Nonparametric Nonstationary Regression

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

vorgelegt von

Melanie Schienle

März 2008

Abteilungssprecher:	Prof. Hans-Peter Grüner, Ph.D.
Referent:	Prof. Dr. Enno Mammen
Korreferent:	Prof. Oliver Linton, Ph.D.

Tag der Verteidigung: Freitag, 13.06.2008

Acknowledgements

First of all, I wish to thank my advisor Enno Mammen for his guidance and supervision of my thesis. He introduced me to the topic of nonparametric regression and my research has benefited a lot from his broad statistical knowledge. I am very grateful for his support and confidence in me.

I also wish to thank Oliver Linton for advice on my thesis and beyond. I benefited a lot from three very encouraging and insightful months in London. Numerous discussions not only broadened my perspective on my own topic but on econometrics in general.

Kyusang Yu was an invaluable support in my struggle with technical details. I benefited a lot not only from his experience with nonparametric statistics but also from uncountably many lunch box apples and clementines.

During the time at the Chair of Statistics I have enjoyed working with my colleagues Christian Conrad, Berthold Haag, Christoph Rothe and Christoph Nagel. It was always pleasant and fruitful to discuss scientific questions or anything else. I also wish to thank my fellows at the CDSEM – in particular the mensa crew. I enjoyed the lively atmosphere and the stimulating research environment.

I am especially grateful to my Stephan for still liking me.

Mannheim, March 28th, 2008

Melanie Schienle

Contents

Chapter 1

Introduction

This thesis studies nonparametric estimation techniques for a general regression set–up under very weak conditions on the covariate process. In particular, regressors are allowed to be high–dimensional stochastically nonstationary processes. The concept of nonstationarity comprises time series observations of random walk or long memory type. Admissible processes might "wander off", but recur any time they do so. We introduce the first kernel type estimation method for such nonstationary regressors without restricting their dimension. This set–up is motivated by and generalizes approaches in parametric econometric time series analysis with nonstationary components. It offers a possible way to extend and test for linear cointegration.

1.1 Relevance and Literature

There is substantial empirical evidence that many important economic factors without a deterministic time trend such as real consumer prices, individual consumption, exchange rates, and real GDP are stochastically nonstationary (See e.g. Meese and Singleton [1982], Kwiatkowski et al. [1992], or Sun and Phillips [2004]). In Econometric time series literature, though, the study of such nonstationary time series has been dominated by parametric models. Most commonly, stochastically nonstationary processes are modeled as integrated or fractionally integrated

(see the extensive literature on unit root processes started with Dickey and Fuller [1979] for purely integrated and Phillips [1987] for general ARIMA models; for fractionally integrated models see Baillie [1996] for a survey and e.g. Diebold and Rudebusch [1999] among others). For valid inference, though, the decision for a nonstationary model as opposed to a stationary one must be correct (tests for unit root can be found in Dickey and Fuller [1981], Phillips and Perron [1988] or for stationarity against unit root in Kwiatkowski et al. [1992]). In regression models, structural relationships between nonstationary variables have been extensively studied in the context of cointegration. Introduced in Engle and Granger [1987], stochastically nonstationary time series are cointegrated if there exists a linear combination such that the residual is I(0) an thus mostly stationary. This linear type of cointegration can be easily tested for , e.g. as proposed by Johansen [1991]. Nonlinear extensions as e.g. in Granger and Hallman [1991], Park and Phillips [1999], Park and Phillips [2001], or de Jong and Wang [2005] are still rare in the applied literature as they appear restrictive in fitting a specific parametric relationship which is hard to test for (see Hong and Phillips [2005]). While economic theory often suggests nonlinear responses (see e.g Lewbel and Ng [2005] in demand), it is often not explicit regarding the functional form (see e.g.Meese and Rose [1991] for exchange rates). Therefore the simplicity of econometric analysis so far is not due to simple true models but due to lack of respective more general tools. There is need for appropriate nonparametric methods in this general setting.

In nonparametric regression the full form of the functional relation between a response variable and observables is determined from the data. This is in contrast to parametric models where a global parametric form is prespecified up to a finite dimensional parameter which is obtained by estimation. In particular, for kernel type nonparametric estimation techniques we derive point estimates of the structural relationship by local weighted averages of observations in the neighborhood of the point of interest. Though for nonparametric estimation to be possible, the vast class of nonstationary processes is too wide and general comprising deterministic trends as well as stochastic nonstationarities. In order to apply local smoothing techniques in the state space, however, the processes cannot “wander off for good”

with a deterministic trend, but they must recur to any point in their range almost surely guaranteeing sufficiently many observations for inference. This intuitive and natural property can be formalized as Harris recurrence – a concept from Markov chain literature. It relaxes usual stationarity and ergodicity assumptions but allows for stochastic nonstationarity as of random walk type. Therefore it is the appropriate framework for nonstationary kernel type inference. While Harris recurrence puts a restriction on the behavior of the time series in the state domain, it is more general than assumptions in the time domain such as local stationarity or mixing, which require a certain alignment of the observed processes in time (For nonparametric nonstationary estimation in the class locally stationary processes via spectral density approximations see Dahlhaus [1997] and the rich literature thereafter)

The idea of Harris recurrence as the key minimal assumption for valid kernel regression techniques with Markov processes was first suggested by Yakowitz [1989]. His analysis, though, was restricted to the positive recurrent case and provided only consistency results for nearest neighbor estimates in this setting. Phillips and Park [1998] were the first to move towards possibly null recurrent processes. They used local time arguments, but their results only applied to one dimensional first order unit autoregressions. Independently Moloche [2001] and Karlsen and Tjøstheim [2001] have introduced an estimation framework for regression with general null Harris recurrent Markov processes. While the first uses embedding techniques, that require restrictive assumptions and employs existing results from probability theory literature, the later is more general with different direct techniques. Within the imperceptibly smaller class of β–null Harris recurrent processes, Karlsen and Tjøstheim [2001] provide results on consistency and derive asymptotic normality by inverting a stable recurrence time process. The type of nonstationarity of the data is captured by a single parameter β, the degree of regular variation in the tails of the recurrence time process. It also represents the polynomial degree of the expected stochastic rates of convergence and therefore offers an important way to compare the nonstationary results to well–known stationary theorems. A comparison of the two strains of literature is contained in Bandi [2004]. In general,

the literature on nonparametric nonstationary estimation is still quite new and therefore scarce. Lately there are some papers following the local time approach such as Bandi and Phillips [2003] and Bandi [2004] studying nonstationary diffusions and Wang and Phillips [2006] with general cointegration type estimation. Though a partial linear model is examined in Chen et al. [2007] under β–null Harris recurrence. We also employ the β–null Harris recurrence framework.

In general, however, high–dimensional nonparametric estimation suffers from standard curse of dimensionality (COD). The more regressors are included the worse the finite sample behavior. For nonstationary data, in particular, this can lead to extremely slow rates of convergence, requiring very large sample sizes for significant results. Furthermore in the nonstationary setting, an additional even more severe nonstationary curse of dimensionality complicates nonparametric estimation. For dimensions larger than two, Harris recurrence of joint regressors is restrictive. In fact, the more regressors are added, the more "unlikely" it is for the compound process to still fit the framework of Harris recurrence. Most prominently, a random walk is Harris recurrent only up to dimension two and transient for any higher dimension. In such cases, the performance of existing procedures of Karlsen et al. [2007] and Moloche [2001] does not only deteriorate, but none of them can be applied at all. There is no existing nonparametric method for such high–dimensional regression in this general setting.

1.2 Model and Approach

In this work, we provide an estimation method which countervails both curses of dimensionality. To overcome the first, ordinary COD, an additive model is estimated. In the stationary mixing case, additive models have provided a powerful technique to overcome this problem and to still maintain high model flexibility.
Denote observations by subscripts and dimension components by superscripts. In the entire thesis we use the short–hand notation $X^{jk} = (X^j, X^k)$. Then given a random design of n joint observations of $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, we estimate an additive conditional mean function $m : \mathbb{R}^d \to \mathbb{R}$ with component functions $m_j : \mathbb{R} \to \mathbb{R}$ for

$j = 1, \ldots, d$ and scalar m_0 by

$$Y_i = m_0 + \sum_{j=1}^{d} m_j(X_i^j) + \varepsilon_i \quad \text{for all } i \in \{1, \ldots, n\} \tag{1.1}$$

under suitable identification conditions for $m_j,\ j = 1, \ldots, d$. We always assume there is no concurvity, i.e. for $m_1, \ldots, m_d$ nontrivial we cannot have $m_1(x^1) + \cdots + m_d(x^d) = 0$ for all $(x^1, \ldots, x^d)$. The response Y and at least all univariate X^j and pairs of bivariate marginal components X^{jk} of the covariate vector X belong to a specific class of Markov processes, β–null Harris recurrent processes, which is wide and general enough to include stationary and a wide range nonstationary processes.
In order to tackle the second nonstationary COD, however, an estimation method for the additive model must rely on low dimensional components only - at best only univariate and bivariate ones to include the widest possible class of processes. In a stationary setting, smooth backfitting introduced by Mammen et al. [1999] fulfills these requirements as it does not need a full–dimensional estimate in any step of the method. On this basis, we develop and introduce the generalized smooth backfitting procedure for a general nonstationary setting.

1.3 Main Results

The main contributions of this thesis are the following. They are stated in order of importance which does not correspond to their order of appearance in the text.

First nonparametric technique for high–dimensional nonstationary regression

With generalized smooth backfitting we introduce the first valid nonparametric estimation method under weakest assumptions on multidimensional covariates. The essential requirement is only pairwise β–null Harris recurrence which comprises a significantly larger class of important practical processes than the class of full β–null Harris recurrent processes where other existing estimation procedures are

restricted to (see Karlsen et al. [2007]). Therefore it offers a first way to countervail both, the standard curse of dimensionality but also the more severe nonstationary curse of dimensionality. The generality in type of underlying data, however, restricts admissible model classes to at most pairwise additive. This implies that generalized smooth backfitting only yields the best additive fit if the true model is additive or pairwise additive as in (3.9). For more general true models this is not guaranteed. In the most general setting, identification is obtained under generalized conditional independence assumptions where the residual can also contain a certain type of stochastic nonstationarity. The exact conditions are stated in Assumptions 4.4 and the most general residual form is mentioned in Remark 4 after Theorem 4.3. We derive the asymptotic expansion of the generalized smooth backfitting estimators in Theorem 4.3. This is the main technical result of this thesis. Components have a rate of convergence and variance of univariate type but governed by the worst case bivariate nonstationary type. Therefore convergence also holds jointly for all component functions. In order to achieve asymptotic normality, the speed of convergence is stochastic due to the nonstationarity of the data. Section 4.4 contains some considerations on oracle efficiency of the procedure.

First nonparametric estimation method for an additive conditional mean function with nonstationary data

Depending on the degree of nonstationarity in the covariate vector, we introduce estimation techniques for additive regression models in this general setting. The more regressors are compoundly β–null Harris recurrent , say γ with $2 \leq \gamma \leq d$ of them, the more general can the true underlying model be, to still obtain the best additive fit in a projection sense via adapted generalized smooth backfitting. Thus allowing for higher model generality, rates of convergence and variances are governed by the worst case γ–wise nonstationary type in each component. The respective method is presented in Section 3.3.2 and its asymptotic expansion is stated in Theorem 4.5 under the weakest assumptions on the spacial correlation structure of covariates and residual. It contains standard smooth backfitting (for $\gamma = d$) and generalized smooth backfitting (for $\gamma = 2$) as extreme subcases. Adapted gener-

alized smooth backfitting can yield the closest additive approximation to a fully general true model only under full β–null Harris recurrence and standard smooth backfitting. Asymptotic results for this case are stated in Section 3.2. Though for generality in the fitted model class, the respective procedures lose in efficiency with methods for increasing γ, by scaling according to higher dimensional types of nonstationary data. This is briefly discussed in Section 4.4.

First nonparametric generalization of cointegration type models without restricting the number of covariates

In the special subcase of ε being stationary mixing in (1.1), we estimate an additive cointegration type relationship. In contrast to the existing more general nonparametric approaches in Karlsen et al. [2007] and in Wang and Phillips [2006], however, the number of cointegrated regressors is not restricted. Furthermore in order to obtain the asymptotic results in Theorem 4.4 and Theorem 4.2 only structurally simple and familiar moment type conditions must be satisfied, if covariates and residual process satisfy a generalized spacial independence assumption.

These are the fundamental contributions of this thesis. Important secondary aspects or extensions of the above results are highlighted in the following.

Stationary and nonstationary processes treated with same procedure

In the chosen framework, the presented estimation methods work irrespective of underlying stationarity or not. While they are stated in the most general form for nonstationary data, they contain the stationary case as a subcase. Thus other than in parametric models, there is no pretesting for stationarity required. Our methods are adaptive to stationarity or nonstationary data. Therefore in this respect, there is no risk of model misspecification causing invalid inference.

Tailored Procedure for Stationary and Nonstationary covariates

In some high–dimensional economic models, in fact only one of the regressors appears as nonstationary while all others can be safely assumed as stationary (See e.g.

the study of Gil-Alaña and Robinson [1997]). With such pre–knowledge about the type of each regressor, we can improve on the efficiency of the general procedures as suggested in Section 3.2 and 3.3. With an adapted method, we can estimate stationary components at stationary rates. Therefore in finite sample studies and in practice, the suggested tailored method offers a significant improvement under feasibility aspects.

Uniform convergence results for β–null Harris recurrent processes

In order to show statistical properties of the generalized smooth backfitting estimator, we establish uniform convergence results for density estimators and regression estimators under the minimal assumption of β–null Harris recurrence. Due to the lack of exponential inequalities of Bernstein or Hoeffding type in this general setting, the proof of exponential tightness is quite lengthy and involved. As they are new to the probability literature, Corollaries A.2.2 and A.3 deserve some attention and might be of interest on their own.

1.4 Outline

This thesis is structured as follows. The second chapter presents the fundamental framework for estimation and the basic form of local smoothers in this general setting. In order to provide a thorough picture of the considered class of processes, certain concepts and notations of Markov theory must be introduced. Though the emphasis is on motivation and intuition how they facilitate our problem, while some technical properties are included only in the Appendix. The focus is on β–null Harris recurrence as the appropriate framework for kernel smoothing. Furthermore form and properties of general multidimensional kernel estimators are discussed, where features and specifics of the nonstationary setting are highlighted. With the basic notions at hand, the subsequent third chapter introduces the new estimation techniques. Depending on the degree of nonstationarity in the covariate process and on how close an additive model is to true model, respective estimation strategies are introduced. With generalized smooth backfitting we provide an estimation

procedure under the weakest assumptions on the covariates. Since it requires only pairwise β–null Harris recurrence as minimal assumption, it is the first valid estimation method for a vast class of high dimensional time series models. Chapter 4 contains the major convergence and asymptotic results. The focus is on the most interesting asymptotic expansions in the case of full β–null Harris recurrence and pairwise β–null Harris recurrence of the covariate vector. For completeness, intermediate cases are briefly treated under Extensions. Furthermore two practically interesting special cases are studied. We conclude the chapter by remarks on efficiency. In the Chapter 5, a simple simulation study shows that the method works in finite samples. the estimated five dimensional random walk model could not be estimated by general existing methods. The last chapter sums up. As the studied field of research is quite new, we conclude with an outlook for further research. All proofs as well as major technicalities are contained in the Appendix, which also comprises the formal statement of some basic notions and tools from Markov theory.

Chapter 2

Motivation and Basic Framework

In this chapter we introduce necessary tools and concepts for conducting nonparametric kernel type smoothing with stochastically nonstationary processes. In particular, in the first section we study β–null Harris recurrent processes as the appropriate framework for estimation. Series of discrete observations in this class of processes comprise all strictly and weakly stationary as well as nonstationary time series with potentially infinite, time dependent variance of unit root or long memory type. For deriving certain stochastic properties and for a thorough understanding, some notions and results from Markov theory are needed. These are presented with an emphasis on meaning and role of the employed techniques, while technical details and exact definitions of most of the Markov chain properties can be found in the Appendix. Comprehensive references for notions from Markov theory are Meyn and Tweedie [1993] and Nummelin [1984]. Furthermore in the second section, form and peculiarities of kernel estimators in this general setting are examined. These are fundamental in Chapter 3 for establishing estimation methods for an additive structural relationship as in (1.1).

2.1 Motivation, Intuition and Some Notation

Let $\{X_i\}_{i=1}^n$ be an aperiodic ϕ–irreducible Markov chain on the state space $\left(\mathcal{R}, \mathcal{B}^d\right)$ with transition probability P. Let the region of estimation $\mathcal{G} = \mathcal{G}^1 \times \ldots \times \mathcal{G}^d \subseteq \mathcal{R} =$

$\mathcal{R}^1 \times \ldots \times \mathcal{R}^d \subseteq \mathbb{R}^d$ be compact. Irreducibility essentially ensures that the Markov chain does not degenerate to a subspace of the original space $\mathcal{R}$. Technically it implies that for any set $A \in \mathcal{R}$ with $\phi(A) > 0$ it is $\sum_n P^n(x, A) > 0$ for any starting point $x \in \mathcal{R}$. Hence ϕ indicates, if a set can be reached by the process or not. For inference, only sets of positive ϕ measure are of interest. Throughout the paper, we assume that supp(ϕ) has non–empty interior[1]. Denote the class of non–negative measurable functions with ϕ–positive support by $\mathcal{E}^+$. Then also a set $A \in \mathcal{R}$ is in $\mathcal{E}^+$ if for the indicator function it is $\mathbf{1}_A \in \mathcal{E}^+$.
It is not intrinsically natural that the process lives in a bounded set - though this is inevitable for technical reasons in the estimation procedure. Note that $\mathcal{G}$ can be chosen sufficiently large such that there are a sufficiently many data points for nonparametric inference in $\mathcal{G}$. Denote by $\partial\mathcal{G}_h$ the h ring boundary of $\mathcal{G}$ in the following sense $x \in \partial\mathcal{G}_h$ iff $\|x - c\| \leq hC_1$ for any c from the boundary $\partial\mathcal{G}$. Furthermore write for the h ring interior $\mathring{\mathcal{G}}_h = \text{int}_h(\mathcal{G}) := \mathcal{G} \backslash \partial\mathcal{G}_h$.

2.1.1 Small Sets and Feller Chains

To ease notation, we use the following short–hand notation: For any non–negative measurable function η and any measure λ the operator kernel $\eta \otimes \lambda$ is defined by: $\eta \otimes \lambda(x, A) := \eta(x)\lambda(A)$, for all $(x, A) \in (\mathbb{R}^n, \mathcal{B}^n)$. For some general operator kernel P denote: $P\eta(x) := \int_A P(x, \mathrm{d}y)\eta(y)$ is a function, $\lambda P(A) := \int_{\mathbb{R}^n} \lambda(\mathrm{d}x)P(x, A)$ is a measure and $\lambda P\eta(x, A) := \int_A \int_{\mathbb{R}^n} \lambda(\mathrm{d}x)P(x, \mathrm{d}y)\eta(y)$ is a real number. Before we can introduce the concept of β–null Harris recurrence, we need some basic notions from Markov theory.

Definition 2.1 (Small Sets and Functions)**.** A function $\eta \in \mathcal{E}^+$ is small if there exist a measure λ, a positive constant $b > 0$ and an integer $m \geq 1$ such that:

$$P^m \geq b\eta \otimes \lambda\,. \tag{2.1}$$

A set A is small if $\mathbf{1}_A$ is small. Then every ϕ–positive subset of this set will also

[1]This condition is formally required to ensure that every Feller chain is a T-chain [Meyn and Tweedie, 1993] Theorem 6.0.1 (iii)

be small. If the measure λ satisfies (2.1) for some η, b and m, then we call λ a small measure.

Every ϕ–irreducible Markov chain has at least one small set (and see (A.1)). Small sets play an important role in describing the stability structure of a Markov chain. In the following, they will be essential for operationalizing estimation procedures. Small sets exist in abundance, in fact the entire $\mathcal{R}$ can be covered by small sets. Though in general for estimation, size and form of small sets are a–priori unknown but depend on the observed but unknown underlying process. And all interesting properties of one small set are not specific to it but also hold for all other small sets. This has been shown in Chen [1999b]. In practice, however, we need to know how to identify small sets, in particular when they come up explicitly in an estimation procedure. For a random walk any compact set is small, but in general this is not the case. However, if X is assumed to be Feller, then every small set is compact[2].

Definition 2.2 (Feller Chains)**.** A chain is called Feller, if $Ph(x) = \int P(x, \mathrm{d}y)h(y)$ is continuous for all h continuous.

Thus Feller processes satisfy a continuity assumption for the transition probability operator. This constraint offers a minimal way of establishing a link between stability of the chain and topology of the space. The Feller property guarantees small sets which are compact - hence "manageable" in practice. Most processes of practical interest in fact satisfy the Feller property. In particular, the random walk or α–stable processes are within the class of Feller processes (see Feller [1971] and Jakob [2001]).

[2]Since the support of the irreducibility measure of the chain is assumed to have non–empty interior, every Feller chain is also a T-chain. The exact definition of a T–chain is not important for our purpose (see Meyn and Tweedie [1993] chapter 6, page 127 ff.), however, we just profit from one important property: For a T-chain every compact set is petite(Theorem 6.2.5.ii in Meyn and Tweedie [1993]), where petite is a generalization of small.

2.1.2 β–null Harris Recurrence

Kernel type estimators consist of weighted local averages yielding pointwise estimates. Therefore intuitively there must be "sufficiently many" observations available in the neighborhood of any point in the range $\mathcal{G}$ to conduct consistent nonparametric inference. More precisely, as sample size increases, locally the number of data points should also grow to infinity at a certain rate. Obviously for evanescent processes, which eventually "wander off to infinity" this is not the case - they cannot be treated with local nonparametric estimation concepts. Thus data from time series with a deterministic trend must be correctly de–trended first to be admissible. An appropriate framework excludes cases of evanescence, but is still general enough to allow for processes having some kind of stochastic nonstationarity. In the Markov chain literature the concept of Harris recurrence captures these desired properties.

Definition 2.3 (Harris Recurrence)**.** A process X is Harris recurrent, if it returns almost surely to any neighborhood $\mathcal{N}_{x,h} = \{y \mid \|y - x\| \leq h\}$ of any $x \in \mathbb{R}^d$ for any h with $\phi(\mathcal{N}_{x,h}) > 0$.

The classes of non-evanescent and Harris recurrent processes are identical. See [Meyn and Tweedie, 1993] Theorem 9.2.2.ii for a formal proof of equivalence between the two concepts. In this sense, Harris recurrence is a minimal requirement for nonparametric Kernel type inference. Note that Harris recurrence only implies that with probability one, the process will recur to any point in its range. The expectation of this recurrence time, however, can be and generally is infinite. In a diffusion setting, Harris recurrence is essentially equivalent to the nonpositivity of the generator of the diffusion semigroup. This is shown in Bandi and Phillips [2004].

Furthermore Harris recurrent processes come with some useful additional properties we will exploit in the following. Harris recurrence allows to construct a split chain which decomposes the original Markov chain into blocks of independent identically distributed parts (see Appendix A.1). The number of these independent parts $T(n)$ corresponds to how often the process regenerates. These

resulting blocks $U_1, \ldots, U_{T(n)}$ play the role of iid observations in sums for asymptotic central limit theorem arguments. Thus for any type of estimation procedure, their stochastic number $T(n) \overset{\text{a.s.}}{\to} \infty$ plays the central role of the effective sample size in asymptotic considerations. Hence rates of convergence are generally path–dependent stochastic determined by $T(n)$ compared to deterministic n for stationary data. Since with probability one, it is $T(n) \leq n$, estimators for nonstationary data will converge slower than in the stationary case, where the relation holds with equality. Though generally, the number of regenerations $T(n)$ of an underlying process is not observable. To compensate for this, we introduce the observable quantity

$$T_C(n) := \sum_{i=0}^{n} \mathbf{1}_C(X_i) \tag{2.2}$$

for $C \in \mathcal{E}^+$, which counts the number of times the process hits a set C. Furthermore for an irreducible Markov chain, small sets play the role of a pseudo–atom, where the process recurs and (3.4) holds with equality and $b = 1$. Therefore if C is small, $T_C(n)$ and $T(n)$ are asymptotically equivalent in the following sense

$$\frac{T_C(n)}{T(n)} \xrightarrow{\text{a.s.}} c \tag{2.3}$$

with $c > 0$ constant (Remark 3.5. in [Karlsen and Tjøstheim, 2001]).

A general Harris recurrent process is nonstationary. Therefore it has no stationary distribution or density function to be estimated nonparametrically. But Harris recurrence ensures the existence of a unique (up to a multiplicative constant) invariant measure π to which the transition probabilities converge to in a certain sense (See Appendix A.1 for its formal construction). If this invariant measure has a density function, it is the object which can be estimated by a kernel type density estimator. It should be noted the invariant measure π and the irreducibility measure ϕ are equivalent in the sense that $\phi = a\pi$ for $a \in \mathbb{R}^+$. Distinguish between two fundamentally different cases: X is positive recurrent if the invariant measure is finite and with some appropriate scaling can be transformed into a probability measure. If the invariant measure is no longer finite but only σ–finite the process is only null recurrent. The later case is technically more intricate. Restricted to

a small set C, however, the invariant measure is always finite, i.e. $\pi(C) < \infty$ (See [Meyn and Tweedie, 1993], Proposition 5.6.page 73). Therefore the invariant measure density on small sets $\pi_C(x) := \frac{\pi(x)}{\pi \mathbf{1}_C}$ is well defined and the exact form of c in (2.3) is determined as $\pi \mathbf{1}_C$. It is assumed throughout the paper that any invariant measure is absolutely continuous with respect to Lebesgue measure. And for convenience, the Radon-Nikodym derivatives are also called densities in the null recurrent case. Furthermore in the following, any support is with respect to the respective invariant measure.

Simple Harris recurrence only yields stochastic rates of convergence for estimators, where distribution and size of $T(n)$ have no a priori known structure but fully depend on the underlying process. Though by imposing a slight regularity condition on the regeneration structure of the process, we get a much simpler and more familiar polynomial form.

Definition 2.4 (β–null Harris recurrence)**.** The chain (X_i) is β–null recurrent if there exists a small non–negative function f, an initial measure λ, a constant $0 < \beta \leq 1$ and a slowly varying at infinity[3] function L_f such that

$$\mathbb{E}_\lambda \left[\sum_{i=0}^{n} f(X_i) \right] \sim \frac{1}{\Gamma(1+\beta)} n^\beta L_f(n) \quad \text{for} \quad n \longrightarrow \infty \,, \tag{2.4}$$

where $\mathbb{E}_\lambda$ denotes the conditional expectation given that the initial distribution of X_0 is λ.

Note that β is a global parameter characterizing the type of nonstationarity of the chain (X_i). In particular it is not specific to the choice of the small function f. This is a simple consequence of Orey's theorem. A detailed proof is given in Karlsen and Tjøstheim [2001] Lemma 3.1. In practice, β–null Harris recurrence does not appear to be a severe constraint, since examples of Harris recurrent but not β–null Harris recurrent processes are still to be found (See Chen [2000] and Darling and Kac [1957]). But the gain of the assumption is substantial. For a small set C and a β–null Harris recurrent chain, we have the asymptotic equivalence

$$\mathbb{E}_\lambda(T(n)) \asymp \mathbb{E}_\lambda(T_C(n)) \asymp n^\beta L(n) \,, \tag{2.5}$$

[3]A function L is slowly varying at infinity if $\lim_{\lambda \to \infty} \frac{L(\lambda x)}{L(\lambda)} = 1$ for all x

with $f = \mathbf{1}_C$ in the above definition. Thus effective sample sizes in estimation are on average of order $n^\beta L(n)$. Furthermore β–null Harris recurrence allows to capture the entire degree of nonstationarity in a single parameter $0 < \beta \leq 1$, where β decreases with increasing nonstationarity of the process. If a process is stationary or positive recurrent, β is 1, for a univariate random walk β is $1/2$, and for two independent random walks the compound β is zero (See Kallianpur and Robbins [1954] and Resnick and Greenwood [1979]). In any higher dimension $d \geq 2$ a random walk is transient.

Assuming β–null Harris recurrence restricts the tail behavior of the recurrence time of the process to be a regular varying function. Therefore β–null Harris recurrence can be equivalently defined as below.

Definition 2.5. Let τ_0 be the recurrence time of the process X. Then X is β–null Harris recurrent if

$$\mathbb{P}_\lambda(\tau_0 > n) = \frac{1}{\Gamma(1-\beta)n^\beta L_s(n)}(1 + o(1)) , \tag{2.6}$$

where L_s is a slowly varying at infinity function depending on s, the function part of the atom kernel in (A.1). The initial measure λ is a dirac point mass at an arbitrary point of regeneration $X_0 = x$.

Furthermore if (2.6) holds, then it is:

$$\sup(p \geq 0 : \mathbb{E}_\lambda \tau_0^p < \infty) = \beta , \tag{2.7}$$

with λ as in (2.6). This is an easy consequence of the definition above (See Proof of Lemma 3.4. in Karlsen and Tjøstheim [2001]). Thus other than for $\beta = 1$, the expectation of the recurrence time is not finite. Though generally for p small enough, $\mathbb{E}_\lambda \tau_0^p$ is finite.

If the tail of the recurrence time is a regular varying at infinity function fulfilling (2.6), this implies the recurrence time process to be a stable increasing process with index β. Inversion yields the asymptotic distribution of $T(n)$. Höpfner and Löcherbach [2000] show, if X is β–null Harris recurrent, we have the asymptotic distribution

$$T(n) \xrightarrow{\mathcal{D}} n^\beta L(n)\, g_\beta \tag{2.8}$$

where g_β is distributed according to a Mittag–Leffler distribution $\mathcal{M}_\beta$. The distribution family $\mathcal{M}_\beta$ generalizes the exponential distribution and is discussed in detail e.g. in Jayakumar and Suresh [2003]. Thus according to the split chain considerations before, it is not surprising that additive functionals of β–null Harris recurrent processes converge to Brownian motion subject to an independent time change according to $\mathcal{M}_\beta$ (See (A.10) and Höpfner and Löcherbach [2000]).

Examples. Besides the random walk up to dimension two, the class of β–null Harris recurrent processes contains other important classes of processes. Linear stationary ARMA and but also ARIMA models fit into the framework. But also nonlinear autoregressive time series are β–null Harris recurrent under certain conditions (See e.g. example 3.1 in Karlsen and Tjøstheim [2001] for a specific case). Furthermore long memory models like all types of fractionally integrated ARFIMA(d) models are contained, irrespective of if they are stationary or nonstationary, i.e. $d \in [0, 1]$ is admissible (see Wang and Phillips [2006]). And general infinitely divisible processes like α–stable processes for $1 < \alpha \leq 2$ and dimension less or equal than α, are β–null Harris recurrent with $\beta = 1 - \frac{1}{\alpha}$ (See Sato [1999]). These include α–stable processes with fat tails and thus infinite variance but finite mean plus Brownian motion. Certain Feller processes as generalizations thereof are also in the considered class (See Schilling [1998]). Another β–null Harris recurrent class of processes of interest for modeling exchange rates or real prices in bubble periods is given by

$$X_t = \mathbf{1}_{\{|X_{t-1}| \leq M\}} g(X_{t-1}) + \mathbf{1}_{\{|X_{t-1}| > M\}} X_{t-1} + e_t \tag{2.9}$$

for some finite $M > 0$ and some measurable function g finite on $|x| \leq M$. This process behaves like a random walk for large X_t's. Furthermore mean reverting processes like the Ornstein–Uhlenbeck process $\mathrm{d}X_t = -aX_t\,\mathrm{d}t + \mathrm{d}W_t$ for $a \geq 0$ are β–null Harris recurrent. Conditions on diffusion models satisfying β–null Harris recurrence are discussed in Höpfner and Löcherbach [2000], Examples 3.5. and Bandi and Phillips [2004]. Other examples of β–null recurrent processes are the first order threshold model studied in Meyn and Tweedie [1993], page 503ff and the exponential autoregressive process looked at in Cline and Pu [1999].

Remarks. Like for standard α–mixing, also β–null Harris recurrence is hard to test for formally. Though as common in time series analysis some plausibility checks, e.g. for no trend, might undermine that β–null Harris recurrence is an appropriate framework for given observations. But essentially it has to be assumed as the minimal abstract framework for nonparametric kernel estimation. Though within the class of β–null Harris recurrent processes, it might sometimes be of interest to determine the type of nonstationarity β from the data. A direct way to estimate β can be derived from the asymptotic equivalence (2.8). For Feller processes, small sets are compact. Thus setting

$$\widehat{\beta} = \frac{\ln(T_C(n))}{\ln(n)}$$

with C small yields a strongly consistent estimator (See Karlsen and Tjøstheim [1998] Lemma 3.1 for a proof). The rate of convergence, however, is quite slow requiring large finite sample sizes for meaningful results. Furthermore according to (2.8), the asymptotic distribution is of log–transformed Mittag–Leffler type $\mathcal{M}_\beta$ depending on the estimated parameter of interest β. Alternatively, since β is the polynomial order of a regular varying function for some tail distribution, a standard Hill estimator (see Hill [1975]) may be applied to estimate β. However, as in its usual domain of application extreme value theory, convergence is extremely slow. This is not improved by the fact that for n observations (Y_i, X_i) there are effectively only $T_C(n) \simeq n^\beta L(n)$ observations for the recurrence time process. Thus unless sample size is huge, such attempts might be of limited practical use. As a third way to estimate β, an empirical version of the expectation in (2.7) could be checked for finiteness with varying values of p.

