Additive Models: Extensions and Related Models.

Enno Mammen* Byeong U. Park† Melanie Schienle‡

August 8, 2012

Abstract

We give an overview over smooth backfitting type estimators in additive models. Moreover we illustrate their wide applicability in models closely related to additive models such as nonparametric regression with dependent error variables where the errors can be transformed to white noise by a linear transformation, nonparametric regression with repeatedly measured data, nonparametric panels with fixed effects, simultaneous nonparametric equation models, and non- and semiparametric autoregression and GARCH-models. We also discuss extensions to varying coefficient models, additive models with missing observations, and the case of nonstationary covariates.

1 Introduction

In this chapter we continue the discussion of last chapter on additive models. We come back to the smooth backfitting approach that was already mentioned there. The basic idea of the smooth backfitting is to replace the least squares criterion by a smoothed version. We now explain its definition in an additive model

$$E(Y|X) = \mu + f_1(X^1) + \cdots + f_d(X^d). \tag{1.1}$$

We assume that n i.i.d. copies $(X_i^1, \ldots, X_i^d, Y_i)$ of $(X^1, \ldots, X^d, Y)$ are observed, or more generally, n stationary copies. Below, in Section 4, we will also weaken the stationarity assumption.

In an additive model (1.1) the smooth backfitting estimators $\widehat{\mu}, \widehat{f}_1, \ldots, \widehat{f}_d$ are defined as the minimizers of the smoothed least squares criterion

$$\int \sum_{i=1}^{n} \left[Y_i - \mu - f_1(x^1) - \cdots - f_d(x^d)\right]^2 K\left(\frac{X_i^1 - x^1}{h_1}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right) \, dx^1 \cdots \, dx^d \tag{1.2}$$

*Department of Economics, Mannheim University, Germany. E-mail: emammen@rumms.uni-mannheim.de. Enno Mammen gratefully acknowledges research support of the German Science Foundation through the Collaborative Research Center 884 "Political Economy of Reforms".

†Department of Statistics, Seoul National University, Korea. E-mail: bupark@stats.snu.ac.kr. Byeong U. Park's research was supported by the NRF Grant funded by the Korea government (MEST)(No. 2010-0017437).

‡School of Business and Economics, Humboldt University Berlin, Germany. E-mail: melanie.schienle@wiwi.hu-berlin.de. Melanie Schienle gratefully acknowledges research support of the German Science Foundation through the Collaborative Research Center 649.

under the constraint

$$\int f_1(x^1)\widehat{p}_{X^1}(x^1)dx^1 = \cdots = \int f_d(x^d)\widehat{p}_{X^d}(x^d)dx^d = 0. \tag{1.3}$$

Here K is a kernel function, i.e., a positive probability density function and $h_1, \ldots, h_d$ are bandwidths. Furthermore, $\widehat{p}_{X^j}$ is the kernel density estimator of the density p_{X^j} of X^j defined by

$$\widehat{p}_{X^j}(x^j) = \frac{1}{nh_j}\sum_{i=1}^{n} K\left(\frac{X_i^j - x^j}{h_j}\right).$$

Below, we will outline that the smooth backfitting estimator can be calculated by an iterative backfitting algorithm. While the estimator got its name from the corresponding algorithm, it could, however, better be described as *smooth least squares estimator* highlighting its statistical motivation.

If there is only one additive component, i.e., if we have $d = 1$, we get a kernel estimator $\widetilde{f}_1(x^1) = \widehat{\mu} + \widehat{f}_1(x^1)$ as the minimizer of

$$f_1 \rightsquigarrow \int \sum_{i=1}^{n} \left[Y_i - f_1(x^1)\right]^2 K\left(\frac{X_i^1 - x^1}{h_1}\right) dx^1. \tag{1.4}$$

The minimizer of this criterion is given as

$$\widetilde{f}_1(x^1) = \left[\sum_{i=1}^{n} K\left(\frac{X_i^1 - x^1}{h_1}\right)\right]^{-1} \sum_{i=1}^{n} Y_i K\left(\frac{X_i^1 - x^1}{h_1}\right).$$

Thus, $\widetilde{f}_1(x^1)$ is just the classical Nadaraya-Watson estimator. We get the smooth backfitting estimator as a natural generalization of Nadaraya-Watson smoothing to additive models.

In this chapter we present a broad discussion of estimators based on minimizing a smoothed least squares criterion. We do this for two reasons. First, we argue that, even for additive models, this method is a powerful alternative to the two-step procedures that were extensively discussed in last chapter and in the chapter "Oracly efficient two-step estimation for additive regression". Furthermore, smooth least squares estimators also work in models that are closely related to the additive model but are not of the form that is directly suitable for two-step estimation. We illustrate this with an example. Suppose that one observes (X_i, Y_i) with $Y_i = f(X_i) + \varepsilon_i$ where ε_i is a random walk, i.e., $\eta_i = \varepsilon_{i+1} - \varepsilon_i$ are zero mean i.i.d. variables that are independent of $X_1, \ldots, X_n$. In this model the Nadaraya-Watson estimator (1.4) is not consistent. Consistent estimators can be based on considering $Z_i = Y_{i+1} - Y_i$. For these variables we get the regression model

$$Z_i = f(X_{i+1}) - f(X_i) + \eta_i.$$

The smooth least squares estimator in this model is based on the minimization of

$$f \rightsquigarrow \int \sum_{i=1}^{n} \left[Z_i - f(x^1) + f(x^2)\right]^2 K\left(\frac{X_{i+1} - x^1}{h_1}\right) K\left(\frac{X_i - x^2}{h_2}\right) dx^1 dx^2.$$

Clearly, an alternative approach would be to calculate estimators $\widehat{f}_1$ and $\widehat{f}_2$ in the model $Z_i = f_1(X_{i+1}) + f_2(X_i) + \eta_i$ and to use $[\widehat{f}_1(x) - \widehat{f}_2(x)]/2$ as an estimator of f. We will come back to related models below.

The additive model is important for two reasons:

(i) It is the simplest nonparametric regression model with several nonparametric components. The theoretical analysis is quite simple because the nonparametric components enter linearly into the model. Furthermore, the mathematical analysis can be built on localization arguments from classical smoothing theory. The simple structure allows for completely understanding of how the presence of additional terms influences estimation of each one of the nonparametric curves. This question is related to semiparametric efficiency in models with a parametric component and nonparametric nuisance components. We will come back to a short discussion of *nonparametric efficiency* below.

(ii) The additive model is also important for practical reasons. It efficiently avoids the curse of dimensionality of a full-dimensional nonparametric estimator. Nevertheless, it is a powerful and flexible model for high-dimensional data. Higher-dimensional structures can be well approximated by additive functions. As lower-dimensional curves they are also easier to visualize and hence to interpret than a higher-dimensional function.

Early references that highlight the advantages of additive modeling are Stone (1985), Stone (1986), Buja, Hastie, and Tibshirani (1989) and Hastie and Tibshirani (1990). In this chapter we concentrate on the discussion of smooth backfitting estimators for such additive structures. For a discussion of two-step estimators we refer to last chapter and the chapter on two-step estimation. For sieve estimators in additive models, see Chen (2006) and the references therein. For the discussion of penalized splines we refer to Eilers and Marx (2002).

In this chapter we only discuss estimation of nonparametric components. Estimation of parametric components such as $\theta = \theta(f_1) = \int f_1(x_1)w(x_1)\ dx_1$ for some given function w requires another type of analysis. In the latter estimation problem natural questions are e.g., whether the plug-in estimator $\widehat{\theta} = \theta(\widehat{f}_1) = \int \widehat{f}_1(x_1)w(x_1)\ dx_1$ for a nonparametric estimator $\widehat{f}_1$ of f_1 converges to θ at a parametric $\sqrt{n}$-rate, and whether this estimator achieves the semiparametric efficiency bound. Similar questions arise in related semiparametric models. An example is the partially linear additive model: $Y_i = \theta^\top Z^i + \mu + f_1(X_1^i) + \cdots + f_d(X_d^i) + \varepsilon^i$. Here, Z is an additional covariate vector. A semiparametric estimation problem arises when $\mu, f_1, \ldots, f_d$ are nuisance components and θ is the only parameter of interest. Then naturally the same questions as above arise when estimating θ. As said, such semiparametric considerations will not be in the focus of this chapter. For a detailed discussion of the specific example we refer to Schick (1996) and Yu, Mammen, and Park (2011).

In this chapter, we concentrate on the description of estimation procedures. Smooth backfitting has been also used in testing problems by Haag (2006), Haag (35) and Lundervold, Tjø stheim, and Yao (2007). For related tests based on kernel smoothing, see also the overview article Fan and Jiang (2007). In Lundervold, Tjø stheim, and Yao (2007) additive models are used to approximate the distribution of spatial Markov random fields. The conditional expectation of the outcome of the random field at a point, given the outcomes in the neighborhood of the point, are modeled as sum of functions of the neighbored outcomes. They propose tests for testing this additive structure. They also discuss the behavior of smooth backfitting if the additive model is not correct. Their findings are also interesting for other

applications where the additive model is not valid but can be used as a powerful approximation.

Another approach that will not be pursued here is parametrically guided nonparametrics. The idea is to fit a parametric model in a first step and then apply nonparametric smoothing in a second step, see Fan, Wu, and Feng (2009) for a description of the general idea. The original idea was suggested by Hjort and Glad (1995) in density estimation. See also Park, Kim, and Jones (2002) for a similar idea.

The next section discusses the smooth backfitting estimator in additive models. In Section 3 we discuss some models that are related to additive models. The examples include nonparametric regression with dependent error variables where the errors can be transformed to white noise by a linear transformation, nonparametric regression with repeatedly measured data, nonparametric panels with fixed effects, simultaneous nonparametric equation models, and non- and semiparametric autoregression and GARCH-models. Other extensions that we will shortly mention are varying coefficient models and additive models with missing observations. In Section 4 we discuss the case of nonstationary covariates. Throughout the chapter we will see that many of the discussed models can be put in a form of noisy Fredholm integral equation of second kind. We come back to this representation in the last section of this chapter. We show that this representation can be used as an alternative starting point for the calculation and also for an asymptotic understanding of smooth least squares estimators.

2 Smooth least squares estimator in additive models

2.1 The backfitting algorithm.

In the additive model (1.1) the smooth backfitting estimator can be calculated by an iterative algorithm. To see this, fix a value of x^1 and define $\widehat{\mu}_1 = \widehat{\mu} + \widehat{f}_1(x^1)$. One can easily see that $\widehat{\mu}_1$ minimizes

$$\begin{aligned} \mu_1 \quad \rightsquigarrow \quad & \int \sum_{i=1}^{n} K\left(\frac{X_i^1 - x^1}{h_1}\right) \left[Y_i - \mu_1 - f_2(x^2) - \cdots - f_d(x^d)\right]^2 \\ & \times K\left(\frac{X_i^2 - x^2}{h_2}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right) \, dx^2 \cdots \, dx^d. \end{aligned} \tag{2.1}$$

This holds because we have no constraint on the function $x^1 \rightsquigarrow \widehat{\mu} + \widehat{f}_1(x^1)$. Thus we can minimize the criterion pointwise in this function and we do not integrate over the argument x^1 in (2.1). Thus, we get

$$\begin{aligned} \widehat{\mu}_1 \quad = \quad & \left[\int \sum_{i=1}^{n} \prod_{j=1}^{d} K\left(\frac{X_i^j - x^j}{h_j}\right) dx^2 \cdots dx^d\right]^{-1} \\ & \times \int \sum_{i=1}^{n} \left[Y_i - f_2(x^2) - \cdots - f_d(x^d)\right] \prod_{j=1}^{d} K\left(\frac{X_i^j - x^j}{h_j}\right) dx^2 \cdots dx^d. \end{aligned}$$

The expression on the right hand side of this equation can be simplified by noting that $\int \frac{1}{h_j} K\left(\frac{X_i^j - x^j}{h_j}\right) \, dx^j = 1$ for $i = 1, \ldots, n; j = 1, \ldots, d$. We get

$$\widehat{\mu}_1 = \widehat{\mu} + \widehat{f}_1(x^1) = \widehat{f}_1^*(x^1) - \sum_{k=2}^{d} \int \frac{\widehat{p}_{X^1, X^k}(x^1, x^k)}{\widehat{p}_{X^1}(x^1)} \widehat{f}_k(x^k) \, dx^k. \tag{2.2}$$

Here, for $1 \leq j \leq d$

$$\widehat{f}_j^*(x^j) = \left[\sum_{i=1}^n K\left(\frac{X_i^j - x^j}{h_j}\right)\right]^{-1} \sum_{i=1}^n K\left(\frac{X_i^j - x^j}{h_j}\right) Y_i = \widehat{p}_{X^j}(x^j)^{-1} \frac{1}{nh_j} \sum_{i=1}^n K\left(\frac{X_i^j - x^j}{h_j}\right) Y_i.$$

This is the marginal Nadaraya-Watson estimator, based on smoothing the response Y_i versus one covariate X_i^j. Furthermore, $\widehat{p}_{X^j,X^k}$ is the two-dimensional kernel density estimator of the joint density p_{X^j,X^k} of two covariates X^j and X^k: for $1 \leq j \neq k \leq d$

$$\widehat{p}_{X^j,X^k}(x^j, x^k) = \frac{1}{nh_jh_k} \sum_{i=1}^n K\left(\frac{X_i^j - x^j}{h_j}\right) K\left(\frac{X_i^k - x^k}{h_k}\right).$$

Similarly to Eq. (2.2) we get for all $j = 1, ..., d$ that

$$\widehat{f}_j(x^j) = \widehat{f}_j^*(x^j) - \widehat{\mu} - \sum_{k \neq j} \int \frac{\widehat{p}_{X^j,X^k}(x^j, x^k)}{\widehat{p}_{X^j}(x^j)} \widehat{f}_k(x^k)\, dx^k. \tag{2.3}$$

One can show that

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^n Y_i. \tag{2.4}$$

A proof of this equation is postponed to the end of this subsection.

