_0,1}| \\ &= o_p\left(\left(n^{2-\delta}h_x^3\right)^{-1/2}\right) + O\left(h_x^p\right) + o_p\left(\theta-\theta_0\right). \end{aligned}$$

Next, we show a stochastic bound for

$$\begin{aligned} |R_{\theta,2}| &= \left| 2\frac{1}{n}\sum_{i=1}^{n} \frac{\nabla f_{\theta^T X}\left(\widehat{f}_{\theta^T X}^{-i} - f_{\theta^T X}\right)^2}{\left(\overline{f}_{\theta^T X}\right)^3} \right| \qquad (2.51) \\ &\leq 2\sup_{x\in\mathbb{S}_X} \left| \frac{\nabla f_{\theta^T X} f_{\theta^T X}^2}{\left(\overline{f}_{\theta^T X}\right)^3} \right| \cdot \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{f}_{\theta^T X}^{-i}}{f_{\theta^T X}} - 1\right)^2. \end{aligned}$$

Note as that as the first term in the RHS is bounded, it is enough to bound the second term, $\frac{1}{n}\sum_{i=1}^{n}(\frac{\widehat{f}_{\theta^T X}^{-i}}{f_{\theta^T X}} - 1)^2$, in probability. Let now

$$\rho_{2,\theta}\left(x_i,x_j,x_k\right) = Z^{(1)}\left(x_i,x_j\right)Z^{(1)}\left(x_i,x_k\right) - Z^{(1)}\left(x_i,x_j\right) - Z^{(1)}\left(x_i,x_k\right) + 1,$$

where $Z^{(1)}\left(\cdot,\cdot\right)$ is defined in (2.36), and let

$$\varsigma_\theta^{(C)}\left(x_i,x_j\right) \equiv \frac{1}{h_x^2}\int K\left(\frac{\theta^T x_j - t}{h_x}\right)K\left(\frac{\theta^T x_i - t}{h_x}\right)f\left(t\right)^{-1}dt - 1.$$

We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{f}^{-i}_{\theta^T X}}{f_{\theta^T X}}-1\right)^2 \\
=\;& \frac{1}{n\left(n-1\right)^2}\sum_{i}\sum_{j\neq i}\sum_{k\neq i}\rho_{2,\theta}\left(x_i,x_j,x_k\right) \\
=\;& \frac{n-2}{n-1}\cdot\underbrace{\frac{1}{n\left(n-1\right)\left(n-2\right)}\sum_{1\leq i\neq j\neq k\leq n}\left(\rho_{2,\theta}\left(x_i,x_j,x_k\right)-\varsigma^{(C)}_{\theta}\left(x_j,x_k\right)\right)}_{U^{(2,A)}} \\
&+\frac{1}{n-1}\cdot\underbrace{\frac{1}{n\left(n-1\right)}\sum_{1\leq i\neq j\leq n}\rho_{2,\theta}\left(x_i,x_j,x_j\right)}_{U^{(2,B)}}+\underbrace{\frac{1}{n\left(n-1\right)}\sum_{1\leq i\neq j\leq n}\varsigma^{(C)}_{\theta}\left(x_i,x_j\right)}_{U^{(2,C)}}.
\end{aligned}$$

Here, $U^{(2,A)}$ is a third order $\mathbb{R}^d$-vector U-statistic, and $U^{(2,B)}$ and $U^{(2,C)}$ are second order $\mathbb{R}^d$-vector U-statistics. Following a similar treatment as above, these three U-statistics are shown to be degenerate up to a $O(h_x^p)$ term, and by Lemma 2.6.4 we have

$$U^{(2,A)}=O_p\left(n^{3-\delta}h_x^3\right)^{-1/2}+O\left(h_x^p\right),\quad U^{(2,B)}=O_p\left(\frac{\ln n}{nh_x}\right)+O\left(h_x^{2p}\right),$$

$$\text{and}\;\; U^{(2,C)}=O_p\left(n^{2-\delta}h_x\right)^{-1/2}+O\left(h_x^p\right).$$

The last arguments imply that

$$R_{\theta,2}=o_p\left(\left(n^{2-\delta}h_x^3\right)^{-1/2}\right)+O\left(h_x^p\right).$$

A similar uniformity argument as in (2.41) yields that $\theta\to_p\theta_0$ implies

$$\left|R_{\theta,2}\right|=o_p\left(\left(n^{2-\delta}h_x^3\right)^{-1/2}\right)+O\left(h_x^p\right)+o_p\left(\theta-\theta_0\right).$$

Finally, bounding $R_{\theta,3}$ is straightforward by the uniform consistency result of Lemma 2.6.2 and the continuous mapping theorem (Amemiya 1985, Theorem 3.2.6) applied to the product function. We have therefore established now (2.30).

Retracing through established results (2.26), (2.29) and (2.30), we have completed the proof of assertion (2.23) and therefore of Theorem 2.3.1. ■

Proof of Theorem 2.3.3. By smoothness condition (A4) and the mean-value theorem applied to the function $\widetilde{f}_{Y|\theta^T X}\left(y|\theta^T x\right)$ with mean value $\overline{\theta} \in \Theta$ such that $\left|\overline{\theta} - \theta_0\right| \leq \left|\widehat{\theta} - \theta_0\right|$ and Theorems 2.6.2 and 2.3.2,

$$\begin{aligned} & \sup_{(y,x)\in\mathbb{S}} \left|\widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\widehat{\theta}^T x\right) - \widetilde{f}_{Y|\theta_0^T X}\left(y|\theta_0^T x\right)\right| \\ \leq \quad & \left\|\widehat{\theta} - \theta_0\right\| \left\|\sup_{(y,x)\in\mathbb{S}} \nabla \widetilde{f}_{Y|\widehat{\theta}^T X}\left(y|\overline{\theta}^T x\right)\right\| \\ = \quad & \left\|\widehat{\theta} - \theta_0\right\| \left\|\sup_{(y,x)\in\mathbb{S}} \nabla f_{Y|\overline{\theta}^T X}\left(y|\overline{\theta}^T x\right) + o_p\left(1\right)\right\| \\ = \quad & o_p\left(\left(\frac{1}{nH_yH_x}\right)^{1/2}\right). \blacksquare \end{aligned}$$

2.6 Appendix B - Technical Lemmas

This section gives some useful technical results that are needed in the proofs of the main theorems.

Recall that for a function $g\left(\theta\right)$ that depends on $\theta \in \Theta$ and possibly also on other variables we denote $\nabla g\left(\theta\right)$ and $\nabla^2 g\left(\theta\right)$ as the vector and matrix of partial derivatives of $g\left(\theta\right)$ with respect to θ. As a convention, we also use $\nabla^0 g\left(\theta\right) = g\left(\theta\right)$.

The following Lemma gives the forms of the partial derivatives of $f_{Y,\theta^T X}\left(y, \theta^T x\right)$ and $f_{\theta^T X}\left(\theta^T x\right)$ with respect to θ. One has to remember that θ affects the value of the probability densities $f_{Y,\theta^T X}\left(y, \theta^T x\right)$ and $f_{\theta^T X}\left(\theta^T x\right)$ not only through the variable $\theta^T x$, but it also defines the density functions $f_{Y,\theta^T X}\left(\cdot,\cdot\right)$ and $f_{\theta^T X}\left(\cdot\right)$ themselves.

Lemma 2.6.1 *Let* $E\left(X|Y = y, \theta^T X = t\right)$, $E\left(XX^T|Y = y, \theta^T X = t\right)$, $E\left(X|\theta^T X = t\right)$, $E\left(XX^T|\theta^T X = t\right)$ *and* $f_{Y,\theta^T X}\left(y, t\right)$ *and* $f_{\theta^T X}\left(t\right)$ *exist and they are twice differentiable* *with respect to $y, t \in \mathbb{R}$. Then*

$$\begin{aligned}
\nabla f_{Y,\theta^T X}\left(y, \theta^T x\right) &= \left.\frac{d}{dt}\right|_{t=\theta^T x} \left\{E\left(x - X|Y = y, \theta^T X = t\right) f_{Y,\theta^T X}\left(y, t\right)\right\}, \\
\nabla^2 f_{Y,\theta^T X}\left(y, \theta^T x\right) &= \left.\frac{d^2}{dt^2}\right|_{t=\theta^T x} \left\{E\left(\left(x - X\right)\left(x - X\right)^T |Y = y, \theta^T X = t\right) f_{Y,\theta^T X}\left(y, t\right)\right\},
\end{aligned}$$

and similarly,

$$\begin{aligned}
\nabla f_{\theta^T X}\left(\theta^T x\right) &= \left.\frac{d}{dt}\right|_{t=\theta^T x} \left\{E\left(x - X|\theta^T X = t\right) f_{\theta^T X}\left(t\right)\right\}, \\
\nabla^2 f_{\theta^T X}\left(\theta^T x\right) &= \left.\frac{d^2}{dt^2}\right|_{t=\theta^T x} \left\{E\left(\left(x - X\right)\left(x - X\right)^T |\theta^T X = t\right) f_{\theta^T X}\left(t\right)\right\}.
\end{aligned}$$

Proof. We prove here only the last two identities of the Lemma as the first two follow similarly. Assume $\theta_d \neq 0$ since otherwise we may reduce the dimension to $d-1$. Denote $\xi_1^{d-1} = \left(\xi_1, ..., \xi_{d-1}\right)$ and let $f_X\left(\xi_1^{d-1}, \xi_d\right) = f_X\left(\xi_1, ..., \xi_d\right)$ be the probability density of X at $(\xi_1, ..., \xi_d)$. We now have with

$$f_{\theta^T X}\left(t\right) = \theta_d^{-1} \int f_X\left(\xi_1^{d-1}, \theta_d^{-1}(t - \sum_{j=1}^{d-1} \theta_j \xi_j)\right) d\xi_1^{d-1},$$

where θ_d^{-1} is the determinant of the Jacobian matrix. Thus, for $t = \theta^T x$,

$$f_{\theta^T X}\left(\theta^T x\right) = \theta_d^{-1} \int f_X(\xi_1^{d-1}, x_d + \theta_d^{-1} \sum_{j=1}^{d-1} \theta_j \left(x_j - \xi_j\right)) d\xi_1^{d-1}.$$

Note also that for $k, l \in \{1, 2, ..., d-1\}$ we have

$$\begin{aligned}
E\left(X_k|\theta^T X = t\right) f_{\theta^T X}(t) &= \theta_d^{-1} \int \xi_k f_X(\xi_1^{d-1}, \theta_d^{-1}(t - \sum_{j=1}^{d-1} \theta_j \xi_j)) d\xi_1^{d-1}, \\
E\left(X_d|\theta^T X = t\right) f_{\theta^T X}(t) &= \theta_d^{-2} \int (t - \sum_{j=1}^{d-1} \theta_j \xi_j) f_X(\xi_1^{d-1}, \theta_d^{-1}(t - \sum_{j=1}^{d-1} \theta_j \xi_j)) d\xi_1^{d-1}, \\
E\left(X_k X_l|\theta^T X = t\right) f_{\theta^T X}(t) &= \theta_d^{-1} \int \xi_k \xi_l f_X(\xi_1^{d-1}, \theta_d^{-1}(t - \sum_{j=1}^{d-1} \theta_j \xi_j)) d\xi_1^{d-1}, \\
E\left(X_k X_d|\theta^T X = t\right) f_{\theta^T X}(t) &= \theta_d^{-2} \int \xi_k (t - \sum_{j=1}^{d-1} \theta_j \xi_j) f_X(\xi_1^{d-1}, \theta_d^{-1}(t - \sum_{j=1}^{d-1} \theta_j \xi_j)) d\xi_1^{d-1}, \\
E\left(X_d^2|\theta^T X = t\right) f_{\theta}(t) &= \theta_d^{-3} \int (t - \sum_{j=1}^{d-1} \theta_j \xi_j)^2 f_X(\xi_1^{d-1}, \theta_d^{-1}(t - \sum_{j=1}^{d-1} \theta_j \xi_j)) d\xi_1^{d-1}.
\end{aligned}$$

Using the above expressions one can use direct differentiation to verify the last two identities of the Lemma. ■

The proofs of Theorems 2.3.1 and 2.3.2 rely heavily on the uniform consistency of the kernel density estimators' derivatives with respect to θ. The next two Lemmas are direct modifications of the results of Hansen (2008), but unlike Hansen's (2008) theory, they concern with partial derivatives of the kernel estimates with respect to θ, rather than with derivatives with respect to the density variables themselves.

Lemma 2.6.2 *Let (A1)-(A4) hold. Then*

$$\begin{aligned}
\sup_{\theta \in \Theta, z \in \mathbb{S}} \left| \widehat{f}_{Y,\theta^T X}\left(y, \theta^T x\right) - f_{Y,\theta^T X}\left(y, \theta^T x\right) \right| &= O_p\left(\left(\frac{\ln n}{n h_y h_x} \right)^{1/2} + h_y^2 + h_x^2 \right), \\
\sup_{\theta \in \Theta, x \in \mathbb{S}_X} \left| \widehat{f}_{\theta^T X}\left(\theta^T x\right) - f_{\theta^T X}\left(\theta^T x\right) \right| &= O_p\left(\left(\frac{\ln n}{n h_x} \right)^{1/2} + h_x^2 \right)
\end{aligned}$$

If, in addition, also (A7) and (A9) hold. Then for $k = 0, 1, 2$,

$$\begin{aligned}
\sup_{\theta\in\Theta, z\in\mathbb{S}} \left|\nabla^k \widehat{f}_{Y,\theta^T X}\left(y, \theta^T x\right) - \nabla^k f_{Y,\theta^T X}\left(y, \theta^T x\right)\right| &= O_p\left(\left(\frac{\ln n}{n h_y h_x^{1+2k}}\right)^{1/2} + h_y^p + h_x^p\right), \\
\sup_{\theta\in\Theta, x\in\mathbb{S}_X} \left|\nabla^k \widehat{f}_{\theta^T X}\left(\theta^T x\right) - \nabla^k f_{\theta^T X}\left(\theta^T x\right)\right| &= O_p\left(\left(\frac{\ln n}{n h_x^{1+2k}}\right)^{1/2} + h_x^p\right).
\end{aligned}$$

Proof of Lemma 2.6.2. We prove here only that under conditions (A1)-(A4) and (A7) and (A9),

$$\sup_{\theta\in\Theta, x\in\mathbb{S}_X} \left|\nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right) - \nabla f_{\theta^T X}\left(\theta^T x\right)\right| = O_p\left(\left(\frac{\ln n}{n h_x^{1+2k}}\right)^{1/2} + h_x^p\right).$$

The proofs for the rest of the arguments are very similar. By Lemma 2.6.3, it is sufficient to prove that $\sup_{\Theta\times\mathbb{S}_X} \left|E\left(\nabla \widehat{f}_{\theta^T X}\left(\theta^T x\right)\right) - \nabla f_{\theta^T X}\left(\theta^T x\right)\right| = O(h_x^p)$. A change of variables, integration by parts, and a Taylor expansion around $h_x = 0$ yield with (A7) and (A9) that uniformly in $x \in \mathbb{S}_X$,

$$\begin{aligned}
& E\left(\nabla^k \widehat{f}_{\theta^T X}\left(\theta^T x\right)\right) \\
= \; & \frac{1}{h_x^2} \int \left(x - E\left(X|\theta^T X = t\right)\right) K'\left(\frac{\theta^T x - t}{h_x}\right) f_{\theta^T X}(t)\, dt \\
= \; & \frac{1}{h_x} \int \left(x - E\left(X|\theta^T X = \theta^T x - h_x u\right)\right) K'(u)\, f_{\theta^T X}\left(\theta^T x - h_x u\right) du \\
= \; & \int \left.\frac{d}{dt}\right|_{t=\theta^T x - h_x u} \left[\left(x - E\left(X|\theta^T X = t\right)\right) f_{\theta^T X}(t)\right] K(u)\, du \\
= \; & \int \left\{\sum_{j=1}^{p-1} \left[\left.\frac{d^{1+j}}{dt^{1+j}}\right|_{t=\theta^T x} \left(x - E\left(X|\theta^T X = t\right)\right) f_{\theta^T X}(t) \left(-h_x u\right)^j\right] + O\left(h_x^p\right)\right\} K(u)\, du \\
= \; & \left.\frac{d}{dt}\right|_{t=\theta^T x} \left[\left(x - E\left(X|\theta^T X = t\right)\right) f_{\theta^T X}(t)\right] + O(h_x^p).
\end{aligned}$$

By Lemma 2.6.1, the last expression is just $\nabla f_{\theta^T X}\left(\theta^T x\right) + O(h^p)$. ■

Lemma 2.6.3 *Let (A1)-(A4) hold. Then*

$$\begin{aligned}\sup_{\theta\in\Theta,z\in\mathbb{S}}\left|\widehat{f}_{Y,\theta^TX}\left(y,\theta^Tx\right)-E\widehat{f}_{Y,\theta^TX}\left(y,\theta^Tx\right)\right| &= O_p\left(\left(\frac{\ln n}{nh_yh_x}\right)^{1/2}\right),\\ \sup_{\theta\in\Theta,x\in\mathbb{S}_X}\left|\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-E\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right| &= O_p\left(\left(\frac{\ln n}{nh_x}\right)^{1/2}\right),\end{aligned}$$

If, in addition, also (A7) holds. Then for $k=0,1,2$*,*

$$\begin{aligned}\sup_{\theta\in\Theta,z\in\mathbb{S}}\left|\nabla^k\widehat{f}_{Y,\theta^TX}\left(y,\theta^Tx\right)-E\nabla^k\widehat{f}_{Y,\theta^TX}\left(y,\theta^Tx\right)\right| &= O_p\left(\left(\frac{\ln n}{nh_yh_x^{1+2k}}\right)^{1/2}\right),\\ \sup_{\theta\in\Theta,x\in\mathbb{S}_X}\left|\nabla^k\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-E\nabla^k\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right| &= O_p\left(\left(\frac{\ln n}{nh_x^{1+2k}}\right)^{1/2}\right).\end{aligned}$$

Proof. We prove here only that under conditions (A1)-(A4) and (A7),

$$\sup_{\Theta\times\mathbb{S}_X}\left|\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-E\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right|=O_p\left(\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right).$$

The proofs for the rest of the arguments in the Theorem are very similar. Let $\overline{\theta}_1\in\Theta$, $\overline{x}_1\in\mathbb{S}_X$ and define

$$A_1=\left\{\theta,x:\left\|\theta-\overline{\theta}_1\right\|\leq\left(\frac{h_x\ln n}{n}\right)^{1/2},\left\|x-\overline{x}_1\right\|\leq\left(\frac{h_x\ln n}{n}\right)^{1/2}\right\}. \tag{2.52}$$

Since $\Theta\times\mathbb{S}_X\in\mathbb{R}^d\times\mathbb{R}^d$ is compact, then it can be covered by $J\left(n\right)=O\left(\left(\frac{n}{h_x\ln n}\right)^d\right)$ such subspaces $A_1,...,A_J$ around centres $\left\{\left(\overline{\theta}_k,\overline{x}_k\right)\right\}_{j=1}^J$. Since

$$\begin{aligned}&P\left(\sup_{\Theta\times\mathbb{S}_X}\left|\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-E\nabla^k\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right|>\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right)\\ \leq\ &J\left(n\right)\max_{j=1,...,J}P\left(\sup_{(\theta,x)\in A_j}\left|\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-E\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right|>\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right),\end{aligned}$$

it is therefore suffice to prove that for any $\left(\overline{\theta}_1,\overline{x}_1\right)\in\Theta\times\mathbb{S}_X$ and A_1 as in (2.52), the following holds

$$P\left(\sup_{(\theta,x)\in A_1}\left|\nabla\widehat{f}_{\theta^T X}\left(\theta^T x\right)-E\nabla\widehat{f}_{\theta^T X}\left(\theta^T x\right)\right|>\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right)=o\left(\left(\frac{h_x\ln n}{n}\right)^d\right), \quad (2.53)$$

where the constant in the $o\left(\cdot\right)$ term is independent of $\left(\overline{\theta}_1,\overline{x}_1\right)$ and n.

Define the functions $\widetilde{K}_j,\ j=1,2,3$, on

$$T=\left\{t\in\mathbb{R}:t=\frac{\theta^T x}{h_x}\text{ for some }\theta\in\Theta\text{ and }\overline{x}_1-x\in\mathbb{S}_X\right\}$$

by

$$\widetilde{K}_1\left(t\right)=\sup_{H(t)}\left\{\|x\|\left|K''\left(\frac{\theta^T x}{h_x}\right)x\right|\right\},\qquad\widetilde{K}_2\left(t\right)=\sup_{H(t)}\left\{\|\theta\|\left|K''\left(\frac{\theta^T x}{h_x}\right)x\right|\right\},$$

and

$$\widetilde{K}_3\left(t\right)=\sup_{H(t)}\left|\ K'\left(\frac{\theta^T x}{h_x}\right)\right|.$$

where all the sups are taken over $\theta\in\Theta$ and $x\in\mathbb{S}_X$ such that $\frac{\theta^T X}{h_x}$ is not too far from t in the sense that

$$H\left(t\right)\equiv\left\{(\theta,x):\|\theta{-}\theta_*\|\le\left(\frac{h_x\ln n}{n}\right)^{1/2},\|x-x_*\|\le\left(\frac{h_x\ln n}{n}\right)^{1/2}\text{ and }\frac{\theta_*^T x_*}{h_x}=t\right\}$$

Note that $\widetilde{K}_j,\ j=1,2,3$, are well-defined, compactly supported and bounded for any $t\in T$ by assumption (A7) and compactness of Θ and $\mathbb{S}_X$. Let x_i denote the i'th X-observation, and for any $(\theta,x)\in A_1$ we have with mean-values θ_*, x_* such that $\left\|\overline{\theta}_1-\theta_*\right\|\le\left\|\overline{\theta}_1-\theta\right\|\le$

$\left(\frac{h_x \ln n}{n}\right)^{1/2}$ and $\|\overline{x}_1 - x_*\| \le \|\overline{x}_1 - x\| \le \left(\frac{h_x \ln n}{n}\right)^{1/2}$ that

$$\begin{aligned}
& \left|\nabla K\left(\frac{\overline{\theta}_1^T\left(\overline{x}_1 - x_i\right)}{h_x}\right) - \nabla K\left(\frac{\theta^T\left(x - x_i\right)}{h_x}\right)\right| \\
\le\ & \frac{1}{h_x}\left|\left[K'\left(\frac{\overline{\theta}_1^T\left(\overline{x}_1 - x_i\right)}{h_x}\right) - K'\left(\frac{\theta^T\left(x - x_i\right)}{h_x}\right)\right]\left(\overline{x}_1 - x_i\right)\right| \\
& + \frac{1}{h_x}\left|K'\left(\frac{\theta^T\left(x - x_i\right)}{h_x}\right)\left(\overline{x}_1 - x\right)\right| \\
\le\ & \frac{1}{h_x^2}\left|\left(\overline{\theta}_1 - \theta\right)^T\left(\overline{x}_1 - x_i\right)K''\left(\frac{\theta_*^T\left(x_* - x_i\right)}{h_x}\right)\left(\overline{x}_1 - x_i\right)\right| \\
& + \frac{1}{h_x^2}\left|\theta^T\left(\overline{x}_1 - x\right)K''\left(\frac{\theta_*^T\left(x_* - x_i\right)}{h_x}\right)\left(\overline{x}_1 - x_i\right)\right| + \frac{1}{h_x}\left|K'\left(\frac{\theta^T\left(x - x_i\right)}{h_x}\right)\left(\overline{x}_1 - x\right)\right| \\
\le\ & \frac{\|\overline{\theta}_1 - \theta\|}{h_x^2}\left|\widetilde{K}_1\left(\frac{\overline{\theta}_1^T\left(\overline{x}_1 - x_i\right)}{h_x}\right)\right| + \frac{\|\overline{x}_1 - x\|}{h_x^2}\left|\widetilde{K}_2\left(\frac{\overline{\theta}_1^T\left(\overline{x}_1 - x_i\right)}{h_x}\right)\right| \\
& + \frac{\|\overline{x}_1 - x\|}{h_x}\left|\widetilde{K}_3\left(\frac{\overline{\theta}_1^T\left(\overline{x}_1 - x_i\right)}{h_x}\right)\right| \\
\le\ & \left(\frac{\ln n}{n h_x^3}\right)^{1/2} \cdot \left(\sum_{j=1}^{3}\left|\widetilde{K}_j\left(\frac{\overline{\theta}_1^T\left(\overline{x}_1 - x_i\right)}{h_x}\right)\right|\right) \qquad (2.54)
\end{aligned}$$

Note that the last term is independent of $(\theta, x) \in A_1$. We now define for any $(\theta, x) \in A_1$ and $j = 1, 2, 3$,

$$\widetilde{f}_{\theta^T X, j}\left(\theta^T x\right) = \frac{1}{n h_x}\sum\nolimits_{i=1}^{n}\widetilde{K}_j\left(\frac{\theta^T\left(x - x_i\right)}{h_x}\right).$$

We have

$$E\left|\widetilde{f}_{\theta^T X, j}\left(\theta^T x\right)\right| \le \sup_{(\theta, x) \in \Theta \times \mathbb{S}_X}\left|f_{\theta^T x}\left(\theta^T x\right)\right| \int\left|\widetilde{K}_j\left(u\right)\right| du < \infty, \qquad (2.55)$$

Also, inequality (2.54) implies

$$\sup_{(\theta, x) \in A_1}\left|\nabla\widehat{f}_{\theta^T X}\left(\overline{\theta}_1^T x\right) - \nabla\widehat{f}_{\theta^T X}\left(\theta^T x\right)\right| \le \left(\frac{\ln n}{n h_x^3}\right)^{1/2} \cdot \left(\sum_{j=1}^{3}\left|\widetilde{f}_{\theta j}\left(\overline{\theta}_1^T \overline{x}_1\right)\right|\right). \qquad (2.56)$$

Thus, the last three inequalities yield for any $(\theta, x) \in A_1$, for some large enough M, independent on θ_1, x_1 and n,

$$\sup_{(\theta,x)\in A_1}\left|E\left\{\nabla\widehat{f}_{\theta^TX}\left(\overline{\theta}_1^Tx\right)-\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right\}\right|\le M\left(\frac{\ln n}{nh_x^3}\right)^{1/2}. \tag{2.57}$$

Next, results (2.55), (2.56), (2.57) and the condition that $\frac{\ln n}{nh_x}=o(1)$ give

$$\begin{aligned}
&\sup_{(\theta,x)\in A_1}\left|\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-E\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right|\\
\le\ &\sup_{(\theta,x)\in A_1}\left|\nabla\widehat{f}_{\theta^TX}\left(\overline{\theta}_1^Tx\right)-\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right|+\left|\nabla\widehat{f}_{\theta^TX}\left(\overline{\theta}_1^Tx\right)-E\nabla\widehat{f}_{\theta^TX}\left(\overline{\theta}_1^Tx\right)\right|\\
&+\sup_{(\theta,x)\in A_1}\left|E\left\{\nabla\widehat{f}_{\theta^TX}\left(\overline{\theta}_1^Tx\right)-\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right\}\right|\\
\le\ &\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\sum_{j=1}^{3}\left\{\left|\widetilde{f}_{\theta^TX,j}\left(\overline{\theta}_1^T\overline{x}_1\right)-E\widetilde{f}_{\theta^TX,j}\left(\overline{\theta}_1^T\overline{x}_1\right)\right|+E\left|\widetilde{f}_{\theta^TX,j}\left(\overline{\theta}_1^T\overline{x}_1\right)\right|\right\}\\
&+\left|\nabla\widehat{f}_{\theta^TX}\left(\overline{\theta}_1^T\overline{x}_1\right)-E\nabla\widehat{f}_{\theta^TX}\left(\overline{\theta}_1^T\overline{x}_1\right)\right|+M\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\\
\le\ &\frac{1}{h_x}\sum_{j=1}^{3}\left|\widetilde{f}_{\theta^TX,j}\left(\overline{\theta}_1^T\overline{x}_1\right)-E\widetilde{f}_{\theta^TX,j}\left(\overline{\theta}_1^T\overline{x}_1\right)\right|+\left|\nabla\widehat{f}_{\theta^TX}\left(\overline{\theta}_1^T\overline{x}_1\right)-E\nabla\widehat{f}_{\theta^TX}\left(\overline{\theta}_1^T\overline{x}_1\right)\right|\\
&+2M\left(\frac{\ln n}{nh_x^3}\right)^{1/2}.
\end{aligned}$$

As a result we get

$$\begin{aligned}
&P\left(\sup_{(\theta,x)\in A_k}\left|\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-E\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right|>5M\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right) \qquad (2.58)\\
\le\ &P\left(\left|\nabla\widehat{f}_{\theta^TX}\left(\overline{\theta}_1^T\overline{x}_1\right)-E\nabla\widehat{f}_{\theta^TX}\left(\overline{\theta}_1^T\overline{x}_1\right)\right|>M\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right)+\\
&+\sum_{j=1}^{3}P\left(\left|\widetilde{f}_{\theta^TX,j}\left(\overline{\theta}_1^T\overline{x}_1\right)-E\widetilde{f}_{\theta^TX,j}\left(\overline{\theta}_1^T\overline{x}_1\right)\right|>M\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right).
\end{aligned}$$

We now bound the four terms in the RHS of (2.58) using the same argument, as all kernels used in the construction of $\widetilde{f}_{\theta^TX,j}$ and $\widehat{f}_{\theta^TX}$ all bounded and compactly supported. We therefore prove the bound only for the term $\left|\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)-E\nabla\widehat{f}_{\theta^TX}\left(\theta^Tx\right)\right|$. Set $m=\left(\frac{nh_x}{\ln n}\right)^{1/2}$, and note that for n sufficiently large, $m<\max\left(n,\frac{\varepsilon}{4b}\right)$ where $b=$

$2\sup_{\Theta\times\mathbb{S}_X}\left|\|x\|\frac{\partial}{\partial u}K(u)\right|<\infty$, and $\varepsilon=M(nh_x\ln n)^{1/2}$. Define for $(\theta,x)\in A_1$,

$$Z_i=(x-x_i)\left\{\left.\frac{\partial}{\partial t}\right|_{t=\frac{\theta^T(x-x_i)}{h_x}}K(t)-E\left(\left.\frac{\partial}{\partial t}\right|_{t=\frac{\theta^T(x-x_i)}{h_x}}K(t)\right)\right\},\quad i=1,...,m.$$

Now, notice that $|Z_i|\le b$, and by Theorem 1 of Hansen (2008),

$$\sigma^2(m)\equiv\sup_{(\theta,x)\in A_1}E\left|\sum_{i=1}^{\lfloor m\rfloor}Z_i\right|^2\le Cmh_x$$

for some large enough $C>0$. By Theorem 2.1 of Liebscher (1996) and condition (A1) we obtain

$$\begin{aligned}
&P\left(\left|\nabla\widehat{f}_{\theta^TX}\left(\overline{\theta}_1^T\overline{x}_1\right)-E\nabla\widehat{f}_{\theta^TX}\left(\overline{\theta}_1^T\overline{x}_1\right)\right|>M\left(\frac{\ln n}{nh_x^3}\right)^{1/2}\right)\\
=\;&P\left(\left|\sum_{i=1}^{n}Z_i\right|>\varepsilon\right)\\
\le\;&4\exp\left(-\frac{\varepsilon^2}{64\frac{n}{m}\sigma^2(m)+\frac{8}{3}\varepsilon mb}\right)+4\frac{n}{m}\alpha_m\\
\le\;&4\exp\left(-\frac{M^2(nh_x\ln n)}{64Cnh_x+3Mnh_xb}\right)+4A\left(\frac{n\ln n}{h_x}\right)^{1/2}\alpha^{\sqrt{nh_x/\ln n}}\\
\le\;&4\exp\left(-\frac{M^2\ln n}{64C+3Mb}\right)+4A\left(\frac{n\ln n}{h_x}\right)^{1/2}\alpha^{\sqrt{nh_x/\ln n}}\\
\le\;&4n^{-M/(64+3b)}+4A\left(\frac{n\ln n}{h_x}\right)^{1/2}\alpha^{\sqrt{nh_x/\ln n}},
\end{aligned}\tag{2.59}$$

where $0<\alpha<1$ and the last inequality is justified by taking $M\ge C$. Now, we have for the first term of (2.59), $n^{-M/(64+3b)}=o\left(\left(\frac{h_x\ln n}{n}\right)^d\right)$ for sufficiently large M. Also, by condition (A3), we get for the the second term of (2.59) with some arbitrarily small $\delta > 0$,

$$\begin{aligned} & 4A\left(\frac{n\ln n}{h_x}\right)^{1/2}\alpha^{\sqrt{nh_x/\ln n}} \\ = & \ O\left(\left(\frac{n\ln n}{h_x}\right)^{1/2}\alpha^{n^{\delta/2}}\right) \\ = & \ o\left(\left(\frac{h_x\ln n}{n}\right)^{d}\right). \end{aligned}$$

This completes the proof of Lemma 2.6.3. ■

The next Lemma is Lemma C.2 of Gao and King (2004) that gives a bound for the stochastic order of second- and third-order degenerate U-statistics of strong-mixing stochastic process.