It should be noted here, that in our estimation methods and their asymptotic expansions in Chapter 3 and 4, the parameter of nonstationarity β does not enter results explicitly. It only appears in the asymptotic choice of bandwidth. But in finite samples, a local choice of bandwidth might be more favorable anyway which requires no pre–knowledge of β (See Chapter 5 for details).

2.1.3 Nonparametric Curse of Dimensionality

As seen in the examples above and in general, with an increasing number of covariates, the compound process becomes transient and is very unlikely to fit the framework of Harris recurrence for $d \geq 2$. Hence in these cases the existing results of Karlsen and Tjøstheim [2001], Karlsen et al. [2007] and Moloche [2001] can no longer be applied and there is no existing method of estimation. So in contrast to the standard curse of dimensionality in nonparametric estimation, this second nonstationary curse of dimensionality does not only deteriorate the performance of nonparametric estimation but does in fact prevent any estimation at all for high dimensional problems. Why local smoothing techniques suffer from the standard curse of dimensionality can be illustratively explained. When increasing the dimensionality of the problem, the kernel windows must be made wider to offset the exponentially sparser density of the data points. This causes slower rates of convergence with increasing d. The nonstationary more severe curse of dimensionality, however, is due to generality in the type of underlying data and not a result of the generality in the applied method of estimation. When adding degrees of freedom by increasing dimensionality for a process without a fixed stationary law, the process can cluster for a very long time in a specific region of the space while leaving others more or less empty. Thus for very low dimensions already, regeneration can no longer be guaranteed almost surely. Thus β–null Harris recurrence cannot be fulfilled anymore. While the standard curse of dimensionality can be circumvented by restricting the structural model class as additive, dealing with the nonstationary curse of dimensionality is rather new to the econometric time series literature. In order to countervail the nonparametric curse of dimensionality, an estimation method should avoid full–dimensional objects. If it is solely built of one–and two dimensional marginal objects, we only need β–null Harris recurrence in these components which is by far less restrictive than requiring β–null Harris recurrence for the full dimensional vector of covariates. This is why pairwise β–null Harris recurrence plays an important role in Chapter 3 and 4.

2.2 Nonparametric Kernel Estimators and Peculiarities for Nonstationary Data

When observing a multivariate nonstationary process X on a fixed bounded set $\mathcal{G}$, available data points of different marginal component processes within $\mathcal{G}$ are generally different - in particular the amount of data points and the actual elements differ asymptotically depending on the type of nonstationarity of the marginal processes. Set

$$J_j = \left\{ i \in \{1, \dots, n\} \, | X_i^j \in \mathcal{G}_j \right\} \quad \text{and} \quad J_{jk} = \left\{ i \in \{1, \dots, n\} \, | X_i^{jk} \in \mathcal{G}_{jk} \right\} \tag{2.10}$$

and J_f analogously for the full dimensional process $X \in \mathcal{G}$. Then in general $|J_f| \leq |J_{jk}| \leq |J_j| \leq n$ and $J_j \neq J_k$ and $J_{jk} \neq J_{jk'}$ for $j \neq k \neq k'$. Thus the amount of data points decreases with increasing dimension and when a X_i^{jk} is in $\mathcal{G}_{jk}$ for $i \in J_{jk}$ this does not at all imply that also X_i^{jl} is in $\mathcal{G}_{jl}$ for $i \in J_{jk}$. This will be important for balancing bias terms in the generalized backfitting procedure presented later on. See Figure A.1 in the Appendix for an illustration. Generally we aim to choose $\mathcal{G}$ large enough – in applications containing as much of the empirical support as possible. If $\beta^j = \beta^k$, then it is asymptotically $|J_j| \asymp |J_k|$ if $\mathcal{G}_j$ and $\mathcal{G}_k$ are small. Actual elements of $|J_j|$ and $|J_k|$, however might in general not coincide. If types of nonstationarities differ, not even the amount of observations will asymptotically be the same. In a stationary setting such complication does not arise since asymptotically speeds for different components are all of the same order n. Hence there it is not problematic to use the index set of the full dimensional process X for all marginal component processes. In our setting this is generally too restrictive as will be explained below. Denote $n_j = T_{\mathcal{G}_j}^j(n) = |J^j|$ and n_{jk} and n_f analogously. If $\mathcal{G}_j$ is small for X^j, then $n_j \asymp n^{\beta^j}$ asymptotically and $\pi_j(\mathbf{1}_{\mathcal{G}_j}) < \infty$.

Furthermore for a multivariate β–null Harris recurrent process X on $\mathcal{G}$, the recurrence frequency can still vary across each univariate component and generally decreases from univariate to bivariate to multivariate subcomponents. This is even true if $\mathcal{G}$ is small, since only asymptotically $T_{\mathcal{G}}(n) \asymp n_f \asymp T(n)$ (See (2.3)).Thus generally, the number of independent blocks of observations and hence the effective asymptotic samples size varies for each one–dimensional direction and tends

to infinity at slower rates for higher dimensions (See Subsection 2.1.2 and the Appendix A.1 for details). Denote by $T(n)$, $T^j(n)$ and $T^{jk}(n)$ the number of recurrence times of the processes $X \in \mathcal{G}$, $X^j \in \mathcal{G}_j$ and $X^{jk} \in \mathcal{G}_{jk}$. Set $(\tau_l^j)_{l=1}^{T^j(n)}$ the sequence of recurrence times for the marginal process X^j, and $(\tau_l^{jk})_{l=1}^{T^{jk}(n)}$ for X^{jk} and $(\tau_l^f)_{l=1}^{T(n)}$ for X respectively. Denote $\tau^j_{T^j(n)+1} = n$. Then define the index sets $I_j(X^j) = J_j(X^j)$ and

$$\begin{aligned} I_{jk}(X^j) &= \bigcup_{l=1}^{T^{jk}(n)} \left\{ i \in J_{jk} | \tau_l^{jk} = \tau_\eta^j \leq i \leq \tau_{\eta+1}^j \leq \tau_{l+1}^{jk}, \text{ for the smallest } \eta \geq l \right\} \\ I_f(X^j) &= \bigcup_{l=1}^{T(n)} \left\{ i \in J_f | \tau_l^f = \tau_\eta^j \leq i \leq \tau_{\eta+1}^j \leq \tau_{l+1}^f, \text{ for the smallest } \eta \geq l \right\} \end{aligned} \quad (2.11)$$

While the formal definitions look quite complicated, the main points are illustrated in Figure 2.1. If type of index set and type of process coincide, the definition of the index sets keeps all observations. Thus for X^j the index set I_j comprises all observations $i = 1, \ldots, n_j$. Tough if the types do not match, some observations might be omitted for coordinating speeds among the involved dimensions. In summations the involved processes for the index set appear as summands - thus for ease of notion we can leave them out in the following. Generally for a fixed process X^j it is $I_f \subseteq I_{jk} \subseteq I_j$, and $I_j \neq I_k$ and $I_{jk} \neq I_{jk'}$ for $j \neq k \neq k'$. If $\beta^j = \beta^k$, then it is asymptotically $|I_j| = |I_k|$. Since in practice recurrence times are not observable, operationalize the choice of index sets by the asymptotically equivalent hitting times $T^j_{\mathcal{C}_j}(n)$ for a small set $\mathcal{C}_j \subseteq \mathcal{G}_j$. Then $T^j_{\mathcal{C}_j}(n) \asymp T^j(n)$ asymptotically. The same holds for all other directions and dimensions. If $\mathcal{G}$ is small for X then any of its coordinatewise projections $\mathcal{G}_j^f$ or $\mathcal{G}_{jk}^f$ are small for respective component directions and generally $\mathcal{G}_j^f \subseteq \mathcal{G}_j$. If we choose $\mathcal{G}$ according to the data, the easiest choice is according to the full dimensional X

$$\mathcal{G}^f = \mathcal{G}_1^f \otimes \ldots \otimes \mathcal{G}_d^f \ . \quad (2.12)$$

Selecting $\mathcal{G}$ is a tradeoff: Choose it big enough not to miss many observations, choose it small enough such that still $\pi_j(\mathbf{1}_{\mathcal{G}_j}) < \infty$. In the sequel, there will be the prominent case, where only pairwise properties between covariates are used

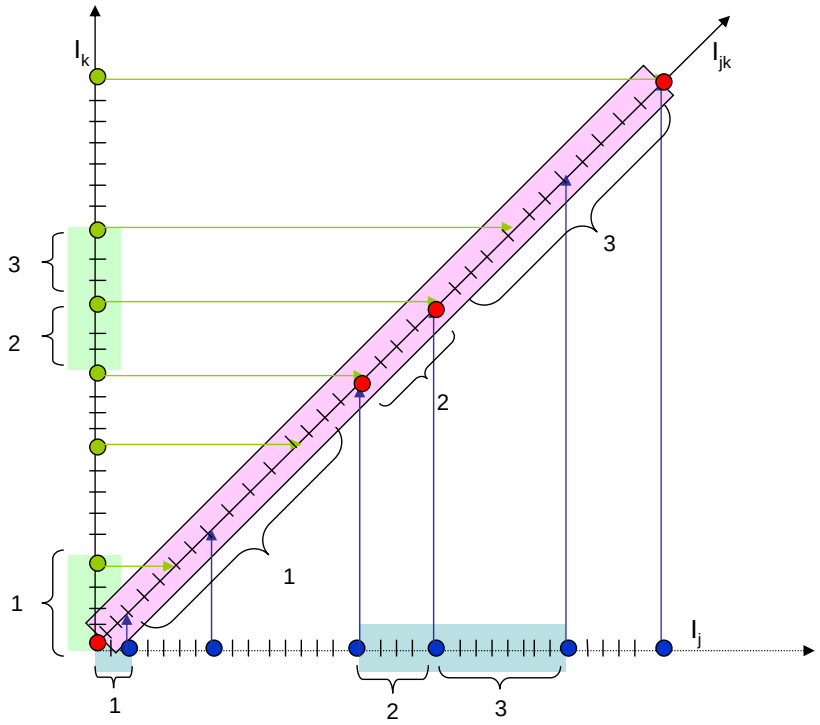


Figure 2.1: The schematic figure shows time points of observations in $\mathcal{G}_{jk}$ or any of its coordinatewise projections, marked by bars for the univariate marginal processes on the axis, for the bivariate process on the diagonal. Thus on all axes they mark the index set J_{jk}. Points of recurrence for the respective process are marked with colored circles. Then observations in the shaded regions comprise the index set I_{jk} in the respective direction. And numbered braces denote the number of independent blocks.

and known. For generalized smooth backfitting we work with only pairwise β–null Harris recurrent processes, where recurrence of higher dimensional components of or the full covariate vector is generally not the case. For this, we need to work in pairwise adapted sets for each $j = 1, \ldots, d$

$$\mathcal{G}^{(j)} = \mathcal{G}_1^{(j)} \otimes \ldots \otimes \mathcal{G}_{j-1}^{(j)} \otimes \mathcal{G}_j \otimes \mathcal{G}_{j+1}^{(j)} \otimes \ldots \otimes \mathcal{G}_d^{(j)} \,, \tag{2.13}$$

where $\mathcal{G}_{jk}$ is chosen according to X^{jk}, and $\mathcal{G}_k^{(j)}$ is its coordinate projection in direction k on $\mathcal{G}_k$ with $k \neq j$. Set $\mathcal{G}_{jk} = \mathcal{G}_k^{(j)} = \mathcal{G}_j$ for $j = k$. Here the scaling occurs according to the highest dimensional available object where the amount of data grows to infinity with increasing sample size. A full dimensional $\mathcal{G}$ with this property is in this general setting not available.

The basic underlying estimation technique will be kernel smoothing with product kernels. Denote

$$K_h\left(X_i - x\right) = \frac{1}{h^d} \prod_{j=1}^{d} K\left(\frac{X_i^j - x^j}{h}\right) \tag{2.14}$$

where $X_i = (X_i^1, \ldots, X_i^d)$ and K is a standard kernel function. The bandwidth h depends on the recurrence frequency of X. It is generally larger than the marginal bandwidth choices h_j for X^j. Due to possible different recurrence structures, not only each univariate direction has in general a different bandwidth $h_j \neq h_k$ for $j \neq k$, but also bivariate and higher components need special bandwidth choices different from the usual product of involved single dimensional bandwidths i.e., $h_{jk} \neq h_j h_k$ in general.

Assumption 2.1. *1. The univariate kernel K is symmetric about 0 and bounded. It has compact support on $S^j = [-c^j, c^j]$ with $S^j \subseteq \mathcal{G}_j$. So the Kernel can in fact depend on j, which, however, will subsequently be suppressed in notation.*

2. Furthermore K as well as $K(u) \cdot u^k$ has to be Lipschitz-continuous for any x and any power $k < 2p + 1$ with Lipschitz constant $L > 0$, where p indicates the number of partial derivatives possible for the conditional mean function m.

Since for general β–Harris recurrent processes no such thing as a stationary density exists, kernel density estimators converge to the corresponding more general object, the density of the invariant measure under suitable assumptions (See Theorem 5.1. Karlsen and Tjøstheim [2001] for exact conditions). In this sense the unique invariant measure serves as a generalization of a stationary distribution. Following the reasoning in the previous section, the appropriate scaling of the usual marginal kernel density estimator $\widehat{\pi}_j(x^j)$ has to be slightly modified. Paying for the potentially nonstationary character of the marginal process X^j, the appropriate scaling in contrast to the usual kernel density estimate is to be adaptively stochastic by the number of effective iid observations, thus by the number of regenerations $(T^j(n))^{-1}$ instead of the usual $(n)^{-1}$. For β–Harris recurrent processes the usual law of large numbers cannot hold anymore, but there exists a more general analogue in the quotient limit theorem (A.11), which guarantees convergence of a quotient of two stochastic components under quite general assumptions.
Kernel density and conditional mean estimators are defined on $\mathcal{G}_j$, $\mathcal{G}_{jk}$ or $\mathcal{G}_f$ bounded as

$$\widehat{\pi}(x) = \frac{1}{T(n_f)} \sum_{i \in I_f} K_h(X_i - x) \tag{2.15}$$

$$\widehat{\pi}_j(x^j) = \frac{1}{T^j(n_j)} \sum_{i \in I_j} K_{h_j}(X_i^j - x^j) \tag{2.16}$$

$$\widehat{m}_j(x^j) = \frac{\sum_{i \in I_j} K_{h_j}(X_i^j - x^j) Y_i}{\sum_{i \in I_j} K_{h_j}(X_i^j - x^j)} = \frac{1}{T^j(n_j)} \cdot \frac{\sum_{i \in I_j} K_{h_j}(X_i^j - x^j) Y_i}{\widehat{\pi}_j(x^j)} \,, \tag{2.17}$$

where the norming function depends on the recurrence frequency of the respective processes. The above definitions can be operationalized according to (2.2) with appropriate small sets. In the univariate case the numerator of the kernel density estimator (2.16) can be regarded as a local time estimator. In higher dimensions local time does not exist any more, but the numerator can still be interpreted as an occupation time like object (see [Phillips and Park, 1998]). Thus occupation time quantities are defined as $\widehat{L}_j(x^j) = \sum_{i \in I_j} K_{x^j,h^j}(X_i^j)$ and

$\widehat{L}_{jk}(x) = \sum_{i\in I_{jk}} K_{x^{jk},h_{jk}}(X_i^{jk})$. Since it is in general

$$\pi_j^{(k)}(x^j) = \int_{\mathcal{G}_k^{(j)}} \pi_{jk}(x^{jk})\mathrm{d}x^k \neq \pi_j(x^j) \tag{2.18}$$

$$\pi_j^{f}(x^j) = \int_{\mathcal{G}_k^{(f)}} \pi(x)\mathrm{d}x^{-j} \neq \pi_j(x^j)\,, \tag{2.19}$$

Define $\pi_j^{(k)} = \pi_{jk} = \pi_j$ for $j = k$. We also need to introduce

$$\widehat{\pi}_j^{(k)}(x^j) = \frac{1}{T^{jk}(n_{jk})}\sum_{i\in I_{jk}} K_{h_{jk}}(X_i^j - x^j) = \frac{\widehat{L}_j^{(k)}(x^j)}{T^{jk}(n_{jk})} = \int_{\mathcal{G}_k^{(j)}} \widehat{\pi}_{jk}(x^{jk})\,\mathrm{d}x^k \tag{2.20}$$

$$\begin{aligned}\widehat{m}_j^{(k)}(x^j) &= \frac{\sum_{i\in I_{jk}} K_{h_{jk}}(X_i^j - x^j)Y_i}{\sum_{i\in I_{jk}} K_{h_{jk}}(X_i^j - x^j)} = \frac{1}{T^{jk}(n_{jk})}\cdot\frac{\sum_{i\in I_{jk}} K_{h_{jk}}(X_i^j - x^j)Y_i}{\widehat{\pi}_j^{(k)}(x^j)} \\ &= \sum_{i\in I_{jk}} \frac{K_{h_{jk}}(X_i^j - x^j)Y_i}{\widehat{L}_j^{(k)}(x^j)}\,.\end{aligned} \tag{2.21}$$

Set $\widehat{\pi}_j^{(k)} = \widehat{\pi}_{jk} = \widehat{\pi}_j$ for $j = k$. Analogously derive $\widehat{\pi}_j^f(x^j))$ and $\widehat{m}_j^f(x^j)$ from the full dimensional process X

$$\widehat{\pi}_j^f(x^j)) = \frac{1}{T(n_f)}\sum_{i\in I_f} K_h(X_i^j - x^j) = \frac{\widehat{L}_j^f(x)}{T(n_f)} = \int_{\mathcal{G}_{-j}^f} \widehat{\pi}_f(x)\,\mathrm{d}x^{-j} \tag{2.22}$$

$$\widehat{m}_j^f(x^j) = \frac{\sum_{i\in I_f} K_h(X_i^j - x^j)Y_i}{\sum_{i\in I_f} K_h(X_i^j - x^j)}\,. \tag{2.23}$$

Note that the nonstationary character for the estimators in (2.20) and (2.21) is determined by the two–dimensional type β^{jk}. The estimators in (2.20) and (2.21) have nonstationary type $\beta = \beta_f$. Hence in their asymptotic behavior the first bivariately governed pair has a univariate rate with bivariate nonstationary character from the data, i.e. $\widehat{\pi}_j^{(k)}(x^j)) - \pi_j^{(k)}(x^j) = \text{bias} + O_P((T^{jk}(n_{jk})h_{jk})^{-1/2}) + o_P(h_{jk}^2)$ where the bias vanishes under suitable technical assumptions with order h_{jk}^2 in the interior (See Karlsen and Tjøstheim [2001]). For (2.20) and (2.21) the statement holds with nonstationary with rate $(T(n_f)h)^{1/2} \asymp (n_f^{\beta}h)^{1/2}$ and bias of order h^2. For the rest of the paper we will suppress indices in n when appearing in recurrence

times and hitting numbers of small sets, e.g. we write $T^j(n)$ instead $T^j(n_j)$ for ease of notation.

As in a stationary setting, since $\mathcal{G}$ or all $\mathcal{G}^{(j)}$ are bounded, near the boundary of the support the standard kernel estimator is poor and has considerable boundary specific bias. This is because the kernel density estimator has no knowledge of the boundary and may, in general, assign probability mass outside the support. Therefore we need to slightly modify the usual kernel without harming its essential nonnegativity. This is common practice in kernel estimation on compact sets. We introduce the modified kernel

$$K_{v,h}(u) = \frac{K_h(u-v)}{\int_{\mathcal{G}_j} K_h(w-v)\mathrm{d}w} \tag{2.24}$$

Note that for the use of modified kernels, extra attention has to be paid to the kernel moments. It is

$$\int_{\mathcal{G}_j} K_{x^j,h_j}(u^j)\mathrm{d}u^j \quad = \quad 1 \quad \text{for all } x^j \in \mathring{\mathcal{G}}^j_{2c^jh_j} \tag{2.25}$$

and depends on x^j otherwise. Thus the corresponding Kernel constants are defined as

$$\kappa_l(u) = \int_{\mathcal{G}_j} (u-v)^l \frac{K_{h_j}(u-v)}{\int_{\mathcal{G}_j} K_{h_j}(w-v)\,\mathrm{d}w}\,\mathrm{d}v.$$

Easy calculations show that there are three different cases

$$\kappa_l(u) = \begin{cases} \int_{\mathcal{G}_j} v^l K(v)\,\mathrm{d}v & \text{for } u \in \mathring{\mathcal{G}}_{j,2c^jh_j} \\ \int_{\mathcal{G}_j} v^l K(v)\,\mathrm{d}v + O({h_j}^{l+1}) & \text{for } u \in \partial\mathcal{G}_{j,2c^jh_j} \backslash \partial\mathcal{G}_{j,c^jh_j} \\ \int_{\mathcal{G}_j} (u-v)^l K_{h_j}(u-v)\,\mathrm{d}v + O({h_j}^{l+1}) & \text{for } u \in \partial\mathcal{G}_{j,c^jh_j} \end{cases}.$$

From now on denote by $\mathring{\mathcal{G}}_{j,h_j} = \mathring{\mathcal{G}}_{j,2c^jh}$ the interior of interest and by $\partial\mathcal{G}_{j,h_j} = \partial\mathcal{G}_{j,2c^jh}$. The modified kernels only have an influence at boundary points $u \in \partial\mathcal{G}_{j,2c^jh}$, where they differ from usual kernel constants. Analogously, kernel constants $\kappa_l^2 = \int_{\mathcal{G}_j}(u-v)^l(K_{h_j}(u,v))^2\,\mathrm{d}v$ are defined. For the rest of this paper all kernels are modified kernels.

Chapter 3

Estimation

In this chapter we introduce nonparametric estimation techniques for a structural additive model (1.1) in a β–null Harris recurrent framework. Countervailing the standard curse of dimensionality, we impose additivity of the unknown function. Aiming to circumvent the nonstationary curse of dimensionality, developed estimation techniques are of smooth backfitting type (See Mammen et al. [1999]). The appropriate estimation method for a given problem must be selected according to the degree of nonstationarity in the covariate vector and according to how close the true model is to an additive structural relationship. From standard smooth backfitting in Section 3.2. to generalized smooth backfitting in Section 3.3 minimal requirements on the compound covariate process can be relaxed remarkably, but also the admissible generality in the true model decreases.

3.1 Choice of the Type of Estimation Technique

To overcome the ordinary curse of dimensionality in nonparametric statistics, the problem is modeled additively. In a usual stationary mixing setting, there are several kernel based techniques how to fit additive models: classical backfitting in Buja et al. [1989] and Tibshirani and Hastie [1990], marginal integration by Linton and Nielsen [1995] and Tjøstheim and Auestad [1994], smooth backfitting by Mammen et al. [1999], and the two–step local partitioned regression (LPR) approach by

Christopeit and Hoderlein [2006]. Though apart from backfitting type estimators, all other procedures are based on a full–dimensional nonparametric regression pilot estimate. In our setting in particular, this would require the full dimensional process X to be Harris recurrent, which is generally too restrictive. In contrast and with hope to countervail the nonparametric course of dimensionality, backfitting avoids fitting a full dimensional regression estimate. The estimation procedure is iterative where for classical backfitting in each step only one component is updated while all others remain fixed. Therefore in fact, only one–dimensional smoothing is applied. Asymptotic theory for classical backfitting, however, suffers from the difficulty that the estimate is defined as the limit of the iterative backfitting algorithm for which there is no explicit closed form available. Although Opsomer and Ruppert [1997] and Opsomer [2000] could show some theoretical results under restrictive conditions on the design densities, asymptotic inference under general assumptions is still an open issue. Furthermore classical backfitting fails to reach the oracle efficiency bound i.e., additive components are not estimated with the same accuracy as if the other components were known. The bias of one additive component depends strongly on all other directions. Even some moderate correlation between covariates may cause the estimator to collapse and diverge. And for classical backfitting to work, rather strong conditions have to be fulfilled. In total, for general β–null Harris recurrent data, it seems more advisible to chose a more robust technique as a starting point for estimation.

Smooth backfitting estimates (SBE) are defined as minimizers of a smoothed least squares criterion. From this, the backfitting iteration algorithm can be derived, according to which the estimates are calculated. Thus asymptotic analysis is simplified, since the estimate is explicitly defined. In view of Mammen et al. [2001], the SBE can also be seen as an orthogonal projection of the data vector onto the space of additive functions. Furthermore, under weak assumptions the SBE reaches efficiency and is furthermore robust, easy to calculate and fast (see Mammen and Park [2005], Haag [2006] and Yu et al. [2007]). As with classical backfitting, the SBE does not need full–dimensional estimates. But in contrast smoothing occurs with respect to all other covariates resulting in a more robust

estimator. Therefore it avoids not only the ordinary stationary curse of dimensionality but also offers a way to countervail the nonstationary curse of dimensionality. Since it requires only one and two–dimensional marginal processes to be pairwise Harris recurrent, a smooth backfitting type estimator appears to be the most suitable choice for a recurrent setting.

3.2 Standard Smooth Backfitting for Nonstationary Covariates

Assume throughout this section that the regression model has additive form as in (1.1). Furthermore all mentioned densities of invariant measures exist and are finite on $\mathcal{G}$ or any of its subspaces. And the regression functions m_j are in the respective weighted $L^2_{\pi_j}(\mathcal{G}_k)$ spaces.

For all stationary data processes identifiability in population is achieved by

$$\int_{\mathcal{G}_j} m_j(x^j)\pi_j(x^j)\mathrm{d}x^j = 0 \;, \tag{3.1}$$

for all $j = 1, \ldots, d$. In this standard stationary case, the smooth backfitting estimators (SBE) for component functions $(\widetilde{m}_0, \ldots, \widetilde{m}_d)$ are then obtained as solutions of the following system of integral equations

$$\widetilde{m}_j(x^j) = \widehat{m}_j(x^j) - \widetilde{m}_{0,j} - \sum_{k \neq j} \int_{\mathcal{G}^k} \widetilde{m}_k(x^k) \frac{\widehat{\pi}_{j,k}(x^j, x^k)}{\widehat{\pi}_j(x^j)} \mathrm{d}x^k \tag{3.2}$$

$$\widetilde{m}_{0,j} = \frac{\int_{\mathcal{G}_j} \widehat{m}_j(x^j)\widehat{\pi}_j(x^j)\mathrm{d}x^j}{\int_{\mathcal{G}_j} \widehat{\pi}_j(x^j)\mathrm{d}x^j} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{3.3}$$

where $\widehat{m}_j$ is a marginal Nadaraya–Watson pilot estimator as defined in (2.17) and $\widehat{\pi}_{jk}$ and $\widehat{\pi}_j$ are standard Kernel density estimators of the respective stationary densities. The form of $\widetilde{m}_{0,j}$ is determined such that $(\widetilde{m}_1, \ldots, \widetilde{m}_d)$ satisfy sample analogue versions of the norming conditions in (3.1). Note that estimates are obtained from univariate and bivariate quantities only. Contrary to ordinary backfitting, smooth backfitting involves some additional smoothing which makes

it more robust. In particular, no restriction on the correlation structure of the covariates is needed in order to obtain estimates with a well–determined asymptotic distribution. In the stationary case, the form of (3.2) can be additionally motivated via a projection argument as the corresponding first order conditions for obtaining the best additive fit to the data in a suitably $\widehat{\pi}$–weighted empirical L_2 norm. Smooth backfitting estimates are the best additive locally weighted least squares approximation to the data

$$(\widetilde{m}_j)_{j=0}^d \;=\; \arg\min_{f_0,\ldots,f_d} \sum_{i\in I} \int_{\mathcal{G}} \Big(Y_i - \widetilde{f}_0 - \widetilde{f}_1(x^1) - \ldots - \widetilde{f}_d(x^d)\Big)^2 K_{x,h}(X_i)\mathrm{d}x \tag{3.4}$$

under the operationalized version of the norming constraint (3.1)

$$\sum_{i=1}^n \int_{\mathcal{G}^j} \widetilde{m}_j(x^j) K_{x^j,h}(X_i^j))\,\mathrm{d}x^j = 0 \quad \text{for } j = 1,\ldots,d\,. \tag{3.5}$$

Solving (3.4) under (3.5) leads to a first order conditions of the following form

$$\sum_{i\in I} \int_{\mathcal{G}_{-j}} \big(Y_i - \widetilde{m}_0 - \widetilde{m}_1(x^1) - \ldots - \widetilde{m}_d(x^d)\big)\, K_{x,h}(X_i)\mathrm{d}x^{-j} \;=\; 0\,,$$

for each component function $\widetilde{m}_j,\ j = 1,\ldots,d$ at $x^j \in \mathring{\mathcal{G}}_j$. With this and standard kernel calculations, the backfitting equations (3.2) are easily derived. Detailed calculations are shown in Mammen et al. [1999].

In a nonstationary setting, however, generally the number of data points in a fixed bounded set is of different order for different covariates of different directions and dimensions. However, if the full–dimensional process X is β–null Harris recurrent, we can restrict the space to $\mathcal{G}^f$ and its coordinatewise projections and still have sufficiently many data points for inference. Then accordingly, all marginal and estimated objects should be constructed from the scale of corresponding full–dimensional objects as in (2.22) and (2.23). When replacing $\widehat{\pi}_j$ by $\widehat{\pi}_j^f$, $\widehat{\pi}_{jk}$ by $\widehat{\pi}_{jk}^f$ and $\widehat{m}_j$ by $\widehat{m}_j^f$, we can use standard backfitting (3.2). In this setting, the projection character remains valid as a projection on the space of functions

$$\begin{aligned}\mathcal{H}^f = \Big\{ m \in L^2_\pi(\mathcal{G}^f) |\ \exists (m_1,\ldots,m_d) \in L^2_{\pi_1^f}(\mathcal{G}_1^f) \times \ldots \times L^2_{\pi_d^f}(\mathcal{G}_d^f) :\\ m(x) = m_1(x^1) + \cdots + m_d(x^d) \text{ for all } x \in \mathcal{G}^f \Big\}\end{aligned}$$

Though it comes at a cost of neglecting a substantial amount of data not in $\mathcal{G}^f$. Furthermore, we will see, that obtained rates of convergence are slow.

3.3 Generalized Smooth Backfitting (GSBE)

If we weaken the assumption on the covariates to only pairwise pairwise β–null Harris recurrence, the class of processes admissible for estimation is substantially larger than the one of full β–null Harris recurrent processes, required for a fully nonparametric regression or standard smooth backfitting. In order to construct a general nonparametric backfitting type procedure, (3.2) shows that β–null Harris recurrence must at least hold for all two–dimensional components. While for all classes of γ–wise β–null Harris recurrent processes with $2 \leq \gamma \leq d$ a smooth backfitting procedure can be introduced, the weakest and most general setting of $\gamma = 2$, pairwise β–null Harris recurrence, is the most interesting setting and deserves the main focus.

3.3.1 Generalized Smooth Backfitting for at least Pairwise β–null Harris Recurrent Covariates

Under pairwise β–null Harris recurrence, only univariate and bivariate components have an invariant measure, higher dimensional objects are generally not recurrent anymore. In order to obtain smooth backfitting type estimates in this setting, we take a corresponding suitable adaptation of the defining integral equations (3.2) as starting point. Then the generalized smooth backfitting estimates $(\widetilde{m}_j)_{j=1}^d$ are defined as solutions to

$$\widetilde{m}_j(x^j) = \sum_{k \neq j} \left(\frac{1}{d-1} \left(\widehat{m}_j^{(k)}(x^j) - \widetilde{m}_{0,j}^{(k)} \right) - \int_{\mathcal{G}_k^{(j)}} \widetilde{m}_k(x^k) \frac{\widehat{\pi}_{j,k}(x^j, x^k)}{\widehat{\pi}_j^{(k)}(x^j)} \mathrm{d}x^k \right) \quad (3.6)$$

$$\widetilde{m}_{0,j}^{(k)} = \int_{\mathcal{G}_j} \widehat{m}_j^{(k)}(x^j) \widehat{\pi}_j^{(k)}(x^j) \mathrm{d}x^j = \frac{1}{T^{jk}(n)} \sum_{i \in I_{jk}} Y_i \quad (3.7)$$

where $\widehat{\pi}_j^{(k)}$, $\widehat{m}_j^{(k)}$ and $\widehat{\pi}_{j,k}$ have the same type of nonstationarity β^{jk} and therefore the same bandwidth and speed of convergence as defined in (2.20), (2.21) and in

the bivariate jk version of (2.16). For component j, the estimation relevant region is $\mathcal{G}^{(j)}$, the product of coordinatewise projections of relevant pairs, defined in (2.13) which might be different for each j. For any $k \neq j$ it still is $\widehat{m}_j^{(k)}(x^j) \xrightarrow{P} m_j(x^j)$ under suitable additional assumptions with speed of bivariate nonstationary type $(n^{\beta_j k}h)^{-1/2}$. The matching norming conditions in population are

$$\sum_{k \neq j} \int_{\mathcal{G}_j} m_j(x^j)\pi_j^{(k)}(x^j)\mathrm{d}x^j = 0 \,, \tag{3.8}$$

for all $j = 1, \ldots, d$. Since the estimator is constructed on the basis of pairwise data, the asymptotic results will confirm the intuition that bivariate types of nonstationarity govern the large sample behavior. Therefore speeds of convergence are significantly faster then in the standard backfitting case. But in general on the other hand, generalized smooth backfitting estimates can no longer yield the best overall additive fit as obtained from (3.4). Since recurrence is only guaranteed for only univariate and bivariate components, the projection character of SBE can only prevail in a weakened pairwise sense. This is not a deficit of the estimator but due to the difficulty of the underlying data. Therefore the underlying model must truly at least be additive in pairs of components, i.e. of the form

$$Y_i = \sum_{j \neq k} m_{jk}(X_i^{jk}) + \varepsilon_i \,, \tag{3.9}$$

for SBE to still yield a sensible approximation to the truth. Then SBE produces the best pairwise additive approximation to (3.9) for each component j by projecting orthogonally via

$$[\mu_{k|j} m_{jk}](x^j) = \int_{\mathcal{G}_k^{(j)}} m_{jk}(x^{jk}) \frac{\pi_{jk}(x^{jk})}{\pi_j^{(k)}(x^j)} \,\mathrm{d}x^k \,, \tag{3.10}$$

for $j \neq k$ on

$$\begin{aligned} \mathcal{H}_{jk} = \Big\{ m \in L^2_{\pi_{jk}}(\mathcal{G}_{jk}) |\; \exists (m_j, m_k) \in L^2_{\pi_j^{(k)}}(\mathcal{G}_j^{(k)}) \times L^2_{\pi_k^{(j)}}(\mathcal{G}_k^{(j)}) : \\ m(x) = m_j(x^j) + m_k(x^k) \text{ for all } x \in \mathcal{G}_{jk} \Big\} \end{aligned} \tag{3.11}$$

for each k, where in the case $k = j$ it is $\mathcal{H}_{jj} = \mathcal{H}_j = L^2_{\pi_j}(\mathcal{G}_j)$, and adding up the results. The corresponding system of population equations to this approximation are for $j = 1, \ldots, d$

$$\mathbb{E}(Y_i|X_i^j) \;\; = \;\; m_j(X_i^j) + \sum_{k \neq j} \mathbb{E}(m_{jk}(X_i^{jk})|X_i^j) \; . \tag{3.12}$$

Therefore, since a best fit is merely achieved in a pairwise sense (3.10), in general, correlation structures beyond pairwise correlation cannot be captured within the estimation procedure and must be regulated through additional assumptions. If all regressors are stationary, (3.6) reduces to (3.2) and the norming constraint (3.8) to (3.1). Thus standard smooth backfitting equations are a subcase of generalized smooth backfitting.