We are now in the position to define the smooth backfitting algorithm. Our main ingredients are Eq. (2.3) and the formula for $\widehat{\mu}$. After an initialization step the backfitting algorithm proceeds in cycles of d steps:

- **Initialization step:** Put $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\widehat{f}_j^{[0]}(x^j) \equiv 0$ for $j = 1, ..., d$.
- **lth iteration cycle:**
  - **jth step of the lth iteration cycle:** in the jth step of the lth iteration cycle one updates the estimator $\widehat{f}_j$ of the jth additive component f_j

$$\begin{aligned} \widehat{f}_j^{[l]}(x^j) &= \widehat{f}_j^*(x^j) - \widehat{\mu} - \sum_{k=1}^{j-1} \int \frac{\widehat{p}_{X^j,X^k}(x^j, x^k)}{\widehat{p}_{X^j}(x^j)} \widehat{f}_k^{[l]}(x^k)\, dx^k \\ &\quad - \sum_{k=j+1}^{d} \int \frac{\widehat{p}_{X^j,X^k}(x^j, x^k)}{\widehat{p}_{X^j}(x^j)} \widehat{f}_k^{[l-1]}(x^k)\, dx^k. \end{aligned} \tag{2.5}$$

We now discuss some computational aspects of the smooth backfitting algorithm. One can show that there exist constants $C > 0$ and $0 < \gamma < 1$ that do not depend on n such that with probability tending to one

$$\int [\widehat{f}_j^{[l]}(x^j) - \widehat{f}_j(x^j)]^2 p_{X^j}(x^j)\, dx^j \leq C\gamma^{2l}. \tag{2.6}$$

For a detailed statement, see Theorem 1 in Mammen, Linton, and Nielsen (1999) where a proof of (2.6) can be also found. The essential argument of the proof is that the approximation error $\sum_{j=1}^d [\widehat{f}_j^{[l]}(x^j) - \widehat{f}_j(x^j)]$ behaves like a function that is cyclically and iteratively projected onto d linear subspaces of a function space. Each cycle of projections reduces the norm of this function by a factor γ, for some fixed $\gamma < 1$, with probability tending to one.

The bound (2.6) allows for two important conclusions.

(i) For a fixed accuracy, the number of iterations of the algorithm can be chosen as constant in n: in particular, it does not need to increase with n.

(ii) Furthermore, for an accuracy of order $n^{-\alpha}$ it suffices that the number of iterations increases with a logarithmic order. This implies, in particular, that the complexity of the algorithm does not explode but increases only slowly in n. We will see in the next subsection that for an optimal choice of bandwidth the rate of $\widehat{f}_j(x^j) - f_j(x^j)$ is of order $O_p(n^{-2/5})$. In that case, a choice of α with $\alpha > 2/5$ guarantees that the numerical error is of smaller order than the statistical error.

When numerically implementing smooth backfitting, estimators $\widehat{f}_j^{[l]}(x^j)$ are only calculated on a finite grid of points and integrals in (2.6) are replaced by discrete approximations. Suppose that the number of grid points is of order n^β for some $\beta > 0$. Then in the initialization step we have to calculate $n^{2\beta}$ two-dimensional kernel density estimators. This results in $O(n^{1+2\beta})$ calculations. Let us briefly discuss this for the case where all functions $f_j(x^j)$ have bounded support and all bandwidths are chosen so that $\widehat{f}_j(x^j) - f_j(x^j)$ is of order $O_p(n^{-2/5})$. It can be shown that one has to choose $\beta > 4/19$ to obtain a numerical error of smaller order than the statistical error. Then the computational complexity of the algorithm is of order $O(n\log(n) + n^{1+2\beta}) = O(n^{1+2\beta}) = O(n^{(27/19)+2\delta})$ with $\delta = \beta - \frac{4}{19}$. This amount of calculations can still be carried out even for large values of n in reasonable time.

Proof of (2.4): To get Eq. (2.4) we multiply both sides of equation (2.3) with $\widehat{p}_{X^j}(x^j)$ and integrate both sides of the resulting equation over x^j. Because of the norming (1.3) this yields:

$$\begin{aligned}
0 &= \int \widehat{f}_j(x^j)\widehat{p}_{X^j}(x^j)\,dx^j \\
&= \int \widehat{f}_j^*(x^j)\widehat{p}_{X^j}(x^j)\,dx^j - \widehat{\mu}\int \widehat{p}_{X^j}(x^j)\,dx^j - \sum_{k\neq j}\int \widehat{p}_{X^j,X^k}(x^j,x^k)\widehat{f}_k(x^k)\,dx^k\,dx^j \\
&= \int \frac{1}{nh_j}\sum_{i=1}^n K\left(\frac{X_i^j - x^j}{h_j}\right) Y_i\,dx^j - \widehat{\mu} - \sum_{k\neq j}\int \widehat{p}_{X^k}(x^k)\widehat{f}_k(x^k)\,dx^k \\
&= \frac{1}{n}\sum_{i=1}^n Y_i - \widehat{\mu},
\end{aligned}$$

where we use the facts that $\int \frac{1}{h_j}K\left(\frac{X_i^j - x^j}{h_j}\right)\,dx^j = 1$ and that $\int \widehat{p}_{X^j,X^k}(x^j,x^k)\,dx^j = \widehat{p}_{X^k}(x^k)$. This completes the proof.

2.2 Asymptotics of the smooth backfitting estimator

Under appropriate conditions, the following result holds for the asymptotic distribution of each component function $\widehat{f}_j(x^j)$, $j = 1,\ldots,d$:

$$\sqrt{nh_j}\left(\widehat{f}_j(x^j) - f_j(x^j) - \beta_j(x^j)\right) \xrightarrow{d} \mathrm{N}\left(0, \int K^2(u)\,du\,\frac{\sigma_j^2(x^j)}{p_{X^j}(x^j)}\right). \tag{2.7}$$

Here the asymptotic bias terms $\beta_j(x^j)$ are defined as minimizers of

$$(\beta_1,\ldots,\beta_d) \rightsquigarrow \int [\beta(x) - \beta_1(x^1) - \cdots - \beta_d(x^d)]^2 p_X(x)\,dx$$

under the constraint that

$$\int \beta_j(x^j) p_{X^j}(x^j) dx^j = \frac{1}{2} h_j^2 \int [2f_j'(x^j) p_{X^j}'(x^j) + f_j''(x^j) p_{X^j}(x^j)]\, dx^j \int u^2 K(u)\, du, \tag{2.8}$$

where p_X is the joint density of $X = (X^1, \ldots, X^d)$ and

$$\beta(x) = \frac{1}{2} \sum_{j=1}^{d} h_j^2 \left[2f_j'(x^j) \frac{\partial \log p_X}{\partial x^j}(x) + f_j''(x^j) \right] \int u^2 K(u)\, du.$$

In Mammen, Linton, and Nielsen (1999) and Mammen and Park (2005) this asymptotic statement has been proved for the case that f_j is estimated on a compact interval I_j. The conditions include a boundary modification of the kernel. Specifically, the convolution kernel $h_j^{-1} K(h_j^{-1}(X_i^j - x^j))$ is replaced by $K_{h_j}(X_i^j, x^j) = h_j^{-1} K(h_j^{-1}(X_i^j - x^j)) / \int_{I_j} h_j^{-1} K(h_j^{-1}(X_i^j - u^j))\, du^j$. Then it holds that $\int_{I_j} K_{h_j}(X_i^j, x^j)\, dx^j = 1$. In particular, this implies $\int_{I_j} \widehat{p}_{X_j, X^k}(x^j, x^k) dx^j = \widehat{p}_{X^k}(x^k)$ and $\int_{I_j} \widehat{p}_{X_j}(x^j) dx^j = 1$ if one replaces $h_j^{-1} K(h_j^{-1}(X_i^j - x^j))$ by $K_{h_j}(X_i^j, x^j)$ in the definitions of the kernel density estimators. In fact, we have already made use of these properties of kernel density estimators in the previous subsection.

Before illustrating how the asymptotic result (2.7) is obtained, we discuss its interpretations. In particular, it is illustrative to compare $\widehat{f}_j$ with the Nadaraya-Watson estimator $\widetilde{f}_j$ in the classical nonparametric regression model $Y_i = f_j(X_i^j) + \varepsilon_i$. Under standard smoothness assumptions it holds that

$$\sqrt{nh_j} \left(\widetilde{f}_j(x^j) - f_j(x^j) - \beta_j^*(x^j) \right) \xrightarrow{d} \mathrm{N}\left(0, \int K^2(u)\, du\, \frac{\sigma_j^2(x^j)}{p_{X^j}(x^j)} \right) \tag{2.9}$$

with the asymptotic bias $\beta_j^*(x^j) = \frac{1}{2} h_j^2 \left[2f_j'(x^j) \frac{\partial \log p_{X_j}(x^j)}{\partial x^j} + f_j''(x^j) \right] \int u^2 K(u)\, du$. We see that $\widetilde{f}_j(x^j)$ has the same asymptotic variance as $\widehat{f}_j(x^j)$ but that the two estimators differ in their asymptotic bias. Thus, as long as one only considers the asymptotic variance, one has not to pay any price for not knowing the other additive components f_k $(k \neq j)$. One gets the same asymptotic variance in the additive model as in the simplified model $Y_i = f_j(X_i^j) + \varepsilon_i$ where all other additive components f_k $(k \neq j)$ are set equal to 0. As said, the bias terms differ. The asymptotic bias of $\widehat{f}_j(x^j)$ may be larger or smaller than that of $\widetilde{f}_j(x^j)$. This depends on the local characteristics of the function f_j at the point x^j and also on the global shape of the other functions f_k $(k \neq j)$. It is a disadvantage of the Nadaraya-Watson smooth backfitting estimator. There may be structures in $\widehat{f}_j(x^j)$ that are caused by other functions. We will argue below that this is not the case for the local linear smooth backfitting estimator. For the local linear smooth backfitting estimator one gets the same asymptotic bias and variance as for the local linear estimator in the classical model $Y_i = f_j(X_i^j) + \varepsilon_i$. In particular, both estimators have the same asymptotic normal distribution. In last chapter this was called oracle efficiency. This notion of efficiency is appropriate for nonparametric models. Typically in nonparametric models there exists no asymptotically optimal estimator, in contrast to parametric models and to the case of estimating the parametric parts of semiparametric models.

We now come to a heuristic explanation of the asymptotic result (2.7). For a detailed proof we refer to Mammen, Linton, and Nielsen (1999) and Mammen and Park (2005). The main argument is based

on a decomposition of the estimator into a *mean part* and a *variance part*. For this purpose one applies smooth backfitting to the "data" $(X^1,\dots,X^d, f_1(X^1)+\cdots+f(X^d))$ and to $(X^1,\dots,X^d,\varepsilon)$. We will argue below that $\widehat{f}_j(x^j)$ is the sum of these two estimators.

Justification of (2.7): We start with a heuristic derivation of the asymptotic bias and variance of the smooth backfitting estimator $\widehat{f}_j(x^j)$. For this purpose note first that the smooth backfitting estimators $\widehat{\mu}, \widehat{f}_1, \dots, \widehat{f}_d$ are the minimizers of

$$(\mu, f_1, \dots, f_d) \rightsquigarrow \int [\widehat{f}(x) - \mu - f_1(x^1) - \cdots - f_d(x^d)]^2 \widehat{p}_X(x)\, dx \tag{2.10}$$

under the constraint (1.3), where $\widehat{p}_X$ is the kernel density estimator of p_X and $\widehat{f}$ is the Nadaraya-Watson estimator of the regression function $f(x) = E(Y|X=x)$:

$$\begin{aligned}
\widehat{p}_X(x) &= \frac{1}{nh_1\cdots h_d} \sum_{i=1}^n K\left(\frac{X_i^1 - x^1}{h_1}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right), \\
\widehat{f}(x) &= \widehat{p}_X(x)^{-1} \frac{1}{nh_1\cdots h_d} \sum_{i=1}^n K\left(\frac{X_i^1 - x^1}{h_1}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right) Y_i.
\end{aligned}$$

One may show that this minimization problem leads to (2.3) and (2.4). We omit the details. For a geometric argument see also Mammen, Marron, Turlach, and Wand (2001).