Lemma 2.6.4 (Gao and King, 2004) *(i) Let $\psi(\cdot,\cdot,\cdot)$ be a symmetric Borel function defined on $\mathbb{R}^r \times \mathbb{R}^r \times \mathbb{R}^r$, and let the process ξ_i be an r-dimensional strictly stationary and strong-mixing stochastic process with mixing coefficients that satisfy $\alpha_t \leq A\alpha^t$ with $0 < A < \infty$ and $0 < \alpha < 1$. Assume that for any fixed $x, y \in \mathbb{R}^r$, $E[\psi(\xi_1, x, y)] = 0$. Then*

$$E\left\{\sum_{1\leq i<j<k\leq T}\psi\left(\xi_i,\xi_j,\xi_k\right)\right\}^2 \leq CT^3M^{1/(1+\delta)},$$

where $0 < \delta < 1$ is a small constant, $C > 0$ is a constant independent of T and the function ψ, $M = \max\{M_1, M_2, M_3\}$, and

$$\begin{aligned} M_1 &= \max_{1\leq i<j\leq T}\max\left\{E\left|\psi\left(\xi_1,\xi_i,\xi_j\right)\right|^{2+\delta}, \int\left|\psi\left(\xi_1,\xi_i,\xi_j\right)\right|^{2+\delta}dP\left(\xi_1\right)dP\left(\xi_i,\xi_j\right)\right\}, \\ M_2 &= \max_{1\leq i<j\leq T}\max\left\{\int\left|\psi\left(\xi_1,\xi_j,\xi_k\right)\right|^{2+\delta}dP\left(\xi_i\right)dP\left(\xi_1,\xi_j\right)\right\}, \\ M_3 &= \max_{1\leq i<j\leq T}\max\left\{\int\left|\psi\left(\xi_1,\xi_j,\xi_k\right)\right|^{2+\delta}dP\left(\xi_1\right)dP\left(\xi_i\right)dP\left(\xi_j\right)\right\}. \end{aligned}$$

(ii) Let $\phi(\cdot,\cdot)$ be a symmetric Borel function defined on $\mathbb{R}^r \times \mathbb{R}^r$, and let the process ξ_i

be defined as in part (i). Assume that for any fixed $x \in \mathbb{R}^r$, $E\left[\phi\left(\xi_1, x\right)\right] = 0$. Then

$$E\left\{\sum_{1\leq i<j<k\leq T} \phi\left(\xi_i, \xi_j\right)\right\}^2 \leq CT^2 M_4^{1/(1+\delta)},$$

where $0 < \delta < 1$ is a small constant, $C > 0$ is a constant independent of T and the function ϕ, and

$$M_4 = \max_{1\leq i\leq T} \max\left\{E\left|\phi\left(\xi_1, \xi_i\right)\right|^{2+\delta}, \int \left|\phi\left(\xi_1, \xi_i\right)\right|^{2+\delta} dP\left(\xi_1\right) dP\left(\xi_i\right)\right\}.$$

We conclude the appendix by proving that the trimming term $\widehat{\rho}_i^{\theta}$, defined in (2.5), is eventually equals to 1 for any sufficiently large n with probability 1.

Lemma 2.6.5 *Let (A1)-(A4) hold and*

$$I_{n,\theta}^i = \begin{cases} \mathbf{1}, & \textit{if } \min\left\{\widehat{f}_{Y,\theta^T X}^{-i}\left(y_i, \theta^T x_i\right), \widehat{f}_{\theta^T X}^{-i}\left(\theta^T x_i\right)\right\} > a_0 n^{-c}, \\ 0, & \textit{otherwise,} \end{cases}$$

for some small constants $a_0, c > 0$ such that $n^c\left(h_y^2 + h_x^2\right) = o\left(1\right)$ and $n^{1-2c-\delta} h_y h_x \to \infty$ for arbitrarily small $\delta > 0$. Then eventually for any sufficiently large n

$$\max_{1\leq i\leq n} \sup_{\theta\in\Theta} \left|I_{n,\theta}^i - 1\right| = 0$$

with probability 1.

Proof. Define

$$\mathbb{T}_\theta = \left\{(y, x) \in \mathbb{R}^{1+d} : \min\left\{f_{Y,\theta^T X}\left(y, \theta^T x\right), f_{\theta^T X}\left(\theta^T x\right)\right\} > 2a_0 n^{-c}\right\}.$$

It is trivial now to show that

$$\sup_{\theta\in\Theta} \left|I_{n,\theta}^i - 1\right| \leq \sup_{\theta\in\Theta} I_{\{(y_i,x_i)\notin\mathbb{T}_\theta\}} + I_{\{Z_n^i > a_0 n^{-c}\}},$$

where

$$Z_n^i = \sup_{\theta\in\Theta} \max\left\{ \left| \widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right) - f_{Y,\theta^T X}\left(y_i, \theta^T x_i\right) \right|, \left| \widehat{f}^{-i}_{\theta^T X}\left(\theta^T x_i\right) - f_{\theta^T X}\left(\theta^T x_i\right) \right| \right\}.$$

By definition of $\mathbb{S}$ there exists some large N such that for any $n \geq N$, we have that $\mathbb{S} \subseteq \bigcap\limits_{\theta\in\Theta} \mathbb{T}_\theta$, and as $(y_i, x_i) \in \mathbb{S}$, we get $\sup_{\theta\in\Theta} I_{\{(y_i,x_i)\notin\mathbb{T}_\theta\}} = 0$ for any $1 \leq i \leq n$. We now show that

$$P\left(\limsup_{n\to\infty} \left\{ \bigcup_{i=1}^{n} \left\{ Z_n > a_0 n^{-c} \right\} \right\} \right) = 0. \tag{2.60}$$

For sake of brevity, we prove here only that

$$\sum_{n=1}^{\infty} P\left(\bigcup_{i=1}^{n} \left\{ \sup_{\theta\in\Theta} \left| \widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right) - f_{Y,\theta^T X}\left(y_i, \theta^T x_i\right) \right| > a_0 n^{-c} \right\} \right) < \infty, \tag{2.61}$$

from which (2.60) follows by the Borel-Cantelli lemma (Gut 2005, Chapter 2, Theorem 18.1). The second term of Z_n^i can be handled in the same way.

For some $C_1, C_2 > 0$ independent of n, we have

$$\sup_{\theta\in\Theta} \left| \widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right) - \widehat{f}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right) \right| \leq \frac{C_1}{n h_y h_x},$$

and from the proof of Lemma 2.6.2,

$$\sup_{\theta\in\Theta, z\in\mathbb{S}} \left| E\widehat{f}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right) - f_{Y,\theta^T X}\left(y, \theta^T x\right) \right| \leq C_2\left(h_y^2 + h_x^2\right).$$

where $z = (y, x)$. The last two results imply that for n large enough,

$$\begin{aligned} &\sum_{n=1}^{\infty} P\left(\bigcup_{i=1}^{n} \left\{ \sup_{\theta\in\Theta} \left| \widehat{f}^{-i}_{Y,\theta^T X}\left(y_i, \theta^T x_i\right) - f_{Y,\theta^T X}\left(y_i, \theta^T x_i\right) \right| > a_0 n^{-c} \right\} \right) \\ \leq\ &\sum_{n=1}^{\infty} P\left(\sup_{z\in\mathbb{S}} \sup_{\theta\in\Theta} \left| \widehat{f}_{Y,\theta^T X}\left(y, \theta^T x\right) - E\widehat{f}_{Y,\theta^T X}\left(y, \theta^T x\right) \right| > a n^{-c} \right), \end{aligned} \tag{2.62}$$

for some $0 < a < a_0$. We can continue to bound the last term as in the proof of Lemma

2.6.3. Let $\{A_k\}_{k=1}^{J}$ form a cover of subspace $\Theta \times \mathbb{S}$, with $J(n) = O\left(h_y^{-1} h_x^{-1} n^{2c}\right)$, and

$$A_k = \left\{\theta, x, y : \left\|\theta - \overline{\theta}_k\right\| \leq \left(h_x n^{-c}\right)^{1/2}, \|x - \overline{x}_k\| \leq \left(h_x n^{-c}\right)^{1/2}, \|y - \overline{y}_k\| \leq h_x n^{-c}\right\},$$

Define for $\left(\overline{\theta}_k, \overline{y}_k, \overline{x}_k\right)$,

$$Z_i = K\left(\frac{\overline{\theta}_k^T\left(\overline{x}_k - x_i\right)}{h_x}\right) K\left(\frac{\overline{\theta}_k^T\left(\overline{y}_k - y_i\right)}{h_y}\right) - E\left(K\left(\frac{\overline{\theta}_k^T\left(\overline{x}_k - x_i\right)}{h_x}\right) K\left(\frac{\overline{\theta}_k^T\left(\overline{y}_k - y_i\right)}{h_y}\right)\right).$$

Now, notice that $|Z_i| \leq b \equiv 2\sup_{\Theta \times \mathbb{S}_X} |K(u)| < \infty$, and by Theorem 1 of Hansen (2008), for any $1 \leq m \leq n$,

$$\sigma^2(m) \equiv \sup_{(\theta,x)} E\left|\sum_{i=1}^{\lfloor m \rfloor} Z_i\right|^2 \leq Cmh_y h_x$$

for some large enough $C > 0$. Set $m = Cn^{1-2c}h_y h_x / a_1$ and $\varepsilon = a_1 n^{1-c} h_y h_x$, and note that $4bm < \varepsilon$ for any sufficiently large n. By Theorem 2.1 of Liebscher (1996) and (A1),

$$\begin{aligned}
& P\left(\left|\widehat{f}_{Y,\theta^T X}\left(\overline{y}_k, \overline{\theta}_k^T \overline{x}_k\right) - E\widehat{f}_{Y,\theta^T X}\left(\overline{y}_k, \overline{\theta}_k^T \overline{x}_k\right)\right| > a_1 n^{-c}\right) \\
= \ & P\left(Z_i > \varepsilon\right) \\
\leq \ & 4\exp\left(-\frac{\varepsilon^2}{64\frac{n}{m}\sigma^2(m) + \frac{8}{3}\varepsilon m b}\right) + 4\frac{n}{m}\alpha_m \\
\leq \ & 4\exp\left(-\frac{a_1^2 n^{2-2c} h_y^2 h_x^2}{64Cn^1 h_y h_x + \frac{8}{3}Cn^{2-3c} h_y^2 h_x^2 b}\right) \\
& +4A\alpha^{Cn^{1-2c}h_y h_x / a_1} h_y^{-1} h_x^{-1} n^{2c} \\
\leq \ & 4\exp\left(-\frac{a_1^2 n^c}{C(64 + 3b)}\right) + 4AJ(n)\,\alpha^{C/a_1 n^{\delta}},
\end{aligned}$$

where $0 < \alpha < 1$ and $J(n) = h_y^{-1} h_x^{-1} n^{2c}$. Thus, we have

$$\sum_{n=1}^{\infty} J(n)\left(\sup_{z \in \mathbb{S}} \sup_{\theta \in \Theta}\left|\widehat{f}_{Y,\theta^T X}\left(y, \theta^T x\right) - E\widehat{f}_{Y,\theta^T X}\left(y, \theta^T x\right)\right| > a_1 n^{-c}\right) < \infty, \qquad (2.63)$$

and (2.61) is established with (2.62), (2.63), and the same arguments as in the proof of Lemma 2.6.3. ■

Chapter 3

Projection Pursuit Conditional Density Estimation

3.1 Introduction

Consider the c.p.d.f. $f_{Y|X}(y|x)$ of a random scalar Y given a random d-vector $X = x$. Even for a small dimension of X, $d \geq 2$, a purely nonparametric approach may suffer from poor performance due to the 'curse of dimensionality' (see Silverman 1986, Section 4.5). In order to overcome this, a vast number of techniques have emerged in the literature for reducing the dimensionality of the problem, without losing too much of the main characteristics of the data. In this chapter, we suggest a projection pursuit approximation of the c.p.d.f. attained by a series of univariate projections of the X-data into a finite number of univariate directions. More precisely, we propose a multiplicative PP approximation of the conditional density that has the form $f(y|x) = f_0(y) \prod_{m=1}^{M} h_m\left(y, \theta_m^T x\right)$, where the projection directions θ_m and the multiplicative elements, h_m, $m = 1, ..., M$, are chosen to minimise a weighted version of the Kullback-Leibler relative entropy between the true and the estimated conditional densities. In particular, the single-index approximation of Chapter 2 can be seen as a private case of the PP approximation where $M = 1$.

The idea of projection pursuit goes back to Kruskal (1969, 1972), Switzer (1970), Switzer and Wright (1971). It was only a few of years later that Friedman and Tukey (1974) successfully implemented the method. Their work led the way to multiple applications, such as projection pursuit classification (Friedman and Stuetzle 1980), projection pursuit regression (Friedman and Stuetzle 1981), and projection pursuit density estimation (henceforth PPDE, by Friedman, Stuetzle and Schroeder 1984). A comprehensive review of the projection pursuit methodology can be found in Huber (1985). The asymptotic properties of the projection pursuit regression method were developed by Hall (1989) for independent data, and by Xia and An (1999) for dependent data, while as far as we are aware, only Touboul (2011) has derived asymptotic properties for the PPDE, although Huber (1985) has already discussed consistency. To the best of our knowledge, no similar projection pursuit approximation has been suggested in the literature to c.p.d.f. estimation.

The main goal of Chapter 3 of the thesis is to develop the projection pursuit methodology and the corresponding asymptotic theory for c.p.d.f. estimation. The method, which we call projection pursuit conditional density estimation (PPCDE), is developed throughout the chapter and both its theoretical and empirical properties are presented.

The PPCDE method also has some links with the work on multiplicative nonparametric correction to an initial density estimator, and the PPCDE can also be applied to achieve a similar goal (see Hjort and Glad 1995, Jones, Linton and Nielsen 1995, Glad, Hjort and Ushakrov 2003 and Buch-Kromann et al 2006). However, the PPCDE differs from the aforementioned approaches in two main aspects. Firstly, it is designed to work in high-dimensional spaces of r.v. X, and as such the method looks for corrections only along some optimal univariate projections of data X. Secondly, the PPCDE is able to continue correcting itself iteratively until a certain optimality criterion is met.

The outline for the rest of the chapter is as follows. Section 3.2 introduces the theoretical approximation, while Section 3.3 states some desired properties to motivate it. In Section 3.4 we move to discuss estimation and the PPCDE algorithm is described. Section 3.5 states asymptotic results for the PPCDE under strong-mixing conditions. Section 3.6 suggests a bootstrap Information Criterion to terminate the estimation algorithm. Section 3.7 illustrates the method using both simulated data and exchange-rate series, and Section 4 concludes. All proofs of the chapter are collected in Section 3.8.

3.2 The Projection Pursuit Approximation

Let Y be a random scalar and X be a random d-vector. Denote the support of Y by $\mathbb{S}_Y \subseteq \mathbb{R}$ and that of X by $\mathbb{S}_X \subseteq \mathbb{R}^d$, and make $\mathbb{S} = \mathbb{S}_Y \times \mathbb{S}_X$. Throughout Chapter 3, we index the probability densities, which are assumed to exist with respect to the underlying Lebesgue measure, with corresponding subscripts to the r.v. they represent, so for example, $f_{Y,X}, f_{Y,\theta^T X}, f_{Y|X}$ and $f_{Y|\theta^T X}$ denote the probability densities of, respectively, (Y,X), $\left(Y, \theta^T X\right)$, $Y|X$ and $Y|\theta^T X$ for some $\theta \in \mathbb{R}^d$, etc.

The common ideology of all of the projection pursuit methods is to approximate multivariate functions by a sequence of univariate functions of linear combinations of the variable. We propose to approximate the c.p.d.f. $f_{Y|X}(y|x)$ by the form

$$f_{Y|X}(y|x) \approx g_{Y|X,M}(y|x) \equiv g_{Y,0}(y)\, h_1\left(y, \theta_1^T x\right) \cdots h_M\left(y, \theta_M^T x\right), \tag{3.1}$$

where M is some positive integer, $h_1, ..., h_{M,}$ are unknown bivariate functions, and $\theta_1, ..., \theta_M$ are unit d-vectors that belong to a parameter space Θ, and which are called the projection pursuit directions. $g_{Y,0}(y)$ is an initial approximation of the unconditional density of Y, and it can be taken as any naive approximation that is positive on $\mathbb{S}_Y$, i.e. a normal density or a histogram. In principle, one can also take an initial approximation that depends on data x to reflect some a prior beliefs about the conditional density.

It is first essential to gain some understanding of the strengths and limitations of the product form approximation. The Projection Pursuit approximation is much more flexible than the single-index model. In particular, it has been shown by Diaconis and Shahshahani (1984) that any smooth function can be approximated to arbitrary precision by a function of the form (3.1). Nonetheless, the projection pursuit representation need not be unique. For example, when $d = M = 2$ note that $g_{Y|X,M}(y|x_1, x_2) = g_{Y|X,M}(y|x_1 \cdot x_2)$ has infinitely many equivalent projection pursuit representations since

$$x_1 x_2 = (1/4ab)[(ax_1 + bx_2)^2 - (ax_1 - bx_2)^2] \tag{3.2}$$

for any real numbers a_1, a_2, b_1, b_2. Diaconis and Shahshahani (1984) provided a necessary condition for non-uniqueness of the product representation (3.1). Therefore, the flexibility of the approximation comes at the cost of interpretability since (3.1) is not necessarily identifiable if M, the θ_m's and the h_m's are left unrestricted. A paper by Yuan (2010) provides an interesting discussion of general conditions under which the closely related additive index model,

$$E\left(Y|X\right) \approx \mu + \textstyle\sum_{m=1}^{M} h_m\left(\theta_m^T x\right),$$

is identifiable. However, as for the PPCDE, it is still an open question whether there are identifying restrictions that yield useful forms of (3.1).

An even more acute issue is that there are well-behaved (say, smooth) functions that cannot be written in the product form (3.1) for any finite M (see Diaconis and Shahshahani 1984). Furthermore, in the nontrivial cases, where $M \geq 2$ and $\theta_1, ..., \theta_M \in \mathbb{R}^d$ are linearly independent, it is not even clear to us whether there are any real c.p.d.f. that follow the form (3.1) without requiring an additional normalisation factor, which is a general function of $x \in \mathbb{R}^d$. This is left as an open question for further research, and it is discussed again in Chapter 4 of the thesis. By the end of this section, we will also allow the inclusion of a normalising factor to (3.1). As a simple example for a real c.p.d.f. that follows the normalised form with $M = 2$, we can consider any parametric family of p.d.f.'s with two parameters. For example, take the Beta distribution with parameters $\alpha = \alpha\left(\theta_1^T x\right)$ and $\beta = \beta\left(\theta_2^T x\right)$. Then

$$f_{Y|X}\left(y|x\right) \approx y^{-\alpha\left(\theta_1^T x\right)}\left(1-y\right)^{-\beta\left(\theta_2^T x\right)} \Big/ Beta\left(\alpha\left(\theta_1^T x\right), \beta\left(\theta_2^T x\right)\right),$$

follows a product form with $g_{Y,0}\left(y\right) = 1$, $h_1\left(y, \theta_1^T x\right) = y^{-\alpha\left(\theta_1^T x\right)}$, $h_2\left(y, \theta_1^T x\right) = \left(1-y\right)^{-\beta\left(\theta_2^T x\right)}$ and the Beta function $Beta\left(\alpha\left(\theta_1^T x\right), \beta\left(\theta_2^T x\right)\right)$ is an additional normalising factor.

It may seem tempting to tackle the proposed approximation with one of the existing projection pursuit techniques. For instance, consider applying the projection pursuit regression approach of Friedman and Stuetzle (1981) to the log-density,

$$\log f_{Y|X}(y|x) \approx \log g_{Y,0}(y) + \textstyle\sum_{m=1}^{M} \log h_m\left(y, \theta_m^T x\right).$$

The application of regression techniques to conditional density estimation is possible due to the double-kernel approach of Fan, Yao and Tong (1996). In the double-kernel approach, however, the dependent variable is taken as $f_{Y|X}(y|x)$ rather than $\log f_{Y|X}(y|x)$, while by reducing the original problem of density estimation to a regression problem, it becomes hard to restrict the estimator to be non-negative and to integrate to 1 (see Hyndman and Yao 2002). Obviously, one can choose to approximate directly the density function by the form

$$f_{Y|X}(y|x) \approx \exp\left(\log g_{Y,0}(y) + \textstyle\sum_{m=1}^{M} \log h_m\left(y, \theta_m^T x\right)\right).$$

This approximation was implemented by Hyndman and Yao (2002) using a local parametric regression model. However, it is not clear whether the projection pursuit regression can be applied to this model, and for our purposes, it does not seem to offer any advantage.

An alternative approximation that maintains the multiplicative nature of the approximation is the PPDE of Friedman, Stuetzle and Schroeder (1984). Consider now an application of this approximation to the joint p.d.f. of Y and X, i.e. approximate

$$f_{Y,X}(y,x) \approx \textstyle\prod_{m=1}^{M} f_m\left(y, \theta_m^T x\right),$$

and then take

$$f_{Y|X}(y|x) \approx \frac{\prod_{m=1}^{M} f_m\left(y, \theta_m^T x\right)}{\int_{y\in\mathbb{R}} \left\{\prod_{m=1}^{M} f_m\left(y, \theta_m^T x\right)\right\} dy}.$$

In this formulation, however, the projections $\theta_1^T x, ..., \theta_M^T x$ are intended to approximate the p.d.f. $f_{Y,X}(y,x)$ efficiently, but they do not necessarily provide effective information with regards to Y. When one is interested in inference about Y given data X, e.g. for making predictions, a dedicated approximation for the c.p.d.f. $f_{Y|X}(y|x)$ is much more preferable.

The projection directions $\theta_1, ..., \theta_M$ and the corresponding functions $h_1, ..., h_M$ pursued are ought to be chosen such that $g_{Y|X,M}(y|x)$ achieves a reasonably good approximation of $f_{Y|X}(y|x)$. At the same time, it is intended to keep the number of projective directions to a minimum so as not to make the PPCDE unwieldy for approximation and computation. To this end, we next define a suitable optimality criterion, by which a general approximation of the c.p.d.f. may be examined.

For the sake of generality, we discard in the rest of the section the subscript M, and for any $x \in \mathbb{S}_X$ let $g_{Y|X}(y|x)$ denote a general measurable and non-negative approximation of the c.p.d.f. of Y given $X = x$ such that

$$g_{Y|X}(y|x) \geq 0 \text{ a.e. for any } (y, x) \in \mathbb{R} \times \mathbb{S}_X.$$

Preferably, of-course, $g_{Y|X}(y|x)$ is itself a c.p.d.f., i.e.,

$$\int g_{Y|X}(y|x)\, dy = 1 \text{ a.e. for any } x \in \mathbb{S}_X. \tag{3.3}$$

A common divergence measure of the difference between $g(y|x)$ and the real c.p.d.f. $f_{Y|X}(y|x)$ is the Kullback-Leibler's relative entropy,

$$D[g_{Y|X}] = \int \log\left(\frac{f_{Y|X}(y|x)}{g_{Y|X}(y|x)}\right) f_{Y,X}(y, x)\, dydx.$$

The Kullback-Leibler's relative entropy has some known desired properties for estimation of probability densities (cf. Huber 1985, section 12), and it has been successfully utilised in several papers for estimation of c.p.d.f. (e.g., Yin and Cook 2005 and Fan et al 2009).

The integrability condition (3.3) is distinctly hard to impose when one is interested in estimating global parameters $\theta_1, ..., \theta_M$. However, if this condition is relaxed, then the relative entropy measure should not be an adequate measure anymore. For example, $D[g_{Y|X}]$ can always get smaller by replacing $g_{Y|X}(y|x)$ by $cg_{Y|X}(y|x)$, $c > 1$, and it tends to $-\infty$ in the limit $c \to \infty$. We therefore define the constrained relative entropy between

$g_{Y|X}$ and $f_{Y|X}$ as

$$D_C[g_{Y|X}] = \int \log\left(\frac{f_{Y|X}\left(y|x\right)}{g_{Y|X}\left(y|x\right)}\right) f_{Y,X}\left(y,x\right) dydx + \left(\int g_{Y|X}\left(y|x\right) f_X\left(x\right) dydx - 1\right). \quad (3.4)$$

Here and in the rest of the chapter we understand

$$\log\left(0\right) = -\infty, \quad \frac{a}{0} = \infty, \quad 0\cdot\left(\pm\infty\right) = 0/0 = 0,$$

for any $a > 0$. The first term in the RHS of (3.4) is the standard relative entropy $D[g_{Y|X}]$, while we show below that the second term in the RHS represents the integrability constraint

$$\int g_{Y|X}\left(y|x\right) f_X\left(x\right) dydx = 1. \quad (3.5)$$

Although this condition is clearly weaker than the desired condition (3.3), it offers a practical and unified measure that can be applied independently of $x \in \mathbb{S}_X$. Another useful interpretation of $D_C[g_{Y|X}]$ is as a standard constrained relative entropy between the approximation of the p.d.f. of (Y,X), $g_{Y,X}\left(y,x\right)$, and the true p.d.f., $f_{Y,X}\left(y,x\right)$, where both are obtained, respectively, by $f_X\left(x\right)$-weighting of the approximated and the true c.p.d.f.,

$$\begin{aligned} g_{Y,X}\left(y,x\right) &= g_{Y|X}\left(y|x\right) f_X\left(x\right), \\ f_{Y,X}\left(y,x\right) &= f_{Y|X}\left(y|x\right) f_X\left(x\right). \end{aligned}$$

Indeed, this viewpoint is expressed by the equality

$$D_C[g_{Y|X}] = \int \log\left(\frac{f_{Y,X}\left(y,x\right)}{g_{Y,X}\left(y,x\right)}\right) f_{Y,X}\left(y,x\right) dydx + \int g_{Y,X}\left(y,x\right) dydx - 1.$$

By this last representation, it is clear that $g_{Y|X}\left(y|x\right)$ indeed seeks to approximate $f_{Y|X}\left(y|x\right)$ in the sense that $g_{Y|X}\left(y|x\right) f_X\left(x\right)$ approximates $f_{Y|X}\left(y|x\right) f_X\left(x\right)$.

This constrained version of the relative entropy for unconditional p.d.f.'s has already been used by Friedman, Stuetzle and Schroeder (1984) in their PPDE, but it appears more explicitly in a likelihood form in the papers of Loader (1996) and of Cule, Samworth and Stewart (2010) for nonparametric p.d.f. estimation. It is therefore not surprising that $D_C[g_{Y|X}]$ enjoys similar properties to those of the standard relative entropy between p.d.f.'s (cf. section 12 of Huber 1985).

In the rest of the chapter we therefore refer to an approximation $g^{opt}_{Y|X}(y|x)$ of the c.p.d.f. $f_{Y|X}(y|x)$ as an optimal approximation if $D_C[g_{Y|X}]$ is minimised by $g^{opt}_{Y|X}(y|x)$ in the relevant function space.

In many cases, it may still be useful to directly impose the integrability condition (3.3) on approximation $g_{Y|X}(y|x)$. This can be done in a straightforward way. Let $g^{opt}_{Y|X}(y|x)$ be an optimal approximation of the c.p.d.f. $f_{Y|X}(y|x)$. Now, instead of $g^{opt}_{Y|X}(y|x)$, consider the normalised form

$$g^{opt}_{Y|X}(y|x) \Big/ \int g^{opt}_{Y|X}(y|x)\, dy \tag{3.6}$$

as the final approximation. However, it bears a significant additional computational cost. While directions $\theta_1, ..., \theta_M$ are global parameters, and functions $h_1, ..., h_M$ can be calculated very quickly, the calculation of the factor$\int g^{opt}_{Y|X}(y|x)\, dy$ may require some heavy computational effort, particularly when one needs to compute the c.p.d.f. instantaneously for numerous values of x. Hence, the normalisation (3.6) should in practice only be applied to problems where the computational burden is not heavy.

The following proposition proves some of the properties satisfied by $D_C[g_{Y|X}]$ that are similar to the properties proved by Huber (1985) for the standard relative entropy. The first part proves that, similarly to the standard relative entropy, the measure is a 'pre-metric' between $g_{Y|X}$ and $f_{Y|X}$, in the sense that $D_C[g_{Y|X}] \geq 0$ with equality iff $g = f_{Y|X}$ a.e.. The second part asserts that normalising any non-normalised approximation $g(y|x)$ to make the integrability condition (3.5) hold will always decrease $D_C[g_{Y|X}]$. In particular, an approximation $g_{Y|X}$ that minimises $D_C[g_{Y|X}]$ must satisfy that integrability constraint. The third part proves that both the L_1 and the Hellinger distance metrics between $g_{Y|X}$ and $f_{Y|X}$ are dominated by $D_C[g_{Y|X}]^{1/2}$. Finally, the fourth part of the proposition shows that the normalisation procedure can only further improve the quality of the estimator in terms of minimising $D_C[g_{Y|X}]$.

Proposition 3.2.1 *Let $f_{Y|X}$ and f_X be the true densities of $Y|X$ and of X respectively, and let $g_{Y|X}(y|x)$ be a non-negative function and integrable w.r.t. $y \in \mathbb{R}$. Then:*

a. $D_C[g_{Y|X}] \geq 0$ with equality iff $g = f_{Y|X}$ a.e. for any $(y,x) \in \mathbb{R} \times \mathbb{S}_X$.

b. $D_C[g_{Y|X}] \geq D_C[g^*_{Y|X}]$ for $g^*_{Y|X}(y|x) = g_{Y|X}(y|x) / \int g_{Y|X}(y|x) f_X(x) dydx$ with equality iff $\int g_{Y|X}(y|x) f_X(x) dydx = 1$.

c. Let $c = \int g_{Y|X}(y|x) f_X(x) dydx$. Then

$$\begin{aligned} & \int \left(\sqrt{f_{Y|X}(y|x)} - \sqrt{g_{Y|X}(y|x)}\right)^2 f_X(x) dydx \\ \leq & \int \left|f_{Y|X}(y|x) - g_{Y|X}(y|x)\right| f_X(x) dydx \\ \leq & \left\{\frac{2}{3}c(1+2c)D_C[g_{Y|X}]\right\}^{1/2}. \end{aligned}$$

d. If $D_C[g_{Y|X}]$ is minimised by $g^{opt}_{Y|X}(y|x)$. Then $D_C[g^{opt}_{Y|X}(y|x)] \geq D_C[g^{*opt}_{Y|X}(y|x)]$ for $g^{*opt}_{Y|X}(y|x) = g^{opt}_{Y|X}(y|x) / \int g^{opt}_{Y|X}(y|x) dy$ for any $x \in \mathbb{S}_X$ with equality iff $\int g^{opt}_{Y|X}(y|x) dy =$ 1 a.e. for any $x \in \mathbb{S}_X$.