We obtain the backfitting estimates as solution to (3.6) via iteration. Start at an arbitrary initial guess $\widetilde{m}_j^{[0]}$, e.g. the Nadaraya–Watson estimator $\widetilde{m}_j^{[0]} = \widehat{m}_j$. Then denote the rth step iterate of the jth component with $\widetilde{m}_j^{[r]}$. Hence iterate according to

$$\begin{aligned} \widetilde{m}_j^{[r]}(x^j) \;\; = \;\; & \frac{1}{d-1} \sum_{k \neq j} \left(\widehat{m}_j^{(k)}(x^j) - \widetilde{m}_{0,j}^{(k)} \right) - \sum_{k<j} \int_{\mathcal{G}_k^{(j)}} \widetilde{m}_k^{[r]}(x^k) \frac{\widehat{\pi}_{j,k}(x^j, x^k)}{\widehat{\pi}_j^{(k)}(x^j)} \mathrm{d}x^k - \\ & - \sum_{k>j} \int_{\mathcal{G}_k^{(j)}} \widetilde{m}_k^{[r-1]}(x^k) \frac{\widehat{\pi}_{j,k}(x^j, x^k)}{\widehat{\pi}_j^{(k)}(x^j)} \mathrm{d}x^k \end{aligned} \tag{3.13}$$

until a convergence criterion is fulfilled. In the simulation study we employ a standard quotient condition for termination.

Note that $\sum_{k \neq j} \widetilde{m}_{0,j}^{(k)}$ is only different from zero, when the norming condition (3.8) is violated. If we set directly

$$m_0 = \sum_{j=1}^{d} \frac{1}{d-1} \sum_{k \neq j} \frac{1}{T^{jk}(n)} \sum_{i \in I_{jk}} Y_i \; , \tag{3.14}$$

an appropriate sample mean type expression, the centering term $\widetilde{m}_{0,j}^{(k)}$ can be omitted from the algorithm.

Remarks. The last equation of (3.2) is only correct on the entire support $\mathcal{G}^{(j)}$ for boundary modified kernels (2.24). If standard kernels are used instead, then the equation still yields the true solution in the interior $\mathring{\mathcal{G}}$, but on the boundary, relation (2.20) does not hold any more. Instead of using boundary modified kernels, we can also directly generalize the defining backfitting equations for standard kernels

$$\begin{aligned}\widetilde{m}_j(x^j) &= \sum_{k\neq j}\left(\frac{1}{d-1}\left(\widehat{m}_j^{(k)}(x^j)-\widetilde{m}_{0,j}^{(k)}\right)\right.\\ &\qquad\left.-\int_{\mathcal{G}_k^{(j)}}\widetilde{m}_k(x^k)\left[\frac{\widehat{\pi}_{j,k}(x^j,x^k)}{\widehat{\pi}_j^{(k)}(x^j)}-\widehat{\pi}_{k,[j+]}(x^k)\right]\mathrm{d}x^k\right)\end{aligned}$$

with

$$\widehat{\pi}_{k,[j+]}(x^k)=\frac{\int_{\mathcal{G}_j}\widehat{\pi}_{j,k}(x^j,x^k)\mathrm{d}x^j}{\int_{\mathcal{G}_j}\widehat{\pi}_j^{(k)}(x^j)\mathrm{d}x^j}\quad\text{and}\quad\widetilde{m}_{0,j}^{(k)}=\frac{\int_{\mathcal{G}_j}\widehat{m}_j^{(k)}(x^j)\widehat{\pi}_j^{(k)}(x^j)\mathrm{d}x^j}{\int_{\mathcal{G}_j}\widehat{\pi}_j^{(k)}(x^j)\mathrm{d}x^j}\ . \quad (3.15)$$

For boundary modified kernels and in the interior, the boundary adaptation $\widehat{\pi}_{k,[j+]}(x^k)$ yields zero contribution in the algorithm due to the norming constraint (3.8) and can be omitted as in (3.6).

If the compact set $\mathcal{G}^{(j)}$ is not a rectangle, relation (2.20) is not fulfilled. Therefore for generally shaped $\mathcal{G}^{(j)}$, the algorithm still works if the norming condition (3.8) is applied after each iteration step. While convergence can still be achieved, the bias behavior, however, is nonstandard and not even determined in the stationary setting. Therefore $\mathcal{G}^{(j)}$ is assumed to be rectangular throughout the paper.

3.3.2 Adapted Generalized Smooth Backfitting for at least γ–wise β–null Harris Recurrent Covariates

Although the extreme cases of full β–null Harris recurrence and pairwise β–null Harris recurrence are of main practical interest, intermediate cases of γ–wise β–null Harris recurrence provide useful insight and complete the picture. Define γ as the maximal number of components in the covariate vector X, such that all

possible permutations of γ dimensional compound component processes are still β–null Harris recurrent. Assume we have a γ–wise β–null Harris recurrent process. If the process is fully β–null Harris recurrent, we have $\gamma = d$, if it is only pairwise β–null Harris recurrent, it is $\gamma = 2$. Intermediate cases have $2 < \gamma < d$. In order to treat all these cases simultaneously, we need to introduce some notation. Set κ as multiindex in $\mathbb{R}^d$ with $\kappa_l \in \{0,1\}$ for all $l = 1,\ldots,d$ and $|\kappa| = \sum_{l=1}^{d} \kappa_l = \gamma$, indicating the dimensions involved by 1 and dimensions left out by 0, where there always are γ dimensions involved. Furthermore put

$$\lambda(\kappa) = \{l|\kappa_l = 1, l = 1,\ldots,d\} \ , \tag{3.16}$$

and $\lambda_j(\kappa) = \{l|\kappa_j = 1 \text{ and } \kappa_l = 1, l \in \{1,\ldots,d\} \setminus \{j\}\}$ and $\lambda_{jk} = \{l|\kappa_{jk} = 1 \text{ and } \kappa_l = 1, l \in \{1,\ldots,d\} \setminus \{j,k\}\}$. When projecting a function $m_k \in L^2_{\pi_k}(\mathcal{G}_k)$ on $L^2_{\pi_j}(\mathcal{G}_j)$, for nonstationary data with γ–wise β–null Harris recurrence, an orthogonal projection in the appropriate conditional expectation sense looks like $[\mu^{(\lambda_{jk})}_{k|j} m_k](x^j) = \int m_k(x^k) \frac{\pi^{(\lambda_{jk})}_{jk}(x^{jk})}{\pi^{(\lambda_j)}_j(x^j)} \mathrm{d}x^k$. Though for $\gamma > 2$ there are $\binom{d}{\gamma-2}$ different versions to construct such a projection since then $\lambda_{jk} \neq \varnothing$ for $j \neq k$. Therefore for each component function generalized smooth backfitting yields $\binom{d}{\gamma-2}$ estimates under γ–wise β–null Harris recurrence indexed by λ_{jk}

$$\begin{aligned} \widetilde{m}^{(\lambda_{jk})}_j(x^j) &= \sum_{k \neq j} \left(\frac{1}{d-1} \left(\widehat{m}^{(k\lambda_{jk})}_j(x^j) - \widetilde{m}^{(k\lambda_{jk})}_{0,j} \right) \right. \\ &\qquad \left. - \int_{\mathcal{G}^{(j\lambda_{jk})}_k} \widetilde{m}^{(\lambda_{jk})}_k(x^k) \frac{\widehat{\pi}^{(\lambda_{jk})}_{j,k}(x^j, x^k)}{\widehat{\pi}^{(k\lambda_{jk})}_j(x^j)} \mathrm{d}x^k \right) \\ \widetilde{m}^{(k\lambda_{jk})}_{0,j} &= \int_{\mathcal{G}^{(\lambda_{jk})}_j} \widehat{m}^{(k\lambda_{jk})}_j(x^j) \widehat{\pi}^{(k)}_j(x^j) \mathrm{d}x^j = \frac{1}{T^{jk\lambda_{jk}}(n)} \sum_{i \in I_{jk\lambda_{jk}}} Y_i \ . \end{aligned} \tag{3.17}$$

To obtain a single final estimate take the pointwise arithmetic mean or median of $(\widetilde{m}^{(\lambda_{jk})}_j)_{\lambda_j k}$ for each component j. The matching norming conditions in population are

$$\sum_{k \neq j} \int_{\mathcal{G}_j} m_j(x^j) \pi^{(k\lambda_{jk})}_j(x^j) \mathrm{d}x^j = 0 \ , \tag{3.18}$$

for all $j = 1, \ldots, d$. If all regressors are stationary, (3.17) reduces to (3.2) and the norming constraint (3.18) to (3.1). Thus ordinary backfitting is also a subcase of generalized backfitting under γ–wise β–null Harris recurrence .

Under γ–wise β–null Harris recurrence , the data allows to control correlation structures up to order γ within the SBE procedure as defined in (3.17). Thus if the underlying covariates show a "nicer" nonstationary behavior than in the pairwise β–null Harris recurrent case, we need less extra assumptions to regulate higher order correlation structures. Furthermore the underlying model can be more general than (3.9). In fact, (3.17) yields a sensible additive approximation for γ–wise additive functions. In population for $j = 1, \ldots, d$ this corresponds to

$$\mathbb{E}(Y_i|X_i^j) \quad = \quad m_j(X_i^j) + \sum_{k \neq j} \mathbb{E}(m_{jk\lambda_{jk}}(X_i^{jk\lambda_{jk}})|X_i^j) \; . \tag{3.19}$$

Chapter 4

Asymptotic Results

In this chapter the main asymptotic results are stated. The focus is on the practically most interesting extreme cases: smooth backfitting for a full–dimensional β–null Harris recurrent vector of covariates and generalized smooth backfitting for at least pairwise β–null Harris recurrent regressors. In a nonstationary setting, generally the difficulty of the problem with different data for different directions and dimensions will affect the result through the type of β to which the backfitting estimation technique is scaled, while keeping the univariate structure in rates and variances as in stationary smooth backfitting. From the construction of generalized smooth backfitting, it is clear that we cannot do better in terms of data and thus in type of nonstationarity than in the worst bivariate case. For standard smooth backfitting even the generally much smaller and thus less favorable full dimensional type of nonstationarity β governs procedure and results. Therefore, although rates and variances have the form as for one–dimensional marginal smoothers, generally we cannot expect full oracle behavior of the estimates as in the stationary case. For a true additive model, the asymptotic expansion of GSBE in Section 4.2 is obtained under the weakest assumptions and yields the most efficient results in an oracle sense even if covariates are γ–wise β–null Harris recurrent with $2 < \gamma \leq d$. Asymptotic results for the generalized smooth backfitting adapted to γ–wise β–null Harris recurrence are mentioned for completeness in the Section Extensions. Here we study how more information about the type of underlying data, can im-

prove efficiency of the estimation procedure. Of practical interest is in particular a tailored procedure for covariates with known stationary and nonstationary components which has nice small sample properties. For all proofs we refer to the Appendix.

4.1 Standard Smooth Backfitting for Nonstationary Covariates

In this subsection we present the base case scenario when the full process is β–null Harris recurrent. In this class of processes we can obtain asymptotic results without further restrictions on the dependence structure among the covariates or for the estimated functions in a nonstationary framework. Furthermore even if the underlying model is not additive, it yields the best additive fit. However, when requiring β–null Harris recurrence for full X, only the standard curse of dimensionality is reduced, while the nonstationary curse of dimensionality remains the same as for standard regression with non–additive m.

Assumption 4.1. *Let X be an irreducible aperiodic Markov chain, which is β–Harris recurrent of type β. The invariant measure π has a Radon–Nikodym derivative with respect to Lebesgue measure which is finite on $\mathcal{G} = \mathcal{G}^f$. The invariant density π is bounded and has continuous second partial derivatives. Furthermore π is bounded away from zero.*

Note that $\pi_f(\mathcal{G}) < \infty$ is obtained if $\mathcal{G}$ is small for X. Therefore in general, the choice of $\mathcal{G}$ is difficult, since size and form of $\mathcal{G}$ is determined by the process X. Though if X is Feller, any compact $\mathcal{G}$ is small.
Under Assumption 4.1, we can construct all relevant objects according to the full dimensional process as in (2.11),(2.12), (2.22), and (2.23). Thus in this setting, when replacing $\widehat{\pi}_j$ by $\widehat{\pi}_j^f$, $\widehat{\pi}_{jk}$ by $\widehat{\pi}_{jk}^f$ and $\widehat{m}_j$ by $\widehat{m}_j^f$, standard smooth backfitting algorithm from (3.2) does not need any modification to directly be a sequence of iterated projections. The generalized smooth backfitting reduces to standard smooth backfitting. With this procedure, however, we can only manage to reduce

the standard stationary type curse of dimensionality, whereas the nonparametric curse of dimensionality remains untouched. This is intuitively clear, when we restrict all marginal univariate and bivariate processes to the full dimensional common set of data $\mathcal{G}^f$ and all estimators have the same but full dimensional type of nonstationarity β. Therefore the speed of convergence for the smooth backfitting estimator is governed by the occupation time $\widehat{L}^f_{x^j,h}$ for the full dimensional process. Other than for stationary data, this implies that through the standardization with X we might lose a substantial amount of data for univariate and bivariate components. Therefore the procedure cannot be oracle as soon as one of the covariates is nonstationary.

For identification of the estimation problem (3.2) the dependence structure between regressors X and the residual ε must be restricted. In our general potentially nonstationary setting usual conditional independence assumptions only have a meaning with respect to an appropriate invariant measure. Furthermore to have a controllable asymptotic behavior of estimators, the compound chain (X, ε) on $\mathcal{G} \times \mathcal{G}_0$ must satisfy certain assumptions.

Assumption 4.2.

1. *The compound chain (X, ε) is a ϕ–irreducible β–null Harris recurrent Markov chain with transition probability operator P and density $\pi_{f\varepsilon}$ of the invariant measure, where $\pi^{\varepsilon}_f(x) = \int_{\mathcal{G}_0} \pi_{f\varepsilon}(x, \varepsilon)\, d\varepsilon > 0$ for all $x \in \mathcal{G}$ and $\pi^{\varepsilon}_f(\mathcal{G}) < \infty$.*

2. *$\mu_{\varepsilon|f}(x) = 0$ and $\sigma^2_{\varepsilon|f}(x) < \infty$ for all $x \in \mathcal{G}$, where both quantities are defined with respect to invariant measures $\mu_{\varepsilon|f}(x) = \int \varepsilon \frac{\pi_{f\varepsilon}(x,\varepsilon)}{\pi^{\varepsilon}_f(x)}\, d\varepsilon$ and $\sigma^2_{\varepsilon|f}(x) = \int \varepsilon^2 \frac{\pi_{f\varepsilon}(x,\varepsilon)}{\pi^{\varepsilon}_f(x)}\, d\varepsilon$.*

3. *The marginal transition probability operator P_x of X is independent of any initial distribution. And for sets $A_h \in \mathcal{B}^{\infty}(\mathbb{R}^{d+1})$ with $\lim_{h\to 0} A_h = \emptyset$ it is for the compound transition probability $\limsup_{\xi \to x} \lim_{h\to 0} \int P((\xi, \varepsilon), A_h)\, |\varepsilon|\, d\varepsilon = 0$ for all $x \in \mathcal{G}$.*

4. *ε has bounded support $\mathcal{G}_0$ and the set $\bar{\mathcal{G}} \otimes \mathcal{G}_0$ is small for (X, ε), where $\mathcal{G}$ is defined as $int_h(\bar{\mathcal{G}}) = \mathcal{G}$.*

5. *The second partial derivatives of the function m exist and are Lipschitz continuous.*

Finiteness of the measure π_f^ε on $\mathcal{G}$ in Assumption 4.2.1 implies that the asymptotic behavior of the compound process (X, ε) is dominated by the β–null structure of the X component (see Karlsen et al. [2007], Lemma 6.1.). It is $\pi_f(x) = c\, \pi_f^\varepsilon(x)$ with c constant. Thus π_f^ε also inherits differentiability properties of π^f from Assumption 4.1. In Assumption 4.2.2 the identifying conditional independence criterion is specified. All subsequent assumptions are needed to control the asymptotic behavior of the compound chain. Assumption 4.2.3 states a local uniform continuity assumption on the transition probability operator P, which allows to control and simplify the variance part in the smoothing as shown in Lemma 5.1. in Karlsen and Tjøstheim [2001]. All these assumptions 4.2.1 - 4.2.3 can be regarded as somehow natural and/or minor technical restrictions. In contrast, however, Assumptions 4.2.4 might appear artificial and technical. Smallness, however, is crucial for controlling stochastic terms of the form $f_x(X_i, \varepsilon_i) = K_{h,x}(X_i)\varepsilon_i$ for $x \in \mathcal{G}^f$ in the estimators. Under Assumption 2.1 on the smoothness of the kernel, f is in particular bounded and therefore small and thus special (see Proposition 5.13. in Nummelin [1984]). This implies

$$\sup_{y \in \mathcal{G} \times \mathcal{G}_0} \mathbb{E}_y \sum_{i=1}^{\tau} K_{h,x}(X_i)\varepsilon_i < \infty \quad \text{for all } x \in \mathcal{G}\,. \tag{4.1}$$

With Assumption 3.5 also $\widetilde{f}_x(X_i) = K_{h,x}(X_i)m(X_i)$ is special for each $x \in \mathcal{G}$ and fulfills (4.1). Note, that weakening the condition towards unbounded support and subexponential tails is not admissible, as trimming techniques in the proofs would fail. Compare that in Karlsen et al. [2007] equivalent pointwise conditions were needed to obtain central limit theorems in such a general framework.

Remark 4.1. Note that Assumptions 4.2 only require a conditional independence condition with respect to invariant measures. Thus short term dependence between residual and covariates is admissible as long as it vanishes asymptotically. This is a much weaker requirement than full independence (see Examples 6.1. and 6.2. in

Karlsen et al. [2007] for examples of asymptotically but not fully independent residuals). Thus in Econometric terms, we can identify the model under endogeneity. The problem remains well–posed as long as dependence vanishes asymptotically. This is opposed to results in the iid case, where any form of endogeneity directly leads to ill–posedness of the problem requiring regularization methods and thus a yielding severely deteriorated small sample behavior (compare Carrasco et al. [2003]).

If ε is ergodic and fully independent of X, we can omit the boundedness and smallness assumption. It is of particular interest to have a closer look at models with ε stationary, since these cases can be regarded as an additive cointegration type model. The more regularity ε offers, the more familiar turn the conditions to the stationary case. Under full independence between ε and X, the catalogue of Assumptions 4.2 simplifies to moment conditions for ε being α–mixing and the smallness condition can be avoided.

Assumption 4.2*

1. *X and ε are independent β–null Harris recurrent Markov chains*

2. *ε is ergodic strongly α–mixing with mixing rate satisfying $\sum_l l^{[2/k]\vee 1}\alpha_l < \infty$, $\mu(\varepsilon) = \int \varepsilon\pi_\varepsilon(\varepsilon)\,d\varepsilon = 0$ and $\int \varepsilon^{4(k+1)}\pi_\varepsilon(\varepsilon)\,d\varepsilon < \infty$ with $k \geq 1$.*

3. *For sets $A_h \in \mathcal{B}^\infty(\mathbb{R}^d)$ with $\lim_{h\to 0} A_h = \emptyset$ the transition probability of X fulfills $\limsup_{\xi\to x}\lim_{h\to 0} P((\xi), A_h) = 0$ for all $x \in \mathcal{G}$.*

4. *The second partial derivatives of the function m exist and are Lipschitz continuous.*

Remark 4.2. If all moments on the residual process are finite, it is sufficient if there exists a $\delta > 0$ such that $\sum_l \alpha_l^{1-\delta} < \infty$ for the mixing coefficients.

Note that in general we need the existence of at least the 8th moment. In the error term. Though if ε is strictly stationary linear, the moment conditions in Assumption 4.2* can be relaxed.

Remark 4.3. If ε is strictly stationary linear, it can be written as $\varepsilon_i = \sum_{k=0}^{\infty} a_k e_{i-k}$ with coefficients $\sum_k |a_k| < \infty$ and e strictly stationary with $\mathbb{E}e_0 = 0$, $\mathbb{E}e_0^4 < \infty$, and ϕ–mixing[1] with $\sum_l \phi_l^{1/2} < \infty$. These conditions can replace Assumption 4.2*.2. Note that these conditions are trivially fulfilled for e_i iid.

We define the Nadaraya-Watson smooth backfitting estimators $\widetilde{m}_j(x^j), j = 1, \ldots, d$ as the iterative solution of the set of equations (3.6) and the normalization (3.8). With $\widetilde{m}_0 = \frac{1}{T(n)} \sum_{i \in I} Y_i$ centering can be omitted in the algorithm. Asymptotic properties of the estimators are stated in the following theorem.

Theorem 4.1. *Let the model be additive as in* (1.1) *and Assumptions 1-3 hold. The bandwidth sequence must satisfy* $n_f^{-(\frac{\beta}{4}+\varepsilon)} \ll h \ll n_f^{-(\frac{\beta}{5}+\varepsilon)}$ *with* $\varepsilon > 0$ *arbitrarily small. Then the algorithm* (3.13) *converges with geometric rate and the smooth backfitting estimators* $\widetilde{m}_j(x^j), j = 1, \ldots, d$ *have the following asymptotic expansion*

$$\sqrt{\widehat{L}_j^f(x^j)h}\left(\widetilde{m}_j(x^j) - m_j(x^j) - B_j(x^j)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_j^2(x^j)\frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2}\right) \quad (4.2)$$

with variance

$$\sigma_j^2(x^j) = \int \varepsilon^2 \frac{\pi_{j\varepsilon}^f(x^j, \varepsilon)}{\pi_j^{f\varepsilon}(x^j)}\, d\varepsilon$$

and bias consisting of two major parts $B_j(x^j) = B_j^A(x^j) + B_j^B(x^j)$. *With* A *the backfitting operator matrix as in* (A.15)*, the bias parts* $B_j^A(x^j)$ *and* $B_j^B(x^j)$ *have the form*

$$B_j^B(x^j) = h\frac{\kappa_1(x^j)}{\kappa_0(x^j)} m_j'(x^j) - b_j + h^2 \frac{\kappa_2(x^j)}{\kappa_0(x^j)} \left((I - A)^{-1}\bar{B}\right)_{(j)}(x^j) \quad (4.3)$$

$$B_j^A(x^j) = \mu_{(j\varepsilon)}^f \left(K_{x^j,h}(X) \otimes \frac{T^{f\varepsilon}(n)}{\widehat{L}^f(x^j)}\right). \quad (4.4)$$

Under the stated choice of bandwidth the asymptotic bias B_j^A *vanishes. For the deterministic bias* B_j^B *it is* $b_j = o_P(h^2)$ *in* $\mathcal{G}$ *and* $\bar{B}(x) = (\bar{b}_1(x^1), \ldots, \bar{b}_d(x^d))^T$ *with*

[1] See Hall and Heyde [1980], page 277 for an exact definition of ϕ–mixing

component functions $\bar{b}_j$ *for* $j \in \{1, \ldots, d\}$ *defined by*

$$\bar{b}_j(x^j) \;=\; \frac{\kappa_2(x^j)}{\kappa_0(x^j)} \left(\frac{1}{2} m_j''(x^j) + \frac{m_j'(x^j)}{\pi(x)} \, \frac{\partial \pi(x)}{\partial x^j} \right)$$

Remarks. 1. The vector of component estimates $(\widetilde{m}_1^A(x^1), \ldots, \widetilde{m}_d^A(x^d))$ converges even jointly to the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \operatorname{diag}(\sigma_1(x^1), \ldots, \sigma_d(x^d)))$ with $\mathbf{0} \in \mathbb{R}^d$. Covariances vanish asymptotically.

2. The unusual restriction from above on the bandwidth is due to the nonstationarity in the data. With this and Assumption 3.3. about the second derivatives, the stochastic bias B_j^A vanishes (See Karlsen and Tjøstheim [2001] Theorem 5.3. and Karlsen et al. [2007] Theorem 5.5.). Furthermore under the upper bound on the bandwidth it is $\sqrt{\widehat{L}_j^f(x^j)h}\; B_j^B(x^j) = o_P(1)$ for all x^j, causing the bias to disappear (See Karlsen and Tjøstheim [2001], proof of Theorem 3.5). Thus the speed of convergence in Theorem 4.1 can be arbitrarily close to but never attains $n^{-\frac{2}{5}\beta}$ under the stated choice of bandwidth.

3. The result holds more generally for a model with transformed error term $g_\varepsilon(\varepsilon)$ when replacing ε in Assumption 4.2 by $g_\varepsilon(\varepsilon)$ and g_ε is bounded and $L^1(\mathbb{R}^+)$. Then Theorem 4.1 holds with asymptotic bias as $B_j^A(x^j) = \mu_{(j\varepsilon)}(K_{x^j,h}(X) \otimes g_\varepsilon(\varepsilon)) \frac{T^{f\varepsilon}(n)}{\widehat{L}^f(x^j)}$ and variance $\sigma_j^2(x^j) = \int g_\varepsilon(\varepsilon)^2 \frac{\pi_{j\varepsilon}(x^j,\varepsilon)}{\pi_j^\varepsilon(x^j)} \, \mathrm{d}\varepsilon$. With considerations in Karlsen et al. [2007], Section 6.4., and Mammen and Nielsen [2003], results can be even further extended to models with heteroscedastic error terms.

4. In this setting, the underlying model does not have to be additive. Even if it is not, $(\widetilde{m}_j)_{j=0}^d$ are optimal in the sense of an additive projection on L_π^2 the best additive fit.

5. In the special case of one nonstationary regressor and all other regressors stationary, the nonstationary rate dominates in all component functions.

The marginal variance $\sigma_j^2(x^j)$ of the jth additive component is exactly the variance of the one–dimensional smoother. It can also be regarded as the projection of σ_ε^2, the variance of the iid-parts in the ε split chain, onto X^j in the following sense: $\mathbb{E}\left[\sigma_\varepsilon^2 | X^j = x^j\right]$.

As in the stationary case, the deterministic bias B_j^B of the Nadaraya–Watson type smooth backfitting estimator consists of three main parts. In addition to the marginal Nadaraya–Watson bias $h\frac{\kappa_1(x^j)}{\kappa_0(x^j)}m_j'(x^j) + \frac{1}{2}h^2\frac{\kappa_2(x^j)}{\kappa_0(x^j)}m_j''(x^j)$ there is the constant shift b_j from a difference in population centering as in (3.8) and centering with the population counterpart. Furthermore there exists a design density dependent part $\left((I-A)^{-1}\bar{B}\right)_{(j)}(x^j)$. The term b_j converges to zero asymptotically since it holds that

$$\begin{aligned}
&\iint m_j(x^j)\pi_j(u)K_{h,x^j}(u)\,\mathrm{d}u\,\mathrm{d}x^j \\
&= \int_{u^j \in \partial\mathcal{G}_{hC_1}^j} \int m_j(x^j)\pi_j(x^j)(K_h(x^j,u^j) - K_h(x^j-u^j))\,\mathrm{d}x^j\,\mathrm{d}u^j + O(h^2) = O(h)\,.
\end{aligned}$$

and the second term is of order $O(h^2)$ because $\kappa_1(x^j)$ is zero at interior points x^j. To stress the projection character of $\widetilde{b}_j(x^j) = ((I-A)^{-1}\bar{B})_{(j)}(x^j)$ with $\iota = (1,\ldots,1) \in \mathbb{R}^{1\times d}$ it is:

$$(\widetilde{b}_0, \widetilde{b}_1(x^1), \ldots, \widetilde{b}_d(x^d)) = \arg\min_{b_1,\ldots,b_d} \int (\iota\bar{B}(x) - b_0 - b_1(x^1) - \cdots - b_d(x^d))^2\pi(x)\,\mathrm{d}x\,.$$

The explicit form of the projection part in the deterministic bias $B_j^B(x^j)$ is

$$\begin{aligned}
\left((I-A)\bar{B}\right)_{(j)}(x^j) &= \frac{1}{2}m_j''(x^j) + \frac{m_j'(x^j)}{\pi_j(x^j)}\,\frac{\partial\pi_j(x^j)}{\partial x^j} \\
&\quad + \sum_{k\neq j}\int_{\mathcal{G}^k}\left(\frac{1}{2}m''(x^k) + \frac{m_j'(x^j)}{\pi_{jk}(x^{jk})}\,\frac{\partial\pi_{jk}(x^{jk})}{\partial x^j}\right)\frac{\pi_{jk}(x^{jk})}{\pi_j(x^j)}\,\mathrm{d}x^k\,.
\end{aligned}$$

In total, the deterministic bias B_j^B is of order $o(h^2)$ in the interior of the estimated compact set as in the stationary case. In the stated form above this is not obvious at first glance. But by symmetry of the kernel, $\kappa_1(x^j)$ is zero for $x^j \in \mathring{\mathcal{G}}_{j,h}$. Therefore all $O(h)$ terms vanish in the interior.

It can be shown that $h^2 b_{j,n} = O(1)$. This is a consequence of the centering constraint in the algorithm, which causes lower order terms to be zero. This term is constant over x^j and does therefore only affect the level and not the shape of the estimator. It originates from the fact that the empirical version of the normalization (3.8) used in the algorithm and the actual theoretical normalization (3.8) are different in finite samples and only asymptotically equivalent.

The stochastic bias B_j^A vanishes with $o_P(h^2)$ under the stated bandwidth assumptions (See Karlsen et al. [2007]). With stronger independence assumptions on the error term as below, it can be omitted.

If we enforce the independence assumption between X^j and ε, from conditional independence to full independence, we can avoid the artificial boundedness and small set assumption and have more familiar moment conditions. If in addition ε is assumed to be stationary, also the variance is no longer only a second moment with respect to an invariant measure but with respect to the stationary density of ε, hence a "real" variance.

Theorem 4.2. *Let the same set of assumptions as in Theorem 4.1 hold, but replace Assumptions 4.2 by Assumptions 4.2*. Choose the bandwidth as* $n_f^{-(\frac{\beta}{4}+\varepsilon)} \ll h \ll n_f^{-(\frac{\beta}{5}+\varepsilon)}$ *with* $\varepsilon > 0$ *arbitrarily small. Then the algorithm* (3.13) *converges and we get the following asymptotic expansion for the smooth backfitting estimates* $(\widetilde{m}_j)_{j=1}^d$

$$\sqrt{\widehat{L}_j^f(x^j)h}\left(\widetilde{m}_j(x^j) - m_j(x^j) - B_j^B(x^j)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_j^2(x^j)\frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2}\right) \quad (4.5)$$

with deterministic bias as in 4.3 and variance

$$\sigma_j^2(x^j) \;=\; \sigma_j^2 = \int \varepsilon^2 \pi_\varepsilon(\varepsilon)\, d\varepsilon \;,$$

where π_ε *is the stationary density of* ε.

Here the stochastic bias B_j^A as in (4.4) is zero under Assumptions 4.2*. The same result also holds under milder moment conditions as in Assumption 4.2* as specified in Remark 4.3.

4.2 GSBE

In the subsection before, we managed to reduce the stationary curse of dimensionality via standard smooth backfitting for the additive model while the nonstationary curse remained untouched. Therefore the resulting speed of convergence is quite slow governed by the full dimensional β. The final version of the generalized backfitting estimation procedure (3.6), however, contains just one–and two dimensional marginal components. Thus under some mild additional assumptions, we cannot only increase the speed of convergence to be of two–dimensional nonstationary β type, but even more importantly increase the treatable class of processes substantially from full β–null Harris recurrent to just pairwise β–null Harris recurrent processes. If the underlying model is additive, we derive the asymptotic results. Naturally, all assumptions are restrictions on pairwise components only.