For heuristics on the asymptotics of $\widehat{f}_j$, $1 \le j \le d$, we now decompose $\widehat{f}$ into its bias and variance component $\widehat{f}(x) = \widehat{f}^A(x) + \widehat{f}^B(x)$, where

$$\begin{aligned}
\widehat{f}^A(x) &= \widehat{p}_X(x)^{-1} \frac{1}{nh_1\cdots h_d} \sum_{i=1}^n K\left(\frac{X_i^1 - x^1}{h_1}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right) \varepsilon^i, \\
\widehat{f}^B(x) &= \widehat{p}_X(x)^{-1} \frac{1}{nh_1\cdots h_d} \sum_{i=1}^n K\left(\frac{X_i^1 - x^1}{h_1}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right) [\mu + f_1(x^1) + \cdots + f_d(x^d)].
\end{aligned}$$

Denote by $(\widehat{\mu}^A, \widehat{f}_1^A, \dots, \widehat{f}_d^A)$ the minimizer of

$$(\mu, f_1, \dots, f_d) \rightsquigarrow \int [\widehat{f}^A(x) - \mu - f_1(x^1) - \cdots - f_d(x^d)]^2 \widehat{p}_X(x)\, dx$$

under the constraint (1.3), and by $(\widehat{\mu}^B, \widehat{f}_1^B, \dots, \widehat{f}_d^B)$ the minimizer of

$$(\mu, f_1, \dots, f_d) \rightsquigarrow \int [\widehat{f}^B(x) - \mu - f_1(x^1) - \cdots - f_d(x^d)]^2 \widehat{p}_X(x)\, dx$$

under the constraint (1.3). Then, we obtain $\widehat{\mu} = \widehat{\mu}^A + \widehat{\mu}^B$, $\widehat{f}_1 = \widehat{f}_1^A + \widehat{f}_1^B, \dots, \widehat{f}_d = \widehat{f}_d^A + \widehat{f}_d^B$. By standard smoothing theory, $\widehat{f}^B(x) \approx \mu + f_1(x^1) + \cdots + f_d(x^d) + \beta(x)$. This immediately implies that $\widehat{f}_j^B(x^j) \approx c_j + f_j(x^j) + \beta_j(x^j)$ with a random constant c_j. Our constraint (2.8) implies that c_j can be chosen equal to zero. This follows by some more lengthy arguments that we omit.

For an understanding of the asymptotic result (2.7) it remains to show that

$$\sqrt{nh_j}\left(\widehat{f}_j^A(x^j) - f_j(x^j)\right) \xrightarrow{d} \mathrm{N}\left(0, \int K^2(u)\, du\, \frac{\sigma_j^2(x^j)}{p_{X^j}(x^j)}\right). \tag{2.11}$$

To see this claim we proceed similarly as in the derivation of (2.3). Using essentially the same arguments as there one can show that

$$\widehat{f}_j^A(x^j) = \widehat{f}_j^{A,*}(x^j) - \widehat{\mu}^A - \sum_{k\neq j} \int \frac{\widehat{p}_{X^j,X^k}(x^j,x^k)}{\widehat{p}_{X^j}(x^j)} \widehat{f}_k^A(x^k)\, dx^k, \tag{2.12}$$

where

$$\widehat{f}_j^{A,*}(x^j) = \left[\sum_{i=1}^n K\left(\frac{X_i^j - x^j}{h_j}\right)\right]^{-1} \sum_{i=1}^n K\left(\frac{X_i^j - x^j}{h_j}\right)\varepsilon_i$$

is the stochastic part of the marginal Nadaraya-Watson estimator $\widehat{f}_j^*(x^j)$. We now argue that

$$\int \frac{\widehat{p}_{X^j,X^k}(x^j,x^k)}{\widehat{p}_{X^j}(x^j)} \widehat{f}_k^A(x^k)\ dx^k \approx \int \frac{p_{X^j,X^k}(x^j,x^k)}{p_{X^j}(x^j)} \widehat{f}_k^A(x^k)\ dx^k \approx 0.$$

The basic argument for the second approximation is that a global average of a local average behaves like a global average, or more explicitly, consider e.g., the local average $\widehat{r}_j(x^j) = (nh_j)^{-1}\sum_{i=1}^n K\left(\frac{X_i^j - x^j}{h_j}\right)\varepsilon_i$. This local average is of order $O_p(n^{-1/2}h_j^{-1/2})$. For a smooth weight function w we now consider the global average $\widehat{\rho}_j = \int_{I_j} w(x^j)\widehat{r}_j(x^j)\ dx^j$ of the local average $\widehat{r}_j(x^j)$. This average is of order $O_p(n^{-1/2}) = o_p(n^{-1/2}h_j^{-1/2})$ because of

$$\begin{aligned}
\widehat{\rho}_j &= \int_{I_j} w(x^j)\widehat{r}_j(x^j)\ dx^j \\
&= \int_{I_j} w(x^j)(nh_j)^{-1}\sum_{i=1}^n K\left(\frac{X_i^j - x^j}{h_j}\right)\varepsilon_i\, dx^j \\
&= n^{-1}\sum_{i=1}^n w_{h_j}(X_i^j)\varepsilon_i
\end{aligned}$$

with $w_{h_j}(X_i^j) = \int_{I_j} w(x^j)h_j^{-1}K\left(\frac{X_i^j - x^j}{h_j}\right)\ dx^j$.

2.3 Smooth backfitting local linear estimator

In the additive model (1.1) the smooth backfitting local linear estimators $\widehat{\mu}, \widehat{f}_1, \widehat{f}_1^\dagger, \ldots, \widehat{f}_d, \widehat{f}_d^\dagger$ are defined as the minimizers of the smoothed least squares criterion

$$\begin{aligned}
&\int \sum_{i=1}^n \left[Y_i - \mu - f_1(x^1) - f_1^\dagger(x^1)(X_i^1 - x^1) - \cdots - f_d(x^d) - f_d^\dagger(x^d)(X_i^d - x^d)\right]^2 \\
&\qquad \times K\left(\frac{X_i^1 - x^1}{h_1}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right)\ dx^1 \cdots dx^d
\end{aligned} \tag{2.13}$$

under the constraint (1.3). This is a natural generalization of the local linear estimator. For the case $d = 1$ the minimization gives the classical local linear estimator as the minimization of (1.4) leads to the classical Nadaraya-Watson estimator. The estimators, $\widehat{f}_j^\dagger$, $1 \le j \le d$, are estimators of the derivatives of the additive components f_j.

The smooth backfitting local linear estimator is given as the solution of a random integral equation. Similarly to Eq. (2.3), the tuples $(\widehat{f}_j, \widehat{f}_j^\dagger)$ fulfill now a two-dimensional integral equation. This integral equation can be used for the iterative calculation of the estimators. For details we refer to Mammen, Linton, and Nielsen (1999). We only mention the following asymptotic result from Mammen, Linton, and Nielsen (1999) for the smooth backfitting local linear estimator that holds under appropriate conditions: for $1 \le j \le d$

$$\sqrt{nh_j}\left(\widehat{f}_j(x^j) - f_j(x^j) - \beta_j(x^j)\right) \xrightarrow{d} \mathrm{N}\left(0, \int K^2(u)\ du\ \frac{\sigma_j^2(x^j)}{p_{X^j}(x^j)}\right), \tag{2.14}$$

where now the asymptotic bias terms $\beta_j(x^j)$ are defined as

$$\beta_j(x^j) = \frac{1}{2}h_j^2 \left[f_j''(x^j) - \int f_j''(u^j) p_{X^j}(u^j)\, du^j \right] \int u^2 K(u)\, du.$$

Up to an additive norming term, the asymptotic bias of $\widehat{f}_j(x^j)$ coincides with the asymptotic bias of local linear estimator $\widetilde{f}_j$ in the classical nonparametric regression model $Y_i = f_j(X_i^j) + \varepsilon_i$. Moreover, we get the same asymptotic distribution for both estimators (up to an additive norming term). Asymptotically one does not lose any efficiency by not knowing the additive components $f_k : k \neq j$ compared to the *oracle model* where these components are known. This is an asymptotic optimality result for the local linear smooth backfitting. It achieves the same asymptotic bias and variance as in the oracle model. As discussed above, the Nadaraya-Watson smooth backfitting estimator achieves only the asymptotic variance of the oracle model. For an alternative implementation of local linear smooth backfitting, see Mammen and Park (2006).

2.4 Smooth backfitting as solution of a noisy integral equation

We write the smooth backfitting estimators as solutions of an integral equation. We discuss this briefly for Nadaraya-Watson smoothing. Put $\widehat{\mathbf{f}}(x_1, \ldots, x_d) = (\widehat{f}_1(x_1), \ldots, \widehat{f}_d(x_d))^\top$ and $\widehat{\mathbf{f}}^*(x_1, \ldots, x_d) = (\widehat{f}_1^*(x_1), \ldots, \widehat{f}_d^*(x_d))^\top$. With this notation we can rewrite (2.3) as

$$\widehat{\mathbf{f}}(x) = \widehat{\mathbf{f}}^*(x) - \int \widehat{\mathcal{H}}(x, z) \widehat{\mathbf{f}}(z)\, dz, \tag{2.15}$$

where for each value of $x, z \in \mathbb{R}$ the integral kernel $\widehat{\mathcal{H}}(x, z)$ is a matrix with element (j, k) equal to $\widehat{p}_{X^j, X^k}(x^j, x^k)/\widehat{p}_{X^j}(x^j)$. This representation motivates an alternative algorithm. One can use a discrete approximation of the integral equation and approximate the integral equation (2.15) by a finite linear equation. This can be solved by standard methods of linear algebra. Eq. (2.15) can also be used as an alternative starting point for an asymptotic analysis of the estimator $\widehat{\mathbf{f}}$. We will come back to this in Section 5 after having discussed further on those models in Section 3 whose estimation can be formulated as solving an integral equation.

2.5 Relations to classical backfitting and two-stage estimation

Smooth backfitting (2.5) is related to classical backfitting and to two-stage estimation. In the classical backfitting, the j-th step of the lth iteration cycle (2.5) of the smooth backfitting is replaced by

$$\widehat{f}_j^{[l]}(X_i^j) = \widehat{p}_{X^j}(x^j)^{-1} \frac{1}{nh_j} \sum_{i=1}^n K\left(\frac{X_i^j - x^j}{h_j} \right) \left[Y_i - \widehat{\mu} - \sum_{k=1}^{j-1} \widehat{f}_k^{[l]}(X_i^k) - \sum_{k=j+1}^{d} \widehat{f}_k^{[l-1]}(X_i^k) \right] \tag{2.16}$$

for $1 \leq j \leq d$ and $1 \leq i \leq n$. This iteration equation can be interpreted as a limiting case of (2.5) where one lets the second bandwidth h_k in the definition of the kernel density estimator $\widehat{p}_{X^j, X^k}(x^j, x^k)$ converge to zero.

If the backfitting algorithm runs through $O(\log n)$ cycles, the algorithm needs $O(n \log n)$ calculation steps. This is slightly faster than the smooth backfitting. In contrast to the smooth backfitting, the backfitting estimator is only defined as the limit of the iterative algorithm (2.16). Note that the smooth backfitting is explicitly defined as minimizer of the smoothed least squares criterion (1.2). The fact that backfiitng estimators are only implicitly defined as limit of an iterative algorithm complicates the asymptotic mathematical analysis. Note also that the algorithm runs in $\mathbb{R}^n$, i.e., in spaces with increasing dimension. An asymptotic treatment of the classical backfitting can be found in Opsomer (2000) and Opsomer and Ruppert (1997). Nielsen and Sperlich (2005) illustrated by simulation that smooth backfitting, in comparison with the classical backfitting, is more robust against degenerated designs and a large number of additive components. The reason behind this is that the iteration equation (2.5) is a smoothed version of (2.16). The smoothing stabilizes the "degenerated integral equation" (2.16). In Opsomer (2000) and Opsomer and Ruppert (1997) stronger assumptions are made on the joint density of the covariates than are needed for the study of the smooth backfitting. This may be caused by the same reasons, but there has been made no direct theoretical argument that supports the empirical finding that the classical backfitting is more sensitive to degenerate designs than smooth backfitting. For another modification of the classical backfitting that takes care of correlated covariates, see Jiang, Fan, and Fan (2010).

Two-stage estimation differs from smooth backfitting in several respects. First of all, only two steps are used instead of an iterative algorithm that runs until a convergence criterion is fulfilled. Furthermore, different bandwidths are used in different steps: undersmoothing is done in the first-step, but an optimal bandwidth is chosen in the second-step. The algorithm of two-step estimation is as simple as that of backfitting. On the other hand, choice of the bandwidth in the first-step is rather complex. Asymptotically, optimal choices will not affect the first order properties of the outcomes of the second-step. But for finite samples the influence of the first-step bandwidth is not clear. The calculation of theoretically optimal values would require a second-order optimal theory that is not available and, as other higher-order theory, may not be accurate for small to moderate sample sizes. In particular, in models with many nonparametric components, backfitting may be preferable because it does not require an undersmoothing step.

Another kernel smoothing method that can be applied to additive models is marginal integration. It has been discussed in last chapter that marginal integration only achieves optimal rates for low dimensional additive models but that it does not work in higher-dimensional models. This drawback is not shared by backfitting, smooth backfitting and two-stage estimation. There is also another aspect in which smooth backfitting and marginal integration differ. If the additive model is not correct smooth backfitting as a weighted least squares estimator estimates the best additive fit to the non-additive model. On the other side, marginal integration estimates a weighted average effect for each covariate. This follows because marginal integration is based on a weighted average of the full dimensional regression function. Thus, the methods estimate quite different quantities if the model is not additive.