3.3 Properties of the Optimal Projections

Approximation (3.1) can be seen as a sequence of modifications to the naive $g_{Y,0}(y)$ such that each modification depends on one linear combination of the coordinates of X. This suggests a recursive stepwise construction of the estimator by

$$g_{Y|X,m}(y|x) = g_{Y|X,m-1}(y|x)\, h_m\left(y, \theta_m^T x\right), \quad m = 1, ..., M. \tag{3.7}$$

Thus, at any iteration $m = 1, ..., M$, given a current model $g_{Y|X,m-1}(y|x)$, we seek an optimal new projection and a corresponding optimal modification function, denoted respectively by $\theta_{0,m}$ and $h_{0,m}\left(y, \theta_{0,m}^T X\right)$, such that model $g_{Y|X,m}(y|x)$ provides an improved approximation to $f_{Y|X}\left(y|x\right)$ in the sense of minimising the constrained relative entropy $D_C[g_{Y|X,m}\left(y|x\right)]$,

$$\left(\theta_{0,m}, h_{0,m}\left(y, \theta_{0,m}^T X\right)\right) = \arg \min_{\theta_m \in \Theta, h_m\left(y, \theta_m^T x\right)} D_C[g_{Y|X,m-1}\left(y|x\right) h_m\left(y, \theta_m^T x\right)].$$

For now, let us assume first that a projection direction θ_m is given, and that the problem is reduced to finding only the corresponding optimal $h_{0,m}\left(y, \theta_m^T x\right)$,

$$h_{0,m}\left(y, \theta_m^T X\right) = \arg \min_{h_m\left(y, \theta_m^T x\right)} D_C[g_{Y|X,m-1}\left(y|x\right) h_m\left(y, \theta_m^T x\right)].$$

Without loss of generality, let direction θ_m be the first coordinate axis, that is, $x_1 = \theta_m^T x$. The density functions of $\theta_m^T X$ and $\left(Y, \theta_m^T X\right)$ are given by

$$\begin{aligned} f_{\theta_m^T X}\left(\theta_m^T x\right) &= \int f_X\left(x\right) dx_2 \cdots dx_d, \\ f_{Y, \theta_m^T X}\left(y, \theta_m^T x\right) &= \int f_{Y|X}\left(y|x\right) f_X\left(x\right) dx_2 \cdots dx_d, \end{aligned} \tag{3.8}$$

and the c.p.d.f. of Y given $\theta_m^T X$ is

$$f_{Y|\theta_m^T X}\left(y|\theta_m^T x\right) = f_{Y, \theta_m^T X}\left(y, \theta_m^T x\right) / f_{\theta_m^T X}\left(\theta_m^T x\right).$$

At the m'th step of the procedure, we can define analogously the $g_{Y|X,m-1}\left(y|x\right)$-based estimators of the density of $\left(Y, \theta_m^T X\right)$ and of the c.p.d.f. of Y given $\theta_m^T X$, respectively, as

$$g_{Y, \theta_m^T X, m-1}\left(y, \theta_m^T x\right) = \int g_{Y|X,m-1}\left(y|x\right) f_X\left(x\right) dx_2 \cdots dx_d, \tag{3.9}$$

and

$$g_{Y|\theta_m^T X, m-1}\left(y|\theta_m^T x\right) = g_{Y, \theta_m^T X, m-1}\left(y, \theta_m^T x\right) / f_{\theta_m^T X}\left(\theta_m^T x\right).$$

The next Proposition states that an explicit solution for the optimal $h_m\left(y, \theta_m^T x\right)$, given the current model $g_{Y|X,m-1}\left(y|x\right)$ and a new direction θ_m (and the real density $f_{Y,X}$), is obtained uniquely by the following expression,

$$h_{0,m}\left(y,\theta_m^T x\right)=\frac{f_{Y,\theta_m^T X}\left(y,\theta_m^T x\right)}{g_{Y,\theta_m^T X,m-1}\left(y,\theta_m^T x\right)}=\frac{f_{Y|\theta_m^T X}\left(y|\theta_m^T x\right)}{g_{Y|\theta_m^T X,m-1}\left(y|\theta_m^T x\right)}, \tag{3.10}$$

so that the m'th step optimal approximation is given by

$$g_{Y|X,m}^{opt}\left(y|x\right)=g_{Y|X,m-1}\left(y|x\right)h_{0,m}\left(y,\theta_m^T x\right).$$

The Proposition further proves some useful expressions for the optimal modification and for the marginal decrease in the constrained relative entropy,

$$D^*[g_{Y|\theta_m^T X,m-1}]\equiv D_C[g_{Y|X,m-1}]-D_C[g_{Y|X,m}^{opt}\left(y|x\right)]. \tag{3.11}$$

Sections (a) and (c) of the Proposition are generalisations of the arguments used by Friedman, Stuetzle and Schroeder (1984) and Huber (1985, section 13) in their PPDE, while Sections (b) provides an identity that is unique to the PPCDE approximation.

Proposition 3.3.1 *Let $g_{Y|X,m-1}\left(y|x\right)$ be a non-negative approximation of the conditional density of Y given X, and consider a new c.p.d.f. approximation $g_{Y|X,m}\left(y|x\right)$ of the form $g_{Y|X,m}\left(y|x\right)=g_{Y|X,m-1}\left(y|x\right)h_m\left(y,\theta_m^T x\right)$, where θ_m is a given direction in $\mathbb{R}^d$, and $h_m\left(\cdot,\cdot\right)$ non-negative bivariate function. Then:*

a. The new approximation $g_{Y|X,m}\left(y|x\right)$ is optimal iff

$$h_m\left(y,\theta_m^T x\right)=h_{0,m}\left(y,\theta_m^T x\right)\ \ a.e.$$

for any $(y,x)\in\mathbb{R}\times\mathbb{S}_X$ such that $g_{Y|X,m-1}\left(y|x\right)>0$.

b. $h_{0,m}\left(y,\theta_m^T x\right)$ satisfies the following equality,

$$h_{0,m}\left(y,\theta_m^T x\right)=\frac{f_{Y,\theta_m^T X}\left(y,\theta_m^T x\right)}{E\left[g_{Y|X,m-1}\left(y|X\right)|\theta_m^T X=\theta_m^T x\right]f_{\theta_m^T X}\left(\theta_m^T x\right)}.$$

c. If $g_{Y|X,m-1}(y|x)$ fulfils $\int g_{Y|X,m-1}(y|x)\, f_X(x)\, dydx = 1$, then the marginal decrease in the constrained relative entropy is equal to the relative entropy between the c.p.d.f.'s $f_{Y|\theta_m^T X}$ and $g_{Y|\theta_m^T X,m}$. That is,

$$\begin{aligned} D^*[g_{Y|\theta_m^T X,m-1}] &= \int \log\left(h_{0,m}\left(y,\theta_m^T x\right)\right) f_{Y,\theta_m^T X}(y,x_1)\, dydx_1 \\ &= \int \log\left(\frac{f_{Y|\theta_m^T X}(y|x_1)}{g_{Y|\theta_m^T X,m-1}(y|x_1)}\right) f_{Y,\theta_m^T X}(y,x_1)\, dydx_1. \end{aligned}$$

According to the last section of Proposition 3.3.1, $D^*[g_{Y|\theta_m^T X,m-1}]$ is the relative entropy, and therefore it is necessarily non-negative. Thus, the constrained relative entropy $D_C[g_{Y|X,m}]$ between $g_{Y|X,m}(y|x)$ and the real c.p.d.f. $f_{Y|X}(y|x)$ is non-increasing with the number of iterations m for any choice of projections θ_m's.

While the last Proposition provides the optimal m'th modification $h_{0,m}\left(y,\theta_m^T x\right)$ for a given direction θ_m, it remains now to find the optimal m'th direction $\theta_{0,m}$. Clearly, by that Proposition, we may now simply replace $g_{Y|X,m}(y|x)$ by $g^{opt}_{Y|X,m}(y|x)$ and obtain $\theta_{0,m}$ as the minimiser of $D_C[g^{opt}_{Y|X,m}(y|x)]$. Moreover, since by definition $g^{opt}_{Y|X,m}(y|x)$ depends on θ_m^T only through the m'th modification function $h_m\left(y,\theta_m^T x\right)$, then $\theta_{0,m}$ can be equivalently characterised as the maximiser of the marginal decrease in the relative entropy, $D^*[g_{Y|\theta_m^T X,m-1}]$.

For ease of presentation, we assume here and below that for each $m = 1, 2, ...,$ $\theta_{0,m} \in \Theta$ is the unique maximiser of $D^*[g_{Y|\theta_m^T X,m-1}]$. In practice, $\theta_{0,m}$ can be any one in the set of solutions to

$$D^*[g_{Y|\theta_{0,m}^T X,m-1}] = \max_{\theta_m \in \Theta}\left\{D^*[g_{Y|\theta_m^T X,m-1}]\right\}.$$

In particular, our asymptotic results, presented in Section 3.5, will still apply as long as this $\theta_{0,m}$ is a local maximiser of $D^*[g_{Y|\theta_m^T X,m-1}]$ in a small neighbourhood. Note that in that case, the choice of $\theta_{0,m}$ within the set of optimum points is not crucial as long as it is finding a modification for the previous approximation $g_{Y|\theta_m^T X,m-1}(y|x_1)$ that we are concerned about. Moreover, the stepwise nature of the approximation allows the procedure to recover other informative projective directions as the procedure keeps on to the next steps.

In summary, the Projection Pursuit approximation procedure is constructed through a recursive formula

$$g_{Y|X,m}\left(y|x\right) = g_{Y|X,m-1}\left(y|x\right) h_{0,m}\left(y,\theta_{0,m}^{T}X\right), \quad m = 1,2,..., \tag{3.12}$$

where $g_{Y|X,0}\left(y|x\right) = g_{Y,0}\left(y\right)$ is a non-negative, and strictly positive on $\mathbb{S}$, initial approximation of the unconditional density of Y,

$$\begin{aligned} h_{0,m}\left(y,\theta_{m}^{T}x\right) &= \frac{f_{Y,\theta_{m}^{T}X}\left(y,\theta_{m}^{T}x\right)}{g_{Y,\theta_{m}^{T}X,m-1}\left(y,\theta_{m}^{T}x\right)} \\ &= \frac{f_{Y,\theta_{m}^{T}X}\left(y,\theta_{m}^{T}x\right)}{E\left[g_{Y|X,m-1}\left(y|X\right)|\theta_{m}^{T}X=\theta_{m}^{T}x\right] f_{\theta_{m}^{T}X}\left(\theta_{m}^{T}x\right)}, \end{aligned} \tag{3.13}$$

and

$$\begin{aligned} \theta_{0,m} &= \arg\max_{\theta_{m}\in\Theta}\left\{D^{*}[g_{Y|\theta_{m}^{T}X,m-1}\right\} \\ &= \arg\max_{\theta_{m}\in\Theta}\left\{\int \log\left(h_{0,m}\left(y,\theta_{m}^{T}x\right)\right) f_{Y,\theta_{m}^{T}X}\left(y,x_{1}\right)dydx_{1}\right\} \\ &= \arg\max_{\theta_{m}\in\Theta}\left\{E\log\left(h_{0,m}\left(y,\theta_{m}^{T}x\right)\right)\right\}. \end{aligned} \tag{3.14}$$

In the proposition stated below we show that $D^{*}[g_{Y|\theta_{0,m}^{T}X,m-1}] \to 0$ as $m \to \infty$. Note that by Propositions 3.2.1(c) and 3.3.1(c), $D^{*}[g_{Y|\theta_{0,m}^{T}X,m-1}]$ dominates the weighted L_1-norm between $f_{Y|\theta_{0,m}^{T}X}$ and $g_{Y|\theta_{0,m}^{T}X,m-1}$. Therefore, since $\theta_{0,m}$ is selected to maximise $D^{*}[g_{Y|\theta_{m}^{T}X,m-1}]$, the asymptotic decay of $D^{*}[g_{Y|\theta_{0,m}^{T}X,m-1}]$ to zero guarantees that the projective (marginal) conditional density approximation $g_{Y|\theta^{T}X}$ converges to $f_{Y|\theta^{T}X}$ in the weak sense for any choice of $\theta \in \Theta$, since

$$\sup_{\theta_{m}\in\Theta}\int\left|f_{Y|\theta_{m}^{T}X}\left(y|\theta_{m}^{T}x\right) - g_{Y|\theta_{m}^{T}X,m-1}\left(y|\theta_{m}^{T}x\right)\right| f_{X}\left(x\right)dydx \to 0 \text{ as } m \to \infty.$$

We further prove that this must imply a convergence of $g_{Y|X,m}\left(y|x\right)$ to $f_{Y|X}\left(y|x\right)$ in the weak sense as $m \to \infty$. For the PPDE, a similar implication was obtained using the Cramér-Wold device (Huber 1985). To prove the weak convergence result in our case, however, we need a conditional version of the Cramér-Wold device where the projections are taken in the space of the r.v. X given in the condition. The next lemma, proved in the appendix, establishes that result. This Lemma states sufficient and necessary conditions for converges in distribution of a random variable Y given random vector X, which hold almost surely for any X. As far as we know, no similar result exists in the literature.

Lemma 3.3.1 (Conditional Cramér-Wold Device) *Let $Y_1, Y_2...$ be a sequence of scalar r.v., Y is a scalar r.v., and X a r.v. in $\mathbb{R}^d$. Then $Y_m|X \to_d Y|X$ a.s. with respect to the probability measure induced by r.v. X as $m \to \infty$ iff for any $\theta \in \mathbb{R}^d$ $Y_m|\theta^T X \to_d Y|\theta^T X$ a.s. with respect to the probability measure induced by r.v. $\theta^T X$ as $m \to \infty$.*

With the Conditional Cramér-Wold Device, we can now obtain the next proposition, which is a generalisation of Proposition 14.2 of Huber (1985).

Proposition 3.3.2 *Let the projection pursuit approximation be defined recursively by (3.12)-(3.14). Then as $m \to \infty$:*

(a) $D^*[g_{Y|\theta_{0,m}^T X,m-1}] \to 0$.

(b) $g_{Y|X,m}\left(y|x\right) \to f_{Y|X}\left(y|x\right)$ in the weak sense.

Note that although $\theta_{0,m}$ is selected to maximise $D^*[g_{Y|\theta_m^T X,m-1}]$ at the m'th step, the decay of $D^*[g_{Y|\theta_{0,m}^T X,m-1}]$ to zero is not necessarily monotonic. In particular, it may be that for some $\theta_1, \theta_2 \in \Theta$

$$D^*[g_{Y|\theta_1^T X,m-1}] > D^*[g_{Y|\theta_2^T X,m-1}],$$

and hence direction θ_1 will be preferable to direction θ_2 in the sense of maximising the $(m-1)$'th marginal decrease in the constrained relative entropy, $D^*[g_{Y|\theta_m^T X,m-1}]$. However, after modifying $g_{Y|X,m-1}$along direction θ_1, a new model $g_{Y|X,m}$ is obtained, for which it is possible that

$$D^*[g_{Y|\theta_2^T X,m}] > D^*[g_{Y|\theta_1^T X,m-1}].$$

Thus, the m'th marginal decrease in the constrained relative entropy, $D^*[g_{Y|\theta_2^T X,m}]$, along direction θ_2 can be larger than the $(m-1)$'th marginal decrease in the constrained relative entropy, $D^*[g_{Y|\theta_1^T X,m-1}]$, along direction θ_1.

Example 1: For illustration of the proposed approximation procedure, consider the following simple example. Let $X = (X_1, X_2)$ where X_1 and X_2 are independent $N(0,1)$ random variables, and

$$Y|X \sim N\left(\sqrt{X_1^2 + X_2^2}, 1\right).$$

Since here $X \in \mathbb{R}^2$, the information entailed by X can be fully specified by any two orthogonal projections, say $\theta^T X$ and $\theta_\perp^T X$, of X. However, $f_{Y|X}\left(y|\theta^T x\right)$ cannot be written as a product $h_1\left(y, \theta_1^T x\right) h_2\left(y, \theta_1^T x\right)$. This fact follows from the argument of Diaconis and Shahshahani (1984, p.176), who established a necessary condition for a representation of a nonlinear function as a sum of nonlinear functions of linear combinations. The following Lemma states their argument.

Lemma 3.3.2 (Diaconis and Shahshahani 1984) *Suppose that* $f \in C^2\left(\mathbb{R}^2\right)$ *has the form*

$$f(x_1, x_2) = g_1(a_1 x_1 + b_1 x_2) + g_2(a_2 x_1 + b_2 x_2), \tag{3.15}$$

for some real numbers a_1, a_2, b_1, b_2*. Then the differential operator*

$$\prod_{i=1}^{2}\left(b_i \frac{\partial}{\partial x_1} - a_i \frac{\partial}{\partial x_2}\right) = b_1 b_2 \frac{\partial^2}{\partial x_1^2} - (a_1 b_2 + a_2 b_1)\frac{\partial^2}{\partial x_1 \partial x_2} + a_1 a_2 \frac{\partial^2}{\partial x_2^2}$$

applied to f *is identically zero.*

In our case, it is easy to check that for any real numbers c_1, c_2, c_3 and fixed $y \in \mathbb{R}$, the differential operator

$$c_1 \frac{\partial^2}{\partial x_1^2} - c_2 \frac{\partial^2}{\partial x_1 \partial x_2} + c_3 \frac{\partial^2}{\partial x_2^2}$$

applied to

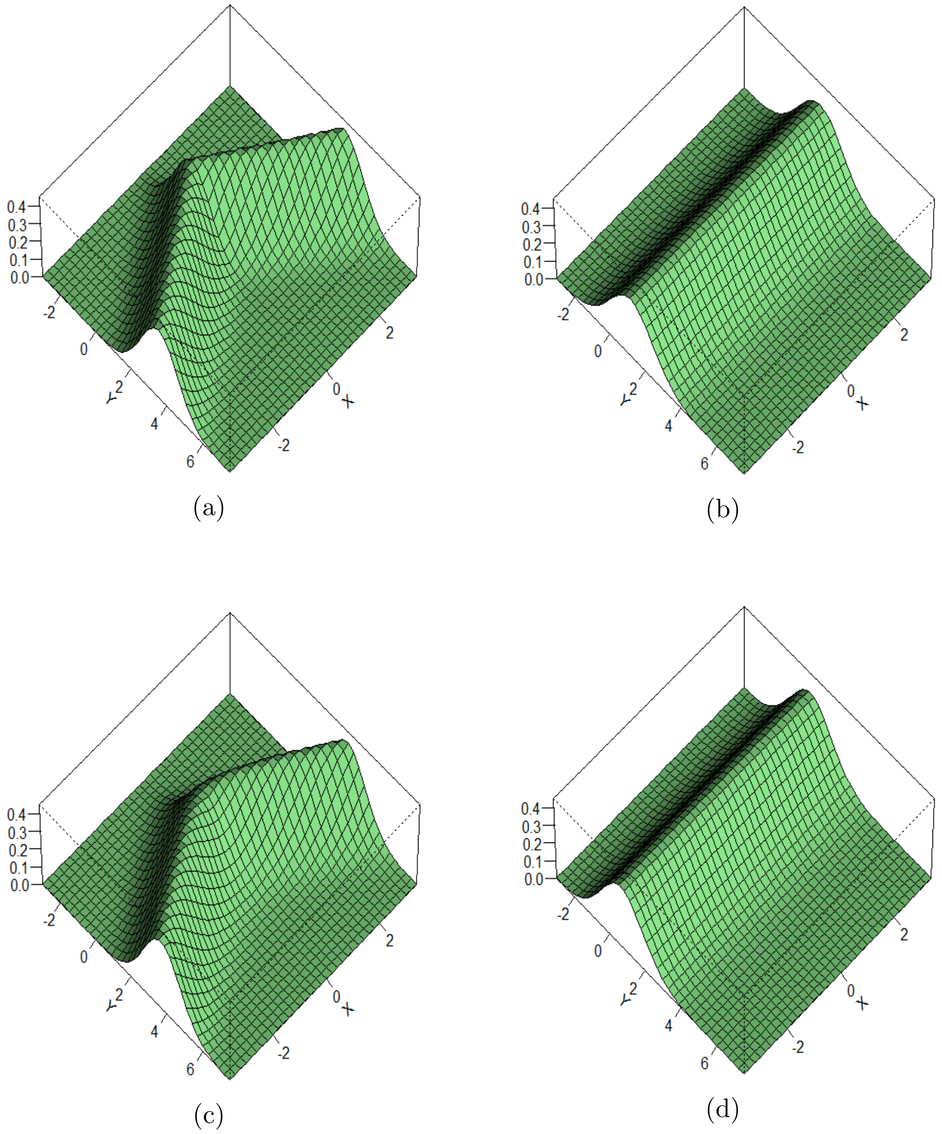
$$\ln f_{Y|X}\left(y|\theta^T x\right) \propto \left(y - \sqrt{X_1^2 + X_2^2}\right)^2$$

cannot be identically zero, and therefore $\ln f_{Y|X}\left(y|\theta^T x\right)$ does not have the additive form 3.15. Nevertheless, in the following we show that the projection pursuit conditional density approximation still achieves a relative high level of accuracy for the model of Example 1 by using two steps of projective corrections to an initial naive approximation. For simplicity of presentation, we choose the projections' directions to be $\theta_1 = (1, 0)$ and $\theta_2 = (0, 1)$.

Figures 3.1(a)-(j) track the progress of the approximation as it attempts to restore the form of the true conditional density. To begin with, Figure 3.1(a) shows the true conditional density $f_{Y|X}(y|X)$ plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$). Note that, by symmetry of the model, the conditional density has the same form as $f_{Y|X}\left(y|\theta^T X = x, \theta_\perp^T X = 0\right)$ for any $\theta \in \Theta$. Figure 3.1(b) presents the unconditional density $f_Y(y)$, which is taken as the initial model $g_{Y|X,0}(y|x)$, and is plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$). Again, $g_{Y|X,0}(y|x)$ has the same form when plotted against any other direction of x. We now use the projection pursuit procedure to modify the current model, $g_{Y|X,0}(y|x)$, along direction $x_1 = \theta_1^T x$, by using the formula

$$\begin{aligned} & g_{Y|X,m}(y|x) \\ = \; & g_{Y|X,m-1}(y|x)\, h_{0,m}(y, x_1) = g_{Y|X,m-1}(y|x) \frac{f_{Y,\theta_m^T X}(y, x_1)}{g_{Y,\theta_m^T X,m-1}(y, x_1)} \\ = \; & g_{Y|X,m-1}(y|x) \frac{\int f_{Y|X}(y|x) f_X(x)\, dx_2}{\int g_{Y|X,m-1}(y, x_1) f_X(x)\, dx_2} \\ = \; & g_{Y|X,m-1}(y|x) \frac{\int f_{Y|X}(y|x) f_{X_2}(x_2)\, dx_2}{\int g_{Y|X,m-1}(y|x) f_{X_2}(x_2)\, dx_2}, \end{aligned}$$

where in the last step above we used $f_X(x) = f_{X_1}(x_1) f_{X_2}(x_2)$. By definition of $h_{0,m}(y, x_1)$, it may behave erratically in regions where both $g_{Y,\theta_1^T X,m-1}(y, x_1)$ and $f_{Y,\theta_m^T X}(y, x_1)$ are



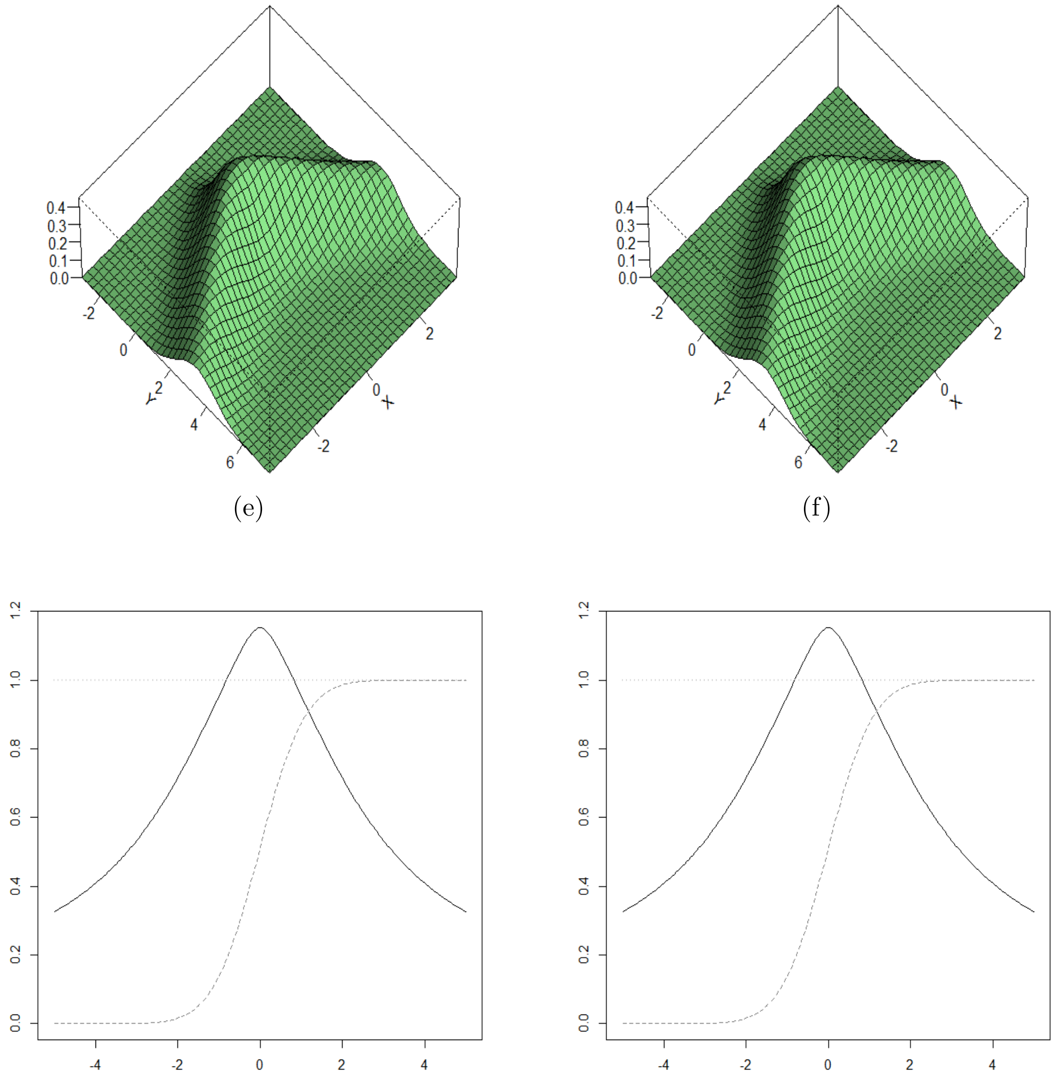
(a)

(b)

(c)

(d)

FIGURE 3.1. (continued on next page) *Example 1*: (a) The true conditional density plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$); (b) Model $g_{Y|X,0}\,(y|x)$ plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$); (c) Model $g_{Y|X,1}\,(y|x)$ plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$); (d) Model $g_{Y|X,1}\,(y|x)$ plotted against y and $\theta_2^T x$ ($\theta_1^T x = 0$).



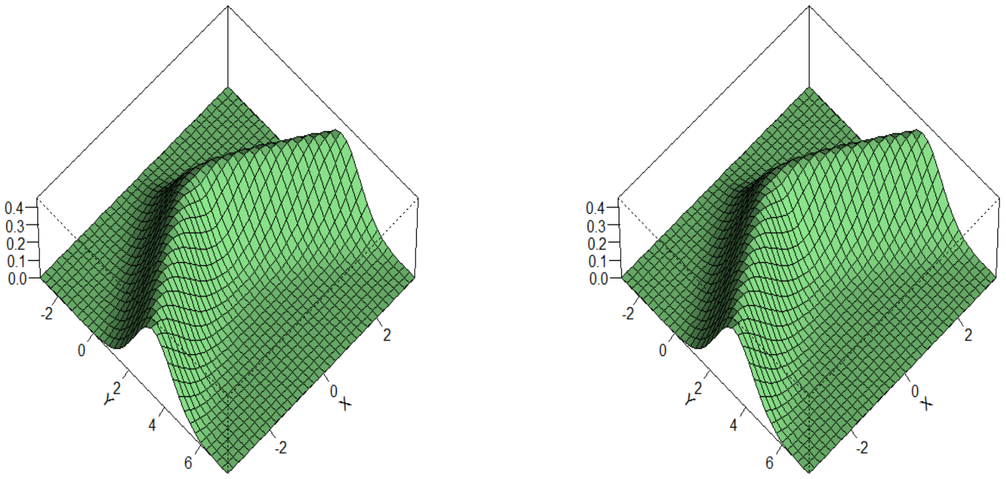
(e)

(f)

(g)

(h)

FIGURE 3.1. (continued on next page) *Example 1:* (e) Model $g_{Y|X,2}\,(y|x)$ plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$); (f) Model $g_{Y|X,2}\,(y|x)$ plotted against y and $\theta_2^T x$ ($\theta_1^T x = 0$); (g) $\int g_{Y|X,2}\,(y|x)\,dy$ (–) and $\int g_{Y|X,2}\,(y|x)\,f_X\,(x)\,dydx$ (...) plotted against $\theta_1^T x$ ($\theta_2^T x = 0$); (h) $\int g_{Y|X,2}\,(y|x)\,dy$ (–) and $\int g_{Y|X,2}\,(y|x)\,f_X\,(x)\,dydx$ (...) plotted against $\theta_2^T x$ ($\theta_1^T x = 0$).



(i)

(j)

FIGURE 3.1. *Example 1*: (i) Final model $g_{Y|X,2}(y|x) / \int g_{Y|X,2}(y|x)\,dy$ plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$); (j) Final model $g_{Y|X,2}(y|x) / \int g_{Y|X,2}(y|x)\,dy$ plotted against y and $\theta_2^T x$ ($\theta_1^T x = 0$).

very low. However, the resulting approximation $g_{Y|X,1}(y|x)$ is not affected by it. In Figures 3.1(c) and 3.1(d) $g_{Y|X,1}(y|x)$ is plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$) or, respectively, y and $\theta_2^T x$ ($\theta_1^T x = 0$). One can see in Figure 3.1(c) that $g_{Y|X,1}(y|x)$ successfully captures the fractured shape of the true conditional density along direction θ_1, albeit smoothing it slightly, while by construction, the shape of $g_{Y|X,1}(y|x)$ along direction θ_2 is still invariable (Figure 3.1(d)). Next, Figures 3.1(e) and 3.1(f) show the new model, $g_{Y|X,2}(y|x)$, obtained after applying the second projective correction along direction θ_2, and plotted, as usual, against y and $\theta_1^T x$ ($\theta_2^T x = 0$) or y and $\theta_2^T x$ ($\theta_1^T x = 0$). These figures indicate that $g_{Y|X,2}(y|x)$ now restores the general shape of the true conditional density along both directions θ_1 and θ_2. Figures 3.1(g) and 3.1(h) show the results of the integrals $\int g_{Y|X,2}(y|x)\,dy$ and $\int g_{Y|X,2}(y|x)\,f_X(x)\,dydx$ plotted against $\theta_1^T x$ ($\theta_2^T x = 0$) and against $\theta_2^T x$ ($\theta_1^T x = 0$). By that it means, for example, that Figure 3.1(g) shows plots of $\int g_{Y|X,2}(y|X = (t, 0))\,dy$ and of $\int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dx_2 \int_{-\infty}^{t} dx_1 g_{Y|X,2}(y|x)\,f_X(x)\,dx$ against t. One can see that $g_{Y|X,2}(y|x)$ is not a ‘proper’ conditional density in the sense that $\int g_{Y|X,2}(y|x)\,dy$ is not 1 for any $x \in S_X$, and it is particularly small in the regions where $f_X(x)$ is very low. Notwithstanding, one can also see that the condition $\int g_{Y|X,2}\left(y|x\right)f_{X}\left(x\right)dx=1$ is kept, and therefore, in practice, the unnormalised $g_{Y|X,2}\left(y|x\right)$ may still serve as an useful 'improper' approximation. Finally, Figures 3.1(i) and 3.1(j) show the normalised model $g_{Y|X,2}\left(y|x\right)/\int g_{Y|X,2}\left(y|x\right)dy$, which has now a very close form to that of $f_{Y|X}\left(y|\theta^{T}x\right)$, apart from some apparent oversmoothness.

3.4 Estimation Methodology and Algorithm

In practice, of course, the true densities of $Y|X$ and of X are unknown and need to be estimated from the data. The researcher is only given a set of observations $\{(x_t,y_t)\}_{t=1}^{n}$, which are assumed to be strictly stationary with the same distribution as (Y,X). Similarly to the theoretical approximation, the estimation procedure is recursive, so that at the m'th iteration, $m=1,...,M$, we assume that the estimate $\widehat{g}_{Y|X,m-1}\left(y|x\right)$ for the conditional density $f_{Y|X}\left(y|x\right)$ is given. The goal is to find an approximation for the optimal multiplicative function $h_{0,m}\left(y,\theta_{0,m}^{T}X\right)$ in order to produce an improved approximation $\widehat{g}_{Y|X,m}\left(y|x\right)$. If $m=1$, one may use any naive density estimator $\widehat{g}_{Y,0}\left(y\right)$ that depends only on y, and which needs to be positive on $\mathbb{S}_{Y}$.