Assumption 4.3. *Let X be an irreducible aperiodic Markov chain. All univariate possible pairs of bivariate marginal processes X^{jk} are β–null Harris recurrent with parameter β^{jk}. Their respective invariant measure has a Radon–Nikodym derivative π_{jk} with respect to Lebesgue measure. Any bivariate invariant density π_{jk} is bounded and has continuous second partial derivatives on $\mathcal{G}_{jk}$. Furthermore any π_{jk} is bounded away from zero.*

Finiteness of bivariate and univariate invariant measures is achieved, if $\mathcal{G}_{jk}$ and $\mathcal{G}_j$ are small. Then for any component j we choose the space composed of pairwise coordinate projections $\mathcal{G}^{(j)}$ as defined in (3.11) as space of estimation.

Please note that theses assumptions do not restrict the pairwise correlation structure of the covariates. Pairwise β–Harris recurrence does not rule out or restrain dependence of regressors. Furthermore, pairwise β–Harris recurrence is truly weaker than full dimensional β–Harris recurrence. For example, a d–dimensional vector of random walks is pairwise β–Harris recurrent independent of the correlation structure between the covariates - while it is not fully β–Harris recurrent for $d \geq 2$ unless covariates are substantially correlated.

Identification and asymptotic expansion of generalized smooth backfitting estimates (3.6) can be obtained by the following assumptions.

Assumption 4.4. *For any bivariate marginal process X^{jk} with $j,k=1,\ldots,d$ we assume:*

1. *The compound chain (X^{jk},ε) is a ϕ–irreducible Harris recurrent Markov chain with transition probability operator $P_{jk\varepsilon}$ and density $\pi_{jk\varepsilon}$ of the invariant measure, where $\pi^{\varepsilon}_{jk}(x^{jk})=\int_{\mathcal{G}_0}\pi_{jk,\varepsilon}(x^{jk},\varepsilon)\,d\varepsilon>0$ for all $x^{jk}\in\mathcal{G}_{jk}$ and $\pi^{\varepsilon}_{jk}(\mathcal{G}^{jk})<\infty$*

2. *$\mu_{\varepsilon|jk}(x^{jk})=0$ and $\sigma^2_{\varepsilon|jk}(x^{jk})<\infty$ for all $x^{jk}\in\mathcal{G}_{jk}$ where both quantities are defined with respect to invariant measures $\mu_{\varepsilon|jk}(x^{jk})=\int\varepsilon\frac{\pi_{jk\varepsilon}(x^{jk},\varepsilon)}{\pi^{\varepsilon}_{jk}(x^{jk})}\,d\varepsilon$ and $\sigma^2_{\varepsilon|jk}(x^{jk})=\int\varepsilon^2\frac{\pi_{jk\varepsilon}(x^{jk},\varepsilon)}{\pi^{\varepsilon}_{jk}(x^{jk})}\,d\varepsilon$.*

3. *The marginal transition function P_{jk} is independent of any initial distribution. And for sets $A_h\in\mathcal{B}^{\infty}(\mathbb{R}^3)$ with $\lim_{h\to0}A_h=\emptyset$ it is for the compound transition probability: $\lim_{h\to0}\limsup_{\xi\to x^{jk}}\int P((\xi,\varepsilon),A_h)\,|\varepsilon|\,d\varepsilon=0$ for all $x^{jk}\in\mathcal{G}_{jk}$.*

4. *ε has bounded support $\mathcal{G}_0$ and the set $\bar{\mathcal{G}}_{jk}\otimes\mathcal{G}_0$ is small for (X^{jk},ε), where $int_h\left(\bar{\mathcal{G}}_{jk}\right)=\mathcal{G}_{jk}$.*

5. *The second partial derivatives of the function m exist and are Lipschitz continuous.*

The same remarks as for Assumption 4.2 apply. In particular Remark 4.1 about well–posedness of the problem under short–run endogeneity also applies. Furthermore it is also in this general setting of particular interest to have a closer look at models with ε stationary, since these cases can be regarded as an additive cointegration type model. If ε is α–mixing and independent of each pair of covariates, identification requirements simplify to moment conditions.

Assumption 4.4*

For every bivariate marginal process X^{jk} we have:

1. *X^{jk} and ε are independent Harris recurrent Markov chains*

2. *ε is ergodic strongly α–mixing with mixing rate satisfying $\sum_l l^{[2/k]\vee 1}\alpha_l < \infty$, $\mu(\varepsilon) = \int \varepsilon\pi_\varepsilon(\varepsilon)\, d\varepsilon = 0$ and $\int \varepsilon^{4(k+1)}\pi_\varepsilon(\varepsilon)\, d\varepsilon < \infty$ with $k \geq 1$.*

3. *For sets $A_h \in \mathcal{B}^\infty(\mathbb{R}^2)$ with $\lim_{h\to 0} A_h = \emptyset$ the transition probability of X^{jk} fulfills $\limsup_{\xi\to x}\lim_{h\to 0} P((\xi), A_h) = 0$ for all $x \in \mathcal{G}^{jk}$.*

4. *The second partial derivatives of the function m exist and are Lipschitz continuous.*

If moments of all order exist Remark 4.2 applies also here. Note that in general we need the existence of at least the 8th moment in the error term. Though if ε is strictly stationary linear, the moment conditions in Assumption 4.4* can be relaxed.

Remark 4.4. If ε is strictly stationary linear, it can be written as $\varepsilon_i = \sum_{k=0}^{\infty} a_k e_{i-k}$ with coefficients $\sum_k |a_k| < \infty$ and e strictly stationary with $\mathbb{E}e_0 = 0$, $\mathbb{E}e_0^4 < \infty$, and ϕ–mixing[2] with $\sum_l \phi_l^{1/2} < \infty$. These conditions can replace Assumption 4.4*.2. These conditions are trivially fulfilled for e_i iid.

Though in contrast to the full dimensional β–null Harris recurrent case, we need an additional assumption for controlling the bias term in order to reach consistency of the estimation procedure.

Assumption 4.5. *Assume that for all X_i^{jl} there exists a X_i^{jk} such that $m_k(X_i^k) = m_l(X_i^l)$ for $i \in I_{jk}$ and $j \neq k \neq l$.*

This implies that on the domain $\mathcal{G}^{(j)}$ all component functions must have the same range. Implicitly, therefore the size of $\mathcal{G}_k^{(j)}$ can be restricted to fit Assumption 4.5. If pairs of covariates are correlated and are close in type of nonstationarity the requirement is mild. If the model entirely consists of trigonometric component functions as in the simulation study, Assumption 4.5 is trivially fulfilled. Alternatively to Assumption 4.5 in some case as e.g. it might advisible to impose a parametric restriction on the component functions outside small sets. See Figure

[2]See Hall and Heyde [1980], page 277 for an exact definition of ϕ–mixing

A.1 in the Appendix for a discussion and Subsection A.2.3 for a discussion with technical details.

For each component function m_j it will be the worst case bivariate nonstationary type jk_0 dominating the asymptotic behavior. Therefore set $\beta^{j+} = \beta^{jk_0} = \min_{k \neq j} \beta^{jk}$, $n_{j+} = n_{jk_0}$ as defined in Section 2.2 and the respective bandwidth h_{j+} for all $j = 1, \ldots, d$. Then we get the following closed form expansion.

Theorem 4.3. *Let the model be additive as in* (1.1) *fulfilling the centering condition* (3.8) *and let Assumptions 2.1, 4.3, 4.4, and 4.5 hold. The bandwidth sequence must satisfy* $n_{j+}^{-(\frac{\beta^{j+}}{4}+\varepsilon)} \ll h_{j+} \ll n_{j+}^{-(\frac{\beta^{j+}}{5}+\varepsilon)}$ *for* $\varepsilon > 0$ *arbitrarily small. Then the algorithm* (3.13) *converges with geometric rate and for the estimators* $\widetilde{m}_j(x^j), j = 1, \ldots, d$ *we find*

$$\sqrt{\widehat{L}_j^{(k_0)}(x^j) h_{j+}} \left(\widetilde{m}_j(x^j) - m_j(x^j) - B_j(x^j) \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_{j+}^2(x^j) \frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2}\right) \tag{4.6}$$

with variance

$$\sigma_{j+}^2(x^j) \;=\; \int \varepsilon^2 \frac{\pi_{j\varepsilon}^{k_0}(x^j, \varepsilon)}{\pi_j^{(k_0)}(x^j)} \, d\varepsilon$$

and bias consisting of two major parts $B_j(x^j) = B_j^A(x^j) + B_j^B(x^j)$ *with*

$$B_j^B(x^j) \;=\; h_{j+} \frac{\kappa_1(x^j)}{\kappa_0(x^j)} m_j'(x^j) + \frac{1}{2} h_{j+}^2 \frac{\kappa_2(x^j)}{\kappa_0(x^j)} m_j''(x^j) + \left((I - A)^{-1} \bar{B} \right)_{(j)} (x^j) - b_{j,n} \tag{4.7}$$

$$B_j^A(x^j) \;=\; \mu_{((j+)\varepsilon)}(K_{x^j,h}(X) \otimes id_\varepsilon) \frac{T^{(j+)\varepsilon}(n)}{\widehat{L}^{j+}(x^j)} \tag{4.8}$$

Under the stated choice of bandwidth the asymptotic bias B_j^A *vanishes. For the deterministic bias* B_j^B, A *is the backfitting operator matrix as in* (A.15) *and* $\bar{B}(x) = (\bar{b}_1(x^1), \ldots, \bar{b}_d(x^d))^T$ *and component functions* $\bar{b}_j$ *for* $j \in \{1, \ldots, d\}$ *defined by*

$$\bar{b}_j(x^j) \;=\; h_{j+}^2 \left[\left(b_j + \sum_{k \neq j} \int_{\mathcal{G}_k^{(j)}} b_{jk}(x^k) + \frac{\pi_{jk}(x^k)}{\pi_j^{(k)}} \mathrm{d}x^k \right) \right] (x^j) \, .$$

The exact form of these bias components is given right below. Most importantly it is $\bar{b}_j = O\left(h_{j+}^2\right)$. *Furthermore the centering constant* $b_{j,n}$ *is given by* $b_{j,n} =$

$\mu_{(j)}\left(\widehat{\Phi}_j \widehat{m}_j\right)$ *where the centering operator is defined in* (A.15)*, and it is* $h_{j+}^2 b_{j,n} = O(1)$.

The exact form of the bias is

$$\begin{aligned} b_j(x^j) &= \frac{\kappa_2(x^j)}{\kappa_0(x^j)} \left(\frac{m_j'(x^j)}{\pi_j^{(k)}(x^j)} \pi_j^{(k)\prime}(x^j) \right) \\ b_{jk}(x^{jk}) &= \frac{\kappa_2(x^j)}{\kappa_0(x^j)} \left(\frac{m_k'(x^k)}{\pi_{jk}(x^{jk})} \frac{\partial \pi_{jk}(x^{jk})}{\partial x^k} \right) \end{aligned}$$

Remark 4.5.

1. Here the underlying model must at least be pairwise additive as in (3.9).

2. The stated result also holds with a slight modification in some constants under some milder assumption than Assumption 6. If we assume instead that $\int_{A_{jk}} m_l(x^l)\pi_l(x^l)\,\mathrm{d}x^l < \infty$ for $j \neq k \neq l$ where the area of integration A_{jk} is defined by the correlation between components (X^{jk}) and X^l. Details are contained in the proof in the appendix.

3. The vector of component estimates $(\widetilde{m}_1^A(x^1), \ldots, \widetilde{m}_d^A(x^d))$ converges jointly to the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathrm{diag}(\sigma_{1+}(x^1), \ldots, \sigma_{d+}(x^d)))$ with $\mathbf{0} \in \mathbb{R}^d$. Covariances vanish asymptotically.

4. The result also holds more generally for a model with transformed error term $g_\varepsilon(\varepsilon)$ when replacing ε in Assumption 5.3 by $g_\varepsilon(\varepsilon)$. Then Theorem 4.3 holds with modified asymptotic bias $B_j^A(x^j)$ and variance $\sigma_j(x^j)$ as described in remark 3 after Theorem 4.1.

5. In the special case of one nonstationary regressor and all other regressors stationary, the nonstationary rate dominates in all component functions. Since such a process can be easily shown to also be fully β–null Harris recurrent, we can proceed with the easier algorithm as in the subsection before and reach the same result.

The marginal variance $\sigma_{j+}^2(x^j)$ of the jth additive component is in form exactly the variance of the one–dimensional smoother. Though the rate of convergence

is of univariate character in its form but governed by the worst case bivariate nonstationarity type β^{j+} for each component function.

As in the stationary case, the deterministic bias B_j^B of the Nadaraya–Watson smooth backfitting estimator consists of three main parts. In addition to the marginal Nadaraya–Watson bias $h_{j+}\frac{\kappa_1(x^j)}{\kappa_0(x^j)}m_j'(x^j) + \frac{1}{2}h_{j+}^2\frac{\kappa_2(x^j)}{\kappa_0(x^j)}m_j''(x^j)$ there is the constant shift $b_{j,n}$ from norming and centering and the design density dependent part $\left((I-A)^{-1}\bar{B}\right)_{(j)}(x^j)$. We pay a price for nonstationarity as the entire deterministic bias B_j^B is of order $o(h_{j+}^2)$ in the interior of $\mathcal{G}_j$ instead of the significantly faster $o(h_j^2)$ in a stationary setting. Furthermore as generally no full dimensional π exists as in the previous subsection, the design dependent deterministic bias part resembles its counterpart in the full–dimensional β–null Harris recurrent case in form, but lacks its projection interpretation. Furthermore, it can be shown that $h_{j+}^2 b_{j,n} = O(1)$. As before, this is a consequence of the centering constraint in the algorithm, which causes lower order terms to be zero. This term is constant over x^j and does therefore only affect the level and not the shape of the estimator. It originates from the fact that the empirical version of the normalization used in the algorithm and the actual theoretical normalization (3.8) are different in finite samples and only asymptotically equivalent.

The stochastic bias B_j^A vanishes with $o_P(h_{j+}^2)$ under the stated bandwidth assumptions (See Karlsen et al. [2007]). With stronger independence assumptions on the error term as below, it can be omitted. If we enforce the independence assumption between X^j and ε, from conditional independence to full independence, we can avoid the artificial boundedness and small set assumption and have more familiar moment conditions. If in addition ε is assumed to be stationary, also the variance is no longer only a second moment with respect to an invariant measure but with respect to the stationary density of ε, hence a "real" variance.

Theorem 4.4. *Let the same set of assumptions as in Theorem 4.3 hold, but replace Assumptions 4.4 by Assumptions 4.4*. Choose the bandwidth as* $n_{j+}^{-(\frac{\beta^{j+}}{4}+\varepsilon)} \ll h_{j+} \ll n_{j+}^{-(\frac{\beta^{j+}}{5}+\varepsilon)}$ *for* $\varepsilon > 0$ *arbitrarily small. Then the algorithm converges and we*

get the following asymptotic expansion for the smooth backfitting estimates $(\widetilde{m}_j)_{j=1}^d$

$$\sqrt{\widehat{L}_j^{(k_0)}(x^j)h_{j+}}\left(\widetilde{m}_j(x^j) - m_j(x^j) - B_j^B(x^j)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_j^2(x^j)\frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2}\right) \quad (4.9)$$

with deterministic bias as in 4.7 and simplified variance

$$\sigma_j^2(x^j) \;=\; \sigma_j^2 = \int \varepsilon^2 \pi_\varepsilon(\varepsilon)\, d\varepsilon \;,$$

where π_ε *is the stationary density of* ε*.*

In this setting the stochastic bias B_j^A is zero. The same result also holds under milder moment conditions as in Assumption 4.4* as specified in Remark 4.4.

4.3 Extensions

If we have more knowledge about the nonstationarity character of the covariates, a more tailored method yields better results.

4.3.1 Adapted GSBE to γ–wise β–null Harris Recurrence

In Assumptions 4.3, 4.4 and 4.4*, replace the pair of components j, k by a set of components λ as defined in (3.16), indicating the γ jointly β–null Harris recurrent covariates, and set (λ_j) as superscript for (k). Assume that this modified set of assumptions holds and that generalized smooth backfitting is conducted according to (3.17). Then the additional assumption for controlling the bias is weaker than in the pairwise β–null Harris recurrent case with Assumption 4.5. Assume that λ_j is fixed for estimation in (3.17), then set.

Assumption 4.6. *Assume that for all* X_i^{jl} *with* $l \notin \lambda_j$ *there exists a* X_i^{jk} *with* $k \in \lambda_j$ *such that* $m_k(X_i^k) = m_l(X_i^l)$ *for* $i \in I_{j\lambda_j}$*.*

This is fulfilled if on $\bigotimes_{k=1}^d \left(\mathcal{G}_k^{(j,\lambda_j)} \cup \bigcup_{l\neq k} A_l^{(jk\lambda_{jk})}\right)$ all component functions have the same range. Here $A_l^{(jk\lambda_{jk})}$ are "outside" of $\mathcal{G}^{(\lambda)}$ and appear due to nonmatching

index sets in the bias expansions. See Figure A.1 in the Appendix for an illustration. An exact definition can be found in Section A.2.3. For each pair (j,k) with $k \in \lambda_j$, the set $\mathcal{A}_k^{(j)} := \bigcup_{l \neq k} A_l^{(jk\lambda_{jk})}$ contains all X_i^l where $l \notin \lambda_j$ with the "wrong" index set $i \in I_{j\lambda_j}$, which are no longer within the small set $\mathcal{G}_l^{(j\lambda_j)}$. Implicitly, the size of $\mathcal{G}_k^{(j\lambda_j)}$ can be restricted to fit Assumption 4.6 via assumptions on the component functions only. If pairs of covariates, however, are correlated and are close in type of nonstationarity, the requirement on the component functions is mild, as X_i^l with $l \notin \lambda_j$ and $i \in I_{j\lambda_j}$ is more likely to be within the small set $\mathcal{G}_l^{(j\lambda_j)}$ and $\mathcal{A}_k^{(j)}$ is comparatively small. If the model entirely consists of trigonometric component function as in the simulation study, Assumption 4.6 is trivially fulfilled.

For each component function m_j, it will be the worst case γ–wise nonstationary type of covariates $j\lambda_{j,0}$ dominating the asymptotic behavior. Therefore set $\beta^{j+} = \min_{\lambda_j} \beta^{j\lambda_j}$, $n_{j+} = n_{j\lambda_{j,0}}$, and the respective bandwidth h_{j+} for all $j = 1, \ldots, d$. Then we get the following closed form expansion.

Theorem 4.5. *Let the model be additive as in* (1.1) *fulfilling the centering condition* (3.18) *and let Assumptions 2.1 and 4.6, and modifications of Assumptions 4.3 and 4.4 as described right above hold. The bandwidth sequence must satisfy* $n_{j+}^{-(\frac{\beta^{j+}}{4}+\varepsilon)} \ll h_{j+} \ll n_{j+}^{-(\frac{\beta^{j+}}{5}+\varepsilon)}$ *for* $\varepsilon > 0$ *arbitrarily small. Then the algorithm* (3.13) *converges with geometric rate and for the estimators* $\widetilde{m}_j^{NW}(x^j), j = 1, \ldots, d$ *we find*

$$\sqrt{\widehat{L}_j^{(\lambda_{j,0})}(x^j) h_{j+}} \left(\widetilde{m}_j(x^j) - m_j(x^j) - B_j(x^j)\right) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\left(0, \sigma_{j+}^2(x^j) \frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2}\right)$$

with variance

$$\sigma_{j+}^2(x^j) \quad = \quad \int \varepsilon^2 \frac{\pi_{j\varepsilon}^{\lambda_{j,0}}(x^j, \varepsilon)}{\pi_j^{\lambda_{j,0}}(x^j)} \, d\varepsilon$$

and bias consisting of two major parts $B_j(x^j) = B_j^A(x^j) + B_j^B(x^j)$ *with*

$$\begin{aligned} B_j^B(x^j) &= h_{j+} \frac{\kappa_1(x^j)}{\kappa_0(x^j)} m_j'(x^j) + \frac{1}{2} h_{j+}^2 \frac{\kappa_2(x^j)}{\kappa_0(x^j)} m_j''(x^j) + \left((I - A)^{-1} \bar{B}\right)_{(j)} (x^j) - b_{j,n} \\ B_j^A(x^j) &= \mu_{((j+)\varepsilon)}(K_{x^j,h}(X) \otimes id_\varepsilon) \frac{T^{(j+)\varepsilon}(n)}{\widehat{L}^{j+}(x^j)} \, . \end{aligned}$$

Under the stated choice of bandwidth the asymptotic bias B_j^A vanishes. For the deterministic bias B_j^B, A is the backfitting operator matrix as in (A.15) *and $\bar{B}(x) = (\bar{b}_1^{(\lambda_1)}(x^1), \ldots, \bar{b}_d^{(\lambda_d)}(x^d))^T$ and component functions $\bar{b}_j^{(\lambda_j)}$ for $j \in \{1, \ldots, d\}$ defined by*

$$\bar{b}_j^{(\lambda_j)}(x^j) \;=\; h_{j+}^2 \left[\left(b_j^{(\lambda_j)} + \sum_{k \neq j} \int_{\mathcal{G}_k^{(j\lambda_{jk})}} b_{jk}^{(\lambda_{jk})}(x^k) + \frac{\pi_{jk}^{(\lambda_{jk})}(x^k)}{\pi_j^{(\lambda_j)}} \mathrm{d}x^k \right) \right] (x^j) \; .$$

The exact form of these bias components is given right below. Most importantly it is $\bar{b}_j^{(\lambda_j)} = O\left(h_{j+}^2\right)$. Furthermore the centering constant $b_{j,n}$ is given by $b_{j,n} = \mu_{(j)}^{(\lambda_j)}\left(\widehat{\Phi}_j^{(\lambda_j)} \widehat{m}_j^{(\lambda_j)}\right)$ where the centering operator is defined in (A.15)*, and it is $h_{j+}^2 b_{j,n} = O(1)$.*

The exact form of the bias is

$$\begin{aligned} b_j^{(\lambda_j)}(x^j) &= \frac{\kappa_2(x^j)}{\kappa_0(x^j)} \left(\frac{m_j'(x^j)}{\pi_j^{(\lambda_j)}(x^j)} \pi_j^{(\lambda_j)\prime}(x^j) \right) \\ b_{jk}^{(\lambda_{jk})}(x^{jk}) &= \frac{\kappa_2(x^j)}{\kappa_0(x^j)} \left(\frac{m_k'(x^k)}{\pi_{jk}^{(\lambda_{jk})}(x^{jk})} \frac{\partial \pi_{jk}^{(\lambda_{jk})}(x^{jk})}{\partial x^k} \right) \; . \end{aligned}$$

Remark 4.6. 1. Remarks 3 and 4 after Theorem 4.3 apply in the same fashion.

2. Here the underlying model must at least be γ–wise additive. Then SBE delivers the best additive fit in the sense of (3.19)

3. The stated result also holds with a slight modification in some constants under some milder assumption than Assumption 4.6. Assume instead that $\int_{A_l^{(jk\lambda_{jk})}} m_l(x^l) \pi_l^{(jl\lambda_j)}(x^l) \, \mathrm{d}x^l < \infty$ for all $l \notin \lambda_j$. This can be achieved if m_l is special on $A_l^{(jk\lambda_{jk})}$, which is e.g. the case for $A_l^{(jk\lambda_{jk})}$ small for X^l. Details are contained in the proof in the appendix.

Structurally the same comments apply as for Theorem 4.3 but in a λ–modified version. As before, if ε is stationary independent to X^λ we get a simplified version of Corollary 4.5.

Theorem 4.6. *Let the same set of assumptions as in Corollary 4.5 hold, but replace Assumptions 4.4 by λ modified Assumptions 4.4*. The bandwidth sequence must satisfy $n_{j+}^{-(\frac{\beta^{j+}}{4}+\varepsilon)} \ll h_{j+} \ll n_{j+}^{-(\frac{\beta^{j+}}{5}+\varepsilon)}$ for $\varepsilon > 0$ arbitrarily small.. Then the algorithm converges and we get the following asymptotic expansion for the smooth backfitting estimates $(\widetilde{m}_j)_{j=1}^d$*

$$\sqrt{\widehat{L}_j^{(\lambda_j)}(x^j)h_{j+}}\left(\widetilde{m}_j(x^j) - m_j(x^j) - B_j^B(x^j)\right) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\left(0, \sigma_{j+}^2(x^j)\frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2}\right)$$

with deterministic bias as in Corollary 4.5 and simplified variance

$$\sigma_{j+}^2(x^j) \quad = \quad \int \varepsilon^2 \pi_\varepsilon(\varepsilon)\, d\varepsilon \;,$$

where π_ε is the stationary density of ε.

4.3.2 Asymptotic Independence – Stationary and Nonstationary Covariates

For stationary data, if covariates are independent, an (additive) regression can be separated, i.e. regressing each covariate separately yields the same marginal result as a joint regression. In a β–null Harris recurrent setting, identification and (generalized) smooth backfitting estimates are obtained in terms of invariant measures. Therefore the appropriate notion of independence in this context should also be with respect to invariant measures. For two β–null Harris recurrent processes, we define weak asymptotic independence as follows.

Definition 4.1 (Weak Asymptotic Independence)**.** Suppose X^{jk} is β–null Harris recurrent . Then X^j and X^k are asymptotically independent if the joint invariant measure factors into the product of the two marginal projections of the joint measure.

$$\pi_{jk} = c_1 \cdot \pi_j^{(k)} \otimes \pi_k^{(j)} \;, \tag{4.10}$$

where $c_1 > 0$ is a constant.

Weak asymptotic independence requires independence between components only on the long run, while in the short term there might be dependence. It is therefore more general than strict independence in every time point.

If one part of the covariates , w.l.o.g. take X^{d_1}, is asymptotically independent of the others X^{d_2} with $X = (X^{d_1}, X^{d_2})$, SBE estimation can be conducted separately. Other than in a stationary setting, using this information for nonstationary data can imply a significant improvement in speed for the estimation of all component functions. In this case, the backfitting projection operators in (3.2) and (3.6) separate for X^{d_1} and X^{d_2}. Therefore estimation can be conducted completely separate the first d_1 components and the second d_2 components according to standard or generalized SBE. Then results for standard SBE in Theorems 4.1 and 4.2 require only X^{d_1} and X^{d_2} to be β–null Harris recurrent with β^{d_1} or β^{d_2} respectively and mixed components out of the two blocks pairwise β–null Harris recurrent. Furthermore the governing type of nonstationarity is β^{d_1} for the first block and β^{d_2} for the second block. This can be a significant improvement in speed while still yielding the best additive approximation to a fully general true model. Under GSBE results in Theorems 4.3 and 4.4 remain unchanged, but Assumption 4.5 is easier to fulfill.

Checking for weak asymptotic independence, however, seems a hard task in general. The concept might be of most practical relevance, in a stricter form as defined below, when one block of covariates is stationary and another low dimensional block is β–null Harris recurrent. Such situations frequently occur in economics, where economic theory delivers plausible guidance about which components can be modeled stationary and which might be nonstationary – for example, in term structure models as described in Tsay [2002] Section 2.9, or if the real price of one good depends on the real price of another good, take gas and oil, and some other stationary factors, say technological change and infrastructure parameters. Asymptotic independence (first defined in Karlsen et al. [2007] Definition 6.1.) captures that asymptotically the impact of stationary parts not only separates from but vanishes from the nonstationary ones.

Definition 4.2 (Asymptotic Independence)**.** Suppose X^{jk} is β–null Harris re-

current . Then X^j and X^k are asymptotically independent if the joint invariant measure factors into the product of the two respective marginal invariant measures.

$$\pi_{jk} = c_2 \cdot \pi_j \otimes \pi_k \ , \tag{4.11}$$

where $c_2 > 0$ is a constant.

Note that asymptotic independence is a weaker assumption than independence if one of the components is stationary. A simple intuition is, if one process X is non-stationary β–null Harris recurrent and another Z is stationary, then asymptotically in the long run, Z cannot have a significant influence on X. But at a specific time point t, there might be dependence between X and Z (See Karlsen et al. [2007] Example 6.1 and 6.2. for examples of asymptotic independence but short–term dependence). Under asymptotic independence, speeds of convergence towards the invariant measures and support of both sides in (4.11) must coincide, which can only be fulfilled if one marginal component is stationary. If $\pi_x^{(z)}(\mathcal{G}_\S) < \infty$, asymptotic independence holds (See Lemma 6.1. in Karlsen et al. [2007]).

Denote all stationary variables by $Z \in \mathbb{R}^{d_2}$ with joint density $0 < p < \infty$ and $p \in C^1(\mathbb{R}^+)$, all other components $X \in \mathbb{R}^{d_1}$ can be nonstationary β–null Harris recurrent with density of the invariant measure π and parameter β. X and Z are asymptotically independent. For ease of exposition all nonstationary component functions are marked as $g : \mathbb{R}^{d_1} \to \mathbb{R}$, all stationary ones are $f : \mathbb{R}^{d_2} \to \mathbb{R}$, and a is a sclar.

$$Y_i = a + \sum_{j=1}^{d_1} g_j(X_i^j) + \sum_{j=1}^{d_2} f_j(Z_i^j) + \varepsilon_i \quad \text{for all } i \in \{1, \ldots, n\} \tag{4.12}$$

under no concurvity, i.e. for $g_1, \ldots, g_{d_1}, h_1 \ldots, h_{d_2}$ nontrivial we cannot have $g_1(x^1) + \cdots + g_{d_1}(x^{d_1}) + f_1(z^1) + \ldots + f_{d_2}(z^{d_2}) = 0$ for all (x, z). We will see that with more detailed knowledge about the data in this sense, we can achieve significantly better results. The standard and generalized smooth backfitting simplify significantly in terms of speed and asymptotic behavior. In practice, the partially nonstationary model (4.12) is of most practical relevance, if the number

of nonstationary covariates is small, i.e. $d_1 \leq 2$. Set the defining system of integral equations for the SBE estimators $(\widetilde{a}, \widetilde{g}_1, \ldots, \widetilde{g}_{d_1}, \widetilde{h}_1, \ldots, \widetilde{h}_{d_2})$ as

$$\begin{aligned}
\widetilde{g}_j(x^j) &= \widehat{g}_j^{(k)}(x^j) - \sum_{\substack{k \neq j \\ 1 \leq k \leq d_1}} \int_{\mathcal{G}_k^{(j)}} \widetilde{g}_k(x^k) \frac{\widehat{\pi}_{j,k}(x^{jk})}{\widehat{\pi}_j^{(k)}(x^j)} \mathrm{d}x^k \quad \text{for } j = 1, \ldots d_1 \\
\widetilde{f}_j(z^j) &= \widehat{f}_j(z^j) - \sum_{\substack{k \neq j \\ d_1 < k \leq d}} \int_{\mathcal{G}_k} \widetilde{f}_k(z^k) \frac{\widehat{p}_{j,k}(z^{jk})}{\widehat{p}_j(z^j)} \mathrm{d}z^k \quad \text{for } j = 1, \ldots d_2 \,, \qquad (4.13)
\end{aligned}$$

for all $j = 1, \ldots, d_1$ in the first line and $j = 1, \ldots, d_2$ in the second line with $\widetilde{a} = \sum_{j=1}^{d} \frac{1}{d-1} \sum_{k \neq j} \frac{1}{T^{jk}(n)} \sum_{i \in I_{jk}} Y_i$ under the identification assumptions

$$\begin{aligned}
\int_{\mathcal{G}_j^{(k)}} g_j(x^j) \pi_j^{(k)}(x^j) \mathrm{d}x^j = 0 \quad \text{for } j = 1, \ldots d_1,\, k \neq j \\
\int_{\mathcal{G}_j} f_j(z^j) p_j(z^j) \mathrm{d}z^j = 0 \quad \text{for } j = 1, \ldots d_2 \,. \qquad (4.14)
\end{aligned}$$

Note that since the dimension of the β–null Harris recurrent component is $d_1 \leq 2$, standard and generalized smooth backfitting coincide. Stationary and nonstationary component functions naturally separate in the backfitting operator because of asymptotic independence. Constant parts are zero due to (4.14). We obtain the backfitting estimates as solution to (4.13) via joint iteration.

Theorem 4.7. *Let the model be as defined in* (4.12) *with* $d_1 \leq 2$ *fulfilling the centering conditions* (4.14) *and let Assumptions 2.1 hold, components* X *satisfy 4.3,* Z *be stationary asymptotically independent of* X*, and* (X, Z) *fulfill 4.4. The nonstationary bandwidth sequence must satisfy* $(n_N^{\beta+\varepsilon})^{-1/4} \ll h_N \ll (n_N^{\beta+\varepsilon})^{-1/5}$ *with* ε *small. For the stationary bandwidth set* $h = n^{-1/5}$*.*
Then the backfitting algorithm converges with geometric rate and for the estimators $(\widetilde{m}_j)_{j=1}^d = ((\widetilde{g}_j)_{j=1}^{d_1}, (\widetilde{h}_j)_{j=1}^{d_2})$ *we find*

$$\begin{aligned}
\sqrt{\widehat{L}_j^{(k)}(x^j) h_N} \left(\widetilde{g}_j(x^j) - g_j(x^j) - B_j^{NS}(x^j) \right) &\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_{j+}^2(x^j) \frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2}\right) \\
\sqrt{nh} \left(\widetilde{f}_j(z^j) - f_j(z^j) - B_j^N(z^j) \right) &\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_j^2(z^j) \frac{\kappa_0^2(z^j)}{\kappa_0(z^j)^2}\right) \,,
\end{aligned}$$

where bias and variance terms are stated right below.