2.6 Bandwidth choice and model selection

Bandwidth selection for additive models has been discussed in Mammen and Park (2005). There, consistency has been shown for bandwidth selectors based on plug-in and penalized least squares criteria. Nielsen and Sperlich (2005) discusses practical implementations of cross validation methods. Because an additive model contains several nonparametric functions there exist two types of optimal bandwidths: bandwidths that are optimal for the estimation of the sum of the additive components and bandwidths that optimize estimation of a single additive component. While the former criterion may be in particular appropriate in prediction, the latter is more motivated in data analytic oriented inference. Whereas all three bandwidth selectors (cross validation, penalized least squares and plug-in) can be designed for the former criterion, only plug-in based approaches can be used. For a further discussion we refer to the above two papers. For the models that will be discussed in the next section bandwidth selection has been only partially studied. The asymptotic results for the estimators that will be discussed can be used to design plug-in methods. For cross validation it is questionable if for all models algorithms can be found that run in reasonable time.

In very high dimensional additive models backfitting methods will suffer from the complexity of the models, in statistical performance and in computational costs. For this reason component selection is an important step to control the size of the model. Recently, some proposals have been made that are influenced by the study of high dimensional models with sparsity constraints. We refer to Lin and Zhang (2006), Meier, van de Geer, and Bühlmann (2009), and Huang, Horowitz, and Wei (2010).

2.7 Generalized additive models

We now discuss nonlinear extensions of the additive models. In a generalized additive model a link function g is introduced and it is assumed that the following equation holds for the regression function $E(Y|X_1,\dots,X_d)$

$$E(Y|X_1,\dots,X_d) = g^{-1}\{\mu + f_1(X^1) + \cdots + f_d(X^d)\}.$$

It has been considered that the link function is known or that it is unknown and has to be estimated. An important example where generalized additive models make sense is the case of binary responses Y. If Y is $\{0,1\}$-valued, the function g^{-1} maps the additive function onto the interval $[0,1]$. In the generalized additive model, the additive functions $f_1,\dots,f_d$ can be estimated by smoothed least squares. An alternative approach for heterogenous errors is a smoothed quasi-likelihood criterion. Quasi-likelihood is motivated for regression models where the conditional variance of the errors is equal to $V(\mu)$ with μ equal to the conditional expectation of Y. Here, V is a specified variance function. Quasi-likelihood coincides with classical likelihood if the conditional error distribution is an exponential family. It also leads to consistent estimators if the conditional variances have another form. The quasi-likelihood criterion $Q(\mu, y)$ is defined as:

$$\frac{\partial}{\partial\mu}Q(\mu,y) = \frac{y-\mu}{V(\mu)}.$$

An early reference to quasi-likelihood approaches in additive models is Hastie and Tibshirani (1990). For the discussion of local linear smoothing in generalized partially linear models see also Fan, Heckman, and Wand (1995). For a discussion of the asymptotics of classical backfitting in generalized additive model, see Kauermann and Opsomer (2003). The Smoothed Quasi-Likelihood criterion is defined as follows: Minimize for $\mathbf{f} = (\mu, f_1, \ldots, f_d)^\top$

$$SQ(\mathbf{f}) = \int \sum_{i=1}^{n} Q(g^{-1}(f^+(x)), Y_i) K\left(\frac{X_i^1 - x^1}{h_1}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right) dx^1 \cdots dx^d.$$

where $f^+(x) = \mu + f_1(x^1) + \cdots + f_d(x^d)$. Minimization of the smoothed quasi-likelihood criterion over $\mathbf{f}$ results in the smoothed maximum quasi-liklihood estimator. Algorithms for the calculation of this estimator were discussed in Yu, Park, and Mammen (2008). In that paper an asymptotic theory for this estimator was also developed. In other applications the quasi-likelihood criterion may be replaced by other M-functionals. We do not discuss this here. An example is quantile regression. For a discussion of backfitting and smooth backfitting in additive quantile models, see Lee, Mammen, and Park (2010).

3 Some models that are related to additive models.

In linear regression, the standard least squares method produces consistent estimators when the errors are uncorrelated. When the errors are correlated, the method may not give consistent or efficient estimators of the regression parameters. In the latter case it is often appropriate to take a linear transformation of the response variable in such a way that it corrects for the correlations between the errors. Linear transformations may be also used to remove some unobserved effects in a regression model that are correlated with the regressors or errors. Taking a linear transformation in parametric linear models does not alter the linear structure of the model, so that conventional methods still work with the transformed data. In nonparametric regression models, however, it often yields an additive model where classical smoothing methods can not be applied, as we illustrate on several cases in this section. Some of the models of this section were also discussed in the overview papers Linton and Mammen (2003) and Mammen and Yu (2009). A general discussion of smooth least squares in a general class of nonparametric models can also be found in Mammen and Nielsen (2003).

3.1 Nonparametric regression with time series errors

Suppose we observe (X_t, Y_t) for $1 \leq t \leq T$ such that $Y_t = f(X_t) + u_t$, where the errors u_t have an AR(1) time series structure so that $\varepsilon_t = u_t - \rho u_{t-1}$ is a sequence of uncorrelated errors. The transformed model $Z_t(\rho) \equiv Y_t - \rho Y_{t-1} = f(X_t) - \rho f(X_{t-1}) + \varepsilon_t$ has uncorrelated errors, but has an additive structure in the mean function. For simplicity, assume that the errors u_t are independent of the covariates X_t. Then, the target function f minimizes

$$Q_T(m) = \frac{1}{T} \sum_{t=1}^{T} E\left[Z_t(\rho) - m(X_t) + \rho m(X_{t-1})\right]^2$$

over m, so that it satisfies

$$\int \left[E(Z_t(\rho)|X_t = x, X_{t-1} = y) - f(x) + \rho f(y)\right] \left[g(x) - \rho g(y)\right] f_{0,1}(x,y)\, dx\, dy = 0 \tag{3.1}$$

for all square integrable functions g. Here $f_{0,1}$ denotes the joint density of (X_t, X_{t-1}) and f_0 is the density of X_t. The equation (3.1) holds for all square integrable functions g if and only if

$$f(x) = f^*_\rho(x) - \int \mathcal{H}_\rho(x,y) f(y)\, dy \tag{3.2}$$

where

$$f^*_\rho(x) = \frac{1}{1+\rho^2}\left[E(Z_t(\rho)|X_t = x) - \rho E(Z_t(\rho)|X_{t-1} = x)\right],$$
$$\mathcal{H}_\rho(x,y) = -\frac{\rho}{1+\rho^2}\left[\frac{f_{0,1}(x,y)}{f_0(x)} + \frac{f_{0,1}(y,x)}{f_0(x)}\right].$$

An empirical version of the integral equation (3.2) may be obtained by estimating f_0, $f_{0,1}$, $E(Z_t(\rho)|X_t = \cdot)$ and $E(Z_t(\rho)|X_{t-1} = \cdot)$. Let $\widehat{f}(\cdot, \rho)$ denotes the solution of the latter integral equation. In case ρ is known, $\widehat{f}(\cdot, \rho)$ can be used as an estimator of f. Otherwise, the parameter ρ can be estimated by $\widehat{\rho}$ that minimizes

$$\frac{1}{T}\sum_{t=1}^{T}\left[Z_t(\rho) - \widehat{f}(X_t, \rho) + \rho\widehat{f}(X_{t-1}, \rho)\right]^2,$$

and then f by $\widehat{f} = \widehat{f}(\cdot, \widehat{\rho})$. We note that the estimator $\widehat{f}(\cdot, \rho)$ is consistent even if the autoregressive coefficient $\rho = 1$. In contrast, smoothing of the original untransformed data (Y_t, X_t) leads to an inconsistent estimator. We mentioned this example already in the introduction.

The above discussion may be extended to a general setting where the errors u_t admit a time series structure such that $\varepsilon_t = \sum_{j=0}^{\infty} a_j u_{t-j}$ is a sequence of uncorrelated errors. In this general case, if we take the transformation $Z_t(a_0, a_1, \ldots) = \sum_{j=0}^{\infty} a_j Y_{t-j}$, then the transformed model $Z_t(a_0, a_1, \ldots) = \sum_{j=0}^{\infty} a_j f(X_{t-j}) + \varepsilon_t$ has an additive structure with uncorrelated errors. For a discussion of this general case, see Linton and Mammen (2008). There weaker assumptions are made on the errors u_t. In particular, it is not assumed that the errors u_t are independent of the covariates X_t.

3.2 Nonparametric regression with repeated measurements.

Suppose that one has J repeated measurements on each of n subjects. Let (X_{ij}, Y_{ij}) be the jth observation on the ith subject. Write $\mathbf{X}_i = (X_{i1}, \ldots, X_{iJ})^\top$ and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^\top$. Assume that $(\mathbf{X}_i, \mathbf{Y}_i), i = 1 \ldots, n$, are i.i.d. copies of $(\mathbf{X}, \mathbf{Y})$. Consider the simple nonparametric regression model

$$Y_{ij} = f(X_{ij}) + \epsilon_{ij}, \tag{3.3}$$

where the errors ϵ_{ij} have zero conditional mean, but are allowed to be correlated within each subject. Let $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iJ})^\top$ and $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{\epsilon}_i)$. The kernel regression estimator based on the ordinary least squares criterion is consistent even in this case where $\boldsymbol{\Sigma}$ is not the identity matrix. However, we may find a better estimator which is based on a weighted least squares criterion. This is in line with parametric

linear regression with repeated measurements, where a weighted least squares estimator outperforms the ordinary least squares estimator. A weighted least squares estimation is equivalent to taking a linear transformation of the response and then applying the ordinary least squares criterion to the transformed model. In contrast to the parametric case, introducing weights in the nonparametric model (3.3) leads to a more complicated estimation problem, as is demonstrated below.

Let $\mathbf{f}(x_1, \ldots, x_J) = (f(x_1), \ldots, f(x_J))^\top$. The regression function f at (3.3) minimizes

$$E[\{\mathbf{Y} - \mathbf{m}(X_1, \ldots, X_J)\}^\top \mathbf{\Sigma}^{-1} \{\mathbf{Y} - \mathbf{m}(X_1, \ldots, X_J)\}] \tag{3.4}$$

over all square integrable functions m, where $\mathbf{m}(x_1, \ldots, x_J) = (m(x_1), \ldots, m(x_J))^\top$. Note that the transformed response vector $\mathbf{\Sigma}^{-1/2}\mathbf{Y}$ admits an additive model and the variance of the transformed error vector $\mathbf{\Sigma}^{-1/2}\boldsymbol{\epsilon}$ equals the identity matrix. The minimizer f satisfies

$$\sum_{j=1}^{J} \sum_{k=1}^{J} \sigma^{jk} E\{[Y_j - f(X_j)]g(X_k)\} = 0$$

for all square integrable functions g, where σ^{jk} denotes the (j,k)th entry of the matrix $\mathbf{\Sigma}^{-1}$. This gives the following integral equation for f;

$$f(x) = f^*(x) - \int \mathcal{H}(x,z) f(z)\, dz, \tag{3.5}$$

where

$$f^*(x) = \left[\sum_{j=1}^{J} \sigma^{jj} p_j(x)\right]^{-1} \sum_{j=1}^{J} \sum_{k=1}^{J} \sigma^{jk} E(Y_k | X_j = x) p_j(x),$$

$$\mathcal{H}(x,z) = \left[\sum_{j=1}^{J} \sigma^{jj} p_j(x)\right]^{-1} \sum_{j=1}^{J} \sum_{k \neq j}^{J} \sigma^{jk} p_{jk}(x,z).$$

Here, p_j and p_{jk} denote the densities of X_j and (X_j, X_k), respectively. The quantities f^*, p_j and p_{jk} can be estimated by the standard kernel smoothing techniques. Plugging these into (3.5) gives an integral equation for estimating f.

One may apply other weighting schemes replacing $\mathbf{\Sigma}^{-1}$ at (3.4) by a weight matrix $\mathbf{W}$. It can be shown the choice $\mathbf{W} = \mathbf{\Sigma}^{-1}$ leads to an estimator with the minimal variance, see Carroll, Maity, Mammen, and Yu (2009) for details. The foregoing weighted least squares regression may be extended to the additive regression model $Y_{ij} = \sum_{d=1}^{D} f_d(X_{ij}^d) + \epsilon_{ij}$ with covariates $\mathbf{X}_{ij} = (X_{ij}^1, \ldots, X_{ij}^D)^\top$. Details are also given in Carroll, Maity, Mammen, and Yu (2009).

3.3 Panels with individual effects

Suppose we have panel data (X_{ij}, Y_{ij}) for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. We assume that

$$Y_{ij} = f(X_{ij}) + \alpha_i + \epsilon_{ij}, \tag{3.6}$$

where α_i are the unobserved random or nonrandom individual effects that are invariant over time j, and ϵ_{ij} are errors such that $E(\epsilon_{ij} | X_{i1}, \ldots, X_{iJ}) = 0$. The individual effect α_i can be uncorrelated or correlated

with the regressors $X_{i1}, \ldots, X_{iJ}$ and the error variables ϵ_{ij}. If $E(\alpha_i | X_{i1}, \ldots, X_{iJ}) = 0$, then the model reduces to the model considered in Subsection 3.2. An interesting case is when the individual effect is correlated with the regressors so that $E(\alpha_i | X_{i1}, \ldots, X_{iJ}) \neq 0$. In this case, the ordinary nonparametric kernel regression fails to obtain a consistent estimator. Recall that the latter is also the case with parametric linear regression.