In order to ensure the stability of the estimation procedure, we need to restrict attention to a compact subset of the support of $Z=(Y,X)$ such that the probability density $f_{Y,\theta^{T}X}\left(y,\theta^{T}x\right)$ is bounded away from 0 for any $\theta\in\Theta$. By abuse of notations, we henceforth denote this subspace by the symbol $\mathbb{S}$, which was used in previous sections to denote the whole support of Z. Accordingly, we redefine $\theta_{0,m}$, the target for our estimation, to be the maximiser of expected log-likelihood conditional on $Z\in\mathbb{S}$, that is,

$$\theta_{0,m}=\arg\max_{\theta_m\in\Theta}E_{\mathbb{S}}\left[\log\left(h_{0,m}\left(Y,\theta_{m}^{T}X\right)\right)\right],$$

where $E_{\mathbb{S}}$ is the conditional expectation given $Z\in\mathbb{S}$, and $h_{0,m}\left(y,\theta_{m}^{T}x\right)$ is the optimal projection given the estimate $\widehat{g}_{Y|X,m-1}(y|x)$,

$$h_{0,m}\left(y,\theta_m^T x\right)=\frac{f_{Y,\theta_m^T X}\left(y,\theta_m^T x\right)}{E\left[\widehat{g}_{Y|X,m-1}(y|X)\,|\theta_m^T X=\theta_m^T x\right] f_{\theta_m^T X}\left(\theta_m^T x\right)}. \tag{3.16}$$

Notice that the condition $Z\in\mathbb{S}$ should not have any significant effect if the subset $\mathbb{S}$ is chosen to be large enough.

For a given θ_m, (3.16) suggest that an estimator of the optimal correction function $h_{0,m}\left(y,\theta_m^T x\right)$ for $g_{Y|X,m-1}(y|x)$ can be obtained by

$$\widehat{h}_m\left(y,\theta_m^T x\right)=\frac{\widehat{f}_{Y,\theta_m^T X}\left(y,\theta_m^T x\right)}{\widehat{g}_{Y,\theta_m^T X,m-1}\left(y,\theta_m^T x\right)},$$

where $\widehat{f}_{Y,\theta_m^T X}\left(y,\theta_m^T x\right)$ and $\widehat{g}_{Y,\theta_m^T X,m-1}\left(y,\theta_m^T x\right)$ are the kernel estimators

$$\begin{aligned}\widehat{f}_{Y,\theta_m^T X}\left(y,\theta_m^T x\right) &= \frac{1}{nh_yh_x}\sum_{j=1}^{n}K\left(\frac{y_j-y}{h_y}\right)K\left(\frac{\theta_m^T(x_j-x)}{h_x}\right),\\ \widehat{g}_{Y,\theta_m^T X,m-1}\left(y,\theta_m^T x\right) &= \frac{1}{nh_x}\sum_{j=1}^{n}\widehat{g}_{Y|X,m-1}(y|x_j)\,K\left(\frac{\theta_m^T(x_j-x)}{h_x}\right).\end{aligned}$$

Here K is a non-negative, boundedly supported and symmetric density function and h_y, h_x are the bandwidths. Notice that $\widehat{g}_{Y,\theta_m^T X,m-1}\left(y,\theta_m^T x\right)$ is evaluated straightforwardly from the observations $\{x_t\}_{t=1}^n$, and there is no need to incorporate cumbersome Monte Carlo sampling as in the PPDE (Friedman, Stuetzle and Schroeder 1984). For the intuition behind this difference, note that according to Proposition 3.3.1(b), $g_{Y,\theta_m^T X,m-1}\left(y,\theta_m^T x\right)$ can be obtained from $g_{Y|X,m-1}(y|x)$ by the relation

$$g_{Y,\theta_m^T X,m-1}\left(y,\theta_m^T x\right)=E_{f_X}\left[g_{Y|X,m-1}(y|X)\,|\theta_m^T X=\theta_m^T x\right] f_{\theta_m^T X}\left(\theta_m^T x\right),$$

where the expectation is taken with respect to the real distribution of random vector X, and $f_{\theta_m^T X}\left(\theta_m^T x\right)$ is the real p.d.f. of $\theta_m^T X$. This can be seen as a demonstration of the fact that $g_{Y|X}(y|x)$ approximates $f_{Y|X}(y|x)$ in the sense that $g_{Y|X}(y|x)\,f_X(x)$ approximates

$f_{Y|X}\left(y|x\right)f_X\left(x\right)$, where $f_X\left(x\right)$ is the real distribution of random vector X (see Section 3.2). In the PPDE of Friedman, Stuetzle and Schroeder (1984), for a given θ_m given a current approximation $g_{X,m-1}\left(x\right)$ for the p.d.f. $f_X\left(x\right)$ of random vector X, their goal was to approximate

$$g_{\theta_m^T X,m-1}\left(\theta_m^T x\right) = E_{g_{X,m-1}}\left[1\left\{\theta_m^T X = \theta_m^T x\right\}\right].$$

Here, the the expectation is taken with respect to the approximated distribution $g_{X,m-1}\left(x\right)$. Therefore, the implementation of the PPDE requires drawing a Monte Carlo sample from the current approximated model $g_{X,m-1}\left(\theta_m^T x\right)$ at each step of the algorithm.

For the same reasons discussed in Section 2.2, we use 'leave-one-out' estimates in our calculations. Let $\widehat{h}_m^{-i}\left(y,\theta_m^T x\right)$ denote the 'leave-one-out' estimate of $h_{0,m}\left(y,\theta_m^T x\right)$ based on all observations other than the i'th, that is,

$$\begin{aligned} \widehat{h}_m^{-i}\left(y,\theta_m^T x\right) &= \frac{\widehat{f}^{-i}_{Y,\theta^T X}\left(y,\theta_m^T x\right)}{\widehat{g}^{-i}_{Y,\theta_m^T X,m-1}\left(y,\theta_m^T x\right)} \\ &= \frac{h_y^{-1}\sum_{j\neq i} K\left(\frac{y_j-y}{h_y}\right) K\left(\frac{\theta_m^T(x_j-x)}{h_x}\right)}{\sum_{j\neq i}\widehat{g}_{Y|X,m-1}\left(y|x_j\right) K\left(\frac{\theta_m^T(x_j-x)}{h_x}\right)}. \end{aligned} \tag{3.17}$$

Given the results of the last section, we can obtain an estimate, $\widehat{\theta}_m$, for the m'th step optimal direction $\theta_{0,m}$ by maximising an approximation of $D^*[g_{Y|\theta_m^T X,m-1}]$ with respect to $\theta_m \in \Theta$. We define

$$\widehat{\theta}_m = \arg\max_{\theta_m\in\Theta}\mathcal{L}\left(\theta_m\right)$$

where $\mathcal{L}\left(\theta_m\right)$ is the empirical m'th step log-likelihood function, obtained by replacing the expectation of expression (3.14) by a sample mean, and plugging-in the appropriate estimator for $h_{0,m}\left(y_i,\theta_m^T x_i\right)$,

$$\mathcal{L}\left(\theta_m\right) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\widehat{h}_m^{-i}\left(y_i,\theta_m^T x_i\right)\right)\widehat{\rho}_i^{\theta}. \tag{3.18}$$

Here, we use an additional trimming term $\widehat{\rho}_i^{\theta}$ as we would like to consider the average over only the observations $(y_i, x_i) \in \mathbb{S}$, and in addition such that $\widehat{h}_m^{-i}\left(y_i, \theta_m^T x_i\right)$ is bounded away from zero and infinity. The later restriction is needed to stabilise the finite-sample performance of the algorithm, but has no asymptotic effect on the method provided that $\widehat{h}_m^{-i}\left(y_i, \theta_m^T x_i\right)$ converges at a sufficient rate to $h_{0,m}\left(y_i, \theta_m^T x_i\right)$. For a given observation (y_i, x_i) and $\theta \in \Theta$ let $I_{\{(y_i,x_i)\in\mathbb{S}\}}$ be the indicator for the event $\{(y_i, x_i) \in \mathbb{S}\}$, while $I_{\{A_{n,\theta}^i\}}$ is the indicator for the event

$$A_{n,\theta}^i = \left\{\text{Both } \widehat{f}_{Y,\theta^T X}^{-i}\left(y_i, \theta^T x_i\right) \text{ and } \widehat{g}_{Y,\theta_m^T X,m-1}^{-i}\left(y_i, \theta^T x_i\right) \text{ lie in } \left(a_0 n^{-\gamma}, a_0^{-1} n^{\gamma}\right)\right\}$$

for some small constants $a_0, \gamma > 0$. The trimming term $\widehat{\rho}_i^\theta$ is then taken as

$$\widehat{\rho}_i^\theta = \frac{I_{\{(y_i,x_i)\in\mathbb{S}\}} \cdot I_{\{A_{n,\theta}^i\}}}{\frac{1}{n}\sum_{i=1}^n I_{\{(y_i,x_i)\in\mathbb{S}\}} \cdot I_{\{A_{n,\theta}^i\}}}.$$

As in Chapter 2, the trimming term $\widehat{\rho}_i^\theta$ is completely data-driven and it depends on the value of the parameter θ, evaluated by the likelihood. However, it does not assume any prior knowledge or applying a pilot estimation of θ_0.

Once $\widehat{\theta}_m$ is obtained, an estimator of $h_{0,m}\left(y, \theta_{0,m}^T X\right)$ can be produced with $\widehat{\theta}_m$ substituting $\theta_{0,m}$. Because the optimal kernel's bandwidths for efficient estimation of θ_m is known to undersmooth the nonparametric estimator of $h_{0,m}(\cdot,\cdot)$, a second stage of estimation is carried out with new bandwidths H_y, H_x. The second stage estimator can now include all observations, and is given by

$$\widetilde{h}_m\left(y, \widehat{\theta}_m^T x\right) = \frac{H_y^{-1}\sum_{j=1}^n K\left(\frac{y_j - y}{H_y}\right) K\left(\frac{\widehat{\theta}_m^T(x_j - x)}{H_x}\right)}{\sum_{j=1}^n \widehat{g}_{Y|X,m-1}\left(y|x_j\right) K\left(\frac{\widehat{\theta}_m^T(x_j - x)}{H_x}\right)}. \tag{3.19}$$

We then take (3.19) as an estimate of the optimal multiplicative correction $h_{0,m}\left(y, \theta_{0,m}^T X\right)$.

A summary of the algorithm for the proposed projection pursuit approximation of the c.p.d.f. $f_{Y|X}(y|x)$ is given as follows.

(1) Set $m = 0$. Initialise the approximation with a naive density estimator $\widehat{g}_{Y,0}(y)$ that

depends only on y and that is positive on $\mathbb{S}_Y$. As an example, one can take the kernel density estimator

$$\widehat{g}_{Y,0}\left(y\right)=n_1^{-1}h_y^{-1}\sum\nolimits_{j=1}^{n}K\left(\frac{y_j-y}{h_y}\right).$$

(2) Set $m\rightarrow m+1$.

(3) For any direction $\theta_m\in\Theta$ and every observation $\left(x_k,y_k\right)$, $k=1,...,n$, let $\widehat{h}_m^{-i}\left(y,\theta_m^T x\right)$ be given by formula (3.17).

(4) Maximise the likelihood function (3.18) with respect to $\theta_m\in\Theta$, and set

$$\widehat{\theta}_m=\arg\max_{\theta_m\in\Theta}\mathcal{L}\left(\theta_m\right)$$

(5) Obtain $\widetilde{h}_m\left(y,\widehat{\theta}_m^T x\right)$ given by formula (3.19).

(6) Let the m'th estimate be

$$\widehat{g}_{Y|X,m}\left(y|x\right)=\widehat{g}_{Y|X,m-1}\left(y|x\right)\widetilde{h}_m\left(y,\widehat{\theta}_m^T x\right).$$

(7) Repeat steps (2)-(6) for $m=1,...,M$ until estimate $\widehat{g}_{Y|X,M}\left(y|x\right)$ is obtained.

(8) Finally, use $\widehat{g}_{Y|X,M}\left(y|x\right)$ to approximate the c.p.d.f. $f_{Y|X}\left(y|x\right)$. If the computational load is not too heavy, it may be beneficial to normalise the estimator for any x-value of interest by taking $\widehat{g}_{Y|X,M}\left(y|x\right)/\int\widehat{g}_{Y|X,M}\left(y|x\right)dy$ as the final approximation.

It can be seen that at the first iteration, $m=1$, $\widehat{g}_{Y|X,1}\left(y|x\right)$ is simply the standard single-index kernel c.p.d.f. estimator of Y given $\theta^T X=\widehat{\theta}_1^T x$ considered by Fan et al (2009) irrespectively of the choice of $\widehat{g}_{Y,0}\left(y\right)$ (unless the 0'th approximation reflects dependency in X).

The PPCDE acts as a greedy algorithm in that at every iteration of the algorithm, $m=1,...,M$, it looks for the optimal orientation $\theta_{0,m}$ and their corresponding multiplicative function $h_{0,m}\left(y,\theta_{0,m}^T X\right)$, given the current estimate $\widehat{g}_{Y|X,m-1}\left(y|x\right)$. At every iteration, m, the algorithm utilises the most recent estimate $\widehat{g}_{Y|X,m-1}\left(y|x\right)$ in order to find the optimal multiplicative modification function. In particular, at every iteration of the algorithm, the previously estimated orientations $\theta_{0,1}, ..., \theta_{0,m-1}$ and their corresponding multiplicative function

$$h_{0,1}\left(y, \theta_{0,1}^T X\right), ..., h_{0,m-1}\left(y, \theta_{0,m-1}^T X\right)$$

need to be known. The greediness of the algorithm implies that the optimal modification function is defined with respect to current estimate $\widehat{g}_{Y|X,m-1}(y|x)$, and therefore there is no 'accumulation' of estimation errors as m goes up in the estimates of the optimal orientation $\theta_{0,m}$ and its corresponding multiplicative function $h_{0,m}\left(y, \theta_{0,m}^T X\right)$. Nevertheless, we would expect the variance of the final c.p.d.f. estimator to increase as m goes up due to the increased flexibility of the approximation when exploiting an increased number of projection directions (see also the numerical results of Section 3.7).

Example 1 (cont.): We apply the PPCDE algorithm to a sample of observations generated from the model of Example 1, introduced in the previous section. The number of observations was selected to $n = 1000$ in order to enable a clear visual illustration of the method performance. Using a smaller number of observations typically leads to appearance of erratic features in the density estimates that make it harder to visualise the progress of the algorithm. In Section 3.7 we consider this model again and we present the results of a Monte Carlo numerical study of the performance of the PPCDE algorithm for a smaller number of observations generated from this model. Figures 3.2(a)-(l) track the progress of the PPCDE, which is compared with the performance of the standard multivariate conditional density kernel estimator. For better comparison with the theoretical approximation, the graphs are plotted again against the previously used x-directions, $\theta_1 = (1, 0)$ and $\theta_2 = (0, 1)$. Figure 3.2(a) is a reminder of the shape of the true conditional density $f_{Y|X}(y|X)$ plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$). Figure 3.2(b) presents the initial naive estimate $\widehat{g}_{Y|X,0}(y|x)$, which is taken as the kernel unconditional density estimator of Y. We also compute the Root Mean Square Percentage Error ($RMSPE$) of the estimate

$\widehat{g}_{Y|X,0}(y|x)$, given by

$$RMSPE\left(\widehat{g}_{Y|X,m}\right) = \sum_{i=1}^{n}\left[\widehat{g}_{Y|X,m}(y_i|x_i) - f_{Y|X}(y_i|x_i)\right]^2 \Big/ \sum_{i=1}^{n} f_{Y|X}(y_i|x_i)^2\,, \tag{3.20}$$

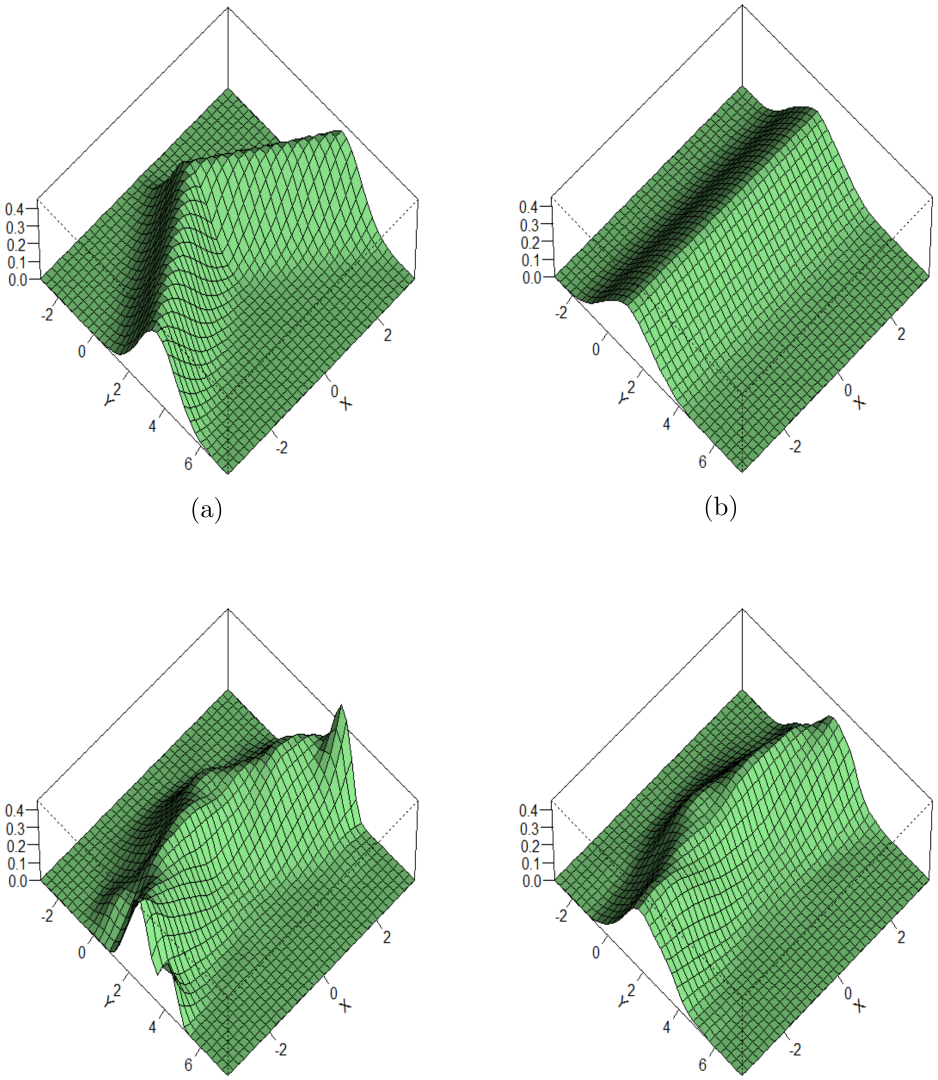
and we get $RMSPE\left(\widehat{g}_{Y|X,0}\right) = 0.104$. Next, maximising the first step empirical log-likelihood function along direction θ_1 yields the estimator $\widehat{\theta}_1 = (0.801, -0.598)$. Figures 3.2(c) and 3.2(d) present the corrected approximation,

$$\widehat{g}_{Y|X,1}(y|x) = \widehat{g}_{Y|X,0}(y|x)\,\widehat{h}_{0,m}\left(y, \widehat{\theta}_1^T x\right).$$

As mentioned above, $\widehat{g}_{Y|X,1}(y|x)$ is simply a single-index conditional density approximation. As usual, the new estimate is plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$) or, respectively, y and $\theta_2^T x$ ($\theta_1^T x = 0$). One can see in these figures that the shape of $\widehat{g}_{Y|X,1}(y|x)$ provides a closer approximation to $f_{Y|X}(y|X)$ relative to the naive one, and apart from a slight jittery effect at the regions where the density of X is very low $\{|X| > 3\}$, it provides a relatively smooth estimate. Indeed, we obtain now $RMSPE\left(\widehat{g}_{Y|X,1}\right) = 0.060$. We continue to estimate a second projective correction. The empirical second step log-likelihood function is maximised now at $\widehat{\theta}_2 = (0.610, 0.792)$. This provides a reasonable estimate, as $\widehat{\theta}_2$ is almost orthogonal to $\widehat{\theta}_1$, $\widehat{\theta}_1^T\widehat{\theta}_2 = 0.015$. Note, however, that generally, the orientation estimates obtained in the different iterations of the algorithm do not need to be orthogonal. Figures 3.2(e) and 3.2(f) show the obtained new estimate, $\widehat{g}_{Y|X,2}(y|x)$. These figures indicate that $\widehat{g}_{Y|X,2}(y|x)$ captures the general curvy form of the true conditional density along both directions θ_1 and θ_2. However, it is immediately apparent that the jittery effect at the regions of very low density of X is much magnified. Nevertheless, because the number of observations that fall inside these regions is very low, $\widehat{g}_{Y|X,2}(y|x)$ provides a better approximation than $\widehat{g}_{Y|X,1}$ with $RMSPE\left(\widehat{g}_{Y|X,2}\right) = 0.0436$. Figures 3.2(g) and 3.2(h) show the results of the integrals $\int \widehat{g}_{Y|X,2}(y|x)\,dy$ and $\int \widehat{g}_{Y|X,2}(y|x)\,f_X(x)\,dydx$ (computed with the true density $f_X(x)$ of X). Similarly to the theoretical approximation, we see that $\widehat{g}_{Y|X,2}(y|x)$ is not a 'proper' conditional density in the sense that $\int g_{Y|X,2}(y|x)\,dy$ is not 1 for any $x \in S_X$, and in particular in the regions where $f_X(x)$ is very low. Nevertheless, it is clear from the figures that the condition $\int \widehat{g}_{Y|X,2}(y|x) f_X(x)\, dx = 1$ is kept with high accuracy. In Figures 3.2(i) and 3.2(j), we show the normalised estimate $\widehat{g}_{Y|X,2}(y|x) / \int \widehat{g}_{Y|X,2}(y|x)\, dy$. The jittery effect is now softened and the normalised estimate has a smooth shape, which generally captures the true structure of $f_{Y|X}(y|X)$. The $RMSPE$ of the normalised estimate is indeed lower than that of the unnormalised version with a value of 0.030. As a benchmark for the performance of the PPCDE, we finally compare the PPCDE with the standard multivariate conditional density kernel estimator. Figures 3.2(k) and 3.2(l) present the estimate obtained with the conditional density kernel estimator applied to the same observations. It can be seen that the standard conditional density kernel estimator is more inclined to suffer from a decreased level of accuracy and erratic features, and especially at the low-density regions. Generally, this is a known feature that many nonparametric estimators tend to suffer from as they attempt to approximate the true model in regions with a very few numbers of observations. This effect is even more magnified in high-dimensions as a result of the 'empty space phenomenon' (see Silverman 1986, Section 4.5). Indeed, one can see that also for low dimensions, $d = 2$, the standard kernel estimator is characterised by some spurious features. The PPCDE, however, rectifies this phenomenon by working on lower-dimensional projections. In this example, the $RMSPE$ of the standard kernel estimator is 0.065, which is even less accurate than that of the single-index model.

3.5 Asymptotic Theory

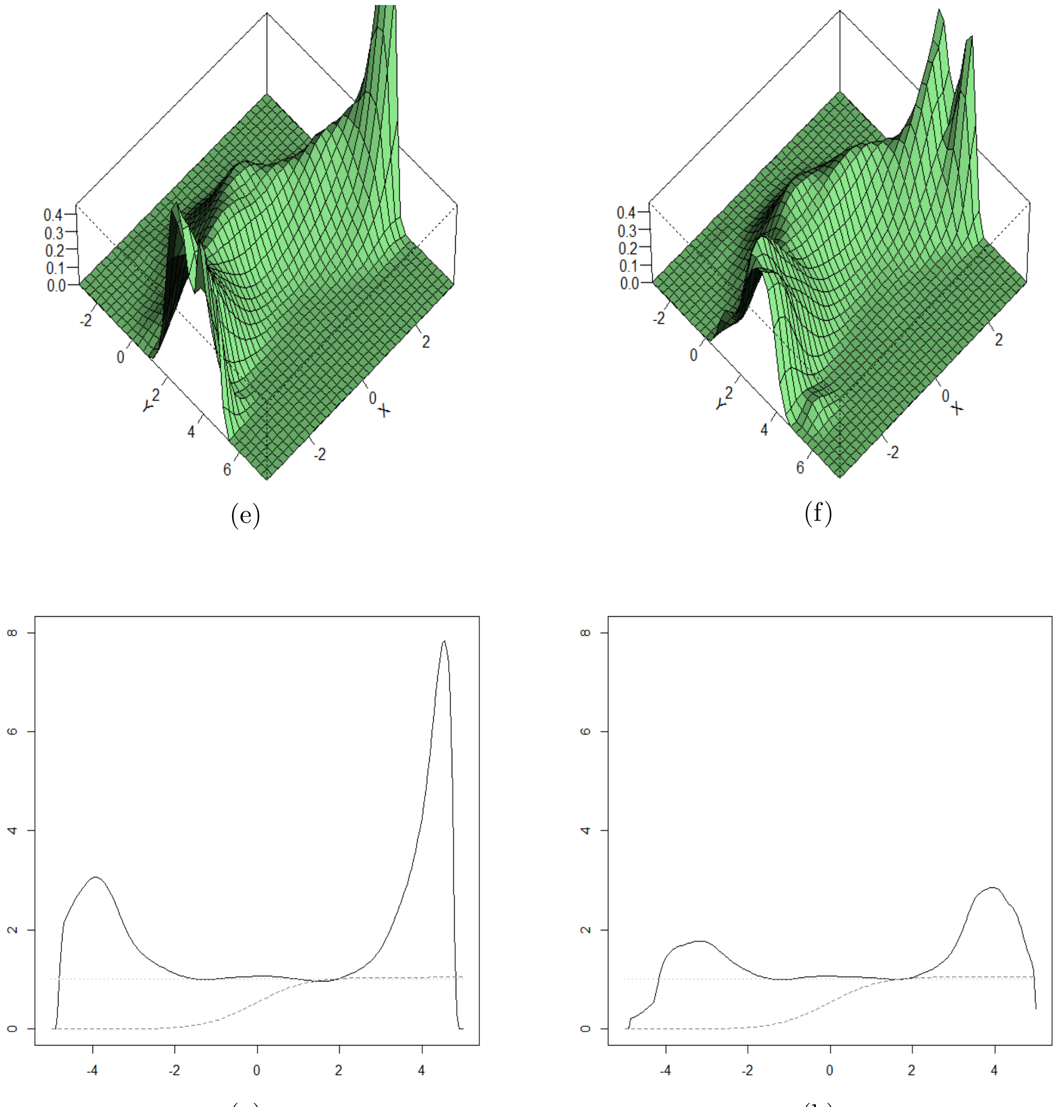
This section outlines some asymptotic results for the PPCDE under strong-mixing conditions. All the results of this section are straightforward generalisations of the asymptotic properties derived for the single-index model in Section 2.3. We shall confine our attention to a single stage of projection pursuit estimation algorithm, estimating the m'th projective approximation, for $m = 1, ..., M$, as all of the projective approximations are estimated similarly and with similar asymptotic properties. Given the initial



(a) (b)

(c) (d)

FIGURE 3.2. (continued on next page) Example 1: (a) The true conditional density plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$); (b) Model $\widehat{g}_{Y|X,0}\,(y|x)$ plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$); (c) Model $\widehat{g}_{Y|X,1}\,(y|x)$ plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$); (d) Model $\widehat{g}_{Y|X,1}\,(y|x)$ plotted against y and $\theta_2^T x$ ($\theta_1^T x = 0$).



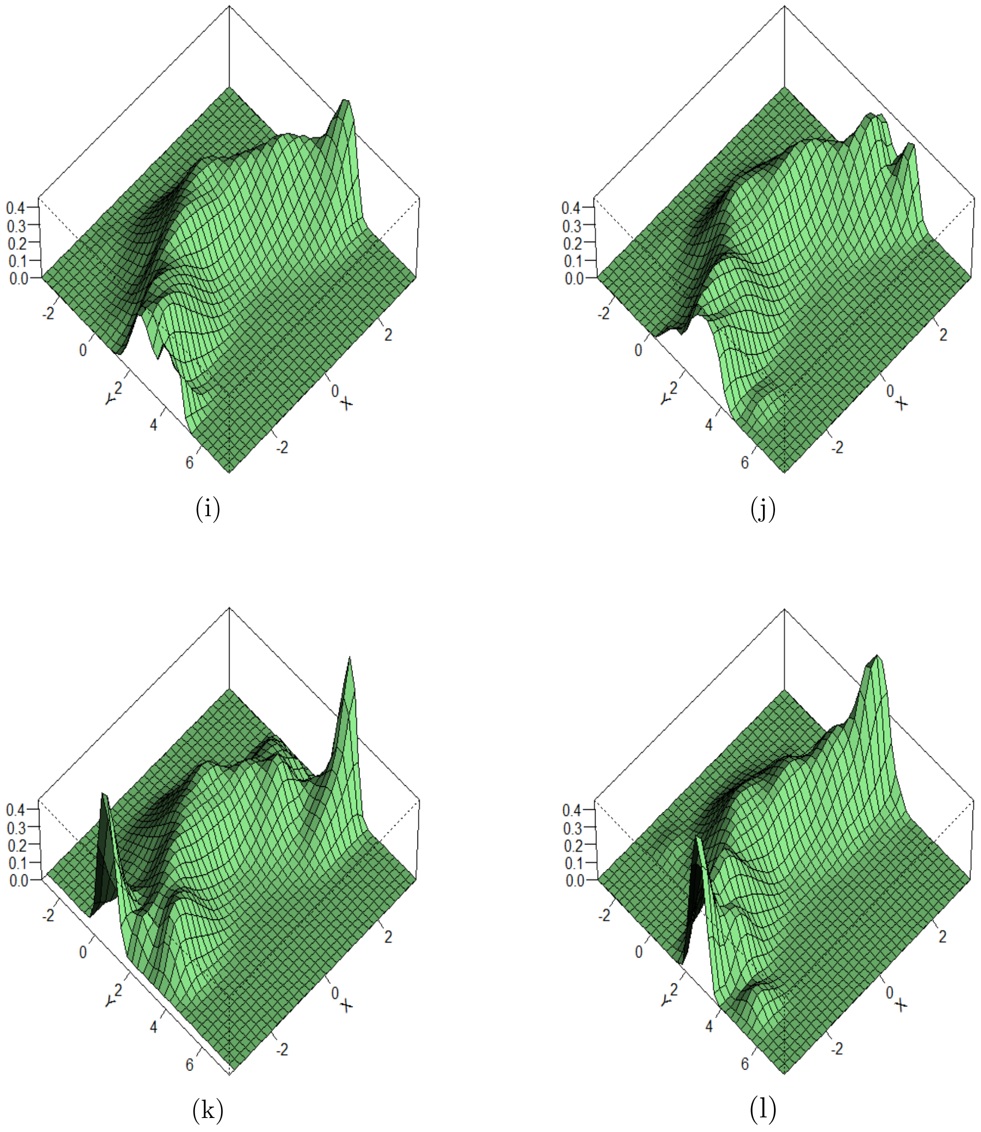
(e)

(f)

(g)

(h)

FIGURE 3.2. (continued on next page) *Example 1*: (e) Model $\widehat{g}_{Y|X,2}\,(y|x)$ plotted against y and $\theta_1^T x$ $(\theta_2^T x = 0)$; (f) Model $g_{Y|X,2}\,(y|x)$ plotted against y and $\theta_2^T x$ $(\theta_1^T x = 0)$; (g) $\int \widehat{g}_{Y|X,2}\,(y|x)\,dy$ (–) and $\int \widehat{g}_{Y|X,2}\,(y|x)\,f_X\,(x)\,dydx$ (...) plotted against $\theta_1^T x$ $(\theta_2^T x = 0)$; (h) $\int \widehat{g}_{Y|X,2}\,(y|x)\,dy$ (–) and $\int \widehat{g}_{Y|X,2}\,(y|x)\,f_X\,(x)\,dydx$ (...) plotted against $\theta_2^T x$ $(\theta_1^T x = 0)$.



(i) (j) (k) (l)

FIGURE 3.2. *Example 1*: (i) Final model $\widehat{g}_{Y|X,2}\,(y|x)\,/\int \widehat{g}_{Y|X,2}\,(y|x)\,dy$ plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$); (j) Final Model $\widehat{g}_{Y|X,2}\,(y|x)\,/\int \widehat{g}_{Y|X,2}\,(y|x)\,dy$ plotted against y and $\theta_2^T x$ ($\theta_1^T x = 0$); (k) Standard bivariate kernel estimator plotted against y and $\theta_1^T x$ ($\theta_2^T x = 0$); (l) Standard bivariate kernel estimator plotted against y and $\theta_2^T x$ ($\theta_1^T x = 0$).

estimate $\widehat{g}_{Y|X,m-1}(y|x)$ obtained in the $(m-1)$'th iteration, our goal now is to estimate $h_{0,m}\left(y, \theta_{0,m}^T x\right)$ by $\widetilde{h}_m\left(y, \widehat{\theta}_m^T x\right)$, according to the procedure described in the previous section, which then may be used to obtain the m'th estimate,

$$\widehat{g}_{Y|X,m}(y|x) = \widehat{g}_{Y|X,m-1}(y|x)\,\widetilde{h}_m\left(y, \widehat{\theta}_m^T x\right).$$

Define the set $\mathbb{S}_\delta$ distant no further than some small $\delta > 0$ from some $\left(y, \theta^T x\right)$ such that $(y,x) \in \mathbb{S}$ and $\theta \in \Theta$. We derive our asymptotic results under the following assumptions.