The variances are

$$\sigma^2_{j+}(x^j) = \int \varepsilon^2 \frac{\pi^{(k)}_{j\varepsilon}(x^j,\varepsilon)}{\pi^{(k)}_j(x^j)}\,\mathrm{d}\varepsilon \qquad \sigma^2_j(z^j) = \int \varepsilon^2 \frac{p_{j\varepsilon}(z^j,\varepsilon)}{p_j(z^j)}\,\mathrm{d}\varepsilon\,.$$

The nonstationary bias consists of two major parts $B^{NS}_j(x^j) = B^A_j(x^j) + B^B_j(x^j)$ with

$$\begin{aligned} B^B_j(x^j) &= h_N \frac{\kappa_1(x^j)}{\kappa_0(x^j)} g_j'(x^j) + \frac{1}{2} h_N^2 \frac{\kappa_2(x^j)}{\kappa_0(x^j)} g_j''(x^j) + \left((I-A)^{-1}\bar{B}^{(1)}\right)_{(j)}(x^j) - b^{(1)}_{j,n} \\ B^A_j(x^j) &= \mu^{(k)}_{((j)\varepsilon)}(K_{x^j,h}(X)\otimes \mathrm{id}_\varepsilon)\frac{T^{(jk)\varepsilon}(n)}{\widehat{L}^{(k)}_j(x^j)}\,. \end{aligned}$$

Under the stated choice of bandwidth the asymptotic bias B^A_j vanishes and also $\sqrt{\widehat{L}^{(k)}_j(x^j)h_N}B^B_j(x^j)$ is negligible for every x^j. The exact form of the deterministic bias B^B_j is implicitly defined via the backfitting operator matrix A as in (A.15). The projected Nadaraya–Watson specific part $\bar{B}^{(1)}(x) = (\bar{b}^{(1)}_1(x^1),\ldots,\bar{b}^{(1)}_d(x^{d_1}))^T$ and component functions $\bar{b}^{(1)}_j$ for $j \in \{1,\ldots,d_1\}$ are given by

$$\bar{b}^{(1)}_j(x^j) = h_N^2\left[\left(b^{(1)}_j + \sum_{k\neq j}\int_{\mathcal{G}^{(j)}_k} b^{(1)}_{jk}(x^k)\frac{\pi_{jk}(x^k)}{\pi_j}\mathrm{d}x^k\right)\right](x^j)\,.$$

These bias components are for $j \in \{1,\ldots,d_1\}$ defined as $b^{(1)}_j(x^j) = \frac{\kappa_2(x^j)}{\kappa_0(x^j)}\left(\frac{g_j'(x^j)}{\pi^{(k)}_j(x^j)}\pi^{(k)\prime}_j(x^j)\right)$ and $b^{(1)}_{jk}(x^{jk}) = \frac{\kappa_2(x^j)}{\kappa_0(x^j)}\left(\frac{g_k'(x^k)}{\pi_{jk}(x^{jk})}\,\frac{\partial\pi_{jk}(x^{jk})}{\partial x^k}\right)$. Most importantly it is $\bar{b}^{(1)}_j = O(h_N^2)$. Furthermore the centering constant $b_{j,n}$ is given by $b^{(1)}_{j,n} = \mu_{(j)}\left(\widehat{\Phi}_j\widehat{g}_j\right)$ where the centering operator is defined in (A.15), and it is $h_N^2 b^{(1)}_{j,n} = O(1)$.

The stationary bias B^S_j only consists of a deterministic part

$$B^S_j(z^j) = h\frac{\kappa_1(z^j)}{\kappa_0(z^j)}f_j'(z^j) + \frac{1}{2}h^2\frac{\kappa_2(z^j)}{\kappa_0(z^j)}f_j''(z^j) + \left((I-A)^{-1}\bar{B}^{(2)}\right)_{(j)}(z^j) - b^{(2)}_{j,n}.$$

It is $\bar{B}^{(2)}(z) = (\bar{b}^{(2)}_1(z^1),\ldots,\bar{b}^{(2)}_d(z^{d_2}))^T$ and component functions $\bar{b}^{(2)}_j$ for $j \in \{1,\ldots,d_2\}$ defined by

$$\bar{b}^{(2)}_j(x^j) = h^2\left[\left(b^{(2)}_j + \sum_{k\neq j}\int_{\mathcal{E}_k} b^{(2)}_{jk}(x^k)\frac{p_{jk}(z^k)}{p_j}\mathrm{d}z^k\right)\right](z^j)\,.$$

These bias components are for $j \in \{1, \ldots, d_2\}$ defined as $b_j^{(2)}(z^j) = \frac{\kappa_2(z^j)}{\kappa_0(z^j)} \left(\frac{f_j'(x^j)}{p_j(z^j)} p_j'(z^j) \right)$ and $b_{jk}^{(2)}(z^{jk}) = \frac{\kappa_2(z^j)}{\kappa_0(z^j)} \left(\frac{f_k'(z^k)}{p_{jk}(z^{jk})} \frac{\partial p_{jk}(z^{jk})}{\partial z^k} \right)$. Most importantly it is $\bar{b}_j = O(h^2)$. Furthermore the centering constant $b_{j,n}^{(2)}$ is given by $b_{j,n}^{(2)} = \mu_{(j)} \left(\widehat{\Phi}_j \widehat{f}_j \right)$ where the centering operator is defined in (A.15), and it is $h^2 b_{j,n}^{(2)} = O(1)$.

Remarks.

1. The vector of of the stochastic parts of component estimates converges jointly to a multivariate normal distribution with only marginal entries on the diagonal of the variance covariance matrix. Covariances vanish asymptotically.

2. The result also holds more generally for a model with transformed error term $g_\varepsilon(\varepsilon)$ when replacing ε in Assumption 4.4.3 by $g_\varepsilon(\varepsilon)$. Then Theorem 4.3 holds with modified asymptotic bias $B_j^A(x^j)$ and variance $\sigma_{j+}^2(x^j) = \int g_\varepsilon(\varepsilon)^2 \frac{\pi_{j\varepsilon}(x^j,\varepsilon)}{\pi_j^\varepsilon(x^j)} \, \mathrm{d}\varepsilon$. With considerations in Karlsen et al. [2007], Section 6.4., and Mammen and Nielsen [2003], results can be even further extended to models with heteroscedastic error terms.

3. If $d_1 > 2$, the results of the above theorem still hold, but a minor additional assumption is needed (See Assumption 4.5 in GSBE). Then the true model must be additive, as for general m, GSBE is no longer guaranteed to yield the best additive approximation.

4. Conditions on components Z via Assumptions 4.4 can also be derived in a nicer moment condition form.

5. The partial nonstationary model can be easily generalized to being partially fully nonparametric in the stationary components.

Here the stationary components are estimated with univariate rates and variances. Compare this to the procedure for fully nonstationary data of in Theorem 4.1 or Theorem 4.3, where estimation of stationary component functions is governed by the worst case univariate nonstationary direction. For the nonstationary part, the rate of convergence is of univariate character in its form, governed by the

bivariate nonstationarity type β for each component function. Asymptotic speeds of convergence are of order smaller than $n^{2/5\beta}$.

As in the stationary case, here it might prove valuable to develop a local linear type version of the estimation procedure in order to obtain oracle efficient bias behavior. with the stated Nadaraya–Watson type estimation method, the deterministic and the stationary bias have a projected design density dependent part $\left((I-A)^{-1}\bar{B}\right)_{(j)}(x^j)$. With a local linear version of the estimation procedure, the obtained bias is directly additive and therefore asymptotically superior. See comments on efficiency in Section 4.4 below.

With enforced independence assumptions between covariates and residual and stationary ε as in Assumption 4.4* we get the analogue to Theorems 4.2 and 4.4 in the partially stationary case.

Theorem 4.8. *Let the same set of assumptions as in Theorem 4.7 hold, but replace Assumptions 4.4 by Assumptions 4.4*. Choose bandwidths as in Theorem 4.7. Then the algorithm converges and we get the following asymptotic expansion for the smooth backfitting estimates* $(\widetilde{m}_j)_{j=1}^d$

$$\sqrt{\widehat{L}_j^{(k)}(x^j)h_N}\left(\widetilde{g}_j(x^j) - g_j(x^j) - B_j^B(x^j)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_{j+}^2(x^j)\frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2}\right)$$
$$\sqrt{nh}\left(\widetilde{h}_j^{NW}(z^j) - f_j(z^j) - B_j^N(z^j)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_j^2(z^j)\frac{\kappa_0^2(z^j)}{\kappa_0(z^j)^2}\right),$$

with simplified variance

$$\sigma_{j+}^2(x^j) \;= \int \varepsilon^2 \pi_\varepsilon(\varepsilon)\, d\varepsilon\,,$$

where π_ε *is the stationary density of* ε*. All other components are as in Theorem 4.7.*

4.4 Remarks on Oracle Efficiency

Efficiency of an estimator can be judged according to the benchmark of an oracle estimator. The infeasible oracle estimator estimates each component function as if all other components were known correctly. An estimator has oracle property or is

oracle, if its asymptotic expansion coincides with the oracle one. In the stationary mixing setting, ordinary smooth backfitting converges for each marginal direction with rate and variance of the one-dimensional smoother. For a local linear version of the estimation procedure, also the oracle bias is reached. Thus for stationary data, smooth backfitting reaches oracle efficiency.

For nonstationary data, however, this is in general not possible – paying for the generality of the underlying data with different data for different directions. Therefore, although rates and variances have the form as for one–dimensional marginal smoothers, in GSBE the worst case bivariate, in adapted GSBE the worst case γ–wise, and in standard smooth backfitting the full dimensional type of nonstationarity β dominates. Thus without further restrictions on the covariate process, we cannot do better in terms of data and oracle type efficiency of the estimation procedure than in GSBE with governing worst case bivariate β^{jk_0}. Hence in this most general case, implementing a local linear version of the proposed estimation methods costs on robustness and increases computing complexity, for asymptotically obtaining an oracle bias. For finite samples, however, a local constant version might still be superior in terms of bias.

Though if the underlying data does not require the full generality of the GSBE framework, more tailored procedures can reach better oracle efficient outcomes. If there is only one known nonstationary component in the vector of covariates while all the others are stationary, then we can reach oracle efficient rates and variances as seen in the section above. In this case, it might prove advisable to even use a local linear version of the suggested backfitting technique to also obtain oracle bias components. Backfitting as suggested based on local constant smoothing suffers from a systematic Nadaraya–Watson bias $(1 - A)^{-1}\bar{B}$ as in the classical setting Mammen et al. [1999] with some additional terms. This results from the fact that $\bar{B}$ is in general not additive but the SBE bias needs to have additive structure. Thus the bias is generally not oracle. In contrast to this, a local linear version in the stationary case directly has an additive bias and is independent of the invariant density of the regressors (see Mammen et al. [1999]). This design independence is specific to the type of estimator and directly carries over from the classical

to the nonstationary setting. For local linear smooth backfitting the underlying projection is not only a projection on the space of additive functions but on an extended additive function space which also includes first order derivatives for the additive components. So the local linear smooth backfitting estimator $\widetilde{m}^{LL}$ has now the form $\left\{ \{\widetilde{m}_j^{LL}\}_{j=0}^n, \left\{\widetilde{m}_j^{LL,1}\right\}_{j=1}^n \right\}$, where $\widetilde{m}_j^{LL}$ estimates m_j and $\widetilde{m}_j^{LL,1}$ its derivative. We under full dimensional β–null Harris recurrence, we obtain them as minimizers of the following criterion with respect to f and f^1

$$\sum_{i=1}^{n} \int \left(Y - f_0 - \sum_{j=1}^{d} f_j(x^j) - \sum_{j=1}^{d} f_j^1(x^j)(X_i^j - x^j) \right)^2 K_h(x - X_i) \mathrm{d}x \; .$$

Chapter 5

Finite Sample Behavior: A Simple Simulation Study

Non–parametric estimation of a general conditional mean function m has already been studied for β–Harris recurrent processes in detail in Karlsen et al. [2007]. Though in many practically relevant cases, models with more than two covariates do not fit the required framework any more. The contribution of this work is to provide a method and its asymptotic theory for theses cases under some mild functional form restriction which still leaves a high degree of modeling flexibility. In order to demonstrate the finite sample power of the proposed procedure, some simulation studies have been performed.

Compared to stationary data, a general β–Harris recurrent process can behave quite "strangely" in finite samples, being clustered in some regions of the space while leaving others almost empty (see Figure 5.1). This results from the fact that the expected time until the process reaches a specific set in the range can in general be infinite (see (2.7)). Therefore in my simulations, we report pointwise (over all 500 replications)–median estimators. In applications, to circumvent the empty space problem, a very large number of observations is needed to reduce number and size of data uncovered regions to a minimum. In line with Theorem 4.3, there are sufficiently many data points required in the local neighborhood of a point, i.e. $h_{j+}\widehat{L}_{j+}(x^j)$ must be large, that estimation at this point is robust. If

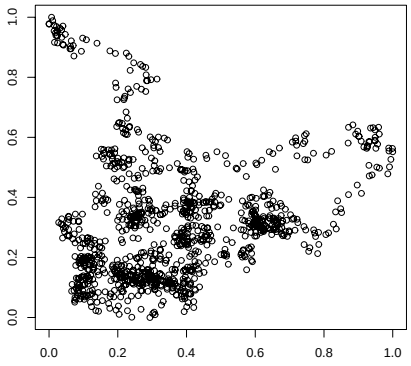


Figure 5.1: Example of a two dimensional random walk for 1000 observations linearly rescaled into $[0,1]^2$ illustrating nonstationary data particular difficulties of inference in finite samples such as clustering and empty parts of the space

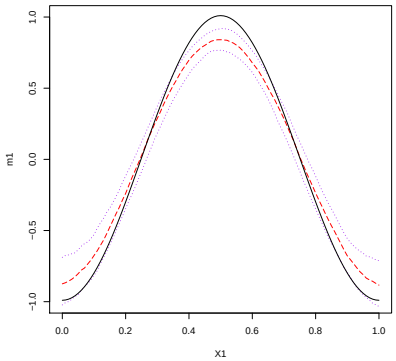
this cannot be achieved for certain points, these local results should be interpreted with care.

In all simulation experiments estimation is repeated 500 times from $n = 1000$ or $n = 910000$ observations in the following model for $i = 1, \ldots, n$:

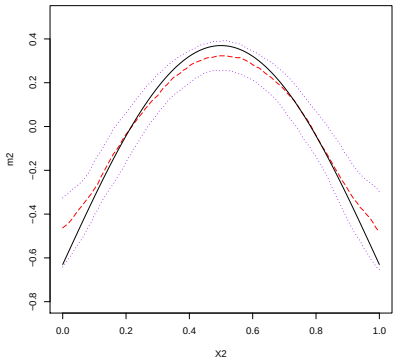
$$\begin{aligned} Y_i &= \sum_{j=1}^{5} m_j(X_i^j) + \varepsilon_i \\ X_i &= X_{i-1} + e_i \end{aligned}$$

where $X_0 = (0,0,0,0,0)^T$ and $m_j(x) = \cos(2\pi(x-0.5))$ for $j \in \{2,4\}$ and $m_j(x) = \sin(\pi x)$ for $j \in \{1,3,5\}$. The residuals are independent $\varepsilon \sim N(0, \sqrt{0.5})$ and $e \sim N(0,\sigma)$ with $\sigma = ((\sigma_{jk}))_{jk} \in \mathbb{R}^{5\times 5}$. To underline the robustness of the method, we simulate settings with independent random walks as well as cases with correlation, where some off–diagonal elements of σ are strictly positive. This model setup is chosen in order to have an easy comparison to the stationary smooth backfitting case, in particular to the extensive simulation study in Nielsen and Sperlich [2005] which focuses on trigonometric relationships. Practically such models are used in macroeconomic business cycle literature.

Figure 5.2: **Local constant fit, 1000 observations**

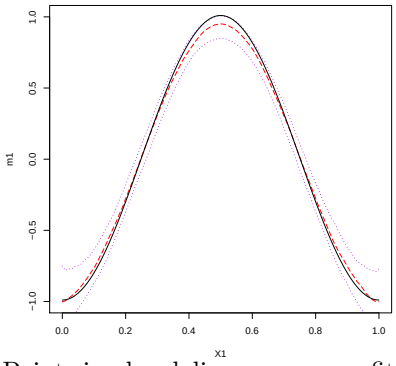


Pointwise local constant average fit (red dashed) in comparison to the true function m_1 (black). The dotted purple lines denote the interquartile range i.e. the 75% and 25% gridpointwise quantiles over all iterations

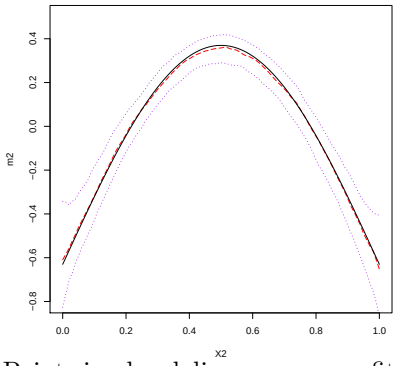


Pointwise local constant average fit (red dashed) in comparison to the true function m_2 (black). The dotted purple lines denote the interquartile range i.e. the 75% and 25% gridpointwise quantiles over all iterations

Figure 5.3: **Local linear fit, 1000 observations**



Pointwise local linear average fit (red dashed) in comparison to the true function m_1 (black). The dotted purple lines denote the interquartile range i.e. the 75% and 25% gridpointwise quantiles over all iterations



Pointwise local linear average fit (red dashed) in comparison to the true function m_2 (black). The dotted purple lines denote the interquartile range i.e. the 75% and 25% gridpointwise quantiles over all iterations

When using kernel smoothing techniques there should be some guideline on how to choose the smoothing parameter h. This is a largely unsolved problem, since the admissible rates as stated in the theorems are only asymptotically true. They are different from the stationary rates, not only with n effectively replaced by n^β but also in an additional speed restriction from above $h_j < n^{-1/5\beta^+ - \varepsilon}$. Due to nonstationarity of the data, the shrinking of the bandwidth h must not only satisfy a maximum speed as usual but also a minimum speed in order to guarantee enough data points in the observation window asymptotically. Cross–validation techniques proved to be useful for finite samples. For simplicity, the bandwidth is chosen via cross validation for the best componentwise fit, which does not necessarily yield the same results as for the best global fit, especially when regressors are correlated. For SBE with stationary data, in Nielsen and Sperlich [2005] a more involved global cross validation procedure is used without proof which seems to induce an additional bias of yet unknown size. We found it favorable to use a data adaptive local choice $h_j(x^j) \sim \left(\sum_{i \in I_{jk_0}} \mathbf{1}_{\mathcal{N}_{x^j,k}}(X_i^j)\right)^{-1/5}$ for fixed small $k << 1$ depending on the number of visits to a k–neighborhood around x^j. For this no preknowledge of β is required, which in practice must be estimated first, e.g. via a Hill type estimator from (2.6) suffering from poor convergence properties. Though deriving formal results for such a local data–driven bandwidth is beyond the scope of this paper. For a local stochastic bandwidth theoretical results are not straightforward, as can be seen from Guerre [2004] in the general non–additive model under the restrictive uniform recurrence assumption. Potentially the theoretical results in Mammen and Park [2005] for mixing processes via plug–in and penalized least squares could be extended for β–null Harris recurrent processes.

The implementation closely follows the strategy by Nielsen and Sperlich [2005] and Haag [2006]. In particular, any steps of the algorithm are performed on a fixed grid in each direction. Thus the data generating processes are linearly transformed to live in the cuboid $[0,1]^d$, for easy comparison. In order to reduce numerical errors in the integrals, $M = 101$ equidistant grid points are chosen. For the algorithm to

stop, the following quotient criterion is employed: If

$$\frac{\sum_{i=1}^{M}\left(\widetilde{m}_j^{[r-1]}(x_i^j)-\widetilde{m}_j^{[r]}(x_i^j)\right)^2}{\sum_{i=1}^{M}(\widetilde{m}_j^{[r]}(x_i^j))^2+0.0001}<0.0001 \tag{5.1}$$

is fulfilled for all $j=1,\ldots,d$ at the M grid points, then end at iteration step r. Besides the local constant type generalized smooth backfitting estimator also the local linear version of generalized smooth backfitting according to (4.15) is implemented for comparison (See Section 4.4). To judge the performance of the estimators, quantiles over the repetitions k of the integrated square error ISE^k for each additive component are compared. For each component $j\in\{1,\ldots,5\}$, ISE^k is defined as

$$ISE^k(m_j)=\frac{1}{101}\sum_{l=1}^{101}(m_j(x_l^j)-\widetilde{m}_j^k(x_l^j))^2 \quad \text{for all } j\in\{1,\ldots,5\}\ , \tag{5.2}$$

on the grid $0=x_0<\ldots<x_l<\ldots<x_{100}=1$ with $x_l=l\cdot 0.01e_5,\ l=\{0,\ldots,100\}$ and e_5 unit vector in $\mathbb{R}^5$, where $\widetilde{m}_j^k$ is the obtained SBE for component j in the k^{th} repetition. For the given data structure, these measures of fit are more appropriate than the more commonly reported $MISE$ – the arithmetic mean of ISE^k.

Figure 5.2 shows the pointwise median estimator (dashed line) in comparison to the true marginal function (solid line) for a local constant fit with 1000 observations for m_1 on the left and m_2 on the right, when all five regressors are independent. It is $\sigma^{jj}=1$ and for the offdiagonal components $\sigma^{kj}=0$ for all $j\in\{1,\ldots,5\}$ and $k\neq j$. The cosine problems are harder and therefore the fit of m_2 must be better than for m_1 due to double the range and due to a larger factor in the second derivative, which governs the leading bias term. The bias at the peak might be specific to the only local constant fit, which has some systematic theoretical bias in comparison to a local linear fit. This is graphically supported by Figure 5.3 which shows a local linear pointwise median estimator in the same scenario with less bias at the peak. The algorithm converges on average after 20.442 iterations in the local linear case and after 15.406 in the local constant case. In the stationary case only about 6 are needed (see Nielsen and Sperlich [2005]). Though given the

type of fit	underlying data			medianISE for the full $[0,1]^5$				
	N	σ^{jj}	σ^{kj}	m_1	m_2	m_3	m_4	m_5
Local linear	10,000	1	0	0.012	0.007	0.009	0.007	0.012
Local constant	10,000	1	0	0.026	0.016	0.031	0.019	0.029
Local linear	10,000	1	0.2	0.013	0.009	0.011	0.008	0.012
Local constant	10,000	1	0.2	0.027	0.016	0.031	0.018	0.027
Local linear	1,000	1	0	0.022	0.018	0.018	0.018	0.021
Local constant	1,000	1	0	0.031	0.021	0.034	0.022	0.033
Local linear	1,000	1	0.2	0.027	0.021	0.019	0.020	0.026
Local constant	1,000	1	0.2	0.030	0.017	0.033	0.020	0.033

type of fit	underlying data			medianISE for the interior $[h,1-h]^5$				
	N	σ^{jj}	σ^{kj}	m_1	m_2	m_3	m_4	m_5
Local linear	10,000	1	0	0.010	0.005	0.008	0.005	0.009
Local constant	10,000	1	0	0.022	0.011	0.024	0.013	0.024
Local linear	10,000	1	0.2	0.012	0.007	0.010	0.006	0.009
Local constant	10,000	1	0.2	0.026	0.011	0.026	0.013	0.024
Local linear	1,000	1	0	0.015	0.009	0.013	0.011	0.013
Local constant	1,000	1	0	0.027	0.014	0.027	0.015	0.027
Local linear	1,000	1	0.2	0.020	0.012	0.013	0.012	0.019
Local constant	1,000	1	0.2	0.026	0.012	0.029	0.014	0.027

Table 5.1: MedianISE as measure of fit with $k \neq j = 1, \ldots 5$

Quantiles	fit	m_1	m_2	m_3	m_4	m_5
50%	LL	0.022	0.018	0.018	0.018	0.021
	NW	0.031	0.021	0.034	0.022	0.033
75%	LL	0.104	0.045	0.050	0.064	0.083
	NW	0.058	0.038	0.061	0.043	0.067
95%	LL	1.159	0.498	0.369	0.857	1.031
	NW	0.103	0.070	0.105	0.089	0.125
97%	LL	2.627	0.972	0.674	1.593	1.458
	NW	0.165	0.104	0.139	0.137	0.216

Table 5.2: Quantiles of ISE^k for local linear (LL) and local constant (NW) fit with 1000 observations and no correlation

increased difficulty of the problem the algorithm performs reasonably well. For 10000 observations convergence is reached on average after 19.164 iterations in the local linear case, after 15.172 iterations in the local constant case. Also the fit is improved as can be seen from Table 5.1. It also shows that for correlated regressors X with $\sigma^{jk} = 0.2$ for $k \neq j$, the problem is easier, thus the fit is better. When omitting regions of sparse data along the boundaries, the overall fit is also improved - especially in the local linear case. Table 5.2 indicates that the local constant estimator is more robust then the local linear version and therefore despite its type–specific bias more preferable in practice.

Chapter 6

Conclusion

6.1 Summary

We have introduced a nonparametric estimation procedure, which allows to estimate a regression problem with more than $d > 2$ potentially nonstationary covariates. As in a stationary setting, estimating an additive model allows to circumvent the ordinary curse of dimensionality. Thus rate of convergence and asymptotic variance are of univariate form. Though the added nonstationary difficulty is reflected by the fact that for generalized smooth backfitting the worst case bivariate type of nonstationarity and the corresponding β govern the rate. For a model with an arbitrary but potentially large finite amount of regressors and nonstationarity as an added difficulty, this is the best to be achieved. Under the other suggested backfitting type methods, we can reach a best additive fit to more general true models, but require more regularity in the data than pairwise β–null Harris recurrence and obtain slower rates. Under full β–null Harris recurrence , standard smooth backfitting scaled according to full–dimensional objects yields the best additive approximation a fully general true model.
Furthermore in the special case of a stationary residual ε, results are obtained which could serve as a starting point for an elegant way of additive nonlinear cointegration. When each component function is monotone, the desired symmetry between response and observables as in linear cointegration can be achieved.

While the more general results of Karlsen et al. [2007] are limited to a very small number of cointegrated components, the method introduced here works for an arbitrary amount of regressors. Therefore in a wide range of applications it might be the only way to determine a general cointegration relationship between variables without prespecifying a parametric form.

6.2 Outlook

The suggested methods can be used to test for linearity in cointegration relationships. While a formal testing procedure might prove difficult to develop, estimation results for a general additive model can already provide a guideline if linearity is appropriate or if not, what kind of nonlinearity should be modeled. In economic models where a cointegration relationship is expected, but could not be detected with existing methods, estimation with GSBE could help to provide empirical evidence for economic intuition. This especially applies to purchasing power parity (PPP) or term structure models as in Tsay [2002] Section 2.9. It might be of interest to develop model specification tests in this general scenario generalizing the cointegration rank test of Johansen [1991]. Furthermore it might be interesting to investigate how detrending of data with a deterministic trend should be done in order to obtain similar results for resulting β–null Harris recurrent observations as presented here. One could also think of deriving least squares type estimators under more smoothness or parametric assumptions and β–null Harris recurrent data. Though the framework is tailored for local smoothing techniques. Thus aiming for a global fit seems somehow unnatural and will always suffer from the nonstationary curse of dimensionality.

Presented techniques can serve to provide nonparametric estimates of the individual marginal utility function in Euler equations. They characterize intertemporal optimization and are driven by nonstationary individual consumption (See e.g. Cochrane [2001]). Up to now estimation in this central economic question has been dominated by parametric GMM methods leading to sometimes contradictory results.

Generally, β–null Harris recurrence delivers a natural blocksize of independent blocks of sums of observations between recurrence times (See Appendix A.1). This might serve as a way to generalize block bootstrap procedures in time series (see Hall et al. [2003]) to the recurrent setting. With this the usually somewhat arbitrary block window has a stochastic meaning and might be estimable for Feller type processes.

If observations are discrete, null–Harris recurrence and positive Harris recurrence coincide. Therefore all measures are finite and estimation is much easier resulting in simpler assumptions. This might be applicable to storage or queuing models.

Appendix A

A.1 Markov Theory

To keep the paper as self–contained as possible, essential notions and results of Markov theory relevant for the understanding of the paper shall be mentioned.

A.1.1 Split Chain and Invariant Measure

Every ϕ irreducible Markov chain $(X_i)_i$ satisfies the minorization inequality. That means, for any such $(X_i)_i$ there exists a small function s, a probability measure ν and an integer $m_0 \geq 1$ such that

$$P^{m_0} \geq s \otimes \nu \ .$$

Without much loss of generality we assume throughout the paper that $m_0 = 1$, i.e. the minorization inequality has the form:

$$P \geq s \otimes \nu \ , \tag{A.1}$$

where s and ν are small and $\nu(\mathbb{R}^n) = 1$. In particular it is $0 \leq s(x) \leq 1$ for $x \in \mathbb{R}^n$. If (A.1) holds, then the pair (s, ν) is called a pseudo–atom for P. Note that ν is independent of x. This is the basis for constructing the corresponding split chain of $(X_i)_i$.

From (A.1) we derive:

$$\begin{aligned} P(x,A) &= (P(x,A) - s(x)\nu(A)) + s(x)\nu(A) \\ &= (1-s(x))\left[\left(\frac{P(x,A)-s(x)\nu(A)}{1-s(x)}\right)\mathbf{1}_{\{s(x)<1\}} + \mathbf{1}_a(x)\mathbf{1}_{\{s(x)=1\}}\right] + s(x)\nu(A) \\ &=: (1-s(x))\,Q(x,A) + s(x)\nu(A) \end{aligned}$$

Hence the transition probability P can be thought of as a convex combination of a transition probability Q and the independent small measure ν. Thus the chain regenerates each time ν is chosen – which occurs with probability $s(x)$. Introducing the split chain (X_i, Y_i) helps to formalize this observation. The auxiliary chain Y_i only takes on values 0 and 1. For $X_i = x$ and $Y_{i-1} = y_{i-1}$, the auxiliary chain $\{Y_t\}$ takes on the value 1 with probability $s(x)$. Thus $\alpha = \mathbf{R}^n \times \{1\}$ is a proper atom for the split chain. Denote by

$$\tau_0 := \min\{i \geq 1 : Y_i = 1\} \tag{A.2}$$

the corresponding recurrence time. We will frequently need the consecutive sequence of recurrence times $(\tau_k)_{k=-1}^{\infty}$ starting in $t = 0$ defined recursively by:

$$\tau_k := \min\{i \geq \tau_{k-1} : Y_i = 1\}\ , \tag{A.3}$$

with starting value set as $\tau_{-1} := -1$ (any negative number would do). Write $\tau_0 = \tau_\alpha = \tau$ for the first recurrence time with respect to the pseudo atom α. Furthermore denote the number of regenerations up to n by

$$T(n) = \max_k \{k : \tau_k \leq n\} \tag{A.4}$$

We need for scaling purposes

$$\begin{aligned} T_C(n) &= \sum_{i=1}^{n} \mathbf{1}_C(X_i) \\ a(n) &= \mathbb{E}_\lambda \frac{(T_C(n))}{\pi_s(C)}\ , \end{aligned} \tag{A.5}$$

with λ any initial distribution and C a so called D–set, such that (A.5) is always well defined. Any small set is a D–set and the definition of a does not depend on

the specific choice of C. It can be easily shown that the invariant measure π_s has a kernel representation in terms of the atom (see [Nummelin, 1984], page 63f)

$$\pi_s := \nu\, G_{s,\nu}\ , \quad \text{with} \quad G_{s,\nu} := \sum_{t=0}^{\infty} (P - s \otimes \nu)^t\ . \tag{A.6}$$

Note that for ease of exposition in the main text, we omit the index s in π_s. Then for $g \in L^1_{\pi_s}(\mathbb{R}^d, \mathbb{R})$ it follows

$$G_{s,\nu} g(x) = \mathbb{E}\left[\sum_{i=0}^{\tau} g(X_i) | X_0 = x\right] = \mathbb{E}_x\left[\sum_{i=0}^{\tau} g(X_i)\right]\ . \tag{A.7}$$

Hence for $g = \mathbf{1}_A$ it is $\pi_s(A) = \nu G_{s,\nu} \mathbf{1}_A$. If the measure π_s is absolutely continuous w.r.t. Lebesgue measure, we also denote the corresponding density by π_s. Then $\pi_s(x)\mathrm{d}x = \pi_s(\mathrm{d}x)$. With this define the density $\pi_C(x) = \frac{\pi_s(x)}{\pi_s \mathbf{1}_C}$ for $x \in C$ with C small.

The minorization inequality and the accompanying split chain permit a decomposition of the chain into separate and identical parts defined by regeneration points.

$$S_n(g) := \sum_{i=0}^{n} g(X_i) = U_0 + \sum_{k=0}^{T(n)} U_k + U_{(n)} \text{ for any } g \in L^1_{\pi_s}(\mathbb{R}^d, \mathbb{R})\ , \tag{A.8}$$

where:

$$U_k = \begin{cases} \sum_{i=\tau_{k-1}+1}^{\tau_k} g(X_i) & \text{when} \quad k \geq 0 \\ \sum_{i=\tau_{T(n)+1}}^{n} g(X_i) & \text{when} \quad k = (n) \end{cases} \tag{A.9}$$

The sequence $\{(U_k, (\tau_k - \tau_{k-1}))\}_{k=1}^{\infty}$ consists of independent identically distributed (iid) random variables. Denote the common marginal distribution of U_k with $U = U(g)$ and respectively $\mu = \mu(g_h) = \mathbb{E}U(g_h) = \pi_s(g_h)$ and $\sigma = \sigma(g_h) = \mathbb{V}U(g_h)$.