Here again, we may use a simple linear transformation to remove the unobserved individual effect from the regression model. Let $Z_i = \sum_{j=1}^J a_j Y_{ij}$ for some constants a_j such that $\sum_{j=1}^J a_j = 0$. Examples include

(i) $a_1 = \cdots = a_{k-2} = 0$, $a_{k-1} = -1$, $a_k = 1$, $a_{k+1} = \cdots = a_J = 0$ for some $1 \leq k \leq J$;

(ii) $a_1 = \cdots = a_{k-1} = -J^{-1}$, $a_k = 1 - J^{-1}$, $a_{k+1} = \cdots = a_J = -J^{-1}$ for some $1 \leq k \leq J$.

For the transformed response variables Z_i, we obtain

$$Z_i = \sum_{j=1}^J a_j f(X_{ij}) + u_i, \tag{3.7}$$

where $u_i = \sum_{j=1}^J a_j \epsilon_{ij}$ has zero conditional mean given $X_{i1}, \ldots, X_{iJ}$. Let Z and X_j denote the generics of Z_i and X_{ij}, respectively. Since f minimizes the squared error risk $E[Z - \sum_{j=1}^J a_j m(X_j)]^2$ over m, it satisfies

$$E\left\{\left[Z - \sum_{j=1}^J a_j f(X_j)\right] \sum_{j=1}^J a_j g(X_j)\right\} = 0 \tag{3.8}$$

for all square integrable functions g. The equation (3.8) is equivalent to

$$\int \left[\sum_{j=1}^J a_j E(Z|X_j = x) p_j(x) - \sum_{j=1}^J \sum_{k \neq j}^J a_j a_k E[f(X_k)|X_j = x] p_j(x) - f(x) \sum_{j=1}^J a_j^2 p_j(x)\right] g(x)\, dx = 0,$$

where p_j and p_{jk} denote the density of X_j and (X_j, X_k), respectively. This gives the following integral equation

$$f(x) = f^*(x) - \int \mathcal{H}(x, z) f(z)\, dz, \tag{3.9}$$

where

$$f^*(x) = \left[\sum_{j=1}^J a_j^2 p_j(x)\right]^{-1} \sum_{j=1}^J a_j E(Z|X_j = x) p_j(x),$$

$$\mathcal{H}(x, z) = \left[\sum_{j=1}^J a_j^2 p_j(x)\right]^{-1} \sum_{j=1}^J \sum_{k \neq j}^J a_j a_k p_{jk}(x, z).$$

As in the additive regression model we need a norming condition for identification of f in the transformed model (3.7). The reason is that in the transformed model we have $\sum_{j=1}^J a_j f(X_{ij}) = \sum_{j=1}^J a_j [c + f(X_{ij})]$ for any constant c since $\sum_{j=1}^J a_j = 0$. We may also see this from the integral equation (3.9) since $\int \mathcal{H}(x, z)\, dz = -1$. For a norming condition, we may define α_i such that $E(Y_{ij}) = Ef(X_{ij})$.

This motivates the normalizing constraint

$$J^{-1}\sum_{j=1}^{J}\int \widehat{f}(x)\widehat{p}_j(x)\,dx = n^{-1}J^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J} Y_{ij}$$

for an estimator $\widehat{f}$ of f. For a kernel estimator based on differencing see also Henderson, Carroll and Li (2008).

The differencing technique we have discussed above may also be applied to a more general setting that allows for discrete response variables. For example, consider a binary response model where each of the n subjects has matched observations (X_{ij}, Y_{ij}) such that the responses Y_{ij}, conditionally on the regressors $X_{i1}, \ldots, X_{iJ}$ and the individual effect α_i, are independent across j and have Bernoulli distributions with success probabilities $p(X_{ij}, \alpha_i)$, respectively. Assume that

$$\log\left[\frac{p(X_{ij}, \alpha_i)}{1 - p(X_{ij}, \alpha_i)}\right] = f(X_{ij}) + \alpha_i$$

and consider the case where $J = 2$ for simplicity. Let $Z_i = I(Y_{i1} = 1)$ and $N_i = Y_{i1} + Y_{i2}$, where I denotes the indicator function. Then, it can be shown that

$$\log\left[\frac{E(Z_i|X_{i1}, X_{i2}, N_i = 1)}{1 - E(Z_i|X_{i1}, X_{i2}, N_i = 1)}\right] = f(X_{i1}) - f(X_{i2}). \tag{3.10}$$

This follows from the equation

$$E(Z_i|X_{i1}, X_{i2}, N_i = 1) = \frac{E\left[p(X_{i1}, \alpha_i)(1 - p(X_{i2}, \alpha_i))\big|X_{i1}, X_{i2}\right]}{E\left[p(X_{i1}, \alpha_i)(1 - p(X_{i2}, \alpha_i)) + p(X_{i2}, \alpha_i)(1 - p(X_{i1}, \alpha_i))\big|X_{i1}, X_{i2}\right]}$$

and the fact that

$$\frac{p(X_{i1}, \alpha_i)[1 - p(X_{i2}, \alpha_i)]}{p(X_{i1}, \alpha_i)[1 - p(X_{i2}, \alpha_i)] + p(X_{i2}, \alpha_i)[1 - p(X_{i1}, \alpha_i)]} = \frac{\exp[f(X_{i1}) - f(X_{i2})]}{1 + \exp[f(X_{i1}) - f(X_{i2})]}$$

does not involve α_i. This generalizes an observation that has been made for parametric conditional maximum likelihood estimation in panel logit models, see Rasch (1960),Rasch (1961), Andersen (1970), and Chamberlain (1994). For extensions of the conditional logit approach see Magnac (2004).

Let Z, X_j, Y_j denote the generics for Z_i, X_{ij}, Y_{ij}, respectively. The function f in the transformed model (3.10) maximizes the expected log-likelihood, so that it satisfies

$$E\, I(N = 1)\,[Z - \eta(X_1, X_2; f)]\,[g(X_1) - g(X_2)] = 0$$

for all square integrable function g, where

$$\eta(x, y; m) = \frac{\exp[m(x) - m(y)]}{1 + \exp[m(x) - m(y)]}.$$

It can be shown that f satisfies $F(f) = 0$, where F is a nonlinear operator defined by

$$F(m)(x) = E\left[I(N = 1)(Z - \eta(X_1, X_2; m))\big|\, X_1 = x\right] p_1(x) - E\left[I(N = 1)(Z - \eta(X_1, X_2; m))\big|\, X_2 = x\right] p_2(x)$$

and p_j denotes the density of X_j, $j = 1, 2$. Here, we also need a norming condition for identifiability of f. The integral equation $F(m) = 0$ is nonlinear, but it can be linearized in the same way as the nonlinear

equation in Section 2. The linear approximation basically puts the problem back to the framework for the model (3.6). To detail this, define $\eta_1(x,y;m) = [1+\exp(m(x)-m(y))]^{-2}$, let $f^{[0]}$ be an a function $f^{[0]}$ close to f. Note that $F(m) \simeq F(f^{[0]}) + F_1(f^{[0]})(m-f^{[0]})$ where $F_1(f^{[0]})$ is a linear operator and $F_1(f^{[0]})(g)$ denotes the Fréchet differential of F at $f^{[0]}$ with increment g. Put $\delta = f - f^{[0]}$ and

$$\begin{aligned}\mathcal{H}_0(x,y) = E\left[I(N=1)|X_1=x, X_2=y\right]\eta_1(x,y;f^{[0]})p_{12}(x,y)\\ + E\left[I(N=1)|X_1=y, X_2=x\right]\eta_1(y,x;f^{[0]})p_{12}(y,x),\end{aligned}$$

where p_{12} denotes the density of (X_1, X_2). Then, the approximating linear integral equation $F(f^{[0]}) + F_1(f^{[0]})(\delta) = 0$ is equivalent to

$$\delta(x) = \delta^*(x) - \int \mathcal{H}(x,y)\delta(y)\,dy, \tag{3.11}$$

where

$$\begin{aligned}\delta^*(x) &= \left[\int \mathcal{H}_0(x,y)\,dy\right]^{-1} F(f^{[0]})(x),\\ \mathcal{H}(x,y) &= -\left[\int \mathcal{H}_0(x,z)\,dz\right]^{-1}\mathcal{H}_0(x,y).\end{aligned}$$

We may estimate F and $\mathcal{H}_0$ by kernel methods. Plugging the estimators $\widehat{F}$ and $\widehat{\mathcal{H}}_0$ into (3.11) gives an integral equation for the update $\widehat{f}^{[1]}$ of the starting estimator $\widehat{f}^{[0]}$. The statistical properties of the resulting backfitting algorithm and the limit of the algorithm $\widehat{f}$ which satisfies $\widehat{F}(\widehat{f}) = 0$ have been studied by Hoderlein, Mammen, and Yu (2011).

3.4 Additive models for panels of time series and factor models

Similar to (3.6), one can consider models with an unobserved time effect η_t instead of an individual effect. We now denote time by t. Suppose that we have panel data $(X_{it}^1, \ldots, X_{it}^d, Y_{it})$ for individuals $1 \le i \le n$ and time points $1 \le T$. We assume that

$$Y_{it} = \sum_{j=1}^{d} m_j(X_{it}^j) + \eta_t + \varepsilon_{it}. \tag{3.12}$$

This model naturally generalizes linear panel data models. It has been studied in Mammen, Støve, and Tjøstheim (2009) for two asymptotic frameworks: $n \to \infty, T$ fixed and $n, T \to \infty$. Their asymptotic analysis includes the case where $\{X_{it}^j\}$, $j = 1, \ldots, p$, are time lagged values of Y_{it}. No assumptions are made on the unobserved temporary effects η_t. They may be deterministic or random, and they may be correlated with covariates or error terms. The basic idea of Mammen, Støve, and Tjøstheim (2009) is to use difference schemes that cancel out the time effects η_t, simliar to the approaches in the last subsection that cancel out individual effects. Here, the values η_t are nuissance parameters.

In Linton and Nielsen (2009) also the model (3.12) is considered, but the statistical aim there is inference on the structure of η_t. It is assumed that η_t is a random process following a parametric specification. A two-step procedure is proposed where the process η_t is fitted in the first-step. In

their mathematics they compare parametric inference based on the fitted values of η_t with an infeasible statistical inference that is based on the unobserved η_t. The main result is that these two approaches are asymptotically equivalent. This can be interpreted as an oracle property and it can be used to construct efficient estimators of the parameters.

Another modification of model (3.12) is the factor model

$$Y_{tl} = m_0(X_{tl}^0) + \sum_{j=1}^{d} Z_t^j m_j(X_{tl}^j) + \varepsilon_{tl} \tag{3.13}$$

for $l = 1, \ldots, L$. Here, the dynamics of the L-dimensional process Y_t is approximated by the unobserved d-dimensional time series Z_t. The basic idea is that elements Y_{tl} of Y_t with similar characteristics $(X_{tl}^j : 1 \leq j \leq d)$ show similar dynamics and that the dynamics of Y_t can be accurately modeled by choices of d that are much smaller than L. This model has been applied in Connor, Hagmann, and Linton (2012) to the analysis of stock returns Y_{tl} with characteristics $(X_{tl}^j : 1 \leq j \leq d)$. Again, a two-step procedure is proposed where in the first-step the unobserved process Z_t is fitted. Also, an oracle property applies: inference based on estimates $\widehat{Z}_t$ of Z_t is asymptotically equivalent to infeasible inference based on the unobserved Z_t.

In Fengler, Härdle, and Mammen (2007) and Park, Mammen, Härdle, and Borak (2009) the following model has been considered

$$Y_{tl} = m_0(X_{tl}) + \sum_{j=1}^{d} Z_t^j m_j(X_{tl}) + \varepsilon_{tl}.$$

This model differs from (3.13) because now the nonparametric components m_j are functions of a single characteristic X_{tl}. As a result, the multivariate time series Z_t is only identified up to linear transformations. Again, an oracle property for parametric inference based on fitted values has been shown in Park, Mammen, Härdle, and Borak (2009). The model has been used in functional principal component analysis. One application in Fengler, Härdle, and Mammen (2007) and Park, Mammen, Härdle, and Borak (2009) is for implied volatility surfaces that develop over time. The surfaces are approximated by a finite-dimensional process and the random movement of the surfaces is then analyzed by a VAR representation of the finite-dimensional process.

3.5 Semiparametric GARCH models

Another example that leads to an additive model is a semiparametric GARCH model. In this model we observe a process Y_t such that $E(Y_t|\mathcal{F}_{t-1}) = 0$, where $\mathcal{F}_{t-1}$ denotes the sigma field generated by the entire past history of the Y process, and $\sigma_t^2 \equiv E(Y_t^2|\mathcal{F}_{t-1})$ assumes a semiparametric model

$$\sigma_t^2 = \theta\sigma_{t-1}^2 + f(Y_{t-1}). \tag{3.14}$$

This model is a natural generalization of the GARCH(1,1) model of Bollerslev (1986) where a parametric assumption is made on f such that $f(x) = \alpha + \beta x$. The generalization was introduced by Engle and Ng

(1993) to allow for more flexibility in the 'news impact curve', i.e., the function f, which measures the effect of news onto volatilities in financial markets.