Assumption 1: (A1) The sequence $\{y_j, x_j\}_{j=1}^n$ is strictly stationary strong-mixing series with mixing coefficients that satisfy $\alpha_t \leq A\alpha^t$ with $0 < A < \infty$ and $0 < \alpha < 1$.

Assumption 2: $K(\cdot)$ is a symmetric, non-negative, compactly supported, three-times boundedly differentiable kernel.

Assumption 3: The bandwidths satisfy $h_y, h_x = o(1)$, $n^{1-\delta}h_y h_x \to \infty$ and $n^{2-\delta}h_y h_x^5 \to \infty$ for some $\delta > 0$.

Assumption 4: For all $\theta \in \Theta$, $\left(Y, \theta^T X\right)$ has probability density $f_{Y,\theta^T X}(y,t)$ with respect to Lebesgue measure on $\mathbb{S}_\delta$ and

$$\inf_{(y,t)\in\mathbb{S}_\delta} f_{Y,\theta^T X}(y,t) > 0 \quad \text{and} \quad \sup_{(y,t)\in\mathbb{S}_\delta} f_{Y,\theta^T X}(y,t) < \infty.$$

In addition, $f_{Y,\theta^T X}(y,t)$ and $E\left[X|Y=y, \theta^T X = t\right]$ and $E\left[XX^T|Y=y, \theta^T X = t\right]$ are four-times continuously differentiable with respect to $(y,t) \in \mathbb{S}_\delta$. Moreover, there is some j^* such that for all $j > j^*$ and $\left(y_1, \theta^T x_1\right), \left(y_j, \theta^T x_j\right) \in \mathbb{S}_\delta$ the joint probability density of $\left(y_1, \theta^T x_1, y_j, \theta^T x_j\right)$ is bounded.

Assumption 5: The initial estimate for the current iteration is non-negative and strictly positive on $\mathbb{S}$, and it satisfies

$$\inf_{(y,t)\in\mathbb{S}_\delta} E\left[\widehat{g}_{Y|X,m-1}(y|X)\,|\theta^T X = t\right] > 0 \text{ and } \sup_{(y,t)\in\mathbb{S}_\delta} E\left[\widehat{g}_{Y|X,m-1}(y|X)\,|\theta^T X = t\right] < \infty.$$

Additionally, $\frac{\partial}{\partial y}\widehat{g}_{Y|X,m-1}(y|x)$ exists and is bounded, and $E\left[\widehat{g}_{Y|X,m-1}(y|X)\,|\theta^T X = t\right]$ is four-times continuously differentiable with respect to t, for any $(y,t)\in\mathbb{S}_{\delta}$.

Assumption 6: For the trimming operator, we require that $a_0,\gamma>0$ and $n^{\gamma}\left(h_y^2+h_x^2\right)=o\left(1\right)$ and $n^{1-2\gamma-\delta}h_yh_x\rightarrow\infty$ for some $\delta>0$.

Assumption 7: $\theta_{0,m}$ is the unique global maximum of $E_{\mathbb{S}}\left[\log\left(h_{0,m}\left(Y,\theta^TX\right)\right)\right]$, and it lies in the interior of Θ.

Assumption 8: H_y and H_x satisfy $H_yH_x/h_yh_x^3=o\left(n^{1-\delta}\right)$, $H_yH_x\left(h_y^2+h_x^2\right)=o\left(n^{-1}\right)$, $H_y,H_x=O\left(n^{-1/6}\right)$ and $n^{1-\delta}H_yH_x\rightarrow\infty$ for some $\delta>0$.

All assumptions (apart from Assumption 5) are due to Chapter 2. The assumptions regarding the bandwidths can be satisfied for bandwidths with optimal asymptotic rate $h_y,h_x\sim n^{-1/4}$ and $H_y,H_x\sim n^{-1/6}$. Assumption 5 is needed to obtain uniform consistency of $\widehat{g}_{Y,\theta_m^TX,m-1}\left(y,\theta^Tx\right)$ and its derivatives with respect to θ. The uniqueness of $\theta_{0,m}$ is assumed merely for theoretical convenience (see the discussion following condition (A6) in Section 2.3). In practice, $\theta_{0,m}$ can be any one from the set of maxima points of $E_{\mathbb{S}}\left[\log\left(h_{0,m}\left(Y,\theta^TX\right)\right)\right]$. Assumption 8 is based on the conditions of Theorem 2.3.3 for the kernel's bandwidths in the second stage of the estimation. It is required in order to keep the orientation estimator's rate of consistency fast enough so that $\widetilde{h}_m\left(y,\widehat{\theta}_m^Tx\right)$ is a consistent estimator of $h_{0,m}\left(y,\theta_{0,m}^Tx\right)$ and it has the same first-order asymptotic properties as if the optimal $\theta_{0,m}$ was known.

The following preliminary Lemma is an adaptation of the results of Hansen (2008), and it shows uniform consistency of the kernel estimators and their derivatives with respect to θ.

Lemma 3.5.1 *Let Assumptions 1-5 hold. Then for* $k=0,1,2$,

$$\sup_{\theta\in\Theta,z\in\mathbb{S}}\left|\frac{\partial^k}{\partial\theta^k}\widehat{f}_{Y,\theta_m^TX}\left(y,\theta^Tx\right)-\frac{\partial^k}{\partial\theta^k}f_{Y,\theta_m^TX}\left(y,\theta^Tx\right)\right|=O_p\left(\left(\frac{\ln n}{nh_yh_x^{1+2k}}\right)^{1/2}+h_y^2+h_x^2\right),$$

and

$$\sup_{\theta\in\Theta,z\in\mathbb{S}}\left|\frac{\partial^k}{\partial\theta^k}\widehat{g}_{Y,\theta_m^T X,m-1}\left(y,\theta^T x\right)-\frac{\partial^k}{\partial\theta^k}g_{Y,\theta^T X,m-1}\left(y,\theta^T x\right)\right|=O_p\left(\left(\frac{\ln n}{nh_x^{1+2k}}\right)^{1/2}+h_x^2\right),$$

where $g_{Y,\theta^T X,m-1}\left(y,t\right)=E\left[\widehat{g}_{Y|X,m-1}\left(y|X\right)|\theta^T X=t\right]f_{\theta^T X}\left(t\right)$*.*

As a consequence of Lemma 3.5.1 it can be shown that for any large enough n, the trimming term $\widehat{\rho}_i^\theta$ is responsible for averaging exactly over those observations that belong to subset $\mathbb{S}$.

Lemma 3.5.2 . *Let Assumptions 1-6 hold. Then for any large enough n,*

$$\max_{1\le i\le n}\sup_{\theta\in\Theta}\left|\widehat{\rho}_i^\theta-\frac{I_{\{(y_i,x_i)\in\mathbb{S}\}}}{\frac{1}{n}\sum_{i=1}^n I_{\{(y_i,x_i)\in\mathbb{S}\}}}\right|=0$$

with probability 1.

With the last two lemmas it is straightforward to establish the uniform consistency of the empirical log-likelihood function.

Lemma 3.5.3 . *Let Assumptions 1-6 hold, and let $\mathcal{L}\left(\theta_m\right)$ be the empirical m'th step log-likelihood function (3.18). Then*

$$\sup_{\theta_m\in\Theta}\left|\mathcal{L}\left(\theta_m\right)-E_{\mathbb{S}}\left[\log\left(h_{0,m}\left(Y,\theta_m^T X\right)\right)\right]\right|=o_p\left(1\right).$$

The consistency of $\widehat{\theta}$ now follows easily.

Proposition 3.5.1 *Let Assumptions 1-7 hold. Then as $n\to\infty$*

$$\widehat{\theta}_m\to_p\theta_{0,m}.$$

By Assumption 4, an application of the mean value theorem applied to $\frac{\partial}{\partial\theta}\mathcal{L}\left(\theta_m\right)$ yields

$$\frac{\partial}{\partial\theta}\mathcal{L}\left(\widehat{\theta}_m\right)-\frac{\partial}{\partial\theta}\mathcal{L}\left(\theta_{0,m}\right)=\frac{\partial^2}{\partial\theta^2}\mathcal{L}_N\left(\overline{\theta}_m\right)\left(\widehat{\theta}_m-\theta_{0,m}\right),\qquad(3.21)$$

where the mean value $\overline{\theta}_m$ satisfies $\left|\overline{\theta}_m - \theta_{0,m}\right| \leq \left|\widehat{\theta}_m - \theta_{0,m}\right|$. Using the above results with the asymptotic theory for U-statistics of strong-mixing observations (Gao and King 2004) we can establish the next Lemma.

Lemma 3.5.4 *Let Assumptions 1-7 hold, and*

$$\Omega\left(\theta_{0,m}\right) = E_{\mathbb{S}}\left[-\frac{\partial^2}{\partial\theta^2}\log\left(h_{0,m}\left(Y, \theta_{0,m}^T X\right)\right)\right].$$

Then

(i) $\mathcal{L}\left(\theta_m\right) = E_{\mathbb{S}}\left[\log\left(h_{0,m}\left(Y, \theta_m^T X\right)\right)\right] + O_p\left(n^{-1/2}\right) + O(h_y^2 + h_x^2),$

(ii) $\frac{\partial}{\partial\theta}\mathcal{L}\left(\theta_{0,m}\right) = O_p\left(n^{2-\delta}h_y h_x^3\right)^{-1/2} + O(h_y^2 + h_x^2)$ *for any* $\delta > 0$*, and*

(iii) $\frac{\partial^2}{\partial\theta^2}\mathcal{L}\left(\overline{\theta}_m\right) = -\Omega\left(\theta_{0,m}\right) + O_p\left(n^{2-\delta}h_y h_x^5\right)^{-1/2} + O(h_y^2 + h_x^2)$ *for any* $\overline{\theta}_m \rightarrow_p \theta_{0,m}$.

The asymptotic expressions given in Lemma 3.5.4 imply the rate of convergence of $\widehat{\theta}_m$. We obtain the next result.

Proposition 3.5.2 *Let Assumptions 1-7 hold. Then*

$$\widehat{\theta}_m - \theta_{0,m} = O_p\left(n^{2-\delta}h_y h_x^3\right)^{-1/2} + O(h_y^2 + h_x^2),$$

for any $\delta > 0$ *arbitrarily small.*

The last proposition suggests that the optimal rate of convergence rate of $\widehat{\theta}_m$ is obtained when bandwidths h_y and h_x both have the asymptotic rate $n^{-1/4}$. This rate is clearly a slower rate than the $\sqrt{n}$-rate achieved for many parametric and semiparametric estimators, and in particular for many single-index regression models where only a univariate density needs to be estimated. Nevertheless, it is arbitrarily close to that parametric rate.

The last result of the section shows that given an appropriate choice of bandwidths, $\widetilde{h}_m\left(y, \widehat{\theta}_m^T x\right)$ can estimate $h_{0,m}\left(y, \theta_{0,m}^T x\right)$ with the same first-order asymptotic properties as if the optimal $\theta_{0,m}$ was known.

Proposition 3.5.3 *Let Assumptions 1-8 hold. Then*

$$\sup_{(y,x)\in\mathbb{S}}\left|\widetilde{h}_m\left(y,\widehat{\theta}_m^T x\right)-h_{0,m}\left(y,\theta_{0,m}^T x\right)\right|=O_p\left(\left(\frac{\ln n}{nH_yH_x}\right)^{1/2}\right).$$

When the optimal asymptotic rate for the bandwidths is chosen, i.e. $H_y, H_x \sim n^{-1/6}$, we then get that $\widetilde{h}_m\left(y,\widehat{\theta}_m^T x\right)$ is a uniformly consistent estimator of $h_{0,m}\left(y,\theta_{0,m}^T x\right)$ with a convergence rate of $n^{\delta-1/3}$ for $\delta > 0$ arbitrarily small.

3.6 Information Criterion Stopping Rule

As with any iterative method, the PPCDE needs a criterion for terminating the algorithm after some finite M'th iteration. Stopping the algorithm too early can increase the bias of the c.p.d.f. estimator, and stopping it too late can increase its variance.

When one has a large amount of data, and it is possible to allocate a suitable validation set, it may be most beneficial to terminate the algorithm based on the out-of-sample performance on the validation set. However, in cases when one does not have a sufficiently large amount of data or when a suitable validation set is hard to define, a different criterion to terminate the algorithm needs to be employed. Friedman, Stuetzle and Schroeder (1984) discussed some alternative heuristic criteria for their iterative PPDE, which are based on comparisons between models obtained in consecutive iterations or simply on graphical inspection. Nevertheless, they did not provide any formal procedure or statistical justification. Cross-validatory techniques were shown to have successful applications to model selection in semiparametric settings (Gao and Tong 2004, Kong and Xia 2007), and they can be used to produce a stopping rule to the PPCDE procedure. However, these computationally intensive techniques are less desirable as at each iteration $\widehat{\theta}_m$ has to found by numerical optimisation. In the following we propose an Information Criterion (IC) stopping rule that is based on bias correction for the estimator of the marginal decrease in the relative entropy.

To motivate our proposal, recall that by the results of Section 3.3 we have that the marginal decrease in the relative entropy satisfies

$$D^*[g_{Y|\theta_{0,m}^T X, m-1}] = E\log\left(h_{0,m}\left(y, \theta_{0,m}^T x\right)\right) \to 0 \quad \text{as } m \to \infty,$$

and as a result, the projection pursuit approximation $g_{Y|X,m}\left(y|x\right)$ is ensured to converge weakly to the true c.p.d.f. $f_{Y|X}\left(y|x\right)$. This suggests that one can terminate the approximation procedure once $E\log\left(h_{0,m}\left(y, \theta_{0,m}^T x\right)\right) \approx 0$. In practice, at the m'th iteration $h_{0,m}\left(\cdot,\cdot\right)$ and $\theta_{0,m}^T$ are replaced by the suboptimal estimates $\widetilde{h}_m\left(\cdot,\cdot\right)$ and $\widehat{\theta}_m$, and the relative entropy may not improve beyond a certain number of iterations. In that case, one would like then to use the approximation obtained at the last iteration before the marginal decrease in the relative entropy becomes non-positive,

$$L(\widetilde{h}_m, \widehat{\theta}_m) \equiv E_{\mathbb{S}}\left[\log\left(h_m\left(Y, \theta^T X\right)\right)\right]\big|_{h_m = \widetilde{h}_m,\ \theta = \widehat{\theta}_m} \leq 0.$$

Note that here, again, we restrict ourselves to a conditional expectation given that $(Y, X) \in \mathbb{S}$ for the same reasons mentioned in Section 3.4.

An obvious estimator for $L(\widetilde{h}_m, \widehat{\theta}_m)$ is the empirical m'th step log-likelihood function, evaluated with $\widetilde{h}_m\left(y_i, \widehat{\theta}_m x_i\right)$,

$$\mathcal{L}(\widetilde{h}_m, \widehat{\theta}_m) = \frac{1}{n_s}\sum_{i=1}^{n} \log\left(\widetilde{h}_m\left(y_i, \widehat{\theta}_m x_i\right)\right) I_{\{(y_i,x_i)\in\mathbb{S}\}},$$

where $n_s = \sum_{i=1}^n I_{\{(y_i,x_i)\in\mathbb{S}\}}$. Although $\mathcal{L}(\widetilde{h}_m, \widehat{\theta}_m)$ is a consistent estimator, it has the tendency to overestimate $L(\widetilde{h}_m, \widehat{\theta}_m)$, since both terms $\widetilde{h}_m\left(\cdot,\cdot\right)$ and $\widehat{\theta}_m$ are estimated using the same observations, used again to approximate the mean in $L(\widetilde{h}_m, \widehat{\theta}_m)$. Let the bias of this estimator be

$$b_m = E_{\mathbb{S}}\left[\mathcal{L}(\widetilde{h}_m, \widehat{\theta}_m) - L(\widetilde{h}_m, \widehat{\theta}_m)\right].$$

Following Akaike (1973), we define an Information Criterion for the m'th step as the bias-corrected log-likelihood function,

$$IC_m = \mathcal{L}(\widetilde{h}_m, \widehat{\theta}_m) - \widehat{b}_m, \tag{3.22}$$

where $\widehat{b}_m$ is an estimator of b_m. This two terms appearing in the Information Criterion IC_m play a similar role as in the standard AIC, with the exception that IC_m captures the marginal decrease in the relative entropy, rather the relative entropy itself. On the RHS of the expression above, the first term $\mathcal{L}(\widetilde{h}_m, \widehat{\theta}_m)$ estimates the relative entropy between the c.p.d.f.'s $f_{Y|\theta_m^T X}$ and $g_{Y|\theta_m^T X,m}$ (see Proposition 3.3.1(c)), and it reflects the marginal model complexity; increasing the number of iterations of the PPCDE procedure from $(m-1)$ to m is likely to yield a positive value $\mathcal{L}(\widetilde{h}_m, \widehat{\theta}_m)$. However, the model's marginal complexity is penalised by the second term, which reflects the model stability. The optimum model, obtained when $IC_m \approx 0$, is a trade-off between the two terms.

We use $IC_m,\ m = 1, 2, ...$, as a 'goodness of model' evaluation tool, in the sense that we can terminate the stepwise algorithm at the first instant that

$$IC_m \leq 0,$$

and we use the approximation obtained at the last iteration before this condition began to hold.

Akaike's Information Criterion (AIC) is derived under somewhat strict parametric settings, and in particular under the assumption that the specified parametric model contains the true distribution. Under these settings, Akaike showed that the bias correction term, b_m, is asymptotically the number of free parameters contained in the model over n. Stone (1977) further showed asymptotic equivalence between the AIC criterion and the leave-one-out cross-validation. Takeuchi (1976) relaxed the later assumption of Akaike (1973), and he allowed the true distribution to lie outside the parameterised family of distributions. Yet, Takeuchi's Information Criterion (TIC) relies on the Fisher efficiency of the parametric MLE, which does not hold in our case (Proposition 3.5.2). An alternative Information

Criterion that works under very weak assumptions is the bootstrap Extended Information Criterion (EIC), proposed by Ishiguro, Sakamoto and Kitagawa (1997), and it can be applied to estimate b_m in our model. Similar to the cross-validation, a substantial drawback of the EIC is that it requires computation of many bootstrap versions of $\widehat{\theta}_m$. Thus, one needs to solve many numerical maximisation problems in $\mathbb{R}^d$ repeatedly for each bootstrap sample.

We therefore propose a semi-analytic bootstrap approach that combines different elements of both the TIC and the EIC. Through expansion of the bias term, we show that the asymptotically dominant terms depend on θ_0 and not on $\widehat{\theta}$, which allows the application of the bootstrap method without producing numerous bootstrap estimates of $\widehat{\theta}$. To that end, let

$$\Omega\left(\theta_{0,m}\right) = \frac{\partial^2}{\partial\theta^2} L(\theta_{0,m}),$$

and

$$\begin{aligned} H\left(\widetilde{h}_m, \theta_{0,m}\right) &= E_{\mathbb{S}}\left[\mathcal{L}\left(\widetilde{h}_m, \theta_{0,m}\right) - L(\widetilde{h}_m, \theta_{0,m})\right], \\ J\left(\widetilde{h}_m, \theta_{0,m}\right) &= E_{\mathbb{S}}\left[\frac{\partial}{\partial\theta}\left\{\mathcal{L}(\widetilde{h}_m, \theta_{0,m}) - L(\widetilde{h}_m, \theta_{0,m})\right\}\frac{\partial}{\partial\theta}\mathcal{L}(\widehat{h}_m, \theta_{0,m})^T\right], \end{aligned}$$

where $\widetilde{h}_m\left(\cdot,\cdot\right)$ and $\widehat{h}_m\left(\cdot,\cdot\right)$ correspond to the estimates (3.17) and (3.19), used for the first and second stage of estimation in the m'th iteration, with the appropriate bandwidths.

In the next proposition we derive an asymptotic bias correction for b_m. The basic argument used in the derivation of the TIC is generalised to the semiparametric case (see Konishi and Kitagawa 1996, Konishi and Kitagawa 2008). As such, this proposition is of interest in its own right. For the sake of simplicity, the error rates derived in the second part of the proposition are presented with the optimal choice of bandwidths, in accordance with the previous section.

Proposition 3.6.1 *Let Assumptions 1-8 hold. Then*

$$b_m = \overline{b}_m + o_p\left(\overline{b}_m\right),$$

where

$$\overline{b}_m = \text{trace}\left[J\left(\widetilde{h}_m, \theta_{0,m}\right)\Omega^{-}\left(\theta_{0,m}\right)\right] + H\left(\widetilde{h}_m, \theta_{0,m}\right).$$

In particular, if $h_y, h_x \sim n^{-1/4}$ and $H_y, H_x \sim n^{-1/6}$, then for any $\delta > 0$

$$\overline{b}_m = O_p\left(n^{-5/6+\delta}\right) + O\left(n^{-1/3}\right).$$

This result clearly suggests that $\overline{b}_m$ should provide a good approximation for the exact bias b_m. The advantage of $\overline{b}_m$ over b_m is that it does not depend on the parameter estimate $\widehat{\theta}_m$, and it can therefore be bootstrapped in reasonable time.

We therefore define the estimator of b_m to be

$$\widehat{b}_m = \text{trace}\left[J^*\left(\widetilde{h}^*_m, \widehat{\theta}_m\right)\widehat{\Omega}^{-}\left(\widehat{\theta}_m\right)\right] + H^*\left(\widetilde{h}^*_m, \widehat{\theta}_m\right). \tag{3.23}$$

Here, $\widehat{\Omega}^{-}\left(\widehat{\theta}_m\right)$ is a simple sample version of $\Omega^{-}\left(\theta_{0,m}\right)$ and $J^*\left(\widetilde{h}^*_m, \widehat{\theta}_m\right)$ and $H^*\left(\widetilde{h}^*_m, \widehat{\theta}_m\right)$ are the bootstrap versions of $J\left(\widetilde{h}_m, \theta_{0,m}\right)$ and $H\left(\widetilde{h}_m, \theta_{0,m}\right)$.

To this end, let the estimator of $\Omega\left(\theta_{0,m}\right)$ be

$$\widehat{\Omega}\left(\widehat{\theta}_m\right) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log\left(\widehat{h}_m\left(Y, \theta_{0,m}^T X\right)\right). \tag{3.24}$$

The reason for using $\widehat{h}_m$ here, instead of $\widetilde{h}_m$, is that it is clear from the proof of Proposition 3.6.1 that the $\Omega\left(\theta_{0,m}\right)$ term is obtained from taking the limit in probability of the term $E_{\mathbb{S}}\left[-\frac{\partial^2}{\partial\theta^2}\log\left(\widehat{h}_m\left(Y, \theta_{0,m}^T X\right)\right)\right]$.

In order to obtain $J^*\left(\widetilde{h}^*_m, \widehat{\theta}_m\right)$ and $H^*\left(\widetilde{h}^*_m, \widehat{\theta}_m\right)$, we produce B bootstrap samples of size n, $\left\{\left\{y_t^{*(i)}, x_t^{*(i)}\right\}_{t=1}^{n}, \ i = 1, ..., B\right\}$. Let the bootstrap versions of $\widehat{h}_m\left(y, \theta_{0,m}^T x\right)$ and $\widetilde{h}_m\left(y, \theta_{0,m}^T x\right)$ based on the bootstrap pseudo sample i be, respectively, $\widehat{h}_m^{*(i)}\left(y, \widehat{\theta}_m^T x\right)$ and

$\widetilde{h}_m^{*(i)}\left(y, \widehat{\theta}_m^T x\right)$. Use also $\widetilde{h}_m\left(y, \widehat{\theta}_m^T x\right)$ as the bootstrap version of $h_{0,m}\left(y, \theta_{0,m}^T x\right)$. We then have

$$H^*\left(\widetilde{h}_m^*, \widehat{\theta}_m\right) = \frac{1}{B}\sum_{i=1}^{B}\left\{\frac{1}{n}\sum_{j=1}^{n}\log\left(\frac{\widetilde{h}_m^{*(i)}\left(y_j^{*(i)}, \widehat{\theta}^T x_j^{*(i)}\right)}{\widetilde{h}_m^{*(i)}\left(y_j, \widehat{\theta}^T x_j\right)}\right)\right\}, \tag{3.25}$$

and

$$\begin{aligned} J^*\left(\widetilde{h}_m^*, \widehat{\theta}_m\right) \quad = \quad & \frac{1}{B}\sum_{i=1}^{B}\left[\left\{\frac{1}{n}\sum_{j=1}^{n}\nabla\log\left(\frac{\widetilde{h}_m^{*(i)}\left(y_j^{*(i)}, \widehat{\theta}^T x_j^{*(i)}\right)}{\widetilde{h}_m^{*(i)}\left(y_j, \widehat{\theta}^T x_j\right)}\right)\right\}\right. \\ & \left.\times\left\{\frac{1}{n}\sum_{k=1}^{n}\nabla\log\left(\widehat{h}_m^{*(i)}\left(y_k^{*(i)}, \widehat{\theta}_m^T x_k^{*(i)}\right)\right)\right\}^T\right]. \end{aligned} \tag{3.26}$$

Equations (3.22)-(3.26) together define a feasible Information Criterion that is relatively easy to implement. Differentiation of $\widehat{\Omega}\left(\widehat{\theta}_m\right)$ and $J^*\left(\widetilde{h}_m^*, \widehat{\theta}_m\right)$ can be performed numerically.

The validity of the bootstrap method was proved for a wide range of statistical nonparametric applications that are close to ours (cf. Hall, Marron and Park 1992, Paparoditis and Politis 2000). However, their theory does not generalise easily to our case. At the same time, as far as we are aware, the asymptotic properties of the AIC (and AIC variants) were theoretically investigated for some particular parametric regression models (see a review by Rao and Wu 2001), but less so in semiparametric and nonparametric situations. Some exceptions include Hurvich, Simonoff and Tsai (1988) and Naik and Tsai (2001), who developed improved versions of the AIC criterion for nonparametric and semiparametric regression models and demonstrated numerically the effectiveness of their criteria. We thus very much regret that we could not show the asymptotic properties of the proposed IC_m-stopping rule, and we leave the theoretical properties of the proposed Information Criterion open for further research. The simulation results reported in the next section demonstrate that the IC_m-criterion performs very well in practice.

3.7 Numerical study

Some reported empirical studies have demonstrated that the PPDE generally outperforms standard kernel estimates (see Hwang, Lay and Lippman 1994). In this section we explore the finite-sample performance of our method using both simulated and real-data examples.

In all of the experiments we used the Triweight kernel,

$$K(u) = \max\left\{35/32 \cdot \left(1 - 3u^2 + 3u^4 - u^6\right), 0\right\}.$$

In order to facilitate the algorithm, we standardised the data by setting $x_j \leftarrow S_x^{-1}\left(x_j - \overline{x}\right)$ and $y_j \leftarrow \left(y_j - \overline{y}\right)/s_y$, where $\overline{x}$ and $\overline{y}$ are the vector and scalar sample means of $\{x_j\}_{j=1}^n$ and $\{y_j\}_{j=1}^n$, and S_x^2 and s_y^2 are the d-matrix and scalar sample variances. Once the PPCDE algorithm is completed, the final estimates of the projective directions and of the conditional density approximation, say $\widehat{f}_{Y|X}\left(y|x\right) = \widehat{g}_{Y|X,M}\left(y|x\right)$, can be transformed back to the original coordinates by setting $\widehat{\theta}_m \leftarrow S_x^{-1}\widehat{\theta}_m / \left\|S_x^{-1}\widehat{\theta}_m\right\|$ for any $m = 1, ..., M$, and

$$\widehat{f}_{Y|X}\left(y|x\right) \leftarrow \widehat{f}_{Y|X}\left(y|x\right)/s_y = \widehat{g}_{Y,0}\left(y\right)\widetilde{h}_1\left(y, \widehat{\theta}_1^T x\right)\cdots\widetilde{h}_M\left(y, \widehat{\theta}_M^T x\right)/s_y.$$

For relatively fast and robust numerical optimisation, we implemented the iterative procedure used in Chapter 2, which in practice also performs automatic bandwidth adjustments.

Step 0. Let $\widehat{\theta}_m^0 \in \Theta$ be any initial guess for $\theta_{0,m}$, for example $\widehat{\theta}_m^0 = (1, 0, ..., 0)$. Set also a finite sequence of decreasing bandwidths $h_y^i = h_x^i = a^i n^{-1/(p+2)}$, $i = 1, ..., I$, where p is the kernel-order and $a^i > 0$ is a decreasing sequence such that the first bandwidths notably oversmooth the conditional density. Our experience suggests that $\left(a^1, a^2, ..., a^I\right) = (9, 8, ..., 3)$ yield good results. Set the iteration number $i = 1$.

Step 1. Apply a multivariate variant of the Newton-Raphson method with starting point $\widehat{\theta}_m^{i-1}$ to find a maximum log-likelihood estimate $\widehat{\theta}_m^i$ numerically based on bandwidths h_y^i and h_x^i. As in Section 2.4, in our simulations we use the Broyden-Fletcher-Goldfarb-

Shanno BFGS method* (see Nocedal and Wright 2006, Chapter 6).

Step 2. Stop the procedure and use the estimate $\widehat{\theta}_m = \widehat{\theta}_m^i$ either if $i = I$ or if a certain convergence criterion is met, i.e. if $\left(\widehat{\theta}_m^i\right)^T \widehat{\theta}_m^{i-1} > 1 - \varepsilon$ for some small $\varepsilon > 0$. Otherwise, set $i \leftarrow i + 1$ and $h_y^i = h_x^i = a^i n^{-1/(p+2)}$, and return to Step 1.

Because a^1 is chosen to oversmooth the conditional density, the corresponding likelihood surface is oversmoothed as well, and the algorithm is insensitive to the choice of the $\widehat{\theta}_m^0$.

For the second-stage estimation during each iteration, when estimating $\widetilde{h}_m\left(y, \widehat{\theta}_m^T x\right)$, we adopted Scott's (1992) normal reference rule or bandwidth selection, which suggests $H_y = H_x = 3n^{-1/6}$ for the Triweight kernel (see discussion in Section 2.4).

The computational complexity of each iteration of the PPCDE algorithm, $m = 1, ..., M$ is of order $O\left(n^2 d^3\right)$ by the same considerations discussed in Section 2.4. However here, we also need to consider the computational complexity of the IC_m-stopping rule. Since computing the likelihood is of computational complexity of $O\left(n^2 d\right)$, the computation of $\widehat{\Omega}\left(\widehat{\theta}_m\right)$, which involves a second derivative matrix of the likelihood w.r. to $\theta \in \mathbb{R}^d$, requires $O\left(n^2 d^3\right)$ computational time. Similar considerations show that the computational complexity of the bootstrap estimates $H^*\left(\widetilde{h}_m^*, \widehat{\theta}_m\right)$ and $J^*\left(\widetilde{h}_m^*, \widehat{\theta}_m\right)$ is $O\left(Bn^2 d\right)$ and $O\left(Bn^2 d^2\right)$, where B is the number of bootstrap samples. Ignoring asymptotically insignificant terms, we thus get that the total complexity of the PPCDE algorithm is of order $O\left(Mn^2 d^2 \left(d + B\right)\right)$.

As in Chapter 2, in the simulations we used R 2.14.1 programme on a computer with 3.4ghz intel core i7-2600 processor. For example, the average computational times of the method (for a single estimation based on Example 3 below) with dimension $d = 4$, number of projection $M = 6$, $B = 50$ number of bootstrap samples and sample sizes $n = 100, 200$ and 400 were $28.4\,\text{sec}, 67.4\,\text{sec}$ and $238.5\,\text{sec}$, respectively.