A.1.2 β–null Harris recurrence

The definitions in (2.4) and (2.6) are equivalent. Under β–Harris recurrence for $n \longrightarrow \infty$ it is asymptotically $a(n) \sim n^{\beta} L_s(n)$ with $L_s(n)$ from the tail condition

of the recurrence time (2.6) slowly varying at infinity. Furthermore the exact asymptotic distribution can be specified (see e.g.Chen [2000] Theorem 1.3.)

$$(a(n))^{-1/2} \sum_{i=1}^{n} \mathbf{1}_C(X_i) f(X_i) \quad \sim \quad \sigma_f \sqrt{g_\beta}\, \mathcal{N}(0,1) \tag{A.10}$$

with $\sigma_f = \int f^2(x)\pi_C(x)\,\mathrm{d}x + 2\sum_{i=0}^{\infty} \int f(x) P^i f(x) \pi_C(x)\,\mathrm{d}x$ where (A.10) is only defined for such functions f for which σ_f exists. The random variable g_β is independent of the normal distribution and is Mittag–Leffler $\mathcal{M}_\beta$ distributed.

A.1.3 The Quotient Limit Theorem

The following result is the appropriate generalization of ergodicity to Harris recurrent Markov chains.

Theorem A.1. *If a discrete Markov process $(X_i)_i$ is Harris recurrent, then for any functions $f, g \in L^1_\pi = \left\{\phi | \int \phi(x)\pi(\mathrm{d}x) < \infty\right\}$ with $\int g(x)\pi(\mathrm{d}x) \neq 0$ it is*

$$\lim_{n\longrightarrow\infty} \frac{\sum_{i=1}^{n} f(X_i)}{\sum_{i=1}^{n} g(X_i)} = \frac{\int f(x)\pi(\mathrm{d}x)}{\int g(x)\pi(\mathrm{d}x)} \quad \mathbb{P} - a.s\ . \tag{A.11}$$

A.2 Proofs

This section is split into three main subsections. In the first part, operator notation for the backfitting procedure is introduced to motivate the proofs in subsection two. In the second part, necessary uniform lemmas are proven which are the main tools for the proofs of the main theorems in subsection three.

A.2.1 On the Structural Form of Generalized Smooth Backfitting

For simplification rewrite the generalized backfitting problem (3.6) in form of an operator equation in the corresponding pairwise Hilbert spaces $L^2_{\widehat{\pi}_{jk}}$. This reveals and emphasizes the underlying structure of the problem and gives fundamental

insights for its understanding and proof. Structurally we obtain an inverse problem — which in contrast to many other situations is well–posed [1]. Componentwise in $j = 1, \ldots, d$ we get with boundary modified kernels (2.24):

$$\widetilde{m}_j(x^j) \;=\; \sum_{k \neq j} \frac{1}{d-1} (\mathbf{1}_k - \widehat{\Phi}_{jk}) \widehat{m}_j(x^j) - [\widehat{A}_{j,k} \widetilde{m}_k](x^j) \tag{A.12}$$

with centering operator operators $\widehat{\Phi}_{jk}$ and projection operators $\widehat{A}_{j,k}$ and $\mathbf{1}_k$ defined as

$$\begin{aligned} \mathbf{1}_k \widehat{m}_j(x^j) &= \widehat{m}_j^{(k)}(x^j) \\ \widehat{\Phi}_{jk} \widehat{m}_j(x^j) &= \frac{\int \widehat{m}_j^{(k)}(x^j) \widehat{\pi}_j^{(k)}(x^j) \mathrm{d}x^j}{\int \widehat{\pi}_j^{(k)}(x^j) \mathrm{d}x^j} = \frac{1}{T^{jk}(n)} \sum_{i \in I_{jk}} Y_i \\ \left[\widehat{A}_{j,k} f_k\right](x^j) &= \int f_k(x^k) \frac{\widehat{\pi}_{j,k}(x^j, x^k)}{\widehat{\pi}_j^{(k)}(x^j)} \mathrm{d}x^k \quad \text{for } j \neq k \,, \end{aligned} \tag{A.13}$$

for $f_k : \mathbb{R} \to \mathbb{R}$. Note that $\sum_{k \neq j} \widehat{\Phi}_{jk} \widehat{m}_j(x^j)$ only differs from zero when (3.8) is not fulfilled. For rectangular $\mathcal{G}^j$ and boundary kernels it is $\sum_{k \neq j} \widehat{\Phi}_{jk} \widehat{m}_j(x^j) = 0$. The operator $\widehat{A}_{j,k}$ projects any $f \in \mathcal{H}_k^{(j)} = L^2_{\pi_k^{(j)}}$ onto $\mathcal{H}_j^{(k)}$. Thus $\widehat{A}_j := \sum_{k \neq j} \widehat{A}_{j,k}$ projects any $f \in \mathcal{H}_{-j}^{(k)}$ onto $\mathcal{H}_j$. Thus for any component j there is a a suitable space of additive functions $\mathcal{H}^{(j)}$ composed of pairwise coordinatewise projections $\mathcal{H}_k^{(j)}$ of (3.11) with $\mathcal{H}^{(j)} = \bigoplus_{j=1}^d \mathcal{H}_k^{(j)}$.

Introducing vector notation $\widetilde{m} = (\widetilde{m}_1(x^1), \ldots, \widetilde{m}_d(x^d))^T \in \mathbb{R}^d$ and analogously $\widehat{m} = (\widehat{m}_1(x^1), \ldots, \widehat{m}_d(x^d))^T \in \mathbb{R}^d$ we obtain the simplest form in matrix notion

$$(I - \widehat{A}) \widetilde{m} \;=\; \frac{1}{d-1} (\mathbf{1} - \widehat{\Phi}) \widehat{m} \tag{A.14}$$

with I the identity and, under Assumptions 1-2 or 1 and 4, compact operator

[1] Compare in contrast the case of ill–posed inverse problems. See Carrasco, Florens and Renault for a survey article on ill–posed inverse problems [Carrasco et al., 2003]

matrices

$$\widehat{\Phi} = \begin{pmatrix} 0 & \widehat{\Phi}_{1,2} & \dots & \widehat{\Phi}_{1,d-1} & \widehat{\Phi}_{1,d} \\ \widehat{\Phi}_{2,1} & 0 & \widehat{\Phi}_{2,3} & \dots & \widehat{\Phi}_{2,d} \\ \vdots & & \ddots & & \vdots \\ \widehat{\Phi}_{d-1,1} & \dots & \widehat{\Phi}_{d-1,d-2} & 0 & \widehat{\Phi}_{d-1,d} \\ \widehat{\Phi}_{d,1} & \widehat{\Phi}_{d,2} & \dots & \widehat{\Phi}_{d,d-1} & 0 \end{pmatrix} \quad \mathbf{1} = \begin{pmatrix} 0 & \mathbf{1}_2 & \dots & \mathbf{1}_{d-1} & \mathbf{1}_d \\ \mathbf{1}_1 & 0 & \mathbf{1}_3 & \dots & \mathbf{1}_d \\ \vdots & & \ddots & & \vdots \\ \mathbf{1}_1 & \dots & \mathbf{1}_{d-2} & 0 & \mathbf{1}_d \\ \mathbf{1}_1 & \mathbf{1}_2 & \dots & \mathbf{1}_{d-1} & 0 \end{pmatrix}$$

$$\widehat{A} = -\begin{pmatrix} 0 & \widehat{A}_{1,2} & \dots & \widehat{A}_{1,d-1} & \widehat{A}_{1,d} \\ \widehat{A}_{2,1} & 0 & \widehat{A}_{2,3} & \dots & \widehat{A}_{2,d} \\ \vdots & & \ddots & & \vdots \\ \widehat{A}_{d-1,1} & \dots & \widehat{A}_{d-1,d-2} & 0 & \widehat{A}_{d-1,d} \\ \widehat{A}_{d,1} & \widehat{A}_{d,2} & \dots & \widehat{A}_{d,d-1} & 0 \end{pmatrix} = \begin{pmatrix} 0 & & & & -\widehat{A}_{up} \\ & & & \ddots & \\ -\widehat{A}_{down} & & & & 0 \end{pmatrix} \quad \text{(A.15)}$$

Simplifying notation in (A.14) we can also write:

$$(I - \widehat{A})\widetilde{m} \;=\; \widehat{m}^{II} \quad \text{(A.16)}$$

with $\widehat{m}^{II} = \frac{1}{d-1}(\sum_{k\neq 1} \mathbf{1}_k \widehat{m}_1(x^1), \dots, \sum_{k\neq d} \mathbf{1}_k \widehat{m}_d(x^d))^T \in \mathbb{R}^d$. And by setting $m_0 = \sum_{j=1}^{d} \frac{1}{d-1} \sum_{k\neq j} \frac{1}{T^{jk}(n)} \sum_{i\in I_{jk}} Y_i$ the centering term can be omitted. In the case of a fully–recurrent vector of covariates and notation as introduced in (3.2), the generalized backfitting equations (A.16) reduce to the standard backfitting operator equation with projections in $L^2_{\widehat{\pi}}$ as in [Mammen et al., 1999].

Since for any sample size n, $\widehat{A}$ is compact and self–adjoint, this is a Fredholm equation of the second kind. Formally, in order to find a solution $\widetilde{m}$, the inverse of $I - \widehat{A}$ must be applied to (A.14). For this the operator must be injective, thus the null space $\mathcal{N}(I - \widehat{A})$ of the operator has to be trivial. This is achieved through the sample counterparts of normalization condition (3.18). According to Fredholm and Riesz theory in functional analysis a solution to (A.14) exists, if $(I-\widehat{\Phi})\widehat{m}$ is in the range of the closure of $(I-\widehat{A})$ which is identical to the orthogonal complement $\mathcal{N}(I - \widehat{A}^*) = \mathcal{N}(I - \widehat{A})$ as $\widehat{A}$ is self adjoint. Since this null space is trivial under the norming constraint, a solution exists and is unique. To obtain a practical solution, however, show that $(I - \widehat{A})$ is a contraction operator. Then according to a generalized version of Banach's fixed–point theorem the unique

solution is reached through an iterative procedure and its rate of convergence will be geometric. In matrix notation the SBE algorithm works as given right below. Instead of iterating the full $\widehat{A}$, the matrix is split into upper and lower triangular part. Then $\widehat{A}_{up}$ projects $\widetilde{m}_j$ from the previous iteration step, while $\widehat{A}_{down}$ treats already updated versions of the estimator components from within the rth iteration step.

$$\widetilde{m}^{[r]} = \widehat{m}^{II} - \begin{pmatrix} 0 & & \widehat{A}_{up} \\ & \ddots & \\ \mathbf{0} & & 0 \end{pmatrix} \widetilde{m}^{[r-1]} - \begin{pmatrix} 0 & & \mathbf{0} \\ & \ddots & \\ \widehat{A}_{down} & & 0 \end{pmatrix} \widetilde{m}^{[r]} \quad \text{(A.17)}$$

However, what complicates the following proof of the asymptotic results is that all operators are estimated depending on the sample size n. Therefore in order to ensure that the obtained $\widetilde{m}$ from (A.17), is also the solution to the original additive regression problem (1.1) under the norming constraint (3.8) some uniform convergence results are needed. Since in the null-recurrent setting these are not available in the existing literature, they are shown in the following section.

A.2.2 Preliminary Lemmata

In order to prove any of the theorems, first, some preliminary technical lemmata with uniform convergence results need to be shown. These are not only essential for the proofs but also of interest on their own.

Uniform consistency of the Kernel invariant density estimate

To our knowledge in the general case of β–null Harris recurrent processes uniform results for consistency have not been established. While for subcases with finite invariant measure, a Hoeffding type exponential inequality exists (see [Glynn and Ormoneit, 2002]), the proof of the general case is more involved.

Although for smooth backfitting only univariate and pairwise bivariate density estimates are of interest, the following proof is given for d covariates. To ease notation, indices and superscripts indicating components to be marginal j or jk specific will be generally omitted. We need the following moment bounds:

Lemma A.1. *Let Assumptions 1-2 hold. Set* $K_x(u) = h^d K_{x,h}(u) = K((u-x)/h)$. *Let the process start in a point of regeneration and set* $U = U_0 = \sum_{i=0}^{\tau_0} K_{x,h}(X_i)$. *Then it is with* $-\infty < \mu, \mu' < \infty$ *and* $0 < \sigma, \sigma' < \infty$

$$\begin{aligned}
\mu(K_{x,h}) &= \mathbb{E}U(K_{x,h}) = \pi(K_{x,h}) = \mu + o(1) \\
\mu(|K_{x,h}|) &= \mathbb{E}U(|K_{x,h}|) = \pi(|K_{x,h}|) = \mu' + o(1) \\
h^d\sigma^2(K_{x,h}) &= h^d\left(\mathbb{E}U^2(K_{x,h}) - \mu^2(K_{x,h})\right) = \sigma^2 + o(1) \\
h^d\sigma^2(|K_{x,h}|) &= h^d\left(\mathbb{E}U^2(|K_{x,h}|) - \mu^2(|K_{x,h}|)\right) = \sigma'^2 + o(1) \\
\sigma^2(K_x) &= \mathbb{E}U^2(K_x) - \mu^2(K_x) = \sigma^2 \ .
\end{aligned}$$

Proof. See Lemma 5.1. and 5.2. in Karlsen and Tjøstheim [2001] for the proof of the bounds. The form of σ follows from Theorem 5.3. $h^d\sigma^2(K_{x,h}) = \int \pi_C(x + hu)K^2(u)\,\mathrm{d}u + 2\int K(u)PG_{s,\nu}K_{x,h}(x+hu) + o(1)$. □

Assume w.l.o.g. that the process starts in a point of regeneration. The proof is on the atomic level, but extends straightforwardly to small sets. The Green function $a(n)$ is defined in (A.5).

Lemma A.2 (Uniform consistency of the Kernel density estimator)**.**
Let Assumptions 1-2 hold. Then choose a bandwidth $h \to 0$ *such that* $\sqrt{a(\frac{n}{L_2 a(n)})}L_2 a(n)h^d \to \infty$ *and set* $h\ l(n)^{1/d} = \sqrt{a\left(\frac{n}{L_2(a(n))}\right)}L_2(a(n))$. *Then*

$$\sup_{x\in\mathring{\mathcal{G}}_h} |\widehat{\pi}(x) - \pi(x)| = O_P\left(h^2 + \frac{1}{h^d\sqrt{a\left(\frac{n}{L_2(a(n))}\right)}L_2(a(n))}\right) \tag{A.18}$$

$$\sup_{x\in\partial\mathcal{G}_h} |\widehat{\pi}(x) - \pi(x)| = O_P\left(h + \frac{1}{h^d\sqrt{a\left(\frac{n}{L_2(a(n))}\right)}L_2(a(n))}\right) \ . \tag{A.19}$$

Proof. It is sufficient to show that for given

$$\begin{aligned} c_n &= h^2 + \frac{1}{h^d\sqrt{a\left(\frac{n}{L_2(a(n))}\right)L_2(a(n))}} \\ c'_n &= h + \frac{1}{h^d\sqrt{a\left(\frac{n}{L_2(a(n))}\right)L_2(a(n))}} \end{aligned}$$

it is that for all $\eta, \eta' > 0$ there exist constants $c, c' > 0$ such that

$$\begin{aligned} \sup_n \mathbb{P}\left(\sup_{x\in\mathring{\mathcal{G}}} |\widehat{\pi}(x) - \pi(x)| \geq c \cdot c_n\right) &= \eta \\ \sup_n \mathbb{P}\left(\sup_{x\in\partial\mathcal{G}} |\widehat{\pi}(x) - \pi(x)| \geq c' \cdot c'_n\right) &= \eta' \,. \end{aligned}$$

In fact we will even show almost sure convergence.

To shorten notation we will write c_n instead of $c \cdot c_n$ and keep in mind that c_n is simply the rate without any constants. The same holds for c'_n.

Split up into variance and bias part. For the interior $\mathring{\mathcal{G}}_h$ it is:

$$\begin{aligned} &\mathbb{P}\left(\sup_{x\in\mathring{\mathcal{G}}_h} |\widehat{\pi}(x) - \pi(x)| \geq c_n\right) \\ &\leq \mathbb{P}\left(\sup_{x\in\mathring{\mathcal{G}}_h} |\widehat{\pi}(x) - \mu(K_{x,h})| + \sup_{x\in\mathring{\mathcal{G}}_h} |\mu(K_{x,h}) - \pi(x)| \geq c_n\right) \\ &\leq \mathbb{P}\left(\sup_{x\in\mathring{\mathcal{G}}_h} |\widehat{\pi}(x) - \mu(K_{x,h})| \geq \frac{c_n}{2}\right) + \mathbb{P}\left(\sup_{x\in\mathring{\mathcal{G}}_h} |\mu(K_{x,h}) - \pi(x)| \geq \frac{c_n}{2}\right) \\ &= S_1^i + S_2^i \,, \end{aligned}$$

Since $\mathcal{G} \subset \mathbb{R}^d$ is compact and hence bounded, we have to be careful at the boundary. For the $C_1 h$-ring-boundary $\partial\mathcal{G}_h$ we get:

$$\begin{aligned} &\mathbb{P}\left(\sup_{x\in\partial\mathcal{G}_h} \left|\widehat{\pi}(x) - \pi(x)\int_{\mathcal{G}} K_{h,x}(u)\mathrm{d}u\right| \geq c_n\right) \\ &\leq \mathbb{P}\left(\sup_{x\in\partial\mathcal{G}_h} |\widehat{\pi}(x) - \mu(K_{x,h})| \geq \frac{c_n}{2}\right) + \mathbb{P}\left(\sup_{x\in\partial\mathcal{G}_h} \left|\mu(K_{x,h}) - \pi(x)\int_{\mathcal{G}} K_{h,x}(u)\mathrm{d}u\right| \geq \frac{c_n}{2}\right) \\ &= S_1^b + S_2^b \,, \end{aligned}$$

For the bias parts S_2^i and S_2^b, standard analysis with the usual kernel arguments carries over. Since we have for x in the interior $\mathring{\mathcal{G}}_{C_1 h}$ that $\int_{\mathcal{G}} K_{h,x}(u)\mathrm{d}u = 1$, we can treat S_2^i and S_2^b together:

$$\begin{aligned}
\mu(K_{x,h}) - \pi(x)\int_{\mathcal{G}} K_{h,x}(u)\mathrm{d}u &= \int_{B_x(C_1 h)\cap\mathcal{G}} (\pi(x+hu)-\pi(x))K(u)\mathrm{d}u \\
&= \frac{\mathrm{d}}{\mathrm{d}x}(\pi(x))\, h \int_{B_x(C_1 h)\cap\mathcal{G}} uK(u)\mathrm{d}u + O(h^2) \\
&= \begin{cases} O(h^2) & \text{for } x \in \mathring{\mathcal{G}}_h \\ O(h) & \text{for } x \in \partial\mathcal{G}_h \end{cases},
\end{aligned}$$

since for $x \in \partial\mathcal{G}_h$ the ball $B_x(C_1 h)$ is not entirely in $\mathcal{G}$. Thus the with symmetry of the kernel the integral is not zero as in the case for x in the interior.

Now treat the stochastic term $S_1 = \mathbb{P}\left(\sup_{x\in\mathcal{G}} |\widehat{\pi}(x) - \mu(K_{x,h})| \geq \frac{c_n}{2}\right)$. Here we do not have to distinguish between cases of x on the boundary or not. As $\mathcal{G}$ is compact, there exists a cover of $l(n)$ open balls $I_1, \ldots, I_k, \ldots, I_{l(n)}$ with radius $\frac{c_1}{l(n)^{1/d}}$ for an appropriate constant c_1 and with centers in x_k and $\bigcup_{k=1}^{l(n)} I_k \supseteq \mathcal{G}$. The maximal distance attainable between elements inside one of the balls is the diameter:

$$\max_{a,b\in I_k} \|a-b\| \leq \frac{2c_1}{l(n)^{1/d}} = \frac{c}{l(n)^{1/d}} \text{ for all } k \in \{1,\ldots,l(n)\} \tag{A.20}$$

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x\in\mathcal{G}} |\widehat{\pi}(x) - \mu(K_{x,h})| \geq \frac{c_n}{2}\right) \\
&= \mathbb{P}\left(\max_{1\leq k\leq l(n)} \sup_{x\in\mathcal{G}\cap I_k} |\widehat{\pi}(x) - \mu(K_{x,h})| \geq \frac{c_n}{2}\right) \\
&\leq \mathbb{P}\left(\max_{1\leq k\leq l(n)} \sup_{x\in\mathcal{G}\cap I_k} |\widehat{\pi}(x) - \widehat{\pi}(x_k)| \geq \frac{c_n}{6}\right) + \mathbb{P}\left(\max_{1\leq k\leq l(n)} |\widehat{\pi}(x_k) - \mu(K_{x_k,h})| \geq \frac{c_n}{6}\right) \\
&\quad + \mathbb{P}\left(\max_{1\leq k\leq l(n)} \sup_{x\in\mathcal{G}\cap I_k} |\mu(K_{x_k,h}) - \mu(K_{x,h})| \geq \frac{c_n}{6}\right) \\
&= Q_1 + Q_2 + Q_3
\end{aligned}$$

The first and the third term, Q_1 and Q_3, are easy to handle and therefore treated

first. Look at Q_1:

$$\begin{aligned}
\sup_{x\in\mathcal{G}\cap I_k}|\widehat{\pi}(x)-\widehat{\pi}(x_k)| &= \frac{1}{T(n)}\sup_{x\in\mathcal{G}\cap I_k}\left|\sum_{i=1}^{n}(K_{h,x}(X_i)-K_{h,x_k}(X_i))(X_i)\right| \\
&= \frac{1}{\pi(\mathcal{G})}\int_{\mathcal{G}}\sup_{x\in\mathcal{G}\cap I_k}|K_{h,x}(u)-K_{h,x_k}(u)|\,\pi(u)\mathrm{d}u \quad \mathbb{P}-a.s. \\
&\leq \sup_{x\in\mathcal{G}\cap I_k}\frac{L}{h_n^{d+1}}\|x-x_k\| \quad \mathbb{P}-a.s. \\
&\leq \frac{Lc_1}{h_n^{d+1}l(n)^{1/d}} \quad \mathbb{P}-a.s.\ .
\end{aligned}$$

The first $\mathbb{P}-a.s$ relation is a consequence of the quotient limit theorem (A.11), while the inequalities thereafter follow directly from (A.20) and the Lipschitz assumption on the kernel.

Since the integral operator and everything inside is continuous, obviously we also get $\max_{1\leq k\leq l(n)}\sup_{x\in\mathcal{G}\cap I_k}|\mu(K_{x_k,h})-\mu(K_{x,h})| = O\left(\frac{1}{h_n^{d+1}l(n)^{1/d}}\right)$. Thus when imposing $c := O\left(\frac{1}{h_n^{d+1}l(n)^{1/d}}\right)$, then Q_1 and Q_3 are $o_P(1)$.

Q_2, the second term, however, needs some extra considerations: On the grid of the x_k-balls we can simplify the expression by the triangle inequality and get an upper bound where "the maximum is outside the measure" and therefore easier tractable:

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq k\leq l(n)}|\widehat{\pi}(x_k)-\mu(K_{x_k,h})|\geq\frac{c_n}{6}\right) \\
&= \mathbb{P}\left(\max_{1\leq k\leq l(n)}\left|\frac{1}{T(n)}\sum_{i=1}^{n}K_{x_k,h}-\mu(K_{x_k,h})\right|\geq\frac{c_n}{6}\right) \\
&\leq l(n)\cdot\sup_{x\in\mathcal{G}}\mathbb{P}\left(\left|\frac{1}{T(n)}\sum_{i=1}^{n}K_{x,h}-\mu(K_{x,h})\right|\geq\frac{c_n}{6}\right) \\
&\leq l(n)\cdot\sup_{x\in\mathcal{G}}\mathbb{P}\left(\left|\frac{1}{T(n)h^d}\left(\sum_{k=0}^{T(n)-1}W_{k,x}+U_{(n)}\right)\right|\geq c_n'\right)
\end{aligned}$$

$$\leq \quad l(n) \cdot \sup_{x \in \mathcal{G}} \mathbb{P}\left(\frac{1}{T(n)h^d} \left| \sum_{k=0}^{T(n)-1} W_{k,x} \right| \geq c_n'' \right) \ ,$$

where the second to last inequality follows because of with $c_n' = \frac{c_n \pi_s(C)}{6}$. Since c_n'' differs from c_n only by a constant, we continue notation with c_n. Furthermore the sum is rewritten in terms of the centered split chain components $W_{k,x}' = U_{x,k} - \mu(K_x)$ where $U_{x,k}$ is the k–th component of the split chain of K_x. Thus it is:

$$\begin{aligned} U_{x,k} &= \sum_{j=\tau_k+1}^{\tau_{k+1}} K_x(X_j) \quad \text{for } k = 0, \ldots, T(n)-1 \\ U_n &= \sum_{j=\tau_{T(n)}+1}^{n} K_x(X_j) \end{aligned}$$

As parts of a split chain all $W_{x,k}$ are iid W_x for a given $x \in \mathcal{G}$. And obviously from the definition it is $\mathbb{E}(W_x) = 0$.

When dealing with $\mathbb{P}\left(\left| \frac{1}{T(n)} \sum_{k=0}^{T(n)-1} W_{k,x} \right| \geq \frac{c_n}{6} \right)$ the main difficulty stems from the fact that the norming $T_C(n)$ is stochastic and not independent of W_x. It is

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{T(n)h^d} \left| \sum_{k=0}^{T(n)-1} W_{k,x} \right| \geq c_n \right) \\ &\leq \quad \mathbb{P}\left(\frac{1}{T(n)} \left| \sum_{k=0}^{T(n)-1} W_{k,x} \right| \geq c_n h^d, 1 \leq T(n) \leq \delta_n \right) \\ &\qquad\qquad + \mathbb{P}\left(\frac{1}{T(n)h^d} \left| \sum_{k=0}^{T(n)-1} W_{k,x} \right| \geq c_n, T(n) > \delta_n \right) \\ &\leq \quad \mathbb{P}\left(\left| \sum_{k=0}^{T(n)-1} W_{k,x} \right| \geq c_n \delta_n h^d \right) \end{aligned}$$

The last inequality follows since for the first term

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{T(n)}\left|\sum_{k=0}^{T(n)-1} W_{k,x}\right| \geq c_n h^d, 1 \leq T(n) \leq \delta_n\right) \\
&\leq \mathbb{P}\left(T(n) \leq \frac{\delta_n \kappa}{h^d c_n}, 1 \leq T(n) \leq \delta_n\right) \\
&\leq \mathbb{P}\left(T(n) \leq \delta_n\right) = \mathbb{P}\left(T(n) \leq l_n \frac{\delta_n}{l_n}\right) \leq \frac{2 l_n a(\frac{\delta_n}{l_n}) - 1}{(l_n+1) a(\frac{\delta_n}{l_n}) + 1} \to 0
\end{aligned}$$

with $\kappa = \max_{1 \leq k \leq \delta_n} |U_k - \mu| < \infty$ and $l_n \to 0$ chosen such that $\frac{2 l_n a(\frac{\delta_n}{l_n})-1}{(l_n+1)a(\frac{\delta_n}{l_n})+1} \to 0$ at polynomial rate $n^{-\alpha}$ with α such that $l(n) n^{-\alpha} \sim n^{-\alpha'}$ and $\alpha' > 1$. The last inequality follows from Theorem 2.1. in [Chen, 1999a].

Treat the remaining second term:

$$\begin{aligned}
\xi_k &= U_k \mathbf{1}_{|U_k| \leq R} - \mathbb{E}(U_k \mathbf{1}_{|U_k| \leq R}) \\
\eta_k &= U_k \mathbf{1}_{|U_k| > R} - \mathbb{E}(U_k \mathbf{1}_{|U_k| > R})
\end{aligned}$$

with $R > 0$ large enough such that $\mathbb{E}\xi^2 := \sigma_R^2 > 0$. Then $W_k = \xi_k + \eta_k$.

For each n, $T(n)$ is a stopping time w.r.t to the iid sequence ξ_k. With a standard maximal inequality for martingales (see e.g. Theorem 2.1., Chapter2, [Hall and Heyde, 1980])

$$\begin{aligned}
\mathbb{P}\left(\left|\sum_{k=0}^{T(n)-1} \xi_k\right| \geq c_n \delta_n h^d\right) &\leq \mathbb{P}\left(\max_{l \leq T(n)} \left|\sum_{k=0}^{l} \xi_k\right| \geq c_n \delta_n h^d\right) \\
&\leq e^{-\lambda L_2 a(n) \theta} \mathbb{E}\left(\exp \theta \left|\sum_{k=0}^{T(n)} \frac{\xi_k}{\sqrt{a(\frac{n}{L_2 a(n)})}}\right|\right) \quad \text{(A.21)}
\end{aligned}$$

For $c_n \delta_n h^d = \lambda \sqrt{a(\frac{n}{L_2 a(n)})} L_2 a(n)$ with $\lambda > 0$ and $\theta > 0$ arbitrary. The separation into weighting and exponential factors seems at this stage somehow arbitrary. But their choice is perfectly balanced in view of the following assessment.

With Lemma 2.2. in [Chen, 2000], which uses that the left hand side of the expression below is a martingale where the optional stopping theorem can be

applied to:

$$\begin{aligned}\mathbb{E}\left(\exp\theta\left|\sum_{k=0}^{T(n)-1}\frac{\xi_k}{\sqrt{a(\frac{n}{L_2a(n)})}}-(T(n)+1)L_n(s,\theta)\right|\right)\\ \leq\ \exp\left(s(1+\varepsilon)L_2a(n)+\frac{R\theta}{\sqrt{a(\frac{n}{L_2a(n)})}}\right)\\ =\ (1+o(1))\exp\left(s(1+\varepsilon)L_2a(n)\right)\end{aligned}\tag{A.22}$$

with $s>\Lambda_\beta(\theta)$, $\varepsilon>0$ where $\Lambda_\beta(\theta)=\left(\Gamma(\beta+1)\frac{\theta^2\sigma_R^2}{2}\right)^{1/\beta}$ for $0<\beta\leq 1$ and

$$\begin{aligned}L_n(s,\theta)\ &=\ \log\mathbb{E}\left(\exp\frac{\theta\xi}{\sqrt{a(\frac{n}{L_2a(n)})}}-\frac{sL_2a(n)}{n}\min\left(\tau_\alpha,n\varepsilon\right)\right)\\ &\leq\ \log\mathbb{E}\left(\exp\frac{\theta\xi}{\sqrt{a(\frac{n}{L_2a(n)})}}-\frac{sL_2a(n)}{n}\tau_\alpha\right)+O(e^{-\varepsilon n}\mathbb{P}\left(\tau_\alpha\geq\varepsilon n\right))\\ &\sim\ \frac{1}{\sqrt{a(\frac{n}{L_2a(n)})}}\left(\frac{\theta^2\sigma_R^2}{2}-\frac{s^\beta}{\Gamma(\beta+1)}\right)\quad\text{for } n\to\infty\,.\end{aligned}\tag{A.23}$$

Thus with (A.22) and (A.23) for n large and $s\to\Lambda_\beta(\theta)$, $\varepsilon\to 0$:

$$\mathbb{E}\left(\exp\theta\left|\sum_{k=0}^{T(n)-1}\frac{\xi_k}{\sqrt{a(\frac{n}{L_2a(n)})}}\right|\right)\ \leq\ \exp\left(\Lambda_\beta(\theta)L_2a(n)\right)\,.$$

The inequality above is also true in the case $\beta=0$ if we set $\Lambda_\beta(0)=\begin{cases}0 & \text{if } \theta^2\sigma_R^2\leq 2\\ \infty & \text{if } \theta^2\sigma_R^2>2\end{cases}$ and take $s\to 0$ in (A.22) and (A.23).