The parameters θ and the function f in the semiparametric model (3.14) are unknown. Since $E(Y_t^2|\mathcal{F}_{t-1}) = \sum_{j=1}^{\infty} \theta^{j-1} f(Y_{t-j})$, the parameter θ and the function $f(\cdot, \theta)$ together minimize $E[Y_0^2 - \sum_{j=1}^{\infty} \theta^{j-1} f(X_{-j})]^2$. For each θ, let f_θ denote the minimizer of the criterion. Then, it satisfies

$$\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \theta^{j+k-2} f_\theta(Y_{-k}) g(Y_{-j}) = \sum_{j=1}^{\infty} E[Y_0^2 \theta^{j-1} g(Y_{-j})]$$

for all square integrable functions g. This gives the following integral equation.

$$f_\theta(x) = f_\theta^*(x) - \int \mathcal{H}_\theta(x, y) f_\theta(y)\, dy, \tag{3.15}$$

where

$$f_\theta^*(x) = (1 - \theta^2) \sum_{j=1}^{\infty} \theta^{j-1} E(Y_0^2 | Y_{-j} = x),$$
$$\mathcal{H}_\theta(x, y) = \sum_{j=1}^{\infty} \theta^j \left[\frac{p_{0,-j}(x, y) + p_{0,j}(x, y)}{p_0(x)} \right],$$

p_0 and $p_{0,j}$ are the densities of Y_0 and (Y_0, Y_j), respectively. For an asymptotic and empirical analysis of the estimators based on the integral equation (3.15), we refer to Linton and Mammen (2005). For a recent extension of the model, see also Chen and Ghysels (2011).

3.6 Varying coefficient models

Suppose we are given a group of covariates $X^1, \ldots, X^d$ and a response Y. The most general form of varying coefficient model was introduced and studied by Lee, Mammen, and Park (2012a). It is given by

$$E(Y|X^1, \ldots, X^d) = g^{-1}\left(\sum_{k \in I_1} X^k f_{k1}(X^1) + \cdots + \sum_{k \in I_p} X^k f_{kp}(X^p) \right), \tag{3.16}$$

where g is a link function and $p \le d$. The index sets I_j may intersect with each other, but each I_j does not include j. It is also allowed that the two groups of covariates, $\{X^j : 1 \le j \le p\}$ and $\{X^k : k \in \cup_{j=1}^p I_j\}$ may have common variables. The coefficient functions are identifiable if we put the following constraints: for nonnegative weight functions w_j, (i) $\int f_{kj}(x^j) w_j(x^j)\, dx^j = 0$ for all $k \in \cup_{j=1}^p I_j$ and $1 \le j \le p$; (ii) $\int x^j f_{kj}(x^j) w_j(x^j)\, dx^j = 0$ for all $j, k \in \{1, \ldots, p\} \cap (\cup_{j=1}^p I_j)$. In this model, the effect of the covariate X^k for $k \in \cup_{j=1}^p I_j$ is set in a nonparametric way as $\sum_{j: I_j \ni k} f_{kj}(X^j)$. The model is flexible enough to include various types of varying coefficient models as special cases. For example, it is specialized to the generalized additive model discussed in Section 2.7 if one takes $I_1 = \cdots = I_p = \{p+1\}$ and set $X^{p+1} \equiv 1$. The model also reduces to the varying coefficient model studied by Lee, Mammen, and Park (2012b) and Yang, Park, Xue, and Härdle (2006) if the two groups, $\{X^j : 1 \le j \le p\}$ and $\{X^k : k \in \cup_{j=1}^p I_j\}$, are disjoint and the sets I_j contain only one element ($1 \le j \le p$). In this case one can rewrite model (3.16)

as

$$Y_i = g^{-1}\left(\sum_{j=1}^{p} Z_i^j f_j(X_i^j)\right) + \varepsilon_i.$$

With an identity link g and with the additional constraint $f_j \equiv f$, this model has been used in Linton, Mammen, Nielsen, and Tanggaard (2001) for nonparametric estimation of yield curves by smoothed least-squares. There, Y_i was the trading price of a coupon bond, Z_i^j denotes the payment returned to the owner of bond i at date X_i^j and f is the discount function. In case $p = 1$ and $I_1 = \{2, \ldots, d\}$, the approach with disjoint sets of covariates results in the model studied, for example, by Fan and Zhang (1999).

For simplicity, suppose that the link g is the identity function. In this case, the coefficient functions f_{kj} minimize $E\left[Y - \sum_{k\in I_1} X^k f_{k1}(X^1) - \cdots - \sum_{k\in I_p} X^k f_{kp}(X^p)\right]^2$. This gives the following system of integral equations for f_{kj}: for $1 \le j \le p$,

$$\begin{aligned}\mathbf{f}_j(x^j) = {} & E(\mathbf{X}_j\mathbf{X}_j^\top|X^j = x^j)^{-1}E(\mathbf{X}_j Y|X^j = x^j) - E(\mathbf{X}_j\mathbf{X}_j^\top|X^j = x^j)^{-1} \\ & \times \sum_{l=1,\neq j}^{p} \int E\left[\mathbf{X}_j\mathbf{X}_l^\top\middle|X^j = x^j, X^l = x^l\right]\mathbf{f}_l(x^l)\frac{p_{jl}(x^j, x^l)}{p_j(x^j)}\,dx^l,\end{aligned}$$

where $\mathbf{X}_j = (X^k : k \in I_j)$ and $\mathbf{f}_j = (f_{kj} : k \in I_j)$. Note that $\mathbf{X}_j$ does not contain X^j as its entry. To get an empirical version of the above integral equations, one may replace the conditional expectations, the joint density p_{jl} of (X^j, X^l) and the marginal density p_j of X^j, by kernel estimators. Lee, Mammen, and Park (2012a) presented complete theory for the estimation of the general model (3.16). Their theory includes sieve and penalized quasi-likelihood estimation as well as the smooth backfitting method described above.

3.7 Missing observations

Additive models can also be consistently estimated if the tuples $(Y_i, X_i^1, ..., X_i^d)$ are only partially observed. We will discuss this for the following simple scheme of missing observations.

Denote

- by $\mathcal{N}_{jk}$ the set of indices i where X_i^j and X_i^k are observed,
- by $\mathcal{N}_j$ the set of indices i where X_i^j is observed,
- by $\mathcal{N}_{0j}$ the set of indices i where X_i^j and Y_i are observed, and
- by $\mathcal{N}_0$ the set of indices i where Y_i is observed.

These sets may be random or nonrandom. We denote the number of elements of these sets by N_{jk}, N_j, N_{0j} or N_0, respectively. We assume that the observations $\{(X_i^j, X_i^k) : i \in \mathcal{N}_{jk}\}$, $\{X_i^j : i \in \mathcal{N}_j\}$, $\{(X_i^j, Y_i) : i \in \mathcal{N}_{0j}\}$ and $\{Y_i : i \in \mathcal{N}_0\}$ are i.i.d. This assumption holds under simple random missingness schemes and also in the case of pooling samples where different subsets of covariates were observed.

Then, under the assumption that $N_{jk} \to \infty$, $N_j \to \infty$, $N_{0j} \to \infty$ and $N_0 \to \infty$, the estimators of p_{X^j,X^k}, p_{X^j}, f_j^* and μ that are based on the subsamples $\mathcal{N}_{jk}$, $\mathcal{N}_j$, $\mathcal{N}_{0j}$ or $\mathcal{N}_0$, respectively, are consistent.

More precisely, for $1 \leq j \neq k \leq d$, put

$$\begin{aligned}
\widetilde{p}_{X^j,X^k}(x^j,x^k) &= \frac{1}{N_{jk}h_jh_k}\sum_{i\in\mathcal{N}_{jk}} K\left(\frac{X_i^j-x^j}{h_j}\right)K\left(\frac{X_i^k-x^k}{h_k}\right),\\
\widetilde{p}_{X^j}(x^j) &= \frac{1}{N_jh_j}\sum_{i\in\mathcal{N}_j} K\left(\frac{X_i^j-x^j}{h_j}\right),\\
\widetilde{f}_j^*(x^j) &= \widetilde{p}_{X^j}(x^j)^{-1}\frac{1}{N_{0j}h_j}\sum_{i\in\mathcal{N}_{0j}} K\left(\frac{X_i^j-x^j}{h_j}\right)Y_i,\\
\widetilde{\mu} &= \frac{1}{N_0}\sum_{i\in\mathcal{N}_0}^{n} Y_i.
\end{aligned}$$

Under appropriate conditions on the bandwidths h_j these estimators converge to $p_{X^j,X^k}(x^j,x^k)$, $p_{X^j}(x^j)$, $f_j^*(x^j)$ and μ, respectively, in probability. Similarly as in Eq. (2.2), we consider the solutions $\widetilde{f}_1,\ldots,\widetilde{f}_d$ of the equations

$$\widetilde{f}_j(x^j) = \widetilde{f}_j^*(x^j) - \widetilde{\mu} - \sum_{k\neq j}\int \frac{\widetilde{p}_{X^j,X^k}(x^j,x^k)}{\widetilde{p}_{X^j}(x^j)}\widetilde{f}_k(x^k)\,dx^k.$$

Using the stochastic convergence of $\widetilde{p}_{X^j,X^k}(x^j,x^k)$, $\widetilde{p}_{X^j}(x^j)$, $\widetilde{f}_j^*(x^j)$ and $\widetilde{\mu}$, one can show that $\widetilde{f}_j(x^j)$ converges in probability to $f_j(x^j)$ for $1 \leq j \leq d$. These consistency proofs can be generalized to more complex missingness schemes. Furthermore, under appropriate conditions one can study normal distribution limits of these estimators. We remark that these identification, consistency and asymptotic normality results are not available for the full-dimensional model specification: $Y = f(X^1,\ldots,X^d)+\varepsilon$.

3.8 Additive diffusion models

Some multivariate diffusion models are based on additive parametric specifications of the mean. Nonparametric generalizations of such models were considered in Haag (2006). There also nonparametric specifications of the volatility term were considered.

3.9 Simultaneous nonparametric equation models

Additive models also naturally occur in economic models, where some covariates are correlated with the disturbance. Despite these so-called endogenous regressors, such models can be identified via a control function approach. In particular, Newey, Powell, and Vella (1999) proposed the following model with additive error terms

$$Y = f(X^1,Z^1)+e\ ,$$

where X^1 and Z^1 are observed covariates and Y is a one-dimensional response. While Z^1 is independent of the error variable e, no assumptions are made on the dependence between X^1 and e at this stage. For identification, however, assume that the following control equation holds for the endogenous variable X^1

$$X^1 = h(Z^1,Z^2)+V,$$

where Z^2 is an observed covariate not contained in the original equation and (Z^1, Z^2) is independent of the joint vector of errors (e, V).

Under the stated independence conditions, it follows that

$$E(Y|X^1, Z^1, Z^2) = f(X^1, Z^1) + \lambda(V) = E[Y|X^1, Z^1, V] \tag{3.17}$$

with $\lambda(V) = E(e|V)$. Thus, we get an additive model where the regressor in the second additive component is not observed but can be estimated as residual of the control equation. This additive model can be also obtained under slightly weaker conditions than the above independence conditions, namely under the assumption that $E(e|Z^1, Z^2, V) = E(e|V)$ and $E(V|Z^1, Z^2) = 0$. The corresponding system of integral equations to be solved for (3.17) is

$$\begin{aligned} f(x^1, z^2) &= f^*(x^1, z^2) - \int \frac{p_{X_1,Z_2,V}(x^1, z^2, v)}{p_{X_1,Z_2}(x^1, z^2)} \lambda(v)\, dv \\ \lambda(v) &= \lambda^*(v) - \int \frac{p_{X_1,Z_2,V}(x^1, z^2, v)}{p_V(v)} f(x^1, z^2)\, d(x^1, z^2) \end{aligned}$$

where $f^*(z^1, z^2) = E[Y|(X^1, Z^1) = (x^1, z^2)]$ and $\lambda^*(v) = E(Y|V = v)$. Note that some ingredients of the smooth backfitting iteration algorithm thus require nonparametric pre-estimates of marginal objects with the nonparametrically generated regressor $\widehat{V} = X^1 - \widehat{h}(Z^1, Z^2)$. The paper Mammen, Rothe, and Schienle (2012) studies how asymptotic theory in nonparametric models has to be adjusted to take care of nonparametrically generated regressors.

4 Nonstationary observations

Additive models are a powerful tool in case of stochastically nonstationary covariates. For this data generality, consistent estimation of a fully nonparametric model requires that the whole compound vector fulfills a specific recurrence condition, i.e., it has to be guaranteed that the full dimensional process X returns infinitely often to local neighborhoods (See e.g. Karlsen and Tjøstheim (2001), Wang and Phillips (2009), and Karlsen, Myklebust, and Tjøstheim (2007)). For an additive model, however, recurrence conditions are only needed for two-dimensional subvectors of X. An illustrative example is a multivariate random walk. A fully nonparametric model cannot be consistently estimated for dimensions greater two, since beyond dimension two random walks become transient and do not fulfill the above recurrence property. For an additive model, however, there is no dimension restriction, as any pair of bivariate random walks is recurrent. Here we briefly outline the main ideas. The detailed theory of additive models for nonstationary covariates is developed in Schienle (2008).

The setting is as follows: Suppose we want to estimate a standard additive model (1.1) where covariates and response are potentially nonstationary Markov chains but satisfy a pairwise recurrence condition, and the residual is stationary mixing. Instead of a stationary data generating process density function, a nonstationary pairwise recurrent Markov chain can be characterized by the densities of pairwise bivariate invariant measures π_{jk} with $j, k \in \{1, \ldots, d\}$. For the specific kind of recurrence

imposed, it is guaranteed that such a bivariate invariant measure exists for each pair and is unique up to a multiplicative constant; but it is generally only finite on so-called small sets and only σ-finite on the full support. Note that e.g., for random walks any compact set is small.