An R code PPCDE.txt for the calculations below is available at

http://personal.lse.ac.uk/rosemari/

*The code for the algorithm was published by Daniel F. Heitjan

In the three first numerical examples listed below, we investigate the performance of the PPCDE for simulated multidimensional data. For each of these examples, 100 replications were generated with sample sizes $n = 100, 200$ and 400. In order to reduce the computational burden, we only produced results for the non-normalised PPCDE. As demonstrated in the previous sections, normalising the final estimates is expected to improve the performance of the PPCDE. The fourth example demonstrates an application of the method to interval predictors for the daily exchange-rate returns between the US Dollar (USD) and the British Pound (GBP). In all four examples, we tested the IC stopping rule with the number of bootstrap samples $B = 50$. In order to evaluate the quality of the performance of the PPCDE, the standard multivariate conditional density kernel estimator, referred to here simply as the kernel estimator, is set as a benchmark.

Example 1 (continued): Consider first the model of Example 1, already employed in previous sections. This model can be written as

$$y_t = \sqrt{x_{1t}^2 + x_{2t}^2} + \varepsilon_t, \quad t = 1, ..., n,$$

where x_1, x_2 and ε_t are independent $N(0,1)$. This model was selected for our first example as it is relatively simple, and because the information entailed by X can be fully specified by no more than two orthogonal projections, say $\theta^T X$ and $\theta_\perp^T X$, of X. As mentioned throughout the Chapter, one needs to bear in mind that there is no 'true' number of projective directions that we expect to be produced by the model. However still, for this simple model, the number of two projective directions can serve as a benchmark for the number of projective directions that should be selected by an efficient approximation. We applied the PPCDE procedure up to a maximum of $m = 5$ iterations. In addition, the Information Criterion (IC) stopping rule was used to select the number of iterations $m = 1, 2, ...,$ for each different realisation of the data, where as usual, the m'th iteration number refers to the m'th estimate $\widehat{g}_{Y|X,m}(y|x)$ obtained by modification of estimate $\widehat{g}_{Y|X,m-1}(y|x)$. Table 3.1 reports the frequency of the selected number of iterations chosen by this criterion out

	$M=0$	$M=1$	$M=2$	$M=3$	$M=4$	$M=5$
$n=100$	0	27	66	7	0	0
$n=200$	0	7	81	12	0	0
$n=400$	0	0	55	39	6	0

TABLE 3.1: *Stopping rule in Example 1*: Frequency of number of iterations (M) selected by the Information Criterion stopping rule (out of 100 Repetitions).

of the 100 Repetitions. For all simulated sample sizes, the most frequently chosen number of iterations is indeed $M = 2$. This may confirm suitability of the PPCDE algorithm and the IC stopping rule. Yet, we also note that the average selected number of iterations by IC is generally higher the more observations are given. At the next stage, we look at the $RMSPE$ of the model, given by

$$RMSPE = \sum_{i=1}^{n} \left[\widetilde{f}_{Y|\widehat{\theta}^T X} \left(y_i | \widehat{\theta}^T x_i \right) - f_{Y|X} \left(y_i | x_i \right) \right]^2 \Big/ \sum_{i=1}^{n} f_{Y|X} \left(y_i | x_i \right)^2 .$$

Thus, the $RMSPE$ is a measure of the fitted error with respect to the real conditional density of the model.

Figure 3.3 displays a box-plot of the $RMSPE$ (see (3.20)) of the PPCDE with the number of iterations ranging from $m = 0$ (unconditional density kernel estimator of Y) to $m = 5$. It also shows box-plots of the $RMSPE$ corresponding to the PPCDE obtained with a varying number of iterations selected by the IC stopping rule for each realisation of the data; to the standard kernel estimator; and to an ‘Oracle’ PPCDE with exactly two projective iterations at the fixed orthogonal directions $\theta_1 = (1, 0)^T$ and $\theta_2 = (0, 1)^T$. We see that for $n = 100$ the optimal number of iterations seems to be $m = 1$, while for $n = 200$ and 400 the optimal number is $m = 2$. As stated in Section 3.2, it is hard to interpret the estimates of of M, θ_m’s and h_m’s with respect to their ‘true’ values, as any such true values are not necessarily unique. Generally speaking, we see from the empirical results that for a small number of observations, using a low number of iterations is preferable in terms of performance to using a high number of iterations. Nonetheless, while increasing the number of observations can gradually decrease the variance of the estimators, the performance of the PPCDE based on too small a number of projective iterations is limited

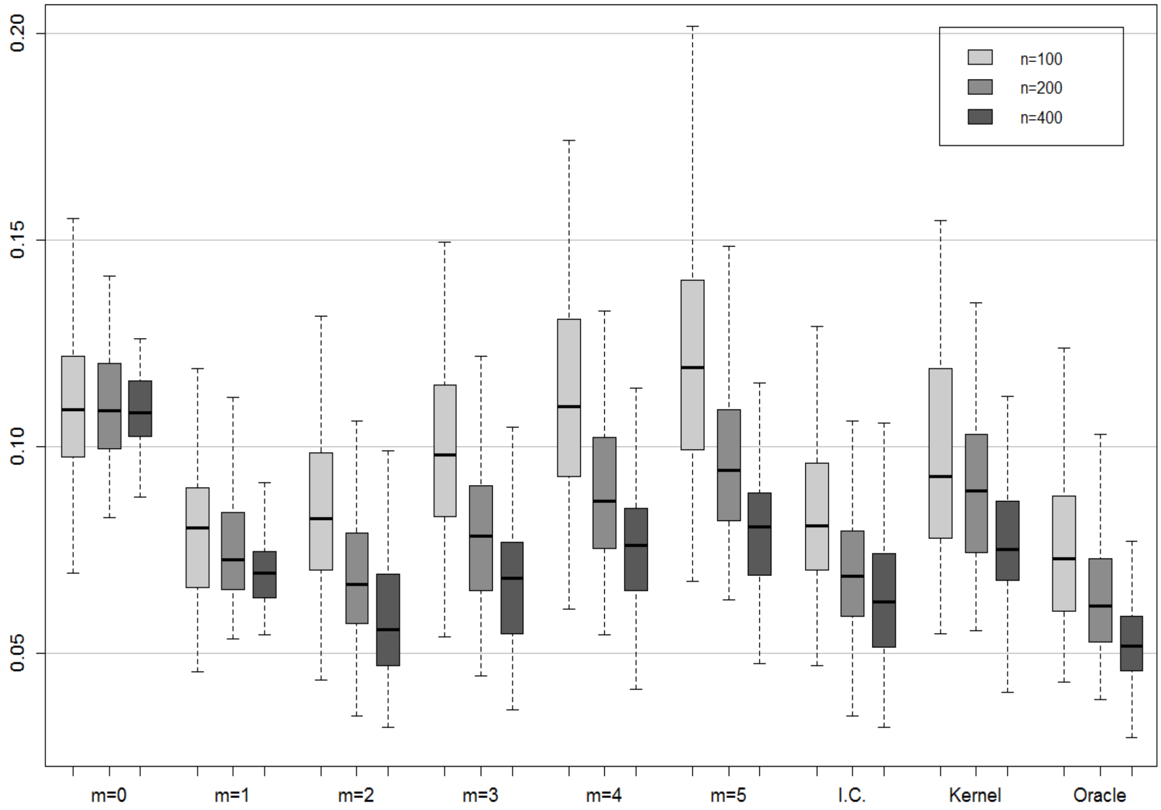


FIGURE 3.3. *Simulation results for Example 1:* Box-plots of the RMSPE obtained by PPCDE with fixed numbers of iterations ($m = 0, ..., 5$); Number of iterations determined by the Information Criterion stopping rule (I.C.); Multivariate conditional density kernel estimator (Kernel); 'Oracle' PPCDE with two iterations with $\theta_1 = (1, 0)^T$ and $\theta_2 = (0, 1)^T$ (Oracle).

by the model bias. On the other hand, the performance of those estimators based on a relative high number of iterations keeps improving with the number of observations, and such estimators would be preferable for large sample sizes. The results of Figure 3.3 also provide some empirical support for the effectiveness of the IC stopping rule. Indeed, when applying the IC stopping rule, the PPCDE generally achieves relatively low levels $RMSPE$. In agreement with our illustrative demonstration for Example 1 in Section 3.4, we see that the standard kernel estimator behaves badly both in terms of the high median and the high variability of the $RMSPE$. The 'Oracle' estimator is clearly the most accurate among all the estimators examined, since it is based on orthogonal projective directions rather than on estimated ones. However, compared to the PPCDE obtained with the same number of iterations, $m = 2$, the increase in accuracy owing to the utilisation of orthogonal projective directions, rather than estimated ones, is relatively small.

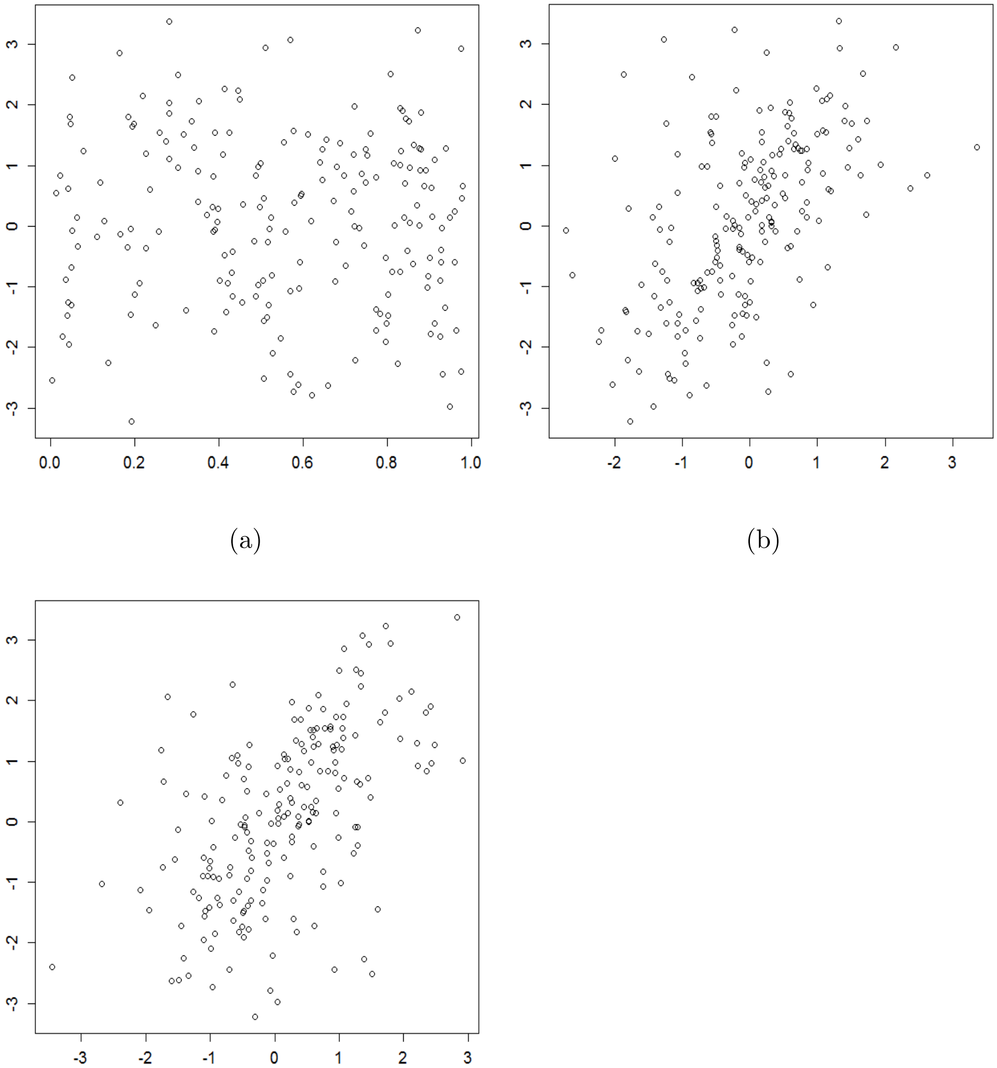
Example 2: We next examine how the PPCDE performs in a more complicated situation. Take $x_t = (x_{1t}, x_{2t}, x_{3t}) \in \mathbb{R}^3$ i.i.d. where x_{1t}, x_{2t} and x_{3t} are independent, $x_{1t} \sim U(0,1)$ and $x_{2t}, x_{3t} \sim N(0,1)$. We generate data y_t, $t = 1, ..., n$, according to the model

$$y_t = \begin{cases} 2\sin\left(\theta_2^T x_t\right) + 0.7\varepsilon_t, & \text{with probability } x_{1t}, \\ 2\sin\left(\theta_3^T x_t\right) + 0.7\varepsilon_t, & \text{with probability } 1 - x_{1t}, \end{cases}$$

where $\theta_2 = (0,2,1)^T/\sqrt{5}$, $\theta_3 = (0,1,-1)^T/\sqrt{2}$ and $\varepsilon_t \sim N(0,1)$ i.i.d. In this example, clearly the distribution of y_t is fully specified given the three projections $\theta_1^T x_t = x_{1t}$ and $\theta_2^T x_t$, $\theta_3^T x_t$. Figure 3.4 shows scatter-plots of y_t against $\theta_1^T x_t$, $\theta_2^T x_t$ and $\theta_3^T x_t$ with $n = 200$.

We now implement the PPCDE algorithm to a maximum of $m = 6$ iterations. Table 3.1 describes the frequency of the selected number of iterations chosen by the IC stopping rule out of the 100 Repetitions. Here, the number of iterations selected for each sample size is spread over a larger range than in Example 1, but as above, the average number of iterations selected generally shifts upwards the more observations are given.

Figure 3.5 gives box-plots for the $RMSPE$ of the PPCDE with number of iterations



(a)

(b)

(c)

FIGURE 3.4. *Scatter plots for Example 2*: (a) y_t plotted against $\theta_1^T x_t$; (b) y_t plotted against $\theta_2^T x_t$; (c) y_t plotted against $\theta_3^T x_t$.

	$M=0$	$M=1$	$M=2$	$M=3$	$M=4$	$M=5$	$M=6$
$n=100$	0	38	47	13	2	0	0
$n=200$	0	21	43	24	11	1	0
$n=400$	0	3	12	27	33	21	4

TABLE 3.1: *Stopping rule in Example 2*: Frequency of number of iterations (M) selected by the Information Criterion stopping rule (out of 100 Repetitions).

	$M=0$	$M=1$	$M=2$	$M=3$	$M=4$	$M=5$	$M=6$
$n=100$	0	55	42	3	0	0	0
$n=200$	0	16	71	11	2	0	0
$n=400$	0	8	61	25	6	0	0

TABLE 3.3: *Stopping rule in Example 3*: Frequency of number of iterations (M) selected by the Information Criterion stopping rule (out of 100 Repetitions).

$m = 0, ..., 6$; the PPCDE based on the IC stopping rule; and for the standard kernel estimator. Here again, the results support the IC stopping rule, and the optimal number of iterations shifts from $m = 2$ for $n = 100$ to $m = 3$ for $n = 400$, while the differences are very small. The PPCDE based on the IC stopping rule performs comparatively well, while the standard kernel estimator's general performance is the second worst after the unconditional kernel estimator.

Example 3: We apply the PPCDE to a time-series model. Consider now the nonlinear AR-ARCH model

$$y_t = g\left(\sum\nolimits_{j=1}^{4} \theta_{1,j} y_{t-j}\right) + h\left(\sum\nolimits_{j=1}^{4} \theta_{2,j} y_{t-j}\right) \varepsilon_t,$$

where $g(u) = 0.3\left(0.8 - u^2\right) / \left(0.2 + u^2\right)$, $h(u) = \sqrt{0.2 + 0.3u^2}$, $\theta_1^T = (1, -2, 1, 0) / \sqrt{6}$, $\theta_{2,j} = \exp(-j) / \|\theta_2\|$ for $j = 1, ..., 4$, and $\varepsilon_t \sim N(0, 1)$ i.i.d. Our goal here is to estimate the predictive density $f_{Y|\theta^T x}(y_t|x_t)$ of y_t given the 4-dimensional lagged observations $x_t = (y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4})$. This model can be verified to be geometrically ergodic by, e.g., Theorem 3.2 of An and Huang (1996). Figure 3.6 shows a plot of one realisation of the time-series model, and the corresponding scatter-plots of y_t against $\theta_1^T x_t$ and $\theta_2^T x_t$.

We implement the PPCDE procedure up to $m = 6$ iterations. Table 3.3 describes the frequency of the selected number of iterations chosen by the IC stopping rule. For

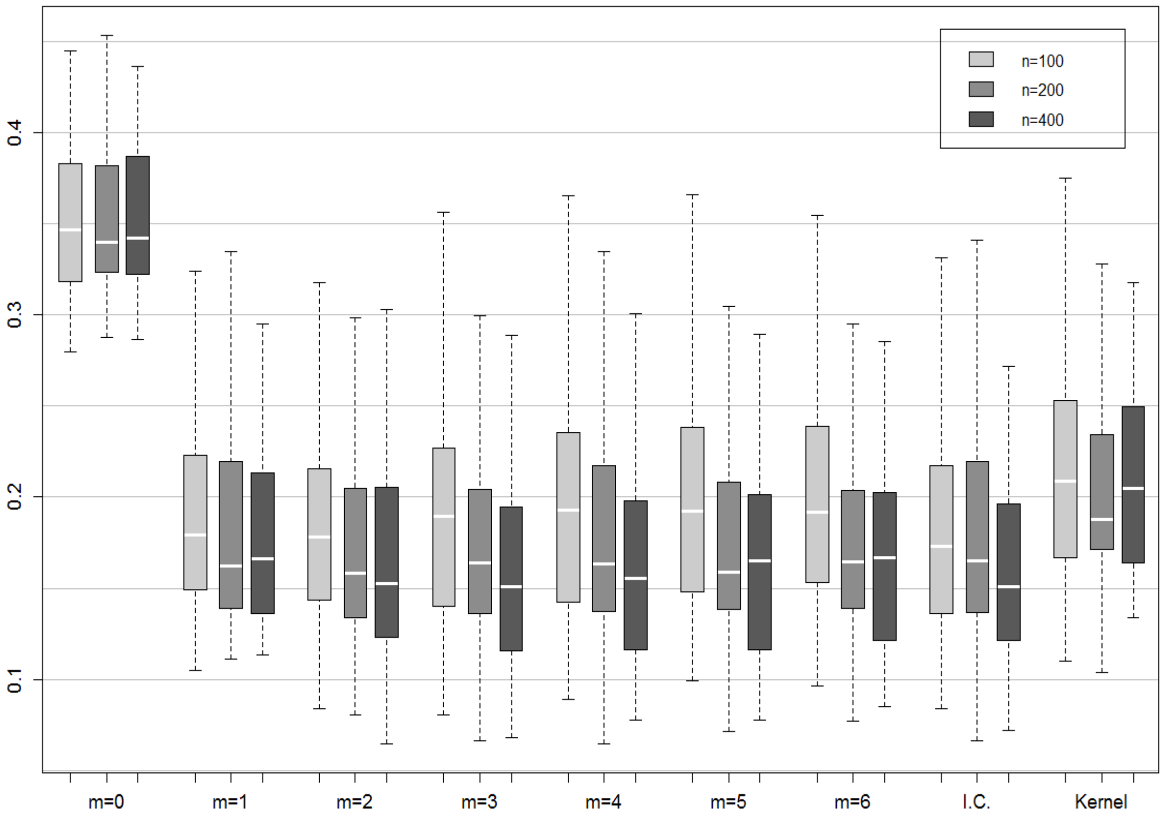
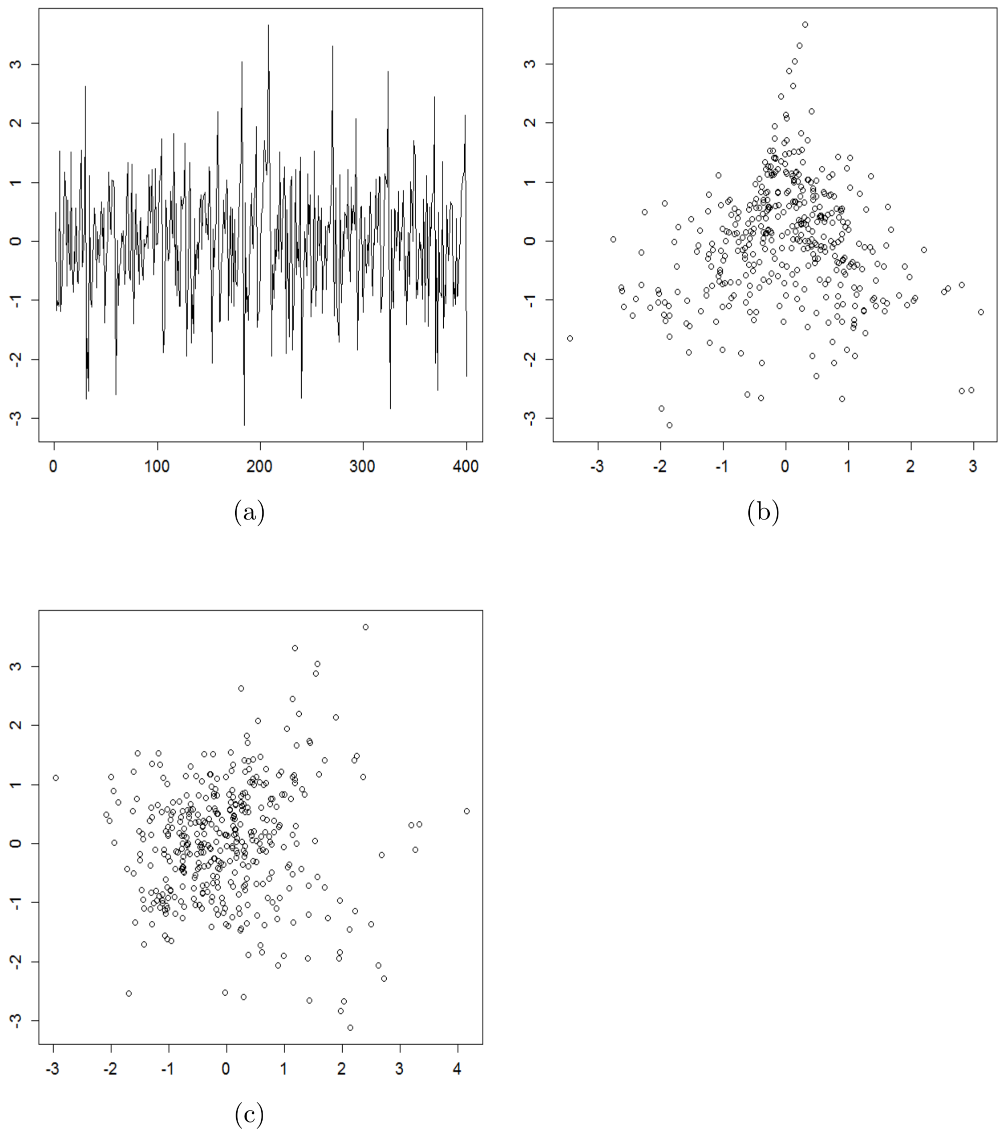


FIGURE 3.5. *Simulation results for Example 2.* Box-plots of the RMSPE obtained by PPCDE with fixed numbers of iterations ($m = 0, ..., 6$); Number of iterations determined by the Information Criterion stopping rule (I.C.); Multivariate Conditional density kernel estimator (Kernel);



(a)

(b)

(c)

FIGURE 3.6. *Plots for Example 3:* (a) y_t plotted against t; (b) Scatter plot of y_t against $\theta_1^T x_t$; (c) Scatter plot of y_t against $\theta_2^T x_t$.

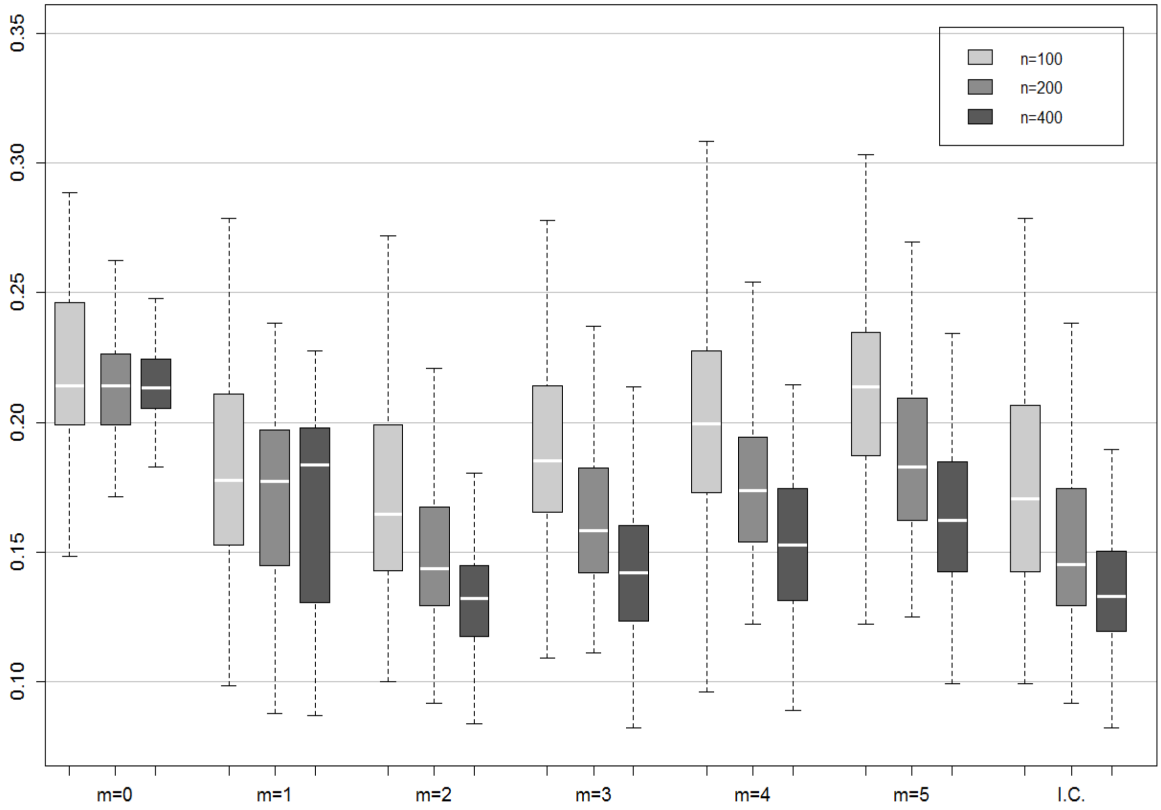


FIGURE 3.7. *Simulation results for Example 3.* Box-plots of the RMSPE obtained by PPCDE with fixed numbers of iterations ($m = 0, ..., 5$); Number of iterations determined by the Information Criterion stopping rule (I.C.).

$n = 100$ the chosen number of iterations is distributed almost equally between $m = 1$ and $m = 2$, where the frequency for $m = 1$ is somewhat higher, while for $n = 200$ and 400 the most frequently chosen number of iterations is $m = 2$. Figure 3.7 gives box-plots for the $RMSPE$ of the PPCDE with number of iterations $m = 0, ..., 6$, and of the PPCDE based on the IC stopping rule. We do not provide here a box-plot for the $RMSPE$ of standard kernel estimator since it is significantly higher than the rest. The pattern here is similar to the last two examples. However, the optimal number of iterations seems to be $m = 2$ for all numbers of observations examined, while the IC stopping rule seemed to favour $m = 1$ for the small number of observations $n = 100$. Notwithstanding, the distribution of the

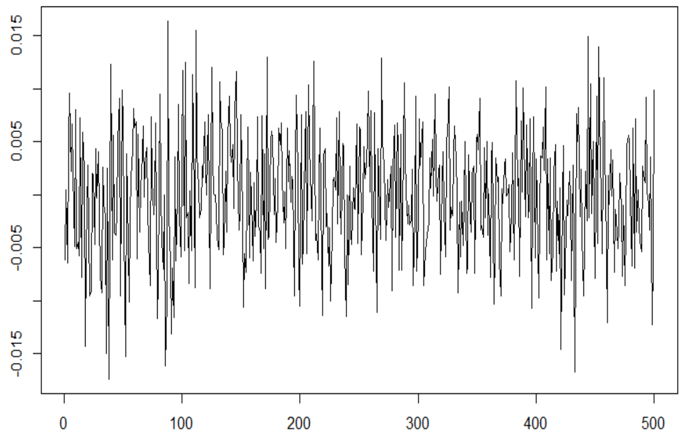


FIGURE 3.8. *Example 4:* Daily exchange rate returns of the USD-GDP between 04/01/2010 and 30/12/2011.

PPCDE based on the IC stopping rule is very close to the optimal one, i.e. when $m = 2$.

Example 4: Finally, we apply the PPCDE to real data. We use a time-series of the daily exchange-rate returns between the US Dollar (USD) and the British Pound (GBP) between 4 January 2010 and 30 December 2011. The data consists of 501 data points, out of which we allocate the last 100 points for prediction. Figure 3.8 presents the time-series data over the full period. As in the time-series Example 3, we implement the PPCDE approach to estimate the predictive density $f_{Y|\theta^T x}(y_t|x_t)$ of y_t given the 4-lagged data $(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4})$. Using only the first 401 data points, we estimate first the projective directions, while the IC stopping rule is applied to determine the number of projections. Two projections are selected, and the estimated projective directions are, in order of selection, $\widehat{\theta}_1 = (0.426, 0.779, -0.325, 0.324)$, $\widehat{\theta}_2 = (0.453, 0.260, 0.852, -0.025)$. For the sake of comparison with the standard multivariate kernel estimator, we continue and also produce the next two estimated projective directions, $\widehat{\theta}_3 = (0.548, -0.456, -0.610, -0.341)$ and $\widehat{\theta}_4 = (0.363, 0.148, -0.623, -0.677)$. Here again, it is hard to interpret the resulted estimates of of θ_m's (see Section 3.2). Next, for any observation y_t that belongs to the last 100 observations, we iteratively construct a predictive density model using the estimated optimal projections $\widehat{\theta}_j^T x_t$, $j = 1, ..., 4$, where all nonparametric functional estimators rely

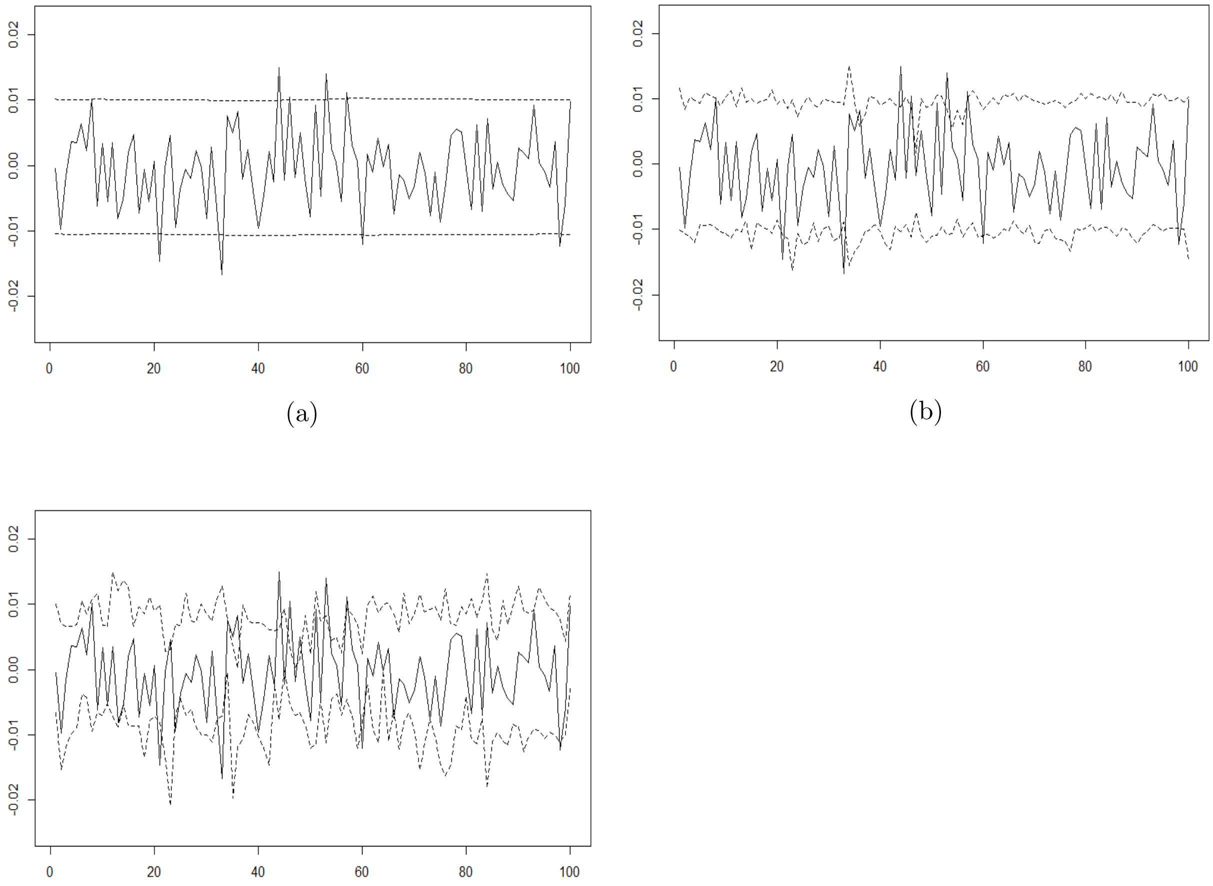
Model	$\alpha = 1\%$		$\alpha = 5\%$		$\alpha = 10\%$		$\alpha = 25\%$	
	Cover.	Length	Cover.	Length	Cover.	Length	Cover.	Length
Uncond.	0.99	32.1	0.96	24.43	0.91	20.61	0.78	14.79
$m = 1$	0.97	31.21	0.94	24.25	0.90	20.48	0.76	14.73
$m = 2$	0.98	30.31	0.93	24.1	0.90	20.21	0.75	14.46
$m = 3$	0.97	29.1	0.93	23.26	0.89	19.67	0.76	14.15
$m = 4$	0.97	29.0	0.93	23.38	0.90	19.88	0.75	14.16
Kernel	0.95	26.89	0.87	19.88	0.81	17.23	0.57	12.74

TABLE 3.4: *Results for Example 4*: Prediction coverage (%) and avg. length ($\cdot 10^3$) of $(1-\alpha)$ −prediction intervals.

on past information $y_1, ..., y_{t-1}$ (that may include some past observations from the last 100 data points). Finally, we also normalise all of our predictive density models such that $\int \widehat{g}_{Y|X}\,(y|x)\,dy = 1$.