In total putting the above inequality into (A.21) and since $\xi_{T(n)}\leq 2R$

$$\mathbb{P}\left(\left|\sum_{k=0}^{T(n)}\xi_k\right|\geq\lambda\sqrt{a(\frac{n}{L_2a(n)})}L_2a(n)\right)\ \leq\ e^{-\Lambda_\beta^\star(\lambda)L_2a(n)}\tag{A.24}$$

with $\Lambda^\star_\beta(\lambda) = \sup_{\theta>0}(\lambda\theta - \Lambda_\beta(\theta)) = (2-\beta)\left(\frac{\beta^\beta\lambda^2}{2\Gamma(\beta+1)\sigma_R^2}\right)^{\frac{1}{2-\beta}}$ by solving the optimization for all $\lambda \in \mathbb{R}$ and with the convention $0^0 = 1$ in the case $\beta = 0$. Set $\nu_k := \inf\left\{n : a\left(\frac{n}{L_2(a(n))}\right) \geq k^{2k}\right\}$. Then $L_2(a(\nu_k)) = \log(2k\log k)$. Thus the right–hand side of (A.24) is summable over k for $\lambda > \sqrt{\frac{2\Gamma(\beta+1)\sigma_R^2}{(2-\beta)^{(2-\beta)}\beta^\beta}}$. With Borel-Cantelli lemma we find :

$$\limsup_{n\to\infty}\left|\sum_{k=0}^{T(n)}\xi_k\right| \leq \lambda\sqrt{a(\frac{n}{L_2a(n)})}L_2a(n) \quad \text{a.s.} \tag{A.25}$$

For η_k the situation is more standard. According to the Hartmann-Winter's law of iterated logarithm it is:

$$\limsup_{n\to\infty}\max_{k\leq n-1}\left|\sum_{k=0}^{l}\eta_k\right| \leq \sqrt{nL_2(n)}\sqrt{\mathbb{E}\eta^2} \quad \text{a.s. .}$$

This implies

$$\limsup_{n\to\infty}\max_{k\leq T(n)-1}\left|\sum_{k=0}^{l}\eta_k\right| \leq \sqrt{T(n)L_2(T(n))}\sqrt{\mathbb{E}\eta^2} \quad \text{a.s. .}$$

But since $\limsup_{n\to\infty} T(n) \leq \kappa\sqrt{a(\frac{n}{L_2a(n)})}L_2a(n)$ a.s. with $\kappa > 0$ (Theorem 2.2 in [Chen, 1999a]), we get:

$$\limsup_{n\to\infty}\max_{k\leq T(n)-1}\left|\sum_{k=0}^{l}\eta_k\right| \leq \sqrt{2\kappa\mathbb{E}\eta^2}\sqrt{a(\frac{n}{L_2a(n)})}L_2a(n) \quad \text{a.s. .}$$

Take $R\to\infty$, which yields $\sigma_R^2 \to \sigma^2$ but $\sigma_R^2 \leq \sigma^2$ and $\mathbb{E}\eta^2 \to 0$. So finally with

$$\begin{aligned} c_n\delta_n h^d &> 6\sqrt{\frac{2\Gamma(\beta+1)\sigma^2}{(2-\beta)^{(2-\beta)}\beta^\beta}}\sqrt{a(\frac{n}{L_2a(n)})}L_2a(n) \\ &= C_\beta\sigma\sqrt{a(\frac{n}{L_2a(n)})}L_2a(n)\ , \end{aligned} \tag{A.26}$$

we find that for n large enough it is

$$l(n)\cdot\sup_{x\in\mathcal{G}}\mathbb{P}\left(\frac{1}{T(n)}\left|\sum_{i=0}^{n}W_{x,i}\right| \geq \frac{c_n}{6}\right) = 0\ ,$$

due to (A.25). In particular the probabilities are summable, i.e.

$$\sum_{k=1}^{\infty} l(n)\mathbb{P}\left(\frac{1}{T(n)}\left|\sum_{i=0}^{n} W_{x,i}\right| \geq \frac{c_n}{6}\right) < \infty \, .$$

Thus with the Borel–Cantelli lemma we can conclude that the entire term S_1 is $o(1)$ for appropriate choices of c_n, δ_n, h in accordance with (A.26).

For all terms including S_2 to vanish, we need additionally $(l(n))^{1/d}h^{d+1} \to \infty$ and $h^d\sqrt{a(\frac{n}{L_2a(n)})}L_2a(n) \to \infty$. Choose $\delta_n = a\left(\frac{n}{L_2a(n)}\right)L_2a(n)^2 < n$. Then the condition for S_2 and (A.26) are simultaneously satisfied for $c_n' = \frac{1}{\sqrt{a(\frac{n}{L_2a(n)})}L_2a(n)h^d}$. Combining this with the bias term, we find the stated final rates. □

Remark A.1.

1. Under the assumption that for any $x \in \mathcal{G}$ there exists a measure ϕ such that

$$\mathbb{P}_x(X_m \in A) \geq \lambda\phi(A) \tag{A.27}$$

for any $A \subset C$, we could also work with a Markov process version of Hoeffding's inequality obtained by Glynn and Ormoneit [Glynn and Ormoneit, 2002]. Together with the usual blocking argument we would find:

$$\mathbb{P}\left(\frac{1}{T_C(n)h^d}\sum_{i=0}^{n} W_{k,x} \geq \frac{c_n}{6}\right) \leq c(a(n)h^d)^{-2} \tag{A.28}$$

with $c > 0$ constant and a faster rate c_n. But the uniformity imposed by (A.27) is quite restrictive. It restricts the set of β–recurrent processes significantly to the positive recurrent ones only.

2. For $\beta = 1$, a refinement of the law of iterated logarithm inequality (A.25) can be found directly in [Chen, 1999a]:

$$\limsup_{n\to\infty} \frac{\sum_{k=1}^{n} f(X_k)}{\sqrt{2nL_2n}} = \sigma_f \quad \text{a.s.} \, . \tag{A.29}$$

Lemma A.2 is for the full dimensional β–null Harris recurrent process. For the SBE algorithm, however, only univariate and bivariate results are of importance.

Therefore the following corollary is stated for these cases in the generalized SBE algorithm, extending the result of A.2 to small sets. Assume w.l.o.g. that $\pi(\mathbf{1}_{\mathcal{G}}) = 1$. Then $\pi_{\mathcal{G}}(x) = \frac{\pi(x)}{\pi(\mathbf{1}_{\mathcal{G}})} = \pi(x)$.

Corollary A.2. *Assume that Assumptions 2.1 and 4.3 hold. Set* $\beta_2 := \beta^{jk} + \varepsilon_2$ *with* ε_2 *very small such that* $\max\{m : L_K(m) \geq m^{\varepsilon_2}\} >> 1$ *with* L_K *the slowly varying function from the* β*–recurrence condition in* (2.4) *with respect to the bivariate Kernel function* K^2*. Let the bandwidth* $h \to 0$ *such that* $n^{\frac{\beta_2}{2}}(L_2 n^{\beta_2})^{1-\beta_2} h^2 \to \infty$ *and set* $h\, l(n)^{1/2} = n^{\frac{\beta_2}{2}}(L_2 n^{\beta_2})^{1-\beta_2}$ *. Then*

$$\sup_{x^{j,k} \in \mathring{\mathcal{G}}_h^{j,k}} \left| \widehat{\pi}_{jk,\mathcal{G}_{jk}}(x^{jk}) - \pi_{jk,\mathcal{G}^{jk}}(x^{jk}) \right| = O_P\left(h^2 + \frac{1}{n^{\frac{\beta_2}{2}}(L_2 n^{\beta_2})^{1-\beta_2} h^2} \right) \quad \text{(A.30)}$$

$$\sup_{x^{j,k} \in \partial \mathcal{G}_h^{j,k}} \left| \widehat{\pi}_{jk,\mathcal{G}_{jk}}(x^{jk}) - \pi_{jk,\mathcal{G}^{jk}}(x^{jk}) \right| = O_P\left(h + \frac{1}{n^{\frac{\beta_2}{2}}(L_2 n^{\beta_2})^{1-\beta_2} h^2} \right) . \quad \text{(A.31)}$$

and

$$\sup_{x^j \in \mathring{\mathcal{G}}_{j,h}^{(k)}} \left| \widehat{\pi}^{(k)}_{j,\mathcal{G}_j^{(k)}}(x^j) - \pi_{j,\mathcal{G}_j^{(k)}}(x^j) \right| = O_P\left(h^2 + \frac{1}{n^{\frac{\beta_2}{2}}(L_2 n^{\beta_2})^{1-\beta_2} h} \right) \quad \text{(A.32)}$$

$$\sup_{x^{j,k} \in \partial \mathcal{G}_{j,h}} \left| \widehat{\pi}^{(k)}_{j,\mathcal{G}_j^{(k)}}(x^j) - \pi^{(k)}_{j,\mathcal{G}_j^{(k)}}(x^j) \right| = O_P\left(h + \frac{1}{n^{\frac{\beta_2}{2}}(L_2 n^{\beta_2})^{1-\beta_2} h} \right) . \quad \text{(A.33)}$$

Proof. The proof is an immediate consequence of the previous lemma and (A.5). □

Remark A.2. Analogous results hold for $\widehat{\pi}_{jk,\mathcal{G}_{jk}^f}$ and $\widehat{\pi}_{j,\mathcal{G}_j^f}$ with full–dimensional type of nonstationarity β under Assumptions 2.1 and 4.1.

Uniform consistency of the regression function

The one–dimensional pilot smoothers can be decomposed into a bias and a stochastic part as our underlying model (1.1) has an additively separable error term.

$$\widehat{m}_j(x^j) \quad = \quad \frac{\sum_{i=1}^n K_{h,x^j}(X_i^j) Y_i}{\sum_{i=1}^n K_{h,x^j}(X_i^j)}$$

$$
\begin{aligned}
&= \left(\frac{\sum_{i=1}^n K_{h,x^j}(X_i^j) m_i(X_i)}{\sum_{i=1}^n K_{h,x^j}(X_i^j)} \right) + \left(\frac{\sum_{i=1}^n K_{h,x^j}(X_i^j) \varepsilon_i}{\sum_{i=1}^n K_{h,x^j}(X_i^j)} \right) \\
&=: \widehat{m}_j^B(x^j) + \widehat{m}_j^A(x^j)
\end{aligned}
$$

Obviously $\widehat{m}_j^A(x^j)$ is the stochastic part whereas $\widehat{m}_j^B(x^j)$ is the bias or expectation part.

When starting the SBE algorithm with these pilot estimates we find that the resulting $\widetilde{m}_j(x^j)$ preserve the additive structure of separate bias and stochastic part. Thus we have

$$\widetilde{m}_j(x^j) = \widetilde{m}_j^B(x^j) + \widetilde{m}_j^A(x^j) \,, \tag{A.34}$$

where each of the parts $\widetilde{m}_j^s(x^j)$ with $s \in \{A, B\}$ separately solves the defining equations (3.2):

$$
\begin{aligned}
\widetilde{m}_j^s(x^j) &= \widehat{m}_j^s(x^j) - \widetilde{m}_{0,j}^s - \sum_{k \neq j} \int \widetilde{m}_k^s(x^k) \frac{\widehat{\pi}_{j,k}(x^j, x^k)}{\widehat{\pi}_j(x^j)} \mathrm{d}x^k \qquad \text{(A.35)} \\
\text{with for } j \neq k: \\
\widetilde{m}_{0,j}^s &= \frac{\int \widehat{m}_j^s(x^j) \widehat{\pi}_j(x^j) \mathrm{d}x^j}{\int \widehat{\pi}_j(x^j) \mathrm{d}x^j} \,.
\end{aligned}
$$

Definition A.1. Instead of the usual conditional expectation, we need an adaptation which only involves one and two dimensional covariates. The notation follows from (A.14)

$$(Am)_j(x^j) := m_j(x^j) + \sum_{k \neq j} \int_{\mathcal{G}^k} m_k(x^k) \frac{\pi_{jk}(x^{jk})}{\pi_j(x^j)} \, \mathrm{d}x^k \tag{A.36}$$

In nonstationary smooth backfitting, Nadaraya–Watson estimates are at least of two dimensional nonstationary type, i.e. $\widehat{m}_j^{(k)}$ or $\widehat{m}_j^f$ are of interest.

Lemma A.3 (Uniform rate of the bias part)**.** *Let either Assumptions 1-3 or Assumptions 1,2, and 3* hold.*

$$\begin{aligned}\sup_{x^j \in \mathring{\mathcal{G}}^f_{j,h}} \left|\widehat{m}^{f,B}_j(x^j) - (Am^f)_j(x^j)\right| &= O_P\left(h^2 + \frac{1}{n^{\beta/2}(L_2 n^\beta)^{1-\beta}h}\right)\\ \sup_{x^j \in \partial\mathcal{G}^f_{j,h}} \left|\widehat{m}^{f,B}_j(x^j) - (Am^f)_j(x^j)\right| &= O_P\left(h + \frac{1}{n^{\beta/2}(L_2 n^\beta)^{1-\beta}h}\right)\end{aligned}$$

Proof. With standard kernel calculations it is $\mathbb{E}\left(\widehat{m}^B_j(x^j)|X^j_1,\ldots,X^j_n\right) = (Am)_j(x^j) + O(h^2)$ uniformly in the interior $\mathring{\mathcal{G}}^j$ and $\mathbb{E}\left(\widehat{m}^B_j(x^j)|X^j_1,\ldots,X^j_n\right) = (Am)_j(x^j) + O(h)$ uniformly in $\partial\mathcal{G}^j$.
For the exponential bound on $\widehat{m}^B_j(x^j) - \mathbb{E}\left(\widehat{m}^B_j(x^j)|X^j_1,\ldots,X^j_n\right)$ we need to show uniform convergence of centered versions $S^{i,l\star}_j$ of the following expressions:

$$S^{i,l}_j(x^j) = K_{h,x^j}(X^j_i)(X^j_i - x^j)^l m^{(l)}_j(x^j)\,\mathrm{d}x^j$$

with $l \in \{1,2\}$. The centering is with respect to the appropriate mean, i.e., $S^{i,l\star}_j(x^j) = S^{i,l}_j(x^j) - \widehat{\pi}_j(x^j)\mu(S^{i,l}_j(x^j))$. Scaled summing of these random variables is denoted by $s^{l\star}_j = (T^j(n))^{-1}\sum_{i=1}^n S^{i,l\star}_j(x^j)$.
Everything follows directly along the steps of the previous lemma if we assume as $N_{h,K}(x) = \{u|\,\|u-x\| \le C_1 h\}$ is small because $\mathcal{G}$ is small. Then m is special (4.1) and the necessary moment bounds follow from Karlsen et al. [2007] Theorem 3.4. □

Remark A.3. Under Assumptions 2.1,4.3 and 4.4 or 2.1,4.3 and 4.4* , we get with $\widehat{m}^{(k)}_j$ rates with bivariate β_{jk} on $\mathcal{G}_{jk}$

Definition A.2. We use the following short hand notation:

$$\mu^f_{(j\varepsilon)}(h\otimes g) := \iint h(u)g(v)\pi^f_{j\varepsilon}(u,v)\,\mathrm{d}u\,\mathrm{d}v \tag{A.37}$$

$$\mu^{(k)}_{(j\varepsilon)}(h\otimes g) := \iint h(u)g(v)\pi^{(k)}_{j\varepsilon}(u,v)\,\mathrm{d}u\,\mathrm{d}v\,. \tag{A.38}$$

And id_ε is the identity on the support of ε, i.e., $\mathrm{id}_\varepsilon(u) = u$ for $u \in \mathcal{G}_0$.

Lemma A.4 (Uniform rate of the variance part of the Nadaraya–Watson–type estimator)**.** *Let Assumptions 2.1 - 4.2 hold.*

$$\sup_{x^j \in \mathring{\mathcal{G}}^f_{j,h}} \left| \widehat{m}^{f,A}_j(x^j) - \mu^f_{(j\varepsilon)}(K_{x^j,h}(\cdot) \otimes id_\varepsilon) \frac{T^f_{j\varepsilon}(n)}{\widehat{L}^f_j(x^j)} \right| = O_P\left(\frac{L_2 n^\beta}{(n^\beta h)^{1/2}} \right)$$

$$\sup_{x^j \in \partial\mathcal{G}^f_{j,h}} \left| \widehat{m}^{f,A}_j(x^j) - \mu^f_{(j\varepsilon)}(K_{x^j,h}(\cdot) \otimes id_\varepsilon) \frac{T^f_{j\varepsilon}(n)}{\widehat{L}^f_j(x^j)} \right| = O_P\left(\frac{L_2 n^\beta}{(n^\beta h)^{1/2}} \right)$$

Proof. Need an exponential bound on $\frac{\sum_{i=0}^n K_{h,x^j}(X_i^j)\varepsilon_i}{\sum_{i=0}^n K_{h,x^j}(X_i^j)}$. Therefore the independent split chain parts for the sum in the numerator are U_k's for the bivariate chain (X^j, ε). Then the argument follows along the lines of lemma A.18 above, where $\beta_{j\varepsilon}$ is for the compound chain (X^j, ε) and the truncation technique will be applied separately to the X^j and ε part. Convergence with β^j instead of $\beta^{j\varepsilon}$ follows with Lemma 6.1. in Karlsen et al. [2007]. □

Remark A.4. In general, the stochastic bias term $\mu^f_{(j\varepsilon)}(K_{x^j,h}(\cdot)\otimes \mathrm{id}_\varepsilon)\frac{T^f_{j\varepsilon}(n)}{\widehat{L}^f_j(x^j)}$ is $o_p(1)$ (see (6.23) in Karlsen et al. [2007]). Under Assumption 4.2.3 and with bandwidth $h < n^{1/5\beta+\varepsilon}$ in Theorem 4.1 the term vanishes.

Remark A.5. Under Assumptions 2.1,4.3 and 4.4, we get with $\widehat{m}^{(k)}_j$ rates with bivariate β_{jk} on $\mathcal{G}_{jk}$ and a stochastic bias $\mu^{(k)}_{(j\varepsilon)}(K_{x^j,h}(\cdot) \otimes \mathrm{id}_\varepsilon)\frac{T^{(k)}_{j\varepsilon}(n)}{\widehat{L}^{(k)}_j(x^j)}$.

If ε and X^j are independent, or only asymptotically independent, then it is $\pi_{j\varepsilon} = \pi_j \cdot \pi_\varepsilon$. Thus $\mu_{(j\varepsilon)}(K_{x^j,h}(X) \otimes \mathrm{id}_\varepsilon)\frac{T^{j\varepsilon}(n)}{\widehat{L}^j(x^j)} = 0$ under Assumption 3*.

Lemma A.5. *[Asymptotic distribution of the variance part] Let Assumptions 2.1,4.1 and 4.2 hold. For $n \to \infty$, $h \to 0$ let $hn^{\beta-\varepsilon} \to \infty$. Then*

$$\sqrt{h\widehat{L}^f_j(x)} \left(\widehat{m}^{f,A}_j(x^j) - \mu^f_{(j\varepsilon)}(K_{x^j,h}(\cdot) \otimes id_\varepsilon) \frac{T^f_{j\varepsilon}(n)}{\widehat{L}^f_j(x^j)} \right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2}\sigma^f_j(x^j)\right).$$

Let Assumptions 2.1,4.3 and 4.4 hold. For $n \to \infty$, $h \to 0$ let $hn^{\beta_{jk}-\varepsilon} \to \infty$. Then

$$\sqrt{h\widehat{L}^{(k)}_j(x)} \left(\widehat{m}^{(k),A}_j(x^j) - \mu^{(k)}_{(j\varepsilon)}(K_{x^j,h}(\cdot) \otimes id_\varepsilon) \frac{T^{(k)}_{j\varepsilon}(n)}{\widehat{L}^{(k)}_j(x^j)} \right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2}\sigma^{(k)}_j(x^j)\right).$$

with $\sigma_j^f(x^j) = \int \varepsilon^2 \frac{\pi_{j\varepsilon}^f(x^j,\varepsilon)}{\pi_j^{f\varepsilon}(x^j)} \mathrm{d}\varepsilon$ *and* $\sigma_j^{(k)}(x^j) = \int \varepsilon^2 \frac{\pi_{j\varepsilon}^{(k)}(x^j,\varepsilon)}{\pi_j^{(k)\varepsilon}(x^j)} \mathrm{d}\varepsilon$.

Proof. The proof directly follows from [Karlsen et al., 2007] Theorem 6.1 and Theorem 5.5. □

As before, if we choose $n^{-(\beta+\varepsilon)} < h < n^{-1/5(\beta+\varepsilon)}$ or $n^{-(\beta_{jk}+\varepsilon)} < h < n^{-1/5(\beta_{jk}+\varepsilon)}$ the bias terms is negligible. If ε is stationary linear there exists a simplified version of CLT which has moment bounds restrictions familiar to the ones in the purely stationary case.

Lemma A.6. *[Asymptotic distribution of the variance part under independence] Let Assumptions 2.1,4.1, and 4.2* hold. For* $n \to \infty$, $h \to 0$ *let* $hn^{\beta\delta-\varepsilon} \to \infty$ *with* $\delta = \frac{1}{2-1/(k+1)}$ *and* k *from the moment conditions in Assumptions 4.2*. Then*

$$\sqrt{h\widehat{L}_j^f(x)} \cdot \widehat{m}_j^{f,A}(x^j) \xrightarrow{d} \mathcal{N}\left(0, \frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2}\sigma_\varepsilon\right)$$

where $\sigma_\varepsilon = \int \varepsilon^2 \pi_\varepsilon(\varepsilon)\, d\varepsilon$.

Proof. The proof directly follows from [Karlsen et al., 2007] Theorem 3.5.. □

Remark A.6. For Assumptions 2.1,4.3, and 4.4* Lemma A.5 holds for $\widehat{m}_j^{(k),A}$ analogously.

A.2.3 Proofs of the Theorems

In the previous subsection the major technical work has been done. With these lemmata, requirements (A1)-(A6), (A8), and (A9) of Mammen et al. in [Mammen et al., 1999] can be shown to be met. For the SBE procedure to work and to lead to a well defined asymptotic distribution and bias behavior these technical conditions have to be fulfilled. For how they fully determine convergence and asymptotic properties of the backfitting operator see [Mammen et al., 1999]. If done so, the proof of any of the stated theorems directly follows from the proofs and the reasoning in [Mammen et al., 1999], page 1470 ff.

For the rest of this section let either the bundle of assumptions for Theorem 4.1 and 4.2 or Theorem 4.3 and 4.4 hold. To treat all nonstationary smooth backfitting cases at once, set $\pi_j = \pi_j^f$ and $\pi_{jk} = \pi_{jk}^f$ or $\pi_j = \pi_j^{(k)}$ and $\pi_{jk} = \pi_{jk}$, depending on whether we have a full or just pairwise Harris recurrent framework.

Assumption (A1)

We want the backfitting projection operator positive self–adjoint and compact. This is why requirement (A1) has to be fulfilled:

(A1) For all $j \neq k$ it holds that:

$$\int \frac{\pi_{j,k}^2(x^j, x^k)}{\pi_j(x^j)\pi_k(x^k)} \mathrm{d}x^j \mathrm{d}x^k \leq \infty \, . \tag{A.39}$$

Proof. With assumption (B1b) we formally get: $\inf \pi_j(x^j) \geq c_1 > 0$ and $\sup \pi_{j,k}(x^j, x^k) \leq c_2 < \infty$. Thus evidently

$$\int \frac{\pi_{j,k}^2(x^j, x^k)}{\pi_j(x^j)\pi_k(x^k)} \mathrm{d}x^j \mathrm{d}x^k \leq \frac{c_2^2}{c_1} < \infty \, .$$

□

In order to establish (A2),(A4) and (A8) we need the uniform convergence result of the estimator of the density of the invariant measure in the uni- and bivariate case as developed above.
For (A3) and (A5) to hold the uniform result for the regression estimator is required.

Assumptions (A2), (A4) and (A8)

We need the stochastic projection operator to converge during iterative applications. This is why assumption (A2) needs to be verified. All three assertions can be shown with the previous lemmata and the resulting corollaries. As simple corollaries we have:

Corollary A.3.

$$\sup_{x^j \in \partial \mathcal{G}^j_{C_1 h}} \left| \widehat{\pi}_j(x^j) \right| = O_P(1) \tag{A.40}$$

$$\sup_{x^{jk} \in \partial \mathcal{G}^j_{C_1 h} \times \mathcal{G}^k} \left| \widehat{\pi}_{jk}(x^{jk}) \right| = O_P(1) \tag{A.41}$$

$$\sup_{x^j \in \partial \mathcal{G}^j_{C_1 h}} \left| \widehat{\pi}_j(x^j)^{-1} \right| = O_P(1) \tag{A.42}$$

Proofs are simple consequences oft the uniform convergence results for boundary and interior.

(A2) has three parts. $\mathcal{G} \in \mathbb{R}^d$ is compact. Denote the finite d–dimensional volume with $|\mathcal{G}| := \int_{\mathcal{G}} \mathrm{d}x$ and the analogously defined one and two dimensional "trace" volumes respectively with $|\mathcal{G}_j|$ and $|\mathcal{G}_{jk}|$. Furthermore name the obtained rates of the uniform consistency in lemma A.2 of the density estimators as c_n^1 in the univariate case and as c_n^2 in the bivariate one:

$$\begin{aligned}
&\int_{\mathcal{G}_j} \left| \frac{\widehat{\pi}_j(x^j) - \pi_j(x^j)}{\pi_j(x^j)} \right|^2 \pi_j(x^j) \mathrm{d}x^j \\
&= \int_{\mathcal{G}_j} \left| \widehat{\pi}_j(x^j) - \pi_j(x^j) \right|^2 \frac{1}{|\pi_j(x^j)|} \mathrm{d}x^j \\
&\leq \max_{x^j \in \mathcal{G}_j} \left| \widehat{\pi}_j(x^j) - \pi_j(x^j) \right|^2 \max_{x^j \in \mathcal{G}_j} \frac{1}{|\pi_j(x^j)|} \int_{\mathcal{G}_j} \mathrm{d}x^j \\
&\leq O_P\left(c_n^1\right)^2 \cdot \frac{|\mathcal{G}_j|}{c_1} \\
&\leq o_P(1)
\end{aligned}$$

$$\begin{aligned}
&\int_{\mathcal{G}_{j,k}} \left| \frac{\widehat{\pi}_{j,k}(x^{j,k})}{\pi_j(x^j)\pi_k(x^k)} - \frac{\pi_{j,k}(x^{j,k})}{\pi_j(x^j)\pi_k(x^k)} \right|^2 \pi_j(x^j)\pi_k(x^k) \mathrm{d}x^j \mathrm{d}x^k \\
&= \int_{\mathcal{G}_{j,k}} \left| \widehat{\pi}_{j,k}(x^{j,k}) - \pi_{j,k}(x^{j,k}) \right|^2 \frac{1}{|\pi_j(x^j)\pi_k(x^k)|} \mathrm{d}x^j \mathrm{d}x^k \\
&\leq \max_{x^{j,k} \in \mathcal{G}_{j,k}} \left| \widehat{\pi}_{j,k}(x^{j,k}) - \pi_{j,k}(x^{j,k}) \right|^2 \max_{x^{j,k} \in \mathcal{G}_{j,k}} \frac{1}{|\pi_j(x^j)\pi_k(x^k)|} \int_{\mathcal{G}^{j,k}} \mathrm{d}x^j \mathrm{d}x^k
\end{aligned}$$

$$\begin{aligned} &\leq O_P\left(c_n^2\right)^2 \cdot \frac{1}{c_1^2} \int_{\mathcal{G}_{j,k}} \mathrm{d}x^j \mathrm{d}x^k \\ &\leq o_P(1) \end{aligned}$$

$$\begin{aligned} &\int_{\mathcal{G}_{j,k}} \left| \frac{\widehat{\pi}_{j,k}(x^{j,k})}{\widehat{\pi}_j(x^j)\pi_k(x^k)} - \frac{\pi_{j,k}(x^{j,k})}{\pi_j(x^j)\pi_k(x^k)} \right|^2 \pi_j(x^j)\pi_k(x^k) \mathrm{d}x^j \mathrm{d}x^k \\ &= \int_{\mathcal{G}_{j,k}} \left| \frac{1}{\widehat{\pi}_j(x^j)} \right| \cdot \frac{\left| \widehat{\pi}_{j,k}(x^{j,k})\pi_j(x^j) - \pi_{j,k}(x^{j,k})\widehat{\pi}_j(x^j) \right|^2}{|\pi_j(x^j)\pi_k(x^k)|} \mathrm{d}x^j \mathrm{d}x^k \\ &\leq \frac{1}{c_1^2} \cdot \int_{\mathcal{G}_{j,k}} \frac{1}{|\widehat{\pi}_j(x^j)|} \left| \widehat{\pi}_{j,k}(x^{j,k})\pi_j(x^j) - \pi_{j,k}(x^{j,k})\widehat{\pi}_j(x^j) \right|^2 \mathrm{d}x^j \mathrm{d}x^k \\ &\leq \frac{|\mathcal{G}_{j,k}|}{c_1^2} \cdot \max_{x^{j,k} \in \mathcal{G}_{j,k}} \left(\underbrace{\frac{1}{|\widehat{\pi}_j(x^j)|}}_{O_P(1)} \left| \underbrace{\left(\widehat{\pi}_{j,k}(x^{j,k}) - \pi_{j,k}(x^{j,k})\right)}_{o_P(1)} \underbrace{\pi_j(x^j)}_{\text{bounded}} - \underbrace{\left(\widehat{\pi}_j(x^j) - \pi_j(x^j)\right)}_{o_P(1)} \underbrace{\pi_{j,k}(x^{j,k})}_{\text{bounded}} \right|^2 \right) \\ &\leq o_P(1) \end{aligned}$$

(A4) is shown by similar arguments. We find:

$$\begin{aligned} &\sup_{x^k \in C^k} \int_{\mathcal{G}_j} \frac{\widehat{\pi}_{j,k}^2(x^{j,k})}{\widehat{\pi}_k^2(x^k)\pi_j(x^j)} \mathrm{d}x^j \\ &= \max_{x^k \in \mathcal{G}^k} \int_{C^j} \left(\frac{\widehat{\pi}_{j,k}(x^{j,k}) - \pi_{j,k}(x^{j,k}) + \pi_{j,k}(x^{j,k})}{\widehat{\pi}_k(x^k)} \right)^2 \frac{1}{\pi_j(x^j)} \mathrm{d}x^j \\ &\leq \max_{x^{jk} \in \mathcal{G}_{jk}} \left(\frac{\widehat{\pi}_{j,k}(x^{j,k}) - \pi_{j,k}(x^{j,k}) + \pi_{j,k}(x^{j,k})}{\widehat{\pi}_k(x^k)} \right)^2 \frac{|C^j|}{c_1} \\ &\leq \max_{x^{jk} \in \mathcal{G}_{jk}} \left[\left(\frac{\widehat{\pi}_{j,k}(x^{j,k}) - \pi_{j,k}(x^{j,k})}{\widehat{\pi}_k(x^k)} \right)^2 + \left(\frac{\pi_{j,k}(x^{j,k})}{\widehat{\pi}_k(x^k)} \right)^2 \right. \\ &\qquad \left. + 2 \frac{(\widehat{\pi}_{j,k}(x^{j,k}) - \pi_{j,k}(x^{j,k}))\pi_{j,k}(x^{j,k})}{\widehat{\pi}_k(x^k)} \right] \frac{|\mathcal{G}_j|}{c_1} \end{aligned}$$

Thus with the same considerations as before in (A2)the terms $\widehat{\pi}_{j,k}(x^{j,k}) - \pi_{j,k}(x^{j,k})$ are a.s. bounded by null sequences according to the uniform convergence lemmata. These sequences are in particular less or equal 1. Furthermore it is $\frac{1}{\widehat{\pi}_k(x^k)} \leq \frac{1}{\pi_k(x^k)}$.

Hence in total we find the desired almost sure bound by adding up:

$$\max_{x^{jk}\in\mathcal{G}_{jk}}\left(\frac{1}{\pi_k(x^k)}+\frac{\pi_{j,k}(x^{j,k})^2}{\pi_k(x^k)^2}+\frac{\pi_{j,k}(x^{j,k})}{\pi_k(x^k)}\right)\frac{|\mathcal{G}_j|}{c_1}\leq(c_1+c_2^2+c_1c_2)\frac{|\mathcal{G}_j|}{c_1^3}=:C_{A4}\qquad\mathbb{P}-\text{a.s.}\,.$$

So in total we find that with probability one it is:

$$\sup_{x^k\in C^k}\int_{\mathcal{G}_j}\frac{\widehat{\pi}_{j,k}^2(x^{j,k})}{\widehat{\pi}_k^2(x^k)\pi_j(x^j)}\mathrm{d}x^j\quad\leq\quad C_{A4}$$

(A8) It is:

$$\sup_{x^j\in\mathcal{G}^j}\int_{\mathcal{G}^k}\left|\frac{\widehat{\pi}_{j,k}(x^{j,k})}{\widehat{\pi}_j(x^j)\widehat{\pi}_k(x^k)}-\frac{\pi_{j,k}(x^{j,k})}{\pi_j(x^j)\pi_k(x^k)}\right|\pi_k(x^k)\mathrm{d}x^k\quad=\quad o_P(1)$$

The result follows through successive application of the triangle inequality by adding and subtracting all missing possible permutations of hats in the term $\frac{\pi_{j,k}(x^{j,k})}{\pi_j(x^j)\pi_k(x^k)}$. Together with uniform convergence for the bivariate density and (A.42) we find that the term is asymptotically negligible.

Assumptions (A3) and (A5)

We want the stochastic projection operator to converge during iterative applications. For this assumption, (A3) is the final condition to be verified.

(A3) It is:

$$\int_{\mathcal{G}_j}(\widetilde{m}_j(x^j))^2\pi(x^j)\mathrm{d}x^j\leq C\,.$$

The proof follows immediately from (A5) below.

(A5) For the variance part it is with lemma A.3:

$$\int_{\mathcal{G}_j}(\widetilde{m}_j^A(x^j))^2\pi(x^j)\mathrm{d}x^j\leq C(\sup_{x^j\in\mathcal{G}_j}\left|\widetilde{m}_j^A(x^j)\right|)^2\,.$$

which is bounded almost surely by an arbitrary positive constant.

For the bias we get with lemma A.3:

$$\int_{\mathcal{G}_j}(\widetilde{m}_j^B(x^j))^2\pi(x^j)\mathrm{d}x^j\leq\int_{\mathcal{G}_j}\mu_j(x^j)\pi_j(x^j)\mathrm{d}x^j+\int_{\mathcal{G}_j}(\widetilde{m}_j^B(x^j)-\mu_j(x^j))^2\pi_j(x^j)\mathrm{d}x^j\,.$$

that the second term is bounded almost surely by a constant. Since any continuous function on compact support is bounded also the first term is bounded.