Furthermore, under the type of pairwise recurrence imposed, bivariate component Markov chains $(X^j, X^k) = (X^{jk})$ can be decomposed into i.i.d. parts of random length depending on the recurrence times of the chain. In particular, the stochastic number of recurrence times $T^{jk}(n)$ characterizes the amount of i.i.d. block observations and thus corresponds to the effective sample size available for inference with the particular pair of components. Thus for different components and pairs of components available effective sample sizes are path dependent and generally vary depending on the recurrence frequency being smaller for more nonstationary processes and closer to the stationary deterministic full sample size n for more stationary processes. Correspondingly, consistent kernel type estimators are weighted averages of $T^{jk}(n)$ i.i.d. block elements

$$\begin{aligned} \widehat{\pi}_{jk}(x^{jk}) &= \frac{1}{T^{jk}(n)} \sum_{i \in I_{jk}} K\left(\frac{X_i^{jk} - x^{jk}}{h_{jk}}\right), \\ \widehat{f}_j(x^j) &= \left[\sum_{i \in I_j} K\left(\frac{X_i^j - x^j}{h_j}\right)\right]^{-1} \sum_{i \in I_j} K\left(\frac{X_i^j - x^j}{h_j}\right) Y_i, \end{aligned} \tag{4.1}$$

$$\begin{aligned} \widehat{\pi}_j^{(k)}(x^j) &= \frac{1}{T^{jk}(n)} \sum_{i \in I_{jk}} K\left(\frac{X_i^j - x^j}{h_j}\right), \\ \widehat{f}_j^{(k)}(x^j) &= \left[\sum_{i \in I_{jk}} K\left(\frac{X_i^j - x^j}{h_j}\right)\right]^{-1} \sum_{i \in I_{jk}} K\left(\frac{X_i^j - x^j}{h_j}\right) Y_i. \end{aligned} \tag{4.2}$$

The estimators in (4.1) provide pointwise consistent estimates of the corresponding bivariate invariant measure density π_{jk} and a general nonparametric link function f_j, respectively (see Karlsen, Myklebust, and Tjøstheim (2007)). Their rates of convergence are driven by respective recurrence frequencies and occupation times $\widehat{L}_{jk}(x^{jk}) = \sum_{i \in I_{jk}} K_{x^{jk},h^{jk}}(X_i^{jk})$ and $\widehat{L}_j$, respectively, which are generally of different order on average over all sample paths. Asymptotically in both cases, they are on average of size $(n^{\beta^{jk}} h)^{-1/2}$ and $(n^{\beta^j} h)^{-1/2}$, respectively, where the global β^{jk}-parameter $\in [0,1]$ characterizes the underlying type of nonstationarity of the corresponding recurrent chain as the tail index on the distribution of recurrence times. For a bivariate random walk we have $\beta^{jk} = 0$, for a stationary process $\beta^{jk} = 1$ recovering standard rates, and generally $\beta^{jk} \leq \beta^j$. The kernel estimators in (4.2) artificially "downgrade" their univariate speed of convergence to the respective bivariate one. Note that the index sets I_{jk} ensure that only $T^{jk}(n)$ i.i.d. sub-blocks are considered of the $T^j(n)$ original ones.

For balancing terms in the empirical version of the smooth backfitting integral equations, such potentially slower than standard estimators $\widehat{\pi}_j^{(k)}$, $\widehat{\pi}_{jl}^{(k)}$ and $\widehat{f}_j^{(k)}$ of bivariate nonstationary type β^{jk} are necessary. Also in the backfitting operator for component j, the impact of other directions on any pair of components containing X^j might now differ depending on respective occupation times of component pairs. Both aspects are reflected by a respectively generalized procedure ensuring consistent estimates.

The generalized smooth backfitting estimates $(\widehat{f}_j)_{j=1}^d$ are defined as

$$\widehat{f}_j(x^j) = \frac{1}{d-1}\left[\sum_{k\neq j}\left(\widehat{f}_j^{(k)*}(x^j) - \widehat{f}_{0,j}^{(k)*}\right) - \sum_{k\neq j}\frac{1}{\widehat{\lambda}_{jk}}\sum_{l\neq j}\int_{\mathcal{G}_l}\widehat{f}_l(x^l)\frac{\widehat{\pi}_{jl}^{(k)}(x^{jl})}{\widehat{\pi}_j^{(k)}(x^j)}\,dx^l\right], \tag{4.3}$$

where $\widehat{f}_j^{(k)*}(x^j)$ are the marginal local constant estimates with bivariate speed of convergence as defined above and constants

$$\widehat{f}_{0,j}^{(k)*} = \frac{\int_{\mathcal{G}_j}\widehat{f}_j^{(k)*}(x^j)\widehat{\pi}_j^{(k)}(x^j)\,dx^j}{\int_{\mathcal{G}_j}\widehat{\pi}_j^{(k)}(x^j)\,dx^j} = \frac{1}{T^{jk}(n)}\sum_{i\in I_{jk}} Y_i, \tag{4.4}$$

which follow from appropriate analogues of the standard norming constraints

$$\sum_{k\neq j}\int_{\mathcal{G}_j} f_j(x^j)\pi_j^{(k)}(x^j)\,dx^j = 0. \tag{4.5}$$

Note that asymptotically in the projection part of (4.3) only those elements $\widehat{\pi}_{jl}$ prevail, where $\beta^{jl} = \beta^{jk}$ while all others vanish. The projection property of standard backfitting only prevails in a generalized sense, since in general an invariant measure for the full-dimensional compound process does not exist for pairwise recurrent X. For each j and k, $\widehat{\lambda}_{jk}$ counts the number of such elements in the sample. In a nonstationary setting also the regions of integration $\mathcal{G}_j$ must be chosen with some care to ensure that integrals exist. Related to small sets, e.g., in a random walk case compact areas are appropriate. If all pairs of components of X have the same type of nonstationarity, the backfitting equations reduce to

$$\widehat{f}_j(x^j) = \frac{1}{d-1}\sum_{k\neq j}\left(\widehat{f}_j^{(k)*}(x^j) - \widehat{f}_{0,j}^{(k)*}\right) - \sum_{k\neq j}\int_{\mathcal{G}_k}\widehat{f}_k(x^k)\frac{\widehat{\pi}_{jk}(x^{jk})}{\widehat{\pi}_j^{(k)}(x^j)}\mathrm{d}x^k\,,$$

since $\lambda_{jk} = d-1$ and $\widehat{\pi}_{jl}^{(k)} = \widehat{\pi}_{jl}$ in this case. In particular, for the special case of identical one- and two-dimensional scales, generalized smooth backfitting reduces to the standard case. This usually occurs for sufficiently stationary data.

Asymptotic results for the generalized backfitting are univariate in form, i.e, the standard curse of dimensionality can be circumvented. However, they are driven by the worst case bivariate type of nonstationarity in the data. In particular, the difference between the true component function f_j and the backfitting estimate $\widehat{f}_j$ is asymptotically normal when inflated with the stochastic occupation time factor $\sqrt{\min_{k\neq j}\widehat{L}_j^{(k)}(x^j)h}$. As $\widehat{L}_j^{(k)}$ is asymptotically of the same order as $T^{jk}(n)$, the rate of convergence is on average of size $\sqrt{n^{\beta^{j+}+\epsilon}h}$, where β^{j+} is the highest degree of nonstationarity and thus the smallest number among the β^{jk}, and $\epsilon > 0$ is very small. That means, if all components are random walks, i.e., $\beta^{jk} = 0$, estimation of each component is possible, but with logarithmic rate. This should be compared to the fact that a fully nonparametric model cannot be estimated in this case where the compound vector is transient. If one component X^{j_0} follows a random walk and all others are stationary, all components are estimated at rate $\sqrt{n^{\beta_{j_0}}h} = \sqrt{n^{1/2}h}$.

5 Noisy Fredholm integral equations of second kind

As outlined in Subsection 2.4, we can define the smooth backfitting estimators in the additive models as solutions of an integral equation $\widehat{\mathbf{f}}(x) = \widehat{\mathbf{f}}^*(x) - \int \widehat{\mathcal{H}}(x,z)\widehat{\mathbf{f}}(z)\,dz$, where $\widehat{\mathbf{f}}(x_1,\ldots,x_d) = (\widehat{f}_1(x_1),\ldots,\widehat{f}_d(x_d))^\top$, $\widehat{\mathbf{f}}^*(x_1,\ldots,x_d) = (\widehat{f}_1^*(x_1),\ldots,\widehat{f}_d^*(x_d))^\top$ and the integral kernel $\widehat{\mathcal{H}}(x,z)$ equals to a matrix with elements $\widehat{p}_{X^j,X^k}(x^j,x^k)/\widehat{p}_{X^j}(x^j)$. We also rewrite this noisy integral equation as

$$\widehat{\mathbf{f}} = \widehat{\mathbf{f}}^* - \widehat{\mathcal{H}}\widehat{\mathbf{f}}.$$

In Section 3 we have also seen that smooth least squares for various models leads to estimators that are given as solutions of such noisy integral equations. There are several approaches to the numerical solution of the integral equation. As already mentioned in Subsection 2.4, one can use a discrete approximation of the integral equation for the numerical solution. This results in a finite system of linear equations that can be solved by standard methods. One approach would be based on a iterative scheme that uses a discrete approximation of the iteration steps:

$$\widehat{\mathbf{f}}^{NEW} = \widehat{\mathbf{f}}^* - \widehat{\mathcal{H}}\widehat{\mathbf{f}}^{OLD}.$$

If $\widehat{\mathbf{f}}$ is a d-dimensional vector of functions with $d \geq 2$, one can also use an iteration scheme that runs cyclically through component-wise updates

$$\widehat{f}_j^{NEW} = \widehat{f}_j^* - \widehat{\mathcal{H}}_j\widehat{\mathbf{f}}^{OLD},\ 1 \leq j \leq d$$

with an obvious definition of $\widehat{\mathcal{H}}_j$. This was the algorithm we discussed in Subsection 2.1. Compare also the Gauss-Seidel method and the Jacobi method in numerical linear algebra.

We now use the definition of the estimators by a noisy integral equation for an asymptotic understanding of the distributional properties of the estimators. We consider the case of one-dimensional $\widehat{\mathbf{f}}$ and $\widehat{\mathbf{f}}^*$ and we rewrite the equation as $\widehat{f} = \widehat{f}^* - \widehat{\mathcal{H}}\widehat{f}$. We now suppose that $\widehat{f}^*$ is a smoothing estimator with

$$\widehat{f}^* \approx \widehat{f}_A^* + f^* + f_B^*,$$

where $\widehat{f}_A^*$ is the *stochastic part* of $\widehat{f}^*$ that is of order $(nh)^{-1/2}$. The function f^* is the stochastic limit of $\widehat{f}^*$ and f_B^* is a bias term that we suppose to be of the standard order h^2. Here, h is a bandwidth that is chosen of order $n^{-1/5}$ so that the stochastic term and the bias term are of order $n^{-2/5}$. A similar discussion applies to $\widehat{\mathcal{H}}f$. This variable has stochastic limit $\mathcal{H}f$ where $\mathcal{H}$ is the stochastic limit of $\widehat{\mathcal{H}}$. We now get

$$\widehat{\mathcal{H}}f \approx (\widehat{\mathcal{H}}f)_A + \mathcal{H}f + (\mathcal{H}f)_B,$$

where $(\widehat{\mathcal{H}}f)_A$ is the *stochastic part* of $\widehat{\mathcal{H}}f$. Again this term is of order $(nh)^{-1/2}$. Although $\widehat{\mathcal{H}}$ is a higher dimensional smoother, all variables up to one are integrated out in $\widehat{\mathcal{H}}f$. Furthermore, $(\mathcal{H}f)_B$ is a bias

term that is of order h^2. By subtracting $f = f^* - \mathcal{H}f$ from $\widehat{f} = \widehat{f}^* - \widehat{\mathcal{H}}\widehat{f}$ we get

$$\begin{aligned}\widehat{f} - f &= \widehat{f}^* - f^* - \widehat{\mathcal{H}}\widehat{f} + \mathcal{H}f \\ &= \widehat{f}^* - f^* - \mathcal{H}(\widehat{f} - f) - (\widehat{\mathcal{H}} - \mathcal{H})f - (\widehat{\mathcal{H}} - \mathcal{H})(\widehat{f} - f) \\ &\approx \widehat{f}^* - f^* - \mathcal{H}(\widehat{f} - f) - (\widehat{\mathcal{H}} - \mathcal{H})f.\end{aligned}$$

Now, simple algebra gives

$$\begin{aligned}\widehat{f} - f &\approx (I + \mathcal{H})^{-1}(\widehat{f}^* - f^* - (\widehat{\mathcal{H}} - \mathcal{H})f) \\ &\approx (I + \mathcal{H})^{-1}(\widehat{f}_A^* + f_B^* - (\widehat{\mathcal{H}}f)_A - (\mathcal{H}f)_B).\end{aligned}$$

We now argue that $(I+\mathcal{H})^{-1}\widehat{f}_A^* \approx \widehat{f}_A^*$ and $(I+\mathcal{H})^{-1}(\widehat{\mathcal{H}}f)_A \approx (\widehat{\mathcal{H}}f)_A$. These claims follow immediately from $(I + \mathcal{H})^{-1} = I - (I + \mathcal{H})^{-1}\mathcal{H}$, $\mathcal{H}\widehat{f}_A^* \approx 0$ and $\mathcal{H}(\widehat{\mathcal{H}}f)_A \approx 0$. Here, the first equality can be easily seen by multiplying both sides of the equation with $(I + \mathcal{H})$. For the two approximations one notes that the integral, over an interval, of the stochastic part of a kernel smoother is typically of order $n^{-1/2}$. For example, one has $\int w(x)n^{-1}\sum_{i=1}^n K_h(x - X_i)\varepsilon_i \, dx = n^{-1}\sum_{i=1}^n w_h(X_i)\varepsilon_i$ with $w_h(u) = \int w(x)K_h(x - u) \, dx$, which is of order $n^{-1/2}$. Using the above approximations we get that

$$\begin{aligned}\widehat{f} - f &\approx (I + \mathcal{H})^{-1}(\widehat{f}_A^* + f_B^* - (\widehat{\mathcal{H}}f)_A - (\mathcal{H}f)_B) \\ &= \widehat{f}_A^* - (\widehat{\mathcal{H}}f)_A - (I + \mathcal{H})^{-1}\mathcal{H}(\widehat{f}_A^* - (\widehat{\mathcal{H}}f)_A) + (I + \mathcal{H})^{-1}(f_B^* - (\mathcal{H}f)_B) \\ &\approx \widehat{f}_A^* - (\widehat{\mathcal{H}}f)_A + (I + \mathcal{H})^{-1}(f_B^* - (\mathcal{H}f)_B)\end{aligned}$$

The expressions on the right hand side of this expansion can be easily interpreted. The first term $\widehat{f}_A^* - (\widehat{\mathcal{H}}f)_A$ is of order $(nh)^{-1/2}$ and asymptotically normal with mean zero. This can be shown as in classical kernel smoothing theory. The second term $(I + \mathcal{H})^{-1}(f_B^* - (\mathcal{H}f)_B)$ is purely deterministic and it is of order h^2 because already $f_B^* - (\mathcal{H}f)_B$ is of this order. For a more detailed discussion of the above arguments we refer to Mammen, Støve, and Tjøstheim (2009) and Mammen and Yu (2009).