In order to examine the predictive capability of the models, we construct the corresponding $(1-\alpha)$ −prediction intervals for any y_t in the last 100 observations. For comparison, we also construct $(1-\alpha)$ −prediction intervals for the standard multivariate kernel estimator. Table 3.4 gives the prediction coverage (% of observations y_t that fall inside the prediction interval) and the average length of the prediction interval over the last 100 observations for all obtained models with $\alpha = 1\%,\ 5\%,\ 10\%$ and 25%. Also, for visual illustration, Figure 3.9 shows plots of the last 100 observations and the corresponding 90%−prediction interval obtained by each model.

For all of the confidence level values examined, the unconditional density estimator produced the widest prediction-intervals on average. In terms of prediction coverage, both the unconditional density estimator, the conditional density based on the most recent lag and the single-index conditional density based on the orientation estimate provide relatively accurate estimates, while the standard conditional density kernel estimator has much less similar to reality. At the same time, the single-index conditional density generally produced narrow prediction-intervals on average. We thus conclude that the single-index approach for c.p.d.f approximation manages to provide increased accuracy and predictive power relative to other standard kernel methods.



(a)

(b)

(c)

FIGURE 3.9. *Example 4*: 90%−prediction intervals for the daily USD-GDP exchange-rate returns between 19/10/2010 and 30/12/2011 based on (a) Unconditional kernel density estimator; (b) PPCDE with two projections; (c) Multivariate conditional density kernel estimator.

3.8 Appendix

Proof of Proposition 3.2.1. The proofs for parts (a) and (b) are straightforward by applying the inequality

$$\log x \leq x - 1 \quad \text{for every } x \geq 0, \tag{3.27}$$

with equality only for $x = 1$. Using $f_{Y,X}(y,x)/f_{Y|X}(y|x) = f_X(x)$ and $\int f_{Y,X}(y,x)\,dydx = 1$, this shows that

$$\begin{aligned}
& -D_C[g_{Y|X}] \\
\leq\ & \int\limits_{\mathbb{S}} \frac{g_{Y|X}(y|x)}{f_{Y|X}(y|x)} f_{Y,X}(y,x)\,dydx - \int\limits_{\mathbb{S}} f_{Y,X}(y,x)\,dydx \\
& + \left(1 - \int\limits_{\mathbb{R}\times\mathbb{S}_X} g_{Y|X}(y|x) f_X(x)\,dydx\right) \\
\leq\ & 0,
\end{aligned}$$

with equalities in both lines iff $g_{Y|X} = f_{Y|X}$ a.e. for any $(y,x) \in \mathbb{R}\times\mathbb{S}_X$. For (b), we have

$$\begin{aligned}
& D_C[g^*_{Y|X}] - D_C[g_{Y|X}] \qquad (3.28) \\
=\ & \int\limits_{\mathbb{S}} \log\left(\frac{g_{Y|X}(y|x)}{g^*_{Y|X}(y|x)}\right) f_{Y,X}(y,x)\,dydx \\
& + \left(\int\limits_{\mathbb{R}\times\mathbb{S}_X} \left\{g^*_{Y|X}(y|x) - g_{Y|X}(y|x)\right\} f_X(x)\,dydx\right) \\
\leq\ & \int\limits_{\mathbb{S}} \frac{g_{Y|X}(y|x)}{g^*_{Y|X}(y|x)} f_{Y,X}(y,x)\,dydx - \int\limits_{\mathbb{S}} f_{Y,X}(y,x)\,dydx \\
& + \left(\int\limits_{\mathbb{R}\times\mathbb{S}_X} \left\{g^*_{Y|X}(y|x) - g_{Y|X}(y|x)\right\} f_X(x)\,dydx\right).
\end{aligned}$$

Using now the properties that

$$\frac{g_{Y|X}\left(y|x\right)}{g^{*}_{Y|X}\left(y|x\right)}=\int\limits_{\mathbb{R}\times\mathbb{S}_X} g_{Y|X}\left(y|x\right)f_X\left(x\right)dydx,$$

and

$$\begin{aligned}\int\limits_{\mathbb{S}} f_{Y,X}\left(y,x\right)dydx &= 1,\\ \int\limits_{\mathbb{R}\times\mathbb{S}_X} g^{*}_{Y|X}\left(y|x\right)f_X\left(x\right)dydx &= 1,\end{aligned}$$

and it is easy to see that (3.28) is non-positive, and is zero iff $g_{Y|X}=g^{*}_{Y|X}$ a.e. for any $(y,x)\in\mathbb{S}$, that is, iff $\int g_{Y|X}\left(y|x\right)f_X\left(x\right)dydx=1$.

For assertion (c), denote $f=f_{Y|X}\left(y|x\right)$, $g=g_{Y|X}\left(y|x\right)$. The left inequality of the assertion is trivial as

$$\left(\sqrt{f}-\sqrt{g}\right)^2\leq\left|\left(\sqrt{f}-\sqrt{g}\right)\left(\sqrt{f}+\sqrt{g}\right)\right|=\left|f-g\right|.$$

For the right inequality of the assertion, denote $G\left(x\right)\equiv\int_{\mathbb{R}}dG_x\left(y\right)=\int_{\mathbb{R}}g\left(y|x\right)dy$ and $k\equiv k\left(y,x\right)=f\left(y|x\right)/g\left(y|x\right).$ We now have, with $F_X\left(x\right)$ the distribution function of r.v. X,

$$\left(\int dG_x\left(y\right)dF_X\left(x\right)\right)=\int g_{Y|X}\left(y|x\right)f_X\left(x\right)dydx=c. \tag{3.29}$$

Following Kemperman (1969, Theorem 6.1), we note that for any $k\geq0$,

$$\left(k-1\right)^2\leq\frac{2}{3}(k+2)\left[k\log k+(1-k)\right], \tag{3.30}$$

with equality iff $k=1$. This last inequality holds since if $\alpha\left(k\right)$ is equal to the RHS minus the LHS, we have by straightforward calculations,

$$\alpha\left(k\right)=\frac{2}{3}(k+2)\left[k\log k+(1-k)\right]-\left(k-1\right)^2,$$

$$\alpha'(k) = \frac{4}{3}(k+1)\log k - \frac{8}{3}(k-1), \quad \text{and} \quad \alpha''(k) = \frac{4}{3}\left(\log k + \frac{1}{k} - 1\right).$$

Now, it is easy to see that $\alpha(1) = 0$, $\alpha'(1) = 0$, and by (3.27) that $\alpha''(k) \geq 0$ for any $k > 0$. Applying the Cauchy-Schwarz inequality in the $L^2\left(\mathbb{R}^2, dG_x(y)\,dF_X(x)\right)$ space and using (3.29), (3.30), yields, for any $(y,x) \in \mathbb{R} \times \mathbb{S}_X$,

$$\begin{aligned}
&\left(\int |f-g|\, f_X dydx\right)^2 \\
\leq\ &\left(\int |k-1|\, dG_x(y)\, dF_X(x)\right)^2 \\
\leq\ &\frac{2}{3}\int (f_{Y|X} + 2g_{Y|X}) dydF_X(x) \int \left[f_{Y|X}\log\left(f_{Y|X}/g_{Y|X}\right) + \left(g_{Y|X} - f_{Y|X}\right)\right] dydF_X(x) \\
=\ &\frac{2}{3}(1+2c)D_C[g_{Y|X}].
\end{aligned}$$

Finally, we prove (d). Since $D_C[g_{Y|X}]$ is minimised by $g^{opt}_{Y|X}(y|x)$, we have by assertion (b) of the Proposition that

$$\int g^{opt}_{Y|X}(y|x)\, f_X(x)\, dydx = 1,$$

and clearly also

$$\int g^{*opt}_{Y|X}(y|x)\, f_X(x)\, dydx = \int f_X(x)\, dx = 1.$$

Then, with (3.27),

$$
\begin{aligned}
& D_C[g^{*opt}_{Y|X}\left(y|x\right)] - D_C[g^{opt}_{Y|X}\left(y|x\right)] \\
= & \int\limits_{\mathbb{S}} \log\left(\int g^{opt}_{Y|X}\left(y|x\right)dy\right) f_{Y,X}\left(y,x\right)dydx \\
= & \int\limits_{\mathbb{S}_X} \log\left(\int g^{opt}_{Y|X}\left(y|x\right)dy\right) f_X\left(x\right)dx \\
\leq & \int\limits_{\mathbb{S}_X} \left(\int g^{opt}_{Y|X}\left(y|x\right)dy - 1\right) f_X\left(x\right)dx \\
= & \int\limits_{\mathbb{R}\times\mathbb{S}_X} g^{opt}_{Y|X}\left(y|x\right) f_X\left(x\right)dydx - 1 \\
= & \ 0,
\end{aligned}
$$

with equality iff $\int g^{opt}_{Y|X}\left(y|x\right)dy = 1$ a.e. for any $x \in \mathbb{S}_X$. ■

Proof of Proposition 3.3.1. Let

$$
g_{Y|X,m}\left(y|x\right) = g_{Y|X,m-1}\left(y|x\right)h_m\left(y,\theta_m^T x\right),
$$

where $h_m\left(y,\theta_m^T x\right)$ is a non-negative bivariate function, and

$$
\begin{aligned}
g^{opt}_{Y|X,m}\left(y|x\right) &= g_{Y|X,m-1}\left(y|x\right)h_{0,m}\left(y,\theta_m^T x\right) \\
&= g_{Y,X,m-1}\left(y,x\right)f_{Y,\theta_m^T X}\left(y,\theta_m^T x\right)/g_{Y,\theta_m^T X,m-1}\left(y,\theta_m^T x\right).
\end{aligned}
$$

Applying inequality (3.27), Fubini's theorem (Gut 2005, Chapter 2,Theorem 9.1) and (3.8)-(3.9) (and keeping the convention that $x_1 = \theta_m^T x$), we have

$$
\begin{aligned}
& D_C[g^{opt}_{Y|X,m}] - D_C[g_{Y|X,m}] \\
& = \int_{\mathbb{S}} \log\left\{h_m\left(y,\theta_m^T x\right)g_{Y,\theta_m^T X,m-1}\left(y,\theta_m^T x\right)/f_{Y,\theta_m^T X}\left(y,\theta_m^T x\right)\right\} f_{Y,X}\left(y,x\right)dydx \\
& + \int_{\mathbb{R}\times\mathbb{S}_X} g_{Y,X,m-1}\left(y,x\right)\left[f_{Y,\theta_m^T X}\left(y,\theta_m^T x\right)/g_{Y,\theta_m^T X,m-1}\left(y,\theta_m^T x\right) - h_m\left(y,\theta_m^T x\right)\right] f_X\left(x\right)dydx \\
& \leq \int_{\mathbb{S}_Y\times\mathbb{S}_{X_1}} h_m\left(y,\theta_m^T x\right)g_{Y,\theta_m^T X,m-1}\left(y,\theta_m^T x\right)dydx_1 - \int_{\mathbb{S}_Y\times\mathbb{S}_{X_1}} f_{Y,\theta_m^T X}\left(y,\theta_m^T x\right)dydx \\
& + \int_{\mathbb{R}\times\mathbb{S}_{X_1}} f_{Y,\theta_m^T X}\left(y,\theta_m^T x\right)dydx_1 - \int_{\mathbb{R}\times\mathbb{S}_{X_1}} h_m\left(y,\theta_m^T x\right)g_{Y,\theta_m^T X,m-1}\left(y,\theta_m^T x\right)dydx_1
\end{aligned}
$$

$\leq 0,$

with equalities iff $h_m\left(y, \theta_m^T x\right) = h_{0,m}\left(y, \theta_m^T x\right)$ a.e. for any $(y, x) \in \mathbb{R} \times \mathbb{S}_X$ such that $g_{Y,\theta_m^T X,m-1}\left(y, \theta_m^T x\right) > 0$. This proves part (a) of the Proposition.

For part (b), we have

$$E\left[g_{Y|X,m-1}\left(y|X\right)|\theta_m^T X = \theta_m^T x\right] = \int g_{Y|X,m-1}\left(y|x\right) f_X\left(x_2, ..., x_d|x_1 = \theta_m^T x\right) dx_2 \cdots dx_d,$$

where $x = \left(\theta_m^T x, x_2, ..., x_d\right)$. Writing now

$$\begin{aligned} f_X\left(x_2, ..., x_d|x_1 = \theta_m^T x\right) &= f_X\left(\theta_m^T x, x_2, ..., x_d\right) / f_{\theta_m^T X}\left(\theta_m^T x\right) \\ &= f_X\left(x\right) / f_{\theta_m^T X}\left(\theta_m^T x\right), \end{aligned}$$

yields

$$\begin{aligned} & E\left[g_{Y|X,m-1}\left(y|X\right)|\theta_m^T X = \theta_m^T x\right] \\ = & \int g_{Y|X,m-1}\left(y|x\right) f_X\left(x\right) dx_2 \cdots dx_d \bigg/ f_{\theta_m^T X}\left(\theta_m^T x\right) \\ = & g_{Y|\theta_m^T X,m-1}\left(y|\theta_m^T x\right), \end{aligned}$$

which implies the equality given in part (b).

Continue with part (c). By Proposition 3.2.1, the optimality of $g_{Y|X,m}^{opt}$ implies that it must satisfy the integrability condition

$$\int g_{Y|X,m}^{opt}\left(y|x\right) f_X\left(x\right) dy dx = 1.$$

Since $g_{Y|X,m-1}\left(y|x\right)$ is assumed to fulfil a similar integrability condition, then we derive by a direct calculation

$$\begin{aligned} D^*[g_{Y|\theta_m^T X,m-1}] &= \int \log\left(\frac{g_{Y|X,m}^{opt}(y|x)}{g_{Y|X,m-1}(y|x)}\right) f_{Y,X}(y,x)\,dydx \\ &= \int \log\left(h_{0,m}\left(y,\theta_m^T x\right)\right) f_{Y,X}(y,x)\,dydx \\ &= \int \log\left(\frac{f_{Y,\theta_m^T X}\left(y,\theta_m^T x\right)}{g_{Y,\theta_m^T X,m-1}\left(y,\theta_m^T x\right)}\right) f_{Y,X}(y,x)\,dydx \\ &= \int \log\left(\frac{f_{Y,\theta_m^T X}(y,x_1)}{g_{Y,\theta_m^T X,m-1}(y,x_1)}\right) f_{Y,\theta_m^T X}(y,x_1)\,dydx_1 \\ &= \int \log\left(\frac{f_{Y|\theta_m^T X}(y|x_1)}{g_{Y|\theta_m^T X,m-1}(y|x_1)}\right) f_{Y,\theta_m^T X}(y,x_1)\,dydx_1. \quad \blacksquare \end{aligned}$$

Proof of Lemma 3.3.1. The 'only if' direction is trivial, since with the standard convention that $\theta_m^T x = x_1$, we have for $x \in \mathbb{S}_X$,

$$F_{Y|\theta^T X}\left(y|\theta^T x\right) = \frac{\int F_{Y|X}(y|x)\, f_X(x)\, dx_2 \cdots dx_d}{\int f_X(x)\, dx_2 \cdots dx_d},$$

where $x = \left(\theta_m^T x, x_2, ..., x_d\right)$. Thus, if $F_{Y_m|X}(y,x) \to F_{Y|X}(y,x)$ a.s. with respect to the probability measure of r.v. X, then

$$\begin{aligned} &\int \left|F_{Y_m|\theta^T X}\left(y|\theta^T x\right) - F_{Y|\theta^T X}\left(y|\theta^T x\right)\right| f_X(x)\, dx_1 \cdots dx_d \\ \leq\ & \frac{\int \left|F_{Y_m|X}(y|x) - F_{Y_m|X}(y|x)\right| f_X(x)\, dx_1 \cdots dx_d}{\int f_X(x)\, dx_2 \cdots dx_d} \to 0, \end{aligned}$$

as $m \to \infty$, which implies that $F_{Y_m|\theta^T X}\left(y|\theta^T x\right) \to F_{Y|\theta^T X}\left(y|\theta^T x\right)$ a.s. with respect to the probability measure of r.v. $\theta^T X$.

We now turn to prove the 'if' direction. Let $\psi_U(\cdot)$ denote a generic characteristic function (ch.f.), or conditional characteristic function (c.ch.f.), of r.v. U. For example the ch.f. of (Y, X) is

$$\psi_{Y,X}(s,v) = E\left[e^{isY + iv^T X}\right]$$

with $s \in \mathbb{R}$, $v \in \mathbb{R}^d$, and for $x \in \mathbb{S}_X$ the c.ch.f. of Y given $X = x$ is

$$\psi_{Y|X}\left(s\right) = E\left[e^{isY}|X = x\right],$$

etc. Assume now that $Y_m|\theta^T X \to_d Y|\theta^T X$ a.s. for any $\theta \in \mathbb{R}^d$. By the Levy-Cramér continuity theorem (cf. Shao 2003, Theorem 1.9) we have for any $\theta \in \mathbb{R}^d$,

$$P_{\theta^T X}\left(\psi_{Y_m|\theta^T X}\left(s\right) - \psi_{Y|\theta^T X}\left(s\right) \to 0 \text{ for all } s \in \mathbb{R}\right) = 1,$$

as $m \to \infty$, where the probability $P_{\theta^T X}\left(\cdot\right)$ denotes here the marginal probability measure induced by r.v. $\theta^T X$. Since $e^{is\theta^T X}$ is bounded, we obtain for any $\theta \in \mathbb{R}^d$ and $s, t \in \mathbb{R}$,

$$\begin{aligned} & \psi_{Y_m,\theta^T X}\left(s,t\right) - \psi_{Y,\theta^T X}\left(s,t\right) \\ = \ & E\left[e^{isY_m + it\theta^T X} - e^{isY + it\theta^T X}\right] \\ = \ & E\left[\left(\psi_{Y_m|X}\left(s\right) - \psi_{Y|X}\left(s\right)\right)e^{it\theta^T X}\right] \to 0. \end{aligned}$$

Using the last result with the identity $\psi_{Y,X}\left(s,v\right) = \psi_{Y,v^T X}\left(s,1\right)$, we obtain for any $s \in \mathbb{R}$ and $v \in \mathbb{R}^d$,

$$\begin{aligned} & \psi_{Y_m,X}\left(s,v\right) - \psi_{Y,X}\left(s,v\right) \\ = \ & E\left[\left(\psi_{Y_m|X}\left(s\right) - \psi_{Y|X}\left(s\right)\right)e^{iv^T x}\right] \\ = \ & E\left[\Psi_m\left(s\right)e^{iv^T x}\right] \to 0. \end{aligned} \tag{3.31}$$

where $\Psi_m\left(s\right) \equiv \psi_{Y_m|X}\left(s\right) - \psi_{Y|X}\left(s\right)$. Now, the proof of the Lemma will be completed by the Levy-Cramér continuity theorem if we establish that (3.31) implies

$$P_X\left(\Psi_m\left(s\right) \to 0 \text{ for all } s \in \mathbb{R}\right) = 1, \tag{3.32}$$

where $P_X\left(\cdot\right)$ is the marginal probability measure induced by r.v. X. The rest of the proof is dedicated to establishing (3.32), which can be seen as a limit version of Lemma 2.1 of Su and White (2007).

Let $\Psi_{m,1}(s) = \max(\operatorname{Re}(\Psi_m(s)), 0)$, $\Psi_{m,2}(s) = \max(-\operatorname{Re}(\Psi_m(s)), 0)$, $\Psi_{m,3}(s) = \max(\operatorname{Im}(\Psi_m(s)), 0)$ and $\Psi_{m,4}(s) = \max(-\operatorname{Im}(\Psi_m(s)), 0)$. Clearly, $\Psi_{m,j}(s)$, $j = 1, ..., 4$, are non-negative, Borel measurable, real functions on $\mathbb{R}$ such that

$$\operatorname{Re}(\Psi_m) = \Psi_{m,1} - \Psi_{m,2}, \quad \operatorname{Im}(\Psi_m) = \Psi_{m,3} - \Psi_{m,4}. \tag{3.33}$$

Set $c_{m,j}(s) = E_X[\Psi_{m,j}(s)]$, and assume for now that $c_{m,j}(s) > 0$, $j = 1, ..., 4$. Thus, for any $s \in \mathbb{R}$, we can define the four conditional probability measures (given $X = x$),

$$\nu_{m,j}(B; s) = \int_B \Psi_{m,j}(s)\, dF_X(x) \bigg/ c_{m,j}(s), \quad j = 1, ..., 4, \tag{3.34}$$

where B is an arbitrary Borel set in $\mathbb{R}^d$. We get for any $s \in \mathbb{R}$, $v \in \mathbb{R}^d$,

$$\begin{aligned}
& E_X\left[\Psi_m(s)\, e^{iv^T x}\right] \\
= & \int (\Psi_{m,1}(s) - \Psi_{m,2}(s))\, e^{iv^T x} dF_X(x) + i \int (\Psi_{m,3}(s) - \Psi_{m,4}(s))\, e^{iv^T x} dF_X(x) \\
= & c_{m,1}(s) \int e^{iv^T x} d\nu_{m,1}(B; s) - c_{m,2}(s) \int e^{iv^T x} d\nu_{m,2}(B; s) \\
& + i\left[c_{m,3}(s) \int e^{iv^T x} d\nu_{m,3}(B; s) - c_{m,4}(s) \int e^{iv^T x} d\nu_{m,4}(B; s)\right] \\
\equiv & c_{m,1}(s)\, \phi_{m,1}(v; s) - c_{m,2}(s)\, \phi_{m,2}(v; s) \\
& + i\left[c_{m,3}(s)\, \phi_{m,3}(v; s) - c_{m,4}(s)\, \phi_{m,4}(v; s)\right],
\end{aligned}$$

where $\phi_{m,j}(v; s) = \int e^{iv^T x} d\nu_{m,j}(B; s)$, $j = 1, ..., 4$, are the c.ch.f.'s generated by the conditional probability measures $\nu_{m,j}(B; s)$, respectively. Then it follows from (3.31) that as $m \to \infty$

$$\begin{aligned}
c_{m,1}(s)\, \phi_{m,1}(v; s) - c_{m,2}(s)\, \phi_{m,2}(v; s) &\to 0, \\
c_{m,3}(s)\, \phi_{m,3}(v; s) - c_{m,4}(s)\, \phi_{m,4}(v; s) &\to 0.
\end{aligned}$$

Substituting $v = \mathbf{0} \in \mathbb{R}^d$, we obtain for any $s \in \mathbb{R}$,

$$c_{m,1}(s) - c_{m,2}(s) \to 0, \text{ and } c_{m,3}(s) - c_{m,4}(s) \to 0, \tag{3.35}$$

and since $\left|\phi_{m,j}(v;s)\right| \leq 1$ for all $s \in \mathbb{R}$, $v \in \mathbb{R}^d$, $j = 1, ..., 4$,

$$\begin{aligned} \int e^{iv^T x} d\nu_{m,1}(B;s) - \int e^{iv^T x} d\nu_{m,2}(B;s) &\to 0, \\ \int e^{iv^T x} d\nu_{m,3}(B;s) - \int e^{iv^T x} d\nu_{m,4}(B;s) &\to 0, \end{aligned}$$

as $m \to \infty$. As a result of Levy-Cramér continuity theorem we get for any $s \in \mathbb{R}$ and Borel set $B \in \mathbb{R}^d$,

$$\nu_{m,1}(B;s) - \nu_{m,2}(B;s) \to 0, \text{ and } \nu_{m,3}(B;s) - \nu_{m,4}(B;s) \to 0. \tag{3.36}$$

From (3.33)-(3.36) we have for all $s \in \mathbb{R}$,

$$\int_B \operatorname{Re}(\Psi_m(s))\, dF_X(x) \to 0, \quad \int_B \operatorname{Im}(\Psi_m(s))\, dF_X(x) \to 0.$$

Applying the left limit result with Borel sets

$$B_1 = \{s \in \mathbb{R} : \operatorname{Re}(\Psi_m(s)) > 0\}, \ B_2 = \{s \in \mathbb{R} : \operatorname{Re}(\Psi_m(s)) < 0\},$$

and the right limit result with Borel sets

$$B_2 = \{s \in \mathbb{R} : \operatorname{Im}(\Psi_m(s)) > 0\}, B_4 = \{s \in \mathbb{R} : \operatorname{Im}(\Psi_m(s)) < 0\},$$

it is clear that as $m \to \infty$,

$$\begin{aligned} P_X(\operatorname{Re}(\Psi_m(s)) \to 0 \text{ for all } s \in \mathbb{R}) &= 1, & (3.37) \\ P_X(\operatorname{Im}(\Psi_m(s)) \to 0 \text{ for all } s \in \mathbb{R}) &= 1. & (3.38) \end{aligned}$$

Finally, if $c_{m,j}(s) = E_X[\Psi_{m,j}(s)] = 0$ for some $m \in \mathbb{N}$, $j = 1, ..., 4$, and $s \in \mathbb{R}$, then because $\Psi_{m,j}(s)$, $j = 1, ..., 4$, are non-negative, the relevant real or imaginary part of $\Psi_m(s)$ must be equal to zero with probability 1. Thus, results (3.37)-(3.38) hold anyway and limit (3.32) is established. ■

Proof of Proposition 3.3.2. Start with part (a). Let $g_{Y|X,0}(y|x) = g_{Y,0}(y)$ be the initial approximation of the unconditional density of Y, and positive on $\mathbb{S}_Y$. Clearly, $D_C[g_{Y|X,0}] \in [0, \infty)$. By (3.11)

$$\begin{aligned} & D_C[g_{Y|X,0}] - \sum_{j=1}^{m} D^*[g_{Y|\theta_{0,j}^T X,j}] \\ = \; & D_C[g_{Y|X,0}] - \sum_{j=1}^{m} \left\{ D_C[g_{Y|X,m-1}] - D_C[g_{Y|X,m}(y|x)] \right\} \\ = \; & D_C[g_{Y|X,m}] \geq 0. \end{aligned}$$

Hence $\sum_{j=1}^{m} D^*[g_{Y|\theta_{0,j}^T X}] \leq D_C[g_{Y|X,0}]$. Since, by Proposition 3.3.1(c), $D^*[g_{Y|\theta_{0,j}^T X}] \geq 0$ for any $j = 1, ..., m$ then we get $D^*[g_{Y|\theta_{0,m}^T X,m-1}] \to 0$, which completes part (a).

Part (b) follows directly from Lemma 3.3.1 and the discussion that preceded it. ■

Proof of Lemma 3.5.1. The proof follows from Theorems 6 and 8 of Hansen (2008). The uniform consistency of the kernel estimators' partial derivatives with respect to θ can be obtained with a straightforward modification of Hansen's Theorem 2 (see the proof of Lemma 2.6.2). ■

Proof of Lemma 3.5.2. See the proof of Lemma 2.6.5.

Proof of Proposition 3.5.1. See the proof of Theorem 2.3.1.