Assumption (A6)

With the definition of $\widehat{m}^A$ and the triangle inequality we get:

$$\begin{aligned}
&\sup_{x^k\in\mathring{\mathcal{G}}_k}\left|\int_{\mathcal{G}_j}\frac{\widehat{\pi}_{j,k}(x^j,x^k)}{\widehat{\pi}_k(x^k)}\widehat{m}^A(x^j)\mathrm{d}x^j\right| \\
&\quad\leq \sup_{x^k\in\mathring{\mathcal{G}}_k}\left|\int_{\mathcal{G}_j}\frac{\pi_{jk}(x^{jk})}{\pi_k(x^k)\pi_j(x^j)}\frac{\widehat{s}_j(x^j)}{T^j(n)}\mathrm{d}x^j\right| \\
&\qquad\qquad + \sup_{x^k\in\mathring{\mathcal{G}}_k}\left|\int_{\mathcal{G}_j}\left[\frac{\widehat{\pi}_{j,k}(x^{jk})}{\widehat{\pi}_k(x^k)}-\frac{\pi_{j,k}(x^{jk})\widehat{\pi}_j(x^j)}{\pi_k(x^k)\pi_j(x^j)}\right]\frac{\widehat{s}_j(x^j)}{T^j(n)}\mathrm{d}x^j\right| \\
&\quad\leq \sup_{x^k\in\mathring{\mathcal{G}}_k}\left|\int_{\mathcal{G}^j}\frac{\pi_{j,k}(x^{jk})}{\pi_k(x^k)\pi_j(x^j)}\frac{\widehat{s}_j(x^j)}{T^j(n)}\mathrm{d}x^j\right| + o_P(h^2)\ ,
\end{aligned}$$

with $\widehat{s}_j(x^j)=\sum_{i=1}^n K_{h,x^j}(X_i^j)\varepsilon_i$ since with Lemmas A.2 and A.3 it holds:

$$\begin{aligned}
&\sup_{x^k\in\mathring{\mathcal{G}}_k}\left|\int_{\mathcal{G}^j}\left[\frac{\widehat{\pi}_{j,k}(x^{jk})}{\widehat{\pi}_k(x^k)}-\frac{\pi_{j,k}(x^j,x^k)\widehat{\pi}_j(x^j)}{\pi_k(x^k)\pi_j(x^j)}\right]\frac{\widehat{s}_j(x^j)}{T(n)}\mathrm{d}x^j\right| \\
&\leq \underbrace{\sup_{x^j\in\mathring{\mathcal{G}}_j}\left|\frac{\widehat{s}_j(x^j)}{T(n)}\right|}_{O_P(h^2+c_n^{(1)})}|\mathcal{G}_j|\left(\underbrace{\sup_{(x^k,x^j)\in\mathring{\mathcal{G}}_{jk}}\left|\frac{\widehat{\pi}_{jk}(x^{jk})-\pi_{jk}(x^{jk})}{\widehat{\pi}_k(x^k)}\right|}_{O_P(h^2+c_n^{(2)})}+\right. \\
&\qquad\left.+\underbrace{\sup_{(x^{jk})\in\mathring{\mathcal{G}}_{k,j}}\left|\frac{\pi_{jk}(x^{jk})}{\pi_k(x^k)\pi_j(x^j)}\left(\widehat{\pi}_j(x^j)-\pi_j(x^j)\right)\right|}_{O_P(h^2+c_n^{(1)})}+\underbrace{\sup_{(x^{jk})\in\mathring{\mathcal{G}}_{k,j}}\left|\frac{\pi_{j,k}(x^{jk})}{\widehat{\pi}_k(x^k)}-\frac{\pi_{jk}(x^{jk})}{\pi_k(x^k)}\right|}_{O_P(h^2+c_n^{(1)})}\right) \\
&= o_P(h^2)
\end{aligned}$$

where the rates $c_n^{(1)}, c_n^{(2)}\to 0$ are as stated in the lemmas.

Now rewrite the remaining first term in the following way:

$$\sup_{x^k\in\mathring{\mathcal{G}}_k}\left|\int_{\mathcal{G}^j}\frac{\pi_{j,k}(x^{jk})}{\pi_k(x^k)\pi_j(x^j)}\widehat{s}_j(x^j)\mathrm{d}x^j\right| = \sup_{x^k\in\mathring{\mathcal{G}}_k}\left|\sum_{i=1}^n\xi_i(x^k)\right|$$

with

$$\xi_i(x^k) = \left(\int_{\mathcal{G}_j} \frac{\pi_{jk}(X_i^j + uh, x^k)}{\pi_k(x^k)\pi_j(X_i^j + uh)} K(u) \mathrm{d}u \right) \frac{\varepsilon_i}{T^j(n)} .$$

where the integral is bounded, since all the densities are bounded from above and from below. Along the lines of the proofs of Lemma A.2 and Lemma A.3 this term can be shown to be $o_P(h^2)$. Hence

$$\sup_{x^k \in \mathring{\mathcal{G}}_k} \left| \int_{\mathcal{G}_j} \frac{\widehat{\pi}_{j,k}(x^{jk})}{\widehat{\pi}_k(x^k)} \widehat{m}^A(x^j) \mathrm{d}x^j \right| = o_P(h^2) ,$$

which implies

$$\left(\int_{\mathcal{G}_k} \left(\int_{\mathcal{G}^j} \frac{\widehat{\pi}_{jk}(x^{jk})}{\widehat{\pi}_k(x^k)} \widehat{m}^A(x^j) \mathrm{d}x^j \right)^2 \widehat{\pi}_k(x^k) \mathrm{d}x^k \right)^{1/2} = o_P(h^2) ,$$

since $\int_{\mathcal{G}_k} \widehat{\pi}_k(x^k) \mathrm{d}x^k$ is $O_P(1)$.

Proof of Theorem 4.1 and 4.2

With (A1)-(A6) and (A8) and the uniform lemmata of the previous subsection, bias expansion and asymptotic distribution are as in [Mammen et al., 1999]. Depending on different independence assumptions on ε, results in 4.1 and 4.2 differ by convergence in ditribution according to lemma A.5 or A.6.

Proof of Theorem 4.3 and 4.4

Since convergence of standard SBE in [Mammen et al., 1999] relies only on up to two dimensional objects, the proof of convergence of generalized backfitting goes along the lines of [Mammen et al., 1999], if we show that the generalized backfitting operator results from a norm. Bias and variance expressions, however, have to be calculated anew as done below.

Bias Expansion

From the operator backfitting equation (A.14), we can deduce:

$$\begin{aligned}\widetilde{m} - m &= (I - \widehat{A})^{-1}\frac{1}{d-1}(\mathbf{1} - \widehat{\Phi})\widehat{m} - (I - \widehat{A})^{-1}(I - \widehat{A})m \\ &= (I - A)^{-1}\left[\frac{1}{d-1}(\mathbf{1} - \widehat{\Phi})\widehat{m} - (I - \widehat{A})m\right] + \\ &\quad + \left((I - \widehat{A})^{-1} - (I - A)^{-1}\right)\left[\frac{1}{d-1}(\mathbf{1} - \widehat{\Phi})\widehat{m} - (I - \widehat{A})m\right] \quad \text{(A.43)}\end{aligned}$$

If we set m_0 as in (3.14), the centering operation with Φ can be omitted. From [Mammen and Linton, 2005] equation (41), it follows that the second summand is negligible for the bias since $\left((I - \widehat{A})^{-1} - (I - A)^{-1}\right) = O_P(\sum_{k\neq j} h_{jk}^2)$ in the interior. Below we focus on the term in squared bracket in order to derive the explicit form of bias of $\widetilde{m}$. The goal is to expand $\mathbf{1}\widehat{m}$ in terms of the projection operator $(I - \widehat{A})$ of the backfitting equations (A.14). With $\widehat{m}_j^{II} = \frac{1}{d-1}(\mathbf{1}\widehat{m})_j$ as in (A.16) it is

$$\sup_{x^j \in \mathring{\mathcal{G}}^j} \left|\widehat{m}_j^{II,B}(x^j) - \widehat{\nu}_{n,j}(x^j)\right| = o_P(\sum_{k\neq j} h_{jk}^2) \quad \text{(A.44)}$$

$$\sup_{x^j \in \partial\mathcal{G}^j} \left|\widehat{m}_j^{II,B}(x^j) - \widehat{\nu}_{n,j}(x^j)\right| = o_P(\sum_{k\neq j} h_{jk}) \quad \text{(A.45)}$$

where

$$\begin{aligned}\widehat{\nu}_{n,j}(x^j) &= m_j(x^j) + \sum_{k\neq j}\int_{\mathcal{G}^k}\left(m_k(x^k)\frac{\widehat{\pi}_{jk}(x^{jk})}{\widehat{\pi}_j^{(k)}(x^j)}\right)\mathrm{d}x^k + \\ &\quad \frac{\kappa_1(x^j)}{\kappa_0(x^j)}\sum_{k\neq j} h_{jk}\left(m_j'(x^j) + \int_{\mathcal{G}^k}\left(m_k'(x^k)\frac{\widehat{\pi}_{jk}(x^{jk})}{\widehat{\pi}_j^{(k)}(x^j)}\right)\mathrm{d}x^k\right) + \\ &\quad \frac{\kappa_2(x^j)}{\kappa_0(x^j)}\sum_{k\neq j} h_{jk}^2\left(\frac{1}{2}m_j''(x^j) + m_j'(x^j)\frac{\pi_j^{(k)\prime}(x^j)}{\pi_j^{(k)}(x^j)}\right. \\ &\quad \left.\int_{\mathcal{G}^k}\left(m_k'(x^k)\frac{\partial\pi_{jk}(x^{jk})}{\partial x^k \pi_{jk}(x^{jk})} + \frac{1}{2}m_k''(x^k)\right)\frac{\pi_{jk}(x^{jk})}{\pi_j^{(k)}(x^j)}\mathrm{d}x^k\right)\end{aligned}$$

Proof. Decompose $\left(\mathbf{1}\widehat{m}^B\right)_j$ in the following way

$$\begin{aligned}
\left(\mathbf{1}\widehat{m}\right)_j^B(x^j) &= \sum_{k\neq j}\widehat{m}_j^{(k),B}(x^j) && \text{(A.46)}\\
&= \sum_{k\neq j}\frac{1}{T_C^{jk}(n)}\sum_{i\in I_{jk}}\frac{K_{h_{jk},x^j}(X_i^j)m(X_i)}{\widehat{\pi}_j^{(k)}(x^j)} && \text{(A.47)}\\
&= \sum_{k\neq j}\frac{1}{T_C^{jk}(n)}\sum_{i\in I_{jk}}\frac{K_{h_{jk},x^j}(X_i^j)\left(m_0+m_1(X_i^1)+\ldots m_d(X_i^d)\right)}{\widehat{\pi}_j^{(k)}(x^j)} && \text{(A.48)}
\end{aligned}$$

In contrast to the stationary case [Mammen et al., 1999] there is no law of large numbers for nonstationary processes. Instead we have to use the quotient limit theorem (A.11) which only works for stochastic denominators.

$$\frac{1}{T_C^{jk}(n)}\sum_{i\in I_{jk}}g(X_i^j)=\frac{\iint g(u)\pi_{jk}(u,v)\mathrm{d}u\,\mathrm{d}v}{\pi_{jk}(C)}=C\mu(g(\cdot))\;. \tag{A.49}$$

Expand (A.48) for each summand separately. The kernel function for dimension j with index set from jk meets a component function m_l for $l=1,\ldots,d$. Distinguish between three cases $l=j$ or $l=k\neq j$ and $l\neq(j\vee k)$ in (A.48). We will see that the last case has some nonstationary peculiarities. For $l=j$ we find with (A.49) and standard kernel calculations:

$$\begin{aligned}
&\sum_{k\neq j}\frac{1}{T_C^{jk}(n)}\sum_{i\in I_{jk}}\frac{K_{h_{jk},x^j}(X_i^j)m_j(X_i^j)}{\widehat{\pi}_j^{(k)}(x^j)}\\
&= \sum_{k\neq j}m_j(x^j)+\frac{\mu_{jk}\left(K_{h_{jk},x^j}(\cdot)m_j(\cdot)\right)-m_j(x^j)\mu_{jk}\left(K_{h_{jk},x^j}(\cdot)\right)}{\mu_{jk}\left(K_{h_{jk},x^j}(\cdot)\right)}+R_{n_{jk},jk}\\
&= \sum_{k\neq j}m_j(x^j)+h_{jk}\frac{\kappa_1(x^j)}{\kappa_0(x^j)}m_j'(x^j)+h_{jk}^2\frac{\kappa_2(x^j)}{\kappa_0(x^j)}\left(\frac{m_j'(x^j)}{\pi_j^{(k)}(x^j)}\pi_j^{(k)\prime}(x^j)+\frac{1}{2}m_j''(x^j)\right)\\
&\qquad +R_{n_{jk},jk}+o_P(h_{jk}^2)\\
&= (d-1)m_j(x^j)+\sum_{k\neq j}h_{jk}\frac{\kappa_1(x^j)}{\kappa_0(x^j)}m_j'(x^j)\\
&\qquad +h_{jk}^2\frac{\kappa_2(x^j)}{\kappa_0(x^j)}\left(\frac{m_j'(x^j)}{\pi_j^{(k)}(x^j)}\pi_j^{(k)\prime}(x^j)+\frac{1}{2}m_j''(x^j)\right)+o_P(h_{jk}^2)
\end{aligned}$$

The last equation is true since for

$$\begin{aligned}
\sup_{x^j\in\mathcal{G}^j}\left|R_{n_{jk},jk}(x^j)\right| &= \sup_{x^j\in\mathcal{G}^j}\left|\frac{1}{T_C^{jk}(n)}\sum_{i\in I_{jk}}\frac{K_{h_{jk},x^j}(X_i^j)m_j(X_i^j)}{\widehat{\pi}_j^{(k)}(x^j)}-\frac{\mu_{jk}\left(K_{h_{jk},x^j}(\cdot)m_j(\cdot)\right)}{\mu_{jk}\left(K_{h_{jk},x^j}(\cdot)\right)}\right| \\
&= \sup_{x^j\in\mathcal{G}^j}\left|\frac{1}{T_C^{jk}(n)}\left[\sum_{i\in I_{jk}}\frac{K_{h_{jk},x^j}(X_i^j)m_j(X_i^j)-\mu(K_{h_{jk},x^j}m_j)}{\widehat{\pi}_j^{(k)}(x^j)}\right]+\right. \\
&\quad \left.+\frac{\mu(K_{h_{jk},x^j}m_j)}{\widehat{\pi}_j^{(k)}(x^j)}-\frac{\mu_{jk}\left(K_{h_{jk},x^j}(\cdot)m_j(\cdot)\right)}{\mu_{jk}\left(K_{h_{jk},x^j}(\cdot)\right)}\right| \\
&= O_P\left(\frac{1}{h_{jk}n^{\beta_{jk}}L_2(n^{\beta^{jk}})^{1-\beta^{jk}}}\right)=o_P(h_{jk}^2)\ .
\end{aligned}$$

The details of this follow exactly from the proof of Lemma A.2 for the stochastic part. For $l=k\neq j$ standard kernel calculations lead to

$$\begin{aligned}
&\sum_{k\neq j}\frac{1}{T_C^{jk}(n)}\sum_{i\in I_{jk}}\frac{K_{h_{jk},x^j}(X_i^j)m_k(X_i^k)}{\widehat{\pi}_j^{(k)}(x^j)} \\
&= \sum_{k\neq j}\frac{1}{T_C^{jk}(n)}\sum_{i\in I_{jk}}\int_{\mathcal{G}^k}\frac{K_{h_{jk},x^j}(X_i^j)K_{h_{jk},x^k}(X_i^k)m_k(X_i^k)}{\widehat{\pi}_j^{(k)}(x^j)}\mathrm{d}x^k \\
&= \sum_{k\neq j}\sum_{i\in I_{jk}}\int_{\mathcal{G}^k}\frac{K_{h_{jk},x^j}(X_i^j)K_{h_{jk},x^k}(X_i^k)}{T_C^{jk}(n)\widehat{\pi}_j^{(k)}(x^j)}\left(m_k(x^k)+\right. \\
&\qquad \left.+m_k'(x^k)(X_i^k-x^k)+\frac{1}{2}m_k''(x^k)(X_i^k-x^k)^2\right)\mathrm{d}x^k+o_P(h_{jk}^2) \\
&= \sum_{k\neq j}\left[\int_{\mathcal{G}^k}\frac{\widehat{\pi}_{jk}(x^{jk})}{\widehat{\pi}_j^{(k)}(x^j)}m_k(x^k)\mathrm{d}x^k+h_{jk}\int_{\mathcal{G}^k}\frac{\widehat{\pi}_{jk}(x^{jk})}{\widehat{\pi}_j^{(k)}(x^j)}\frac{\kappa_1(x^j)}{\kappa_0(x^j)}m_k'(x^k)\mathrm{d}x^k+\right. \\
&\quad \left.+h_{jk}^2\frac{\kappa_2(x^j)}{\kappa_0(x^j)}\int_{\mathcal{G}^k}\left(\frac{\partial\pi_{jk}(x^{jk})}{\pi_{jk}(x^{jk})\partial x^k}m_k'(x^k)+\frac{1}{2}m_k''(x^k)\right)\frac{\pi_{jk}(x^{jk})}{\pi_j^{(k)}(x^j)}\mathrm{d}x^k\right] \\
&\qquad +R_{n_{jk},jk}(x^j)+o_P(h_{jk}^2)
\end{aligned}$$

For the second to last equation, standard kernel arguments are applied together with (A.18) and A.3. Exact details follow [Mammen et al., 1999] equations (118)-(122). In particular we need to show uniform convergence in h_{jk}^2 of the following

expressions against their respective means, which can be expanded into the terms occurring above:

$$t_j^l(x^j) = K_{h_{jk},x^j}(X_i^j) \int_{\mathcal{G}^k} K_{h_{jk},x^k}(X_i^k)(X_i^k - x^k)^l m_k^{(l)}(x^k)\,\mathrm{d}x^k$$

with $l \in \{1,2\}$. This is achieved along the lines of Lemma A.3. The last equation is true since $\sup_{x^j \in \mathcal{G}^j} |R_{n,j}(x^j)| = o_P(h_{jk}^2)$. This is shown as before.

For $l \neq (j \vee k)$ the fact that we might use different data in different directions complicates the expansion and might therefore add an additional term paying for the different characters of nonstationarities involved. For the index set I_{jk}, data of the marginal X^l might also be found outside $\mathcal{G}_l$. To control these outside happenings, pairwise β–null Harris recurrence is not sufficient. Under Assumption 6, however, we can control each term

$$\sum_{k \neq j} \frac{1}{T_C^{jk}(n)} \sum_{i \in I_{jk}} \frac{K_{h_{jk},x^j}(X_i^j) m_l(X_i^l)}{\widehat{\pi}_j^{(k)}(x^j)}$$

in the same way as seen above, sum up and we are done.

Otherwise, expand the $(d-2)$ summands for each pair (j,k) in the following way

$$\begin{aligned}
&\sum_{k \neq j} \frac{1}{T^{jk}(n)} \sum_{i \in I_{jk}} \frac{K_{h_{jk},x^j}(X_i^j) m_l(X_i^l)}{\widehat{\pi}_j^{(k)}(x^j)} \\
&= \sum_{k \neq j} \frac{\widehat{L}_j^{(lk)}(x^j)}{\widehat{L}_j^{(k)}(x^j)} \left[\sum_{i \in I_{jk} \cap I_{jl}} \frac{K_{h_{jk},x^j}(X_i^j) m_l(X_i^l)}{\widehat{L}_j^{(lk)}(x^j)} \right] + \\
&\qquad + \mathbf{1}\left(I_{jk} \backslash I_{jl} \neq \varnothing\right) \frac{\widehat{L}_j^{(k)}(x^j) - \widehat{L}_j^{(lk)}(x^j)}{\widehat{L}_j^{(k)}(x^j)} \left[\sum_{i \in I_{jk} \backslash I_{jl}} \frac{K_{h_{jk},x^j}(X_i^j) m_l(X_i^l)}{\widehat{L}_j^{(k)}(x^j) - \widehat{L}_j^{(lk)}(x^j)} \right] \\
&= \sum_{k \neq j} \frac{\widehat{L}_j^{(lk)}(x^j)}{\widehat{L}_j^{(k)}(x^j)} \widehat{b}_{jl}^{(k),S}(x^{jl}) + \left(1 - \frac{\widehat{L}_j^{(lk)}(x^j)}{\widehat{L}_j^{(k)}(x^j)}\right) \widehat{b}_{jl}^{(k),NS}(x^{jl})
\end{aligned}$$

As long as X_i^{jl} is in $\mathcal{G}_{jl}$ for the index set I_{jk}, only the stationary bias $\widehat{b}_{jl}^{(k),S}(x^{jl})$ occurs. This coincides with the case where $\frac{\widehat{L}_j^{(lk)}(x^j)}{\widehat{L}_j^{(k)}(x^j)} = O_p(1)$. But since I_{jk} is tai-

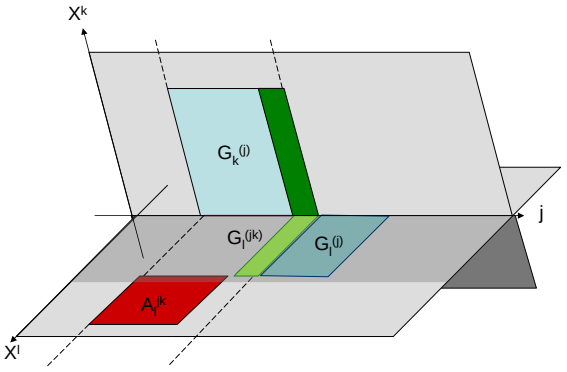


Figure A.1: The schematic figure shows that index sets in direction j for the compound processes X^{jk} being in $\mathcal{G}_{jk}$ and for the compound processes X^{jl} being in $\mathcal{G}_{jl}$, differ not only in size but also in actual elements. In general, X_i^l for $i \in I_{jk}$ will no longer be within the respective small set $\mathcal{G}_l^{(jk)} \subseteq \mathcal{G}_l^{(j)}$ but outside in $\mathcal{A}_l^{jk}$.

lored only to X_i^{jk} being in $\mathcal{G}_{jk}$, it can occur that X_i^{jl} lies outside of the controllable region $\mathcal{G}_{jl}$ in $\mathcal{A}_{jl,n_{jk}} \in \mathbb{R}^2$ with $\mathcal{G}_{jl} \cap \mathcal{A}_{jl,n_{jk}} = \varnothing$ for $i \in I_{jk} \backslash I_{jl}$. See Figure A.1 for an illustration of the problem. To control this issue, it is not sufficient that the X^{jk} and X^{jl} have the same type of nonstationarity, yielding asymptotically the same amount of data within small sets. Even in this case elements can still differ. For $n \to \infty$ it can even be that $\mathcal{A}_{jl,n_{jk}} = \mathbb{R}^2 \backslash \mathcal{G}_{jl}$. The stationary bias $\widehat{b}_{jl}^{(k),S}$ can be expanded as seen before.

$$
\begin{aligned}
\widehat{b}_{jl}^{(k),S}(x^{jl}) &= \frac{1}{T^{jlk}(n)} \sum_{i \in I_{jk} \cap I_{jl}} \int_{\mathcal{G}_l^{jk}} \frac{K_{h_{jk},x^j}(X_i^j) K_{h_{jl},x^l}(X_i^l) m_l(X_i^l)}{\widehat{\pi}_j^{(lk)}(x^j)} \mathrm{d}x^l \\
&= \left[\int_{\mathcal{G}_l^{jk}} \frac{\widehat{\pi}_{jl}^{(k)}(x^{jl})}{\widehat{\pi}_j^{(lk)}(x^j)} m_l(x^l) \mathrm{d}x^l + h_{jl} \int_{\mathcal{G}_l^{jk}} \frac{\widehat{\pi}_{jlk}(x^{jl})}{\widehat{\pi}_j^{(lk)}(x^j)} \frac{\kappa_1(x^j)}{\kappa_0(x^j)} m_l'(x^l) \mathrm{d}x^l + \right. \\
&\quad \left. + h_{jl}^2 \frac{\kappa_2(x^j)}{\kappa_0(x^j)} \int_{\mathcal{G}_l^{jk}} \left(\frac{\partial \pi_{jl}^{(k)}(x^{jl})}{\pi_{jl}(k)(x^{jl}) \partial x^l} m_l'(x^l) + \frac{1}{2} m_l''(x^l) \right) \frac{\pi_{jl}^{(k)}(x^{jl})}{\pi_j^{(lk)}(x^j)} \mathrm{d}x^l \right] \\
&\quad + R_{n,jl}(x^j) + o_P(h_{jl}^2)
\end{aligned}
$$

For the second nonstationary bias term $\widehat{b}_{jk}^{NS}$ a direct expansion in terms of $\widehat{A}_{jk}$ as done above is not possible. If $I_{jk} \backslash I_{jl} \neq \varnothing$ standard kernel calculations lead to:

$$\sum_{i \in I_{jk} \backslash I_{jl}} \frac{K_{h_{jk},x^j}(X_i^j) m_l(X_i^l)}{\widehat{L}_j^{(k)}(x^j) - \widehat{L}_j^{(l)}(x^j)} = \frac{\mu_{jl}^a(K_{h_{jk},x^j}(X^j) m_l(X^l))}{\pi_j^{a(l)}(x^j)} + o_P(h_{jl}^2)$$

with $\mathcal{A}_j \in \mathcal{R}_j$ and $\mathcal{A}_{l,n_{jk}} \in \mathcal{R}_l$ the coordinatewise projections of $\mathcal{A}_{jl,n_{jk}}$ it is

$$\begin{aligned}
&\mu_{jl}^a(K_{h_{jk},x^j}(X^j) m_l(X^l)) \\
&= \iint_{\mathcal{A}_{jl,n_{jk}}} K_{h_{jk},x^j}(u) m_l(v) \pi_{jl}(u,v) \,\mathrm{d}u \,\mathrm{d}v \\
&= \int_{\mathcal{A}_{l,n_{jk}}} m_l(x^l) \pi_{jl}(x^{jl}) \,\mathrm{d}x^l + h_{jk} \frac{\kappa_1(x^j)}{\kappa_0(x^j)} \int_{\mathcal{A}_{l,n_{jk}}} m_l(x^l) \frac{\partial \pi_{jl}(x^{jl})}{\partial x^j} \,\mathrm{d}x^l + \\
&\quad + \frac{h_{jk}^2}{2} \frac{\kappa_2(x^j)}{\kappa_0(x^j)} \int_{\mathcal{A}_{l,n_{jk}}} m_l(x^l) \frac{\partial^2 \pi_{jl}(x^{jl})}{\partial (x^j)^2} \,\mathrm{d}x^l + o_p(h_{jl}^2)
\end{aligned}$$

Then consistency of generalized smooth backfitting can be achieved via parametric form assumptions in the outside regions, e.g. $\int_{\mathcal{A}_{l,n_{jk}}} m_l(x^l) \pi_{jl}(x^{jl}) \,\mathrm{d}x^l = 0$. A weaker sufficient condition is $T^l_{\mathcal{A}_{l,n_{jk}}}(n) / T^l_{\mathcal{G}_l^{(j)}}(n) = o_P(1)$ for X_i^l with $i \in I_{jk}$. Without additional assumptions though, it can generally not be expected that the necessary condition $\mu_{jl}^a(K_{h_{jk},x^j}(X^j) m_l(X^l)) < \infty$ is fulfilled. Most elegantly these conditions could be phrased in probabilistic terms trough restrictions on the correlation structures among the three involved dimensions.

In total addding up in a clever way with $o_P(\sum_{k \neq j} h_{jk}^2) = o_P(h_{j+}^2)$, claims (A.44) and (A.45) have been proven. □

Thus the exact form of the bias of the generalized backfitting estimator can now be derived by putting (A.44) or (A.45) into (A.43).

Proof of Theorems 4.3 and 4.4

The asymptotic distribution follows directly from [Mammen et al., 1999], as everything carries over for the pairwise case. The bias has been developed right above. It is $o_P(\sum_{k \neq j} h_{jk}^2) = o_P(h_{j+}^2)$ and the corresponding bias component dominates.

Proof of Theorems 4.5 and 4.6

The proof of Theorems 4.5 and 4.6 directly follows along the lines of the proof of Theorem 4.3 by replacing (k) by (λ_j) in the superscript of all objects.

Proof of Theorems 4.7 and 4.8

In order to obtain Theorems 4.7 and 4.8, observe that the backfitting operator separates stationary and nonstationary components through the asymptotic independence assumption, where the constant parts are zero through the identification assumptions (4.14). Then results are proven analogously to Theorems 4.3 and 4.4.

Bibliography

Baillie, R. 1996. Long memory processes and fractional integration in econometrics. *Journal of Econometrics 73,* 1, 5–59.

Bandi, F. and Phillips, P. 2003. Fully Nonparametric Estimation of Scalar Diffusion Models. *Econometrica 71,* 1, 241–283.

Bandi, F. M. 2004. On persistence and nonparametric estimation (with an application to stock return predictability). Tech. rep., Graduate School of Business, University of Chicago.

Bandi, F. M. and Phillips, P. C. 2004. *Handbook of Financial Econometrics.* Chapter Nonstationary Continuous-Time Processes, 1–63.

Buja, A., Tibshirani, R., and Hastie, T. 1989. Linear smoothers and additive models. *Annals of Statistics 17*, 453 – 555.

Carrasco, M., Florens, J.-P., and Renault, E. 2003. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. Tech. rep., IDEI working paper. September.

Chen, J., Gao, J., and Li, D. 2007. Semiparametric regression estimation in null recurrent time series. submitted to annals of Statistics.

Chen, X. 1999a. How often does a harris recurrent markov chain recur? *Annals of Probability 27,* 3, 1324–1346.

Chen, X. 1999b. Some dichotomy results for functionals of harris recurrent markov chains. *Stochastic Processes and their Applications 83*, 211–236.

Chen, X. 2000. On the limit laws of the second order for additive functionals of harris recurrent markov chains. *Probab. Theory and Relat.Fields 116*, 89–123.

Christopeit, N. and Hoderlein, S. G. N. 2006. Local partitioned regression. *Econometrica 74,* 3 (05), 787–817.

Cline, D. and Pu, U. 1999. Stability of nonlinear AR(1) time series with delay. *Stochastic Processes and their Applications 82*, 307–333.

Cochrane, J. 2001. *Asset Pricing.* Princeton University Press.

Dahlhaus, R. 1997. Fitting Time Series Models to Nonstationary Processes. *The Annals of Statistics 25,* 1, 1–37.

Darling, D. A. and Kac, M. 1957. On occupation times of markoff processes. *Transactions of the American Mathematical Society 84*, 444–458.

de Jong, R. and Wang, C.-H. 2005. Further results on the asymptotics for nonlinear transformations of integrated time series. *Econometric Theory 21,* 02 (March), 413–430.

Dickey, D. and Fuller, W. 1979. Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association 74,* 366, 427–431.

Dickey, D. and Fuller, W. 1981. Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica 49,* 4, 1057–1072.

Diebold, F. and Rudebusch, G. 1999. Long Memory and Persistence in Aggregate Output. *Business Cycles: Durations, Dynamics, and Forecasting*.

Engle, R. and Granger, C. 1987. Cointegration and error correction: representation, estimation and testing. *Econometrica 55,* 2, 251–276.

Feller, W. 1971. *An Introduction to Probability Theory and Its Applications*. Vol. II. Wiley.

Gil-Alaña, L. and Robinson, P. 1997. Testing of unit root and other nonstationary hypotheses in macroeconomic time series. *Journal of Econometrics 80,* 2, 241–268.

Glynn, P. W. and Ormoneit, D. 2002. Hoeffding's inequality for uniformly ergodic markov chains. *Statist. Probab. Lett. 56*, 143 –146.

Granger, C. and Hallman, J. 1991. Nonlinear transformations of integrated time series. *Journal of Time Series Analysis* 12, 207–224.

Guerre, E. 2004. Design–adaptive pointwise nonparametric regression estimation for recurrent markov time series.

Haag, B. 2006. Model choice in structured nonparametric regression and diffusion models. Ph.D. thesis, Mannheim University.

Hall, P. and Heyde, C. 1980. *Martingale Limit Theorems and its Application*. Academic Press, New York.

Hall, P., Horowitz, J., and Jing, B. 2003. On blocking rules for the bootstrap with dependent data. *Biometrika 82,* 3, 561–574.

Hill, B. 1975. A Simple General Approach to Inference About the Tail of a Distribution. *The Annals of Statistics 3,* 5, 1163–1174.

Hong, S. H. and Phillips, P. C. B. 2005. Testing linearity in cointegrating relations with an application to purchasing power parity. 1541 (Dec.).

Höpfner, R. and Löcherbach, E. 2000. Limit theorems for null recurrent markov processes.

Jakob, N. 2001. *Pseudo Differential Operators and Markov Processes*. Vol. 1. Imperial College Press, London.

Jayakumar, K. and Suresh, R. 2003. Mittag–Leffler distributions. *J. Ind. Soc. Probab. Statist. 7*, 51–71.

Johansen, S. 1991. Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica 59,* 6, 1551–1580.

Kallianpur, G. and Robbins, H. 1954. The sequence of sums of independent random variables. *Duke Math. J 21,* 2, 285–307.

Karlsen, H., Myklebust, T., and Tjøstheim, D. 2007. Nonparametric estimation in a nonlinear