We conclude this section by noting that the above noisy integral equations are quite different from integral equations of the form

$$0 = \widehat{\mathbf{f}}^* - \widehat{\mathcal{H}}\widehat{\mathbf{f}}.$$

This is called an ill-posed inverse problem because, typically, the eigenvalues of the operator $\widehat{\mathcal{H}}$ accumulate at 0. For this reason the inverse of the operator $\widehat{\mathcal{H}}$ is not continuous. The integral equation studied in this chapter leads to the inversion of the operator $(I+\widehat{\mathcal{H}})$. The eigenvalues of this operator accumulate around 1 and allow for a continuous inverse of $(I+\widehat{\mathcal{H}})$. Thus our set-up is quite different from ill-posed problems. For a discussion of ill-posed problems we refer to Carrasco, Florens, and Renault (2006), Chen and Reiss (2011), Darolles, Florens, and Renault (2011), Donoho (1995), Engl and Neubauer (1996), Johnstone and Silverman (1990) and Newey and Powell (2003).

References

Andersen, E. (1970): "Asymptotic Properties of Conditional Maximum Likelihood Estimators.," *Journal of the Royal Statistical Society Series B*, 32, 283–301.

Bollerslev, T. (1986): "Generalized autoregressive conditional heteroscedasticity," *J. Econometrics*, 31, 307–327.

Buja, A., T. Hastie, and R. Tibshirani (1989): "Linear smoothers and additive models (with discussion)," *Ann. of Statist.*, 17, 453–510.

Carrasco, M., J. Florens, and E. Renault (2006): "Linear Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics Vol. 6*, ed. by J. Heckman, and E. Leamer. Elsevier.

Carroll, R. J., A. Maity, E. Mammen, and K. Yu (2009): "Nonparametric additive regression for repeatedly measured data," *Biometrika*, 96, 383–398.

Chamberlain, G. (1994): "Panel Data," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. Intriligator, pp. 1247–1318. Elsevier.

Chen, X. (2006): "Large Sample Sieve Estimation of Semi-nonparametric Models," in *Handbook of Econometrics Vol. 6*, ed. by J. Heckman, and E. Leamer, pp. 5549–5632. Elsevier.

Chen, X., and E. Ghysels (2011): "News - good or bad - and its impact on volatility predictions over multiple horizons," *Rev. Financ. Stud*, 24, 46–81.

Chen, X., and M. Reiss (2011): "On rate optimality for ill-posed inverse problems in econometrics," *Econometric Theory*, 27, 497–52.

Connor, G., M. Hagmann, and Linton (2012): "Efficient estimation of a semiparametric characteristic-based factor model of security returns 18 730-754.," *Econometrica*, 18, 730–754.

Darolles, S., J. P. Florens, and E. Renault (2011): "Nonparametric instrumental regression," *Econometrica*, 79, 1541–1565.

Donoho, D. L. (1995): "Nonlinear solutions of linear inverse problems by wavelet-vaguelette decomposition," *J. Applied and Comput. Harmonic Anal.*, 2, 101–126.

Eilers, P. H. C., and B. D. Marx (2002): "Generalized linear additive smooth structures," *J. Comput. Graph. Statist.*, 11, 758–783.

Engl, H.and Hanke, M., and A. Neubauer (1996): *Regularization of Inverse Problems*. Kluwer Academic Publishers, London.

Engle, R. F., and V. K. Ng (1993): "Measuring and testing the impact of news on volatility," *J. Finance*, 48, 987–1008.

Fan, J., N. Heckman, and M. P. Wand (1995): "Local polynomial kernel regression for generalized linear models and quasi-likelihood functions," *J. Amer. Statist. Assoc.*, 90, 141–150.

Fan, J., and J. Jiang (2007): "Nonparametric inference with generalized likelihood ratio tests (with discussion)," *Test,*, 16, 409–478.

Fan, J., Y. Wu, and Y. Feng (2009): "Local quasi-likelihood with a parametric guide," *Ann. Statist.*, 27, 4153–4183.

Fan, J., and W. Zhang (1999): "Statistical estimation in varying coefficient models," *Ann. Statist.*, 27, 1491–1518.

Fengler, M., W. Härdle, and E. Mammen (2007): "A semiparametric factor model for implied volatility surface dynamics," *J. Financial Econometrics*, 5, 189–218.

Haag, B. (2006): "Model Choice in Structured Nonparametric Regression and Diffusion Models," Ph.D. thesis, Mannheim University.

Haag, B. (35): "Non-parametric regression tests using dimension reduction techniques," *Scand. J. Statist.*, 2008, 719–738.

Hastie, T. J., and R. J. Tibshirani (1990): *Generalized Additive Models*. Chapman and Hall, London.

Hjort, N. L., and I. K. Glad (1995): "Nonparametric density estimation with a parametric start," *Ann. of Statist.*, 23, 882–904.

Hoderlein, S., E. Mammen, and K. Yu (2011): "Nonparametric models in binary choice fixed effects panel data," *Econometrics J.*, 14, 351–367.

Huang, J., J. L. Horowitz, and F. Wei (2010): "Variable selection in nonparametric additive models," *Ann. Statist.*, 38, 2282–2313.

Jiang, J., Y. Fan, and J. Fan (2010): "Estimation of additive models with highly or nonhighly correlated covariates," *Ann. Statist.*, 38, 1403–1432.

Johnstone, I. M., and B. W. Silverman (1990): "Speed of estimation in positron emission tomography and related inverse problems," *Ann. Statist.*, 18, 251–280.

Karlsen, H. A., T. Myklebust, and D. Tjøstheim (2007): "Nonparametric estimation in a nonlinear cointegration type model," *Ann.Statist.*, 35, 1–57.

Karlsen, H. A., and D. Tjøstheim (2001): "Nonparametric Estimation in Null–Recurrent Time Series," *Annals of Statistics*, 29, 372–416.

Kauermann, G., and J. D. Opsomer (2003): "Local likelihood estimation in generalized additive models," *Scand. J. Statistics*, 30, 317–337.

Lee, Y. K., E. Mammen, and B. U. Park (2010): "Backfitting and smooth backfitting for additive quantile models," *Ann. Statist.*, 38, 2857–2883.

——— (2012a): "Flexible generalized varying coefficient regression models," *Forthcoming in Ann. Statist.*

——— (2012b): "Projection-type estimation for varying coefficient regression models," *Bernoulli*, 18, 177–205.

Lin, Y., and H. Zhang (2006): "Component selection and smoothing in multivariate nonparametric regression," *Ann. Statist.*, 34, 2272–2297.

Linton, O., and E. Mammen (2003): "Nonparametric smoothing methods for a class of non-standard curve estimation problems," in *Recent advances and trends in nonparametric statistics*, ed. by M. Akritas, and D. N. Politis. Elsevier.

——— (2005): "Estimating semiparametric ARCH(∞) models by kernel smoothing methods," *Econometrica*, 73, 771–836.

——— (2008): "Nonparametric transformation to white Noise," *J. Econometrics*, 142, 241–264.

Linton, O., E. Mammen, J. Nielsen, and C. Tanggaard (2001): "Estimating yield curves by kernel smoothing methods," *J. Econometrics*, 105, 185–223.

Linton, O., and J. Nielsen (2009): "Nonparametric regression with a latent time series," *Econometrics J.*, 12, 187–207.

Lundervold, L., D. Tjø stheim, and Q. Yao (2007): "Exploring spatial nonlinearity using additive approximation," *Bernoulli*, 13, 447–472.

Magnac, T. (2004): "Panel binary variables and sufficiency: generalizing conditional logit," *Econometrica*, 72, 1859–1876.

Mammen, E., O. Linton, and J. Nielsen (1999): "The existence and asymptotic properties of a backfitting projection algorithm under weak conditions," *Ann. Statist.*, 27, 1443 – 1490.

Mammen, E., J. S. Marron, B. A. Turlach, and M. P. Wand (2001): "A general framework for constrained smoothing," *Statist. Sci.*, 16, 232–248.

Mammen, E., and J. Nielsen (2003): "Generalised structured models," *Biometrika*, 90, 551–566.

Mammen, E., and B. U. Park (2005): "Bandwidth selection for smooth backfitting in additive models," *Ann. Statist.*, 33, 1260–1294.

——— (2006): "A simple smooth backfitting method for additive models," *Ann. Statist.*, 34, 2252–2271.

Mammen, E., C. Rothe, and M. Schienle (2012): "Nonparametric Regression with Nonparametrically Generated Covariates," *the Annals of Statistics*, 40,, 1132–1170.

Mammen, E., B. Støve, and D. Tjøstheim (2009): "Nonparametric additive models for panels of time series," *Econometric Theory*, 25, 442–481.

Mammen, E., and K. Yu (2009): "Nonparametric estimation of noisy integral equations of the second kind," *J. Korean Stat. Soc.*, 38, 99–110.

Meier, L., S. van de Geer, and P. Bühlmann (2009): "High-dimensional additive modeling," *Ann. Statist.*, 37, 3779–3821.

Newey, W., J. Powell, and F. Vella (1999): "Nonparametric estimation of triangular simultaneous equations models," *Econometrica*, 67(3), 565–603.

Newey, W. K., and J. L. Powell (2003): "Instrumental variables estimation for nonparametric models," *Econometrica*, 71, 1565–1578.

Nielsen, J., and S. Sperlich (2005): "Smooth backfitting in practice," *J. Roy. Statist. Soc.B*, 67, 43–61.

Opsomer, J. D. (2000): "Asymptotic properties of backfitting estimators," *J. Mult. Anal.*, 73, 166–179.

Opsomer, J. D., and D. Ruppert (1997): "Fitting a bivariate additive model by local polynomial regression," *Ann. Statist.*, 25, 186 – 211.

Park, B. U., W. C. Kim, and M. Jones (2002): "On local likelihood density estimation," *Ann. Statist.*, 30, 1480–1495.

Park, B. U., E. Mammen, W. Härdle, and S. Borak (2009): "Time series modelling with semiparametric factor dynamics," *J. Amer. Statist. Assoc.*, 104, 284–298.

Rasch, G. (1960): *Probabilistic Models for some Intelligence and Attainment Tests.* University of Chicago Press.

——— (1961): "On General Law and the Meaning of Measurement in Psychology," in *Proceeding of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 4. UC Press, Berkeley and Los Angeles.

Schick, A. (1996): "Root-n-consistent and efficient estimation in semiparametric aditive regression models," *Statist. Probab. Lett.*, 30, 45–51.

Schienle, M. (2008): "Nonparametric Nonstationary Regression," Ph.D. thesis, Universität Mannheim.

Stone, C. J. (1985): "Additive regression and other nonparametric models," *Ann. Statist.*, 13, 689–705.

——— (1986): "The dimensionality reduction principle for generalized additive models," *Ann. Statist.*, 14, 590–606.

Wang, Q., and P. C. B. Phillips (2009): "Structural Nonparametric Cointegrating Regression," *Econometrica*, 77, 1901–1948.

Yang, L., B. U. Park, L. Xue, and W. Härdle (2006): "Estimation and testing for varying coefficients in additive models with marginal integration," *J. Amer. Statist. Assoc.*, 101, 1212–1227.

Yu, K., E. Mammen, and B. U. Park (2011): "Semiparametric regression: efficiency gains from modeling the nonparametric part," *Bernoulli*, 17, 736–748.

Yu, K., B. U. Park, and E. Mammen (2008): "Smooth backfitting in generalized additive models," *Ann. Statist.*, 36, 228–260.