Proof of Lemma 3.5.4. Parts (ii) and (iii) of the Lemma are contained in the proof of Theorem 2.3.2. Part (i) is also proved in a similar way. Write

$$\mathcal{L}(\theta_m) = \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\widehat{f}_{Y,\theta^T X}^{-i}(y_i, \theta_m^T x_i)}{\widehat{g}_{Y,\theta_m^T X,m-1}^{-i}(y_i, \theta_m^T x_i)} \right) \widehat{\rho}_i^{\theta}.$$

Let $g_{Y,\theta^T X,m-1}(y,t) = E\left[\widehat{g}_{Y|X,m-1}(y|X) \,|\theta^T X = t\right] f_{\theta^T X}(t)$. By a Taylor expansion of

$\log(x)$ and Lemma 3.5.1,

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\widehat{f}^{-i}_{Y,\theta^T X}\left(y_i,\theta_m^T x_i\right)}{\widehat{g}^{-i}_{Y,\theta_m^T X,m-1}\left(y_i,\theta_m^T x_i\right)}\right)\widehat{\rho}_i^{\theta} - \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f_{Y,\theta^T X}\left(y_i,\theta_m^T x_i\right)}{g_{Y,\theta^T X,m-1}\left(y_i,\theta_m^T x_i\right)}\right)\widehat{\rho}_i^{\theta} \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{f}^{-i}_{Y,\theta^T X}\left(y_i,\theta_m^T x_i\right)}{f_{Y,\theta^T X}\left(y_i,\theta_m^T x_i\right)} - 1\right)\widehat{\rho}_i^{\theta} - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{g}^{-i}_{Y,\theta_m^T X,m-1}\left(y_i,\theta_m^T x_i\right)}{g_{Y,\theta^T X,m-1}\left(y_i,\theta_m^T x_i\right)} - 1\right)\widehat{\rho}_i^{\theta} \\
& + o_p\left(n^{-1/2} + h_y^2 + h_x^2\right) \\
= \; & U_1\left(Z_i,Z_j\right) + U_2\left(Z_i,Z_j\right) + o_p\left(n^{-1/2} + h_y^2 + h_x^2\right),
\end{aligned}$$

where

$$\begin{aligned}
U_1\left(Z_i,Z_j\right) &= \frac{1}{n\left(n-1\right)}\sum_{i\neq j}^{n}\frac{1}{h_y h_x}\frac{1}{f_{Y,\theta^T X}\left(y_i,\theta_m^T x_i\right)}K\left(\frac{y_j-y_i}{h_y}\right) \\
& \quad \times K\left(\frac{\theta_m^T\left(x_j-x_i\right)}{h_x}\right)\widehat{\rho}_i^{\theta} - 1, \\
U_2\left(Z_i,Z_j\right) &= \frac{1}{n\left(n-1\right)}\sum_{i\neq j}^{n}\frac{1}{h_x}\frac{\widehat{g}_{Y|X,m-1}\left(y_i|x_j\right)}{g_{Y,\theta^T X,m-1}\left(y_i,\theta_m^T x_i\right)}K\left(\frac{\theta_m^T\left(x_j-x_i\right)}{h_x}\right)\widehat{\rho}_i^{\theta} - 1.
\end{aligned}$$

Moreover, a standard calculation shows that $E\left[U_j\left(Z_i,Z_j\right)|Z_i\right]$ and $E\left[U_j\left(Z_i,Z_j\right)|Z_j\right]$, $j = 1, 2$, are of order $O\left(h_y^2 + h_x^2\right)$. Hence, up to an error term of the order of $O\left(h_y^2 + h_x^2\right)$, the terms U_1, U_2 can be expressed as symmetric second-order degenerate U-statistics. An application of Chebyshev's inequality (Gut 2005, Chapter 3, Theorem 1.4) and Lemma C.2 of Gao and King (2004) then yields (see the proof of Theorem 2.3.2)

$$U_1, U_2 = O_p\left(n^{2-\delta}h_y h_x\right)^{-1/2} + O\left(h_y^2 + h_x^2\right),$$

for any $\delta > 0$. Thus, we have showed

$$\mathcal{L}\left(\theta_m\right) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f_{Y,\theta^T X}\left(y_i,\theta_m^T x_i\right)}{g_{Y,\theta^T X,m-1}\left(y_i,\theta_m^T x_i\right)}\right)\widehat{\rho}_i^{\theta} + o_p\left(n^{-1/2}\right) + O\left(h_y^2 + h_x^2\right).$$

Part (i) is then proved as a result of the central limit theorem (CLT) for strong-mixing processes (cf. Fan and Yao 2003, Theorem 2.21). ■

Proof of Proposition 3.5.2. Immediate from Lemma 3.5.4. ■

Proof of Proposition 3.5.3. See the proof of Theorem 2.3.3. ■

Proof of Proposition 3.6.1. Decompose $\mathcal{L}(\widetilde{h}_m, \widehat{\theta}_m) - L(\widetilde{h}_m, \widehat{\theta}_m)$ into the three terms

$$\begin{aligned} \mathcal{L}(\widetilde{h}_m, \widehat{\theta}_m) - L(\widetilde{h}_m, \widehat{\theta}_m) &= \underbrace{\{\mathcal{L}(\widetilde{h}_m, \widehat{\theta}_m) - \mathcal{L}(\widetilde{h}_m, \theta_{0,m})\}}_{D_1} \\ +\underbrace{\{\mathcal{L}(\widetilde{h}_m, \theta_{0,m}) - L(\widetilde{h}_m, \theta_{0,m})\}}_{D_2} &+ \underbrace{\{L(\widetilde{h}_m, \theta_{0,m}) - L(\widetilde{h}_m, \widehat{\theta}_m)\}}_{D_3}. \end{aligned}$$

Starting with the term D_2, we have by definition

$$E_{\mathbb{S}}\left[D_2\right] = H\left(\widetilde{h}_m, \theta_{0,m}\right),$$

which, by the same arguments as in the proof of Lemma 3.5.4, is of order $O_p\left(n^{2-\delta} H_y H_x\right)^{-1/2} + O\left(H_y^2 + H_x^2\right)$ for any $\delta > 0$. In particular, if $H_y, H_x \sim n^{-1/6}$, then

$$H\left(\widetilde{h}_m, \theta_{0,m}\right) = O_p\left(n^{-5/6+\delta}\right) + O\left(n^{-1/3}\right).$$

Continuing with D_1, we obtain with a mean value $\overline{\theta}_m$ such that $\left|\widehat{\theta}_m - \theta_{0,m}\right| < \left|\overline{\theta}_m - \theta_{0,m}\right|$,

$$\begin{aligned} \mathcal{L}(\widetilde{h}_m, \widehat{\theta}_m) &= \mathcal{L}(\widetilde{h}_m, \theta_{0,m}) + \frac{\partial}{\partial \theta}\mathcal{L}(\widetilde{h}_m, \theta_{0,m})^T\left(\widehat{\theta}_m - \theta_{0,m}\right) \\ &\quad + \frac{1}{2}\left(\widehat{\theta}_m - \theta_{0,m}\right)^T \frac{\partial^2}{\partial \theta^2}\mathcal{L}(\widetilde{h}_m, \theta_{0,m})\left(\widehat{\theta}_m - \theta_{0,m}\right). \end{aligned}$$

An application of (3.21) and Lemma 3.5.4 yields

$$\widehat{\theta}_m - \theta_{0,m} = \Omega^{-}\left(\theta_{0,m}\right)\frac{\partial}{\partial \theta}\mathcal{L}(\widehat{h}_m, \theta_{0,m}) + o_p\left(\widehat{\theta}_m - \theta_{0,m}\right),$$

where $\Omega^{-}(\theta_{0,m})$ is the generalised inverse of $\Omega(\theta_0)$. The last results imply

$$\begin{aligned} D_1 &= \frac{\partial}{\partial\theta}\mathcal{L}(\widetilde{h}_m,\theta_{0,m})^T\Omega^{-}(\theta_{0,m})\frac{\partial}{\partial\theta}\mathcal{L}(\widehat{h}_m,\theta_{0,m}) \\ &\quad -\frac{1}{2}\mathcal{L}(\widehat{h}_m,\theta_{0,m})^T\Omega^{-}(\theta_{0,m})\frac{\partial}{\partial\theta}\mathcal{L}(\widehat{h}_m,\theta_{0,m}) \\ &\quad +(\text{terms of smaller order in probability}). \end{aligned}$$

A similar argument also applies for term D_3. Now, expanding $L(\widetilde{h}_m,\widehat{\theta}_m)$ around $L(\widetilde{h}_m,\theta_{0,m})$, we get with a mean value $\overline{\theta}_m$,

$$\begin{aligned} D_3 &= -\left\{L(\widetilde{h}_m,\widehat{\theta}_m)-L(\widetilde{h}_m,\theta_{0,m})\right\} \\ &= -\left\{\frac{\partial}{\partial\theta}L(\widetilde{h}_m,\theta_{0,m})^T\left(\widehat{\theta}_m-\theta_{0,m}\right)+\frac{1}{2}\left(\widehat{\theta}_m-\theta_{0,m}\right)^T\frac{\partial^2}{\partial\theta^2}L\left(\widetilde{h}_m,\overline{\theta}_m\right)\left(\widehat{\theta}_m-\theta_{0,m}\right)\right\} \\ &= -\frac{\partial}{\partial\theta}L(\widetilde{h}_m,\theta_{0,m})^T\Omega^{-}(\theta_{0,m})\frac{\partial}{\partial\theta}\mathcal{L}(\widehat{h}_m,\theta_{0,m}) \\ &\quad +\frac{1}{2}\mathcal{L}\left(\widehat{h}_m,\theta_{0,m}\right)^T\Omega^{-}(\theta_{0,m})\frac{\partial}{\partial\theta}\mathcal{L}\left(\widehat{h}_m,\theta_{0,m}\right) \\ &\quad +(\text{terms of smaller order in probability}). \end{aligned}$$

We thus have

$$\begin{aligned} E_{\mathbb{S}}\left[D_1+D_3\right] &= E_{\mathbb{S}}\left[\frac{\partial}{\partial\theta}\left\{\mathcal{L}(\widetilde{h}_m,\theta_{0,m})-L(\widetilde{h}_m,\theta_{0,m})\right\}^T\Omega^{-}(\theta_{0,m})\frac{\partial}{\partial\theta}\mathcal{L}(\widehat{h}_m,\theta_{0,m})\right] \\ &\quad +(\text{terms of smaller order in probability}) \\ &= \operatorname{trace}\left[J\left(\widetilde{h}_m,\theta_{0,m}\right)\Omega^{-}(\theta_{0,m})\right] \\ &\quad +(\text{terms of smaller order in probability}). \end{aligned}$$

From Lemma 3.5.4 it is clear that for any $\delta>0$,

$$\begin{aligned} w \qquad & \operatorname{trace}\left[J(\theta_{0,m})\,\Omega^{-}(\theta_{0,m})\right] \\ &= \left\{O_p\left(n^{2-\delta}h_yh_x^3\right)^{-1/2}+O(h_y^2+h_x^2)\right\}\left\{O_p\left(n^{2-\delta}H_yH_x^3\right)^{-1/2}+O(H_y^2+H_x^2)\right\}, \end{aligned}$$

and if $h_y,h_x\sim n^{-1/4}$ and $H_y,H_x\sim n^{-1/6}$, then $\operatorname{trace}\left[I(\theta_{0,m})\,\Omega^{-}(\theta_{0,m})\right]=O_p\left(n^{-5/6+\delta}\right).$

Hence, the proposition is established. ■

Chapter 4

Discussion

In recent decades, the advances in computational power and the accumulation of vast amounts of data led to much interest in statistical methods that perform well in high dimensions. Amongst these methods, the single-index and the projection pursuit methodologies work by projecting high-dimensional data into lower dimensions that retain the most useful information. In this work we extended the applicability of these methodologies to estimation of the c.p.d.f. $f_{Y|X}(y|x)$ of a random scalar Y given a random d-vector $X = x$, where $d \geq 2$.

In Chapter 2 of the thesis, we suggested a 'single-index' approximation of the conditional density $f_{Y|X}(y|x)$ by $f_{Y|\theta^T X}\left(y|\theta^T x\right)$. We showed that similar asymptotic properties of the method, as were established for the i.i.d. case by Delecroix, Härdle and Hristache (2003), Yin and Cook (2005) and Fan et al (2009), still hold under strong-mixing conditions. In so doing, the suggested method was shown to be applicable for dependent data, and in particular to the estimation of predictive densities in time-series. As a second contribution, we derived a general second-order asymptotic representation for the orientation estimator $\widehat{\theta}$ that holds for kernels of any order, while the asymptotically dominant terms are determined by the order of kernels in use and the choice of kernel bandwidths. These two theoretical contributions were justified by appealing to a result by Gao and King (2004), who established a moment inequality for degenerate U-statistics of strong-mixing processes.

The performance single-index model was illustrated in simulations with nonlinear time-series models. Our simulation results demonstrated that the method generally works very well in various different settings. Our results also indicate that despite having better asymptotic properties, orientation estimators obtained with fourth-order kernels perform poorly relative to those obtained with only second-order non-negative kernels.

In Chapter 3 of the thesis, we generalised the result of Chapter 2 to a ‘multi-index’ approximation using a Projection Pursuit type approximation. More precisely, motivated by Projection Pursuit Density Estimation (PPDE) of Friedman, Stuetzle and Schroeder (1984), we proposed a multiplicative projection pursuit approximation of the conditional density that has the form $f\left(y|x\right) = f_0\left(y\right)\prod_{m=1}^{M} h_m\left(y, \theta_m^T x\right)$. The proposed PPCDE was shown to share many of the properties of the previous projection pursuit models, and in particular those of the PPDE of Friedman, Stuetzle and Schroeder (1984). The implementation of the algorithm is relatively straightforward, and due to the nature of the problem, there is no need to incorporate cumbersome Monte Carlo samplings (as in the PPDE), which renders our method simple and computationally undemanding even for very large datasets. In addition, we provided asymptotic justification for the proposed procedure under general stationary conditions that include dependent data, and we offered a bootstrap Information Criterion to terminate the algorithm.

Our simulation results demonstrated that the PPCDE out-performs the unconditional kernel density estimator as well as the single-index and the multivariate conditional density kernel estimators in various different settings.

Of course, the effectiveness of the PPCDE depends on the correctness of the approximation, that is, on the ability to achieve a parsimonious representation of the true conditional density using only relatively few projections. Additionally, the amount of data provided should be relatively large in order for the method to achieve accuracy in high-dimensions. A theoretical discussion by Huber (1985) generally suggests taking n/d in the range of several hundreds to a few thousand. Indeed, Hwang, Lay and Lippman’s (1994) simulation study of problems with 2-5 dimensions concludes that although projection pursuit is more robust to the curse of dimensionality relative to other existing methods, it may require a minimum number of several hundreds before it can perform reasonably. In our simulations of the PPCDE, we used n in the range of several hundreds and the dimension d in the range of $2-5$, which generally produces a relatively low ratio of n/d. However, we still conclude that even for lower ratios of n/d the PPCDE algorithm performs better than the multivariate kernel estimator. Nevertheless, when moving to higher dimensions or to a higher complexity of data generating processes, the performance of the PPCDE may deteriorate and produce many undesired spurious features (see Section 3.4). In such cases, it is therefore sensible to apply first the methods for estimation of dimension reduction subspace at the first step (for example, the dMAVE or dOPG methods of Xia 2007), and then at the second step apply the PPCDE in the reduced dimension reduction subspace such that n/d is in the appropriate range.

There is much room for future research on the theory and applications of the methods discussed in this thesis. Below we highlight some open questions, as well as some interesting possible extensions, for left for further research in the area:

1. The single-index model is expected to work particularly well when the approximated conditional density indeed depends mainly on a single projective direction of the X-data. However, in many cases we expect that the single-index model will be oversimplistic and lead to model bias. As demonstrated in this thesis, the PPCDE approach discussed in Chapter 3 generalises the single-index model, and for example it can provide some evidence for or against the validity of the single-index approximation, depending whether the PPCDE algorithm is stopped after a single-iteration by applying the bootstrap Information Criterion. Nevertheless, in a broader view, developing appropriate goodness-of-fit tests for the c.p.d.f. single-index model is an important and relevant problem. Related goodness-of-fit tests have been already developed for the single-index regression model by Xia et al (2004) and by Stute and Zhu (2005), and generalisations of these methods to a single-index c.p.d.f. estimation seem desirable.

2. Related to the last topic, in some cases we may expect that a nonlinear transforma-

tion of the X-data to a lower dimensional space would be more adequate than using simple linear projections in the form of $\theta^T X$. This is again especially true for the single-index model, when the approximated conditional density depends only on a single and fixed projective direction of the X-data. However, this is also relevant for the PPCDE. As demonstrated in Proposition 3.3.2, even in highly nonlinear situations, the projection pursuit approximation converges weakly to the true conditional density as the number of projections approaches infinity. However, in practice, using a large number of projective directions can make the PPCDE unwieldy for approximation and computation. In recent years, many new methods for nonlinear dimension reduction have been developed. These methods are usually ad-hoc in nature, and aimed mainly for visualisation of high-dimensional data rather than for inference (see Lee and Verleysen 2007). Notwithstanding, many of these methods are based on principles that can likely be generalised to problems of inference like ours. Some of these methods are based on Local-Linear Embedding (LLE) of the highly dimensional variable. As a simple example where a local-linear modelling can be desired, we may envisage situations where different projective directions are needed in order to approximate a c.p.d.f in its tails or in the centre of the distribution*. A possible starting point toward achieving a nonlinear local-linear model for dimension reduction might be considering a local modelling of the orientation vector, that is having $\theta = \theta(x)$ (see Loader 1999). Where c.p.d.f. estimation is concerned, as in our thesis, this requires developing a 'localised' version of the proposed method (see research direction 4 below). As an alternative approach towards nonlinear dimensionality reduction for statistical inference, one may consider applying a linear projection of the data after being transformed first. This approach requires identifying an efficient way of transforming the X-data in order to allow the gain of more information on the c.p.d.f. of Y given the linear projections of the transformed X. Some results of Huber (1985, Section 13) imply that the PPDE aims to find the least normal projections of

*I thank Professor Piotr Fryzlewicz for the example.

the probability density of interest. For our PPCDE, we currently do not have any equivalent result. However, identifying a similar result may assist in developing a constructive criterion for the optimal transformation of the X-data prior to applying the PPCDE algorithm.

3. In the thesis we suggested two methods for approximation of a c.p.d.f. However, once an approximation is obtained it is not clear to what extent we can trust the approximation obtained. While the theory provides uniform convergence rates along the univariate projective direction, it does not provide any confidence intervals for the c.p.d.f. estimates obtained for various values of $y \in \mathbb{R}$ and $x \in \mathbb{R}^d$. In particular, the theory can provide asymptotic confidence intervals for the univariate c.p.d.f. in the ingle-index model, or for the univariate multiplicative modification function in the PPCDE method, by appealing to the CLT result for kernel c.p.d.f. estimation (see Robinson 1983, Theorem 6.1, and Chen, Linton and Robinson 2001, Theorem 3). However, it is not clear how one can construct confidence intervals for the obtained multivariate approximation, as it requires some inference in high-dimensions, which is what the suggested approximation methods are meant to avoid.

4. As discussed in detail in Section 3.2, our understanding of the strengths and limitations of the projection pursuit product representation suggested in this thesis is still lacking. The works of Diaconis and Shahshahani (1984) and Yuan (2010) provides some necessary and sufficient conditions for the projection pursuit regression model to hold and to have a unique representation. However, as far as the PPCDE approximation is concerned, there are still some open questions, particularly whether there are identifying restrictions that yield identifiable unique representations of the optimal projection pursuit product approximation, and whether there are any real c.p.d.f.'s that follow the projection pursuit product representation without requiring an additional normalisation factor.

5. As demonstrated in Section 3.4, even for low dimensions, the standard kernel es-

timator is characterised by some spurious features. The single-index model and the PPCDE rectifies this phenomenon by working on lower-dimensional projections. However, it is interesting to examine possible implementations of these methods with semiparametric or even fully parametric estimation in high-dimensions attained by lower-dimensional projections. For example, copula models may provide a convenient and flexible framework, as they allow separate modelling of the marginal distributions and dependence structure in the multivariate distribution.

6. The applicability of the proposed constrained relative entropy, $D_C[g_{Y|X}]$ may extend beyond the PPCDE methodology to other problems that involve the estimation of conditional densities. For example, local parametric modelling offers a general form of nonparametric model. Local likelihood models have been examined in the literature, for instance, by Tibshirani and Hastie (1987), Hjort and Jones (1996), Loader (1996) and Fan, Farmen and Gijbels (1998). In a similar manner, the proposed constrained relative entropy may be 'localised' in order to produce locally parametric conditional densities in a natural way.

7. Many lines of similarity exist between the PPDE of Friedman, Stuetzle and Schroeder (1984) and the PPCDE approach offered in this thesis. Developing a unifying procedure for the PPDE and PPCDE may enable the achievement of a useful approximation of the c.p.d.f. $f_{Y|X}(y|x)$ where both Y and X are random vectors of high dimension.

8. For both approximation methods presented in the thesis, it may happen that not all explanatory variables in X contain useful information to predict Y. If irrelevant variables are included, which is very likely in high-dimensional environments, the precision of parameter estimation as well as the accuracy of forecasting will suffer (Altham, 1984). Therefore, it makes sense to exclude irrelevant variables from the approximations. In particular, in a time-series setting, the researcher has to choose optimal number of lagged observations to be included in the model by considering

$x_t = (y_{t-1}, ..., y_{t-d})$. In the literature on single-index regression, some variable selection methods have been considered. Naik and Tsai (2001) developed a variant of the AIC criterion for single-index regression models. Their criterion, however, is adequate for regression problems, and can not be generalised easily to probability density estimation. Kong and Xia (2007) proposed a cross-validatory model selection method for the single-index regression model. Nevertheless, cross-validatory methods are usually computationally intensive techniques, and therefore they are less desirable in our case, as $\widehat{\theta}$ is obtained by numerical optimisation. We believe that the Bootstrap Information Criterion, proposed in Section 3.6, may offer a general applicable model selection criterion in semiparametric setting, and in particular for the purpose of variable selection in the single-index and the PPCDE approximations. However, in order to confirm the validity of the Bootstrap Information Criterion, its performance should be examined in various different settings and models, and its theoretical properties should be further explored.

References

Aït-Sahalia, Y. (1999). Transition Densities for Interest Rate and Other Nonlinear Diffusions. The Journal of Finance, 54, 1361–1395.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. B. N. Petrov and F. Csaki (eds.). 2nd International Symposium on Information Theory, 267-81.

Altham, P.M.E. (1984). Improving the precision of estimation by fitting a generalized linear model and quasi-likelihood. Journal of the Royal Statistical Society B, 46, 118–9.

Andersen, T.G., Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. International Economic Review, 39, 885-905.

An, H.Z., Huang, F.C. (1996). The geometric ergodicity of nonlinear autoregressive models. Statistica Sinica, 6, 943-956.

Andrews, D.W.K. (1992). Generic Uniform Convergence. Econometric Theory, 8, 241-257.

Buch-Kromann, T., Guillen, M., Linton, O., Nielsen, J.P. (2011). Multivariate Density Estimation using Dimension Reducing Information and Tail Flattening Transformations. Insurance: Mathematics and Economics, 48, 99-110.

Bickel, P.J., Rosenblatt, M. (1973). On some global measures of the deviation of density function estimates. The Annals of Statistics, 1, 1071–1095.

Breima, L.,Meisel,W., Purcell, E. (1977). Variable kernel estimates of multivariate densities. Technornetrics, 19, 135-144.

Amemiya, T. (1985). Advanced Econometrics. Harvard University Press, Harvard, Boston.

Carrasco, M., Chen, X. (2002). Mixing and moment properties of various GARCH and stochastic volatility models. Econometric Theory, 18, 17-39.

Chen, X., Linton, O., Robinson, P. M., (2001). The estimation of conditional densities, in: M. L.. Puri (ed.), Asymptotics in Statistics and Probability, Festschrift for George Roussas (M.L. Puri, ed.), VSP International chience Publishers, the Netherlands, 71-84.

Cook, R. D. (1998) Regression Graphics. New York, Wiley.

Cook, R. D., Weisberg, S. (1991). Discussion of Li (1991). Journal of the American Statistical Association, 86, 328–332.

Common, P. (1984). Independent Component Analysis, a new concept?. Signal Processing, 36, 287–314.

Davis, R.A., Mikosch, T. (2009). Probabilistic Properties of Stochastic Volatility Models. In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (eds.): Handbook of Financial time-series, 255-267. Springer, New York.

Delecroix, M., Härdle, W., Hristache, M. (2003). Efficient estimation in conditional single-index regression. Journal of Multivariate Analysis, 86, 213–226.

Delecroix, M., Hristache, M., Patilea, V. (2006). On semiparametric M-estimation in single-index regression., Journal of Statistical Planning and Inference, 136, 730–769.

Diaconis, P., Shahshahani, M. (1984). On nonlinear functions of linear combinations. SIAM Journal of Schientific Computing 5, 175-191.

Efromovich, S. (2010). Dimension Reduction and Adaptation in Conditional Density Estimation. Journal of the American Statistical Association, 105, 761-774.

Engle, R.F. (2001). Financial econometrics a new discipline with new methods. Journal of Econometrics, 100, 53-56.

Engle, R. F., Manganelli, S. (2004). CAViaR: conditional autoregressive value at risk by regression quantiles. Journal of Business and Economic Statistics, 22, 367–381.

Fan, J., Farmen, M. and Gijbels, I. (1998). Local maximum likelihood estimation. Journal of the Royal Statistical Society B, 60, 591-608.

Fan, J., Peng, L., Yao, Q., Zhang, W. (2009). Approximating conditional density functions using dimension reduction. Acta Mathematica Applicatae Sinica, 25, 445-456.

Fan J., Yao, Q. (2003). Nonlinear time-series: Nonparametric and Parametric Methods. Springer-Verlag.

Fan, J., Yao, Q., Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Biometrika, 83, 189-206.

Friedman, J.H., Stuetzle, W. (1980). Projection pursuit classification. unpublished manuscript.

Friedman, J.H., Stuetzle, W. (1981). Projection pursuit regression. Journal of the American Statistical Association, 76, 817–823.

Friedman, J. H. and W. Stuetzle (1982). Smoothing of scatterplots. Technical Report Orion 3, Department of Statistics, Stanford University.

Friedman, J.H., Stuetzle, W. Schroeder, A. (1984). Projection pursuit density estimation. Journal of the American Statistical Association, 79, 599–608.

Friedman, J.H., Tukey, J.W. (1974). A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers C, 23, 881-890.

Gao, J., King, M.L. (2004). Adaptive testing in continuous-time models. Econometric Theory, 20 , 844-882.

Gao, J., Tong, H. (2004). Semiparametric nonlinear time-series model selection. Journal of the Royal Statistical Society B, 66, 321–36.

Glad, I.K., Hjort, N.L., Ushakrov, N.G. (2003). Correction of density estimators that are not densities. Scandinavian Journal of Statistics, 30, 415- 427.

Gorsuch, R. L. (1983). Factor analysis. Hillsdale, NJ: Lawrence Erlbaum.

Gut, A. (2005). Probability: A Graduate Course. Springer Series in Statistics.

Hall, P. (1989). On projection pursuit regression. Annals of Statistics, 17, 573–588.

Hall, P. Marron, J.S. Park, B.U. (1992). Smoothed cross-validation. Probability Theory and Related Fields, 92, 1–20.

Hall, P., Racine, J. Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. Journal of the American Statistical Association, 99, 1015–1026.

Hall, P., Yao, Q. (2005). Estimation for conditional distribution functions via dimension reduction. The Annals of Statistics, 33, 1404-1421.

Hansen, B.E. (2008). Uniform convergence rates for kernel estimation with dependent data. Econometric Theory, 24 , 726-748.

Härdle, W., Hall, P., Ichimura, H. (1993). Optimal smoothing in single-index models. Annals of Statistics, 21, 157–178.

Härdle, W., Stoker, T.M. (1989). Investigating smooth multiple regression by the method of average derivatives. Journal of the American Statistical Association, 84, 986–995.

Hastie, T.J., Tibshirani, R.J. (1990). Generalized additive models. London: Chapman and Hall.

Hjort, N. L., Glad, I.K. (1995). Nonparametric density estimation with a parametric start., The Annals of Statistics, 23 882-904.

Hjort, N.L. and Jones, M.C. (1996) Locally parametric nonparametric density estimation.Ann. Statist., 24, 1619–1647.

Horowitz, J.L., Mammen, E. (2007). Oracle-efficient nonparametric estimation of an additive model with an unknown link function. working paper.

Huber, P.J. (1985). Projection pursuit, The Annals of Statistics, 13, 435 - 525.

Hurvich, C.M., Simonoff, J.S., Tsai, C.L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. Journal of the Royal Statistical Society B, 60, 271–293.

Hwang, J.N., Lay, S.R., Lippman A. (1994). Nonparametric multivariate density estimation: a comparative study. IEEE Transactions of Signal Processing, 42, 2795-2810.

Hyndman, R.J. (1995). Highest density forecast regions for non-linear and non-normal time-series models. Journal of Forecasting, 14, 431–441.

Hyndman, R.J., Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. Journal of Nonparametric Statistics, 14, 259-278.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. Journal of Econometrics, 58, 71-120.

Ichimura, H. (1995). Asymptotic distribution of non-parametric and semi-parametric estimators with data dependent smoothing parameters. Working Papers, University of California at Berkeley.

Ichimura, H., Todd, P.E. (2006). Implementing nonparametric and semiparametric estimators. In Handbook of Econometrics, Volume 6.

Ishiguro, M., Sakamoto, Y., Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. Annals of Institute of Statistical Mathematics. 49, 411-434.

Jolliffe, I. (2002). Principal component analysis. Springer Series in Statistics, 2nd edition.

Jones, M.C., Linton, O., Nielsen, J.P. (1995). A simple and effective bias reduction method for density and regression estimation. Biometrika, 82, 327-338.

Kemperman, J.H.B. (1969). On the Optimum Rate of Transmitting Information. The Annals of Mathematical Statististics, 40, 2156-2177.

Kong, E., Xia, Y. (2007). Variable selection for the single-index model. Biometrika, 94, 217-229.

Konishi, S., Kitagawa, G. (1996). Generalised information criteria in model selection. Biometrika, 83, 875-890.

Konishi, S., Kitagawa, G. (2008). Information criteria and statistical modelling. Springer.

Kruskal, J.B. (1969). Toward a practical method which helps uncover the structure of a set of observations by finding the line tranformation which optimizes a new "index of condensation", in R. C. Milton and J. A. Nelder (eds), Statistical Computation, Academic Press, New York, 427-440.

Kruskal, J.B. (1972). Linear transformation of multivariate data to reveal clustering, in R. N. Shepard, A. K. Romney and S. B. Nerlove (eds), Multidimensional scaling: Theory and Applications in the Behavioural Sciences, Vol. 1, Seminar Press, London, 179-191.

Lee, J. and Verleysen, M. (2007). Nonlinear Dimensionality Reduction. Springer, Berlin.

Li, K.C. (1991). Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 86, 316–342.

Liebscher, E. (1996). Strong convergence of sums of α-mixing random variables with applications to density estimation. Stochastic Processes and their Applications, 65, 69-80.

Linton, O.B., Nielsen, J.P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. Biometrika, 82, 93-100.

Loader, C.R. (1996). Local likelihood density estimation. The Annals of Statistics, 24, 1602–1618.

Loader, C. R. (1999). Local Regression and Likelihood. Springer, New York.

Marron, J.S. (1992). Graphical understanding of higher order kernels. North Carolina Inst. Statistics Mimeo Series 2082.

Marron, J.S., Wand, M.P. (1992). Exact mean integrated squared error. Annals of Statistics, 20, 712-736.

Moore, D.S., Yackel, J.W. (1977). Consistency Properties of Nearest Neighbor Density Function Estimators. The Annals of Statistics, 5, 143-154

Müller, H.G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. Annals of Statistics, 12, 766–774.

Naik, P.A., Tsai, C.L. (2001). Single-index model selections. Biometrika, 88, 821–32.

Nocedal, J., Wright, S.J. (2006). Numerical Optimization (2nd ed.). Springer-Verlag, Berlin, New York.

Nolan, D., Pollard, D. (1987). U-processes: Rates of convergence. Annals of Statistics, 15, 780-799.

Paparoditis, E., Politis, D.N. (2000) The local bootstrap for kernel estimators under general dependence conditions. Annals of the Institute of Statistical Mathematics, 52, 139-159.

Pham, D.T., Tran, L.T. (1985). Some mixing properties of time-series models. Stochastic Processes and their Applications, 19, 297–303.

Powell, J.L., Stock, J.M., Stoker, T.M. (1989). Semiparametric estimation of index coefficients. Econometrica, 57, 1403–1430.

Robinson, P.M. (1983). Nonparametric estimation for time-series. Journal of time-series Analysis, 4, 185-207.

Scott, D.W. (1992). Multivariate density estimation: Theory, Practice, and Visualization. John Wiley and Sons, New York, Chichester.

Serfling, R.J. (1980). Approximation Theorems of Mathematical Statistics. Wiley, New York.

Shao, J. (2003). Mathematical statistics (Second edition). Springer Texts in Statistics, Springer-Verlag, New York.

Sherman, R.S. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. The Annals of Statistics, 22, 439-459.

Silverman, B.W. (1986). Density estimation. London: Chapman and Hall.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society B, 39, 44-47.

Stute, W. and Zhu, L.X. (2005). Nonparametric checks for single-index models. Annals of Statistics, 33, 1048-1083.

Su, L. White, H. (2007). A consistent characteristic function-based test for conditional independence. Journal of Econometrics, 141, 807-834.

Switzer, P. (1970). Numerical classification. Geostatistics, Plenum, New York, 1970.

Switzer, P., Wright, R.M. (1971). Numerical classification applied to certain Jamaican eocene nummulitids. Mathematical Geology, 3, 297-311.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. Mathematical Sciences, 153, 12-18 (in Japanese).

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. Journal of the American Statistical Association 82, 559-567.

Touboul, J. (2011). Projection pursuit through relative entropy minimization. Communications in Statistics - Simulation and Computation, 40, 854-878.

Windsor, C., Thyagaraja A. (2001). The prediction of periods of high volatility in exchange markets. The European Physical Journal B, 20, 581-584.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. Econometrica, 50, 1-25.

White, H. (1984). Asymptotic theory for econometricians. Orlando, Academic Press, Inc.

Wu, W., Yu, K., Mitra, G. (2008). Kernel conditional quantile estimation for stationary processes with application to conditional Value-at-Risk. Journal of Financial Econometrics, 6, 253-270.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. The Annals of Statistics, 35, 2654-2690

Xia, Y. (2008). A multiple-index model and dimension reduction. Journal of the American Statistical Association, 103, 1631-1640.

Xia, Y., An, H.Z. (1999). Projection pursuit autoregression in time-series. Journal of time-series Analysis, 20, 693–714

Xia Y., Härdle, W., Linton, O. (2012). Optimal Smoothing for a Computationally and Statistically Efficient Single Index Estimator, within "Exploring Research Frontiers in Contemporary Statistics and Econometrics" edited by Van Keilegom I. and Wilson P.W. Springer-Verlag, Berlin.

Xia, Y., Li, W. K., Tong, H. and Zhang, D. (2004). A goodness-of-fit test for single-index models. Statistica Sinica, 14, 1-39.

Xia Y., Tong, H., Li, W.K. (1999). On extended partially linear single-index models. Biometrika, 86, 831–842.

Xia, Y., Tong, H., Li, W.K., Zhu, L. (2002). An adaptive estimation of dimension reduction space. Journal of the Royal Statisticial Society B, 64, 363-410.

Yao, Q., Tong, H. (1994). On prediction and chaos in stochastic systems. Philosophical Transactions of the Royal Society A, 348, 357-369.

Yin, X., Cook, R.D. (2002). Dimension reduction for the conditional k-th moment in regression. Journal of the Royal Statistical Society Series B, 64, 159–175.

Yin, X., Cook, R.D. (2005). Direction estimation in single-index regressions. Biometrika, 92, 371–384.

Yin, X., Li, B., Cook, R.D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. Journal of Multivariate Analysis, 99, 1733-1757.

Zheng, J.X. (1998). A consistent nonparametric test of parametric regression models under conditional quantile restrictions. Econometric Theory 14, 123-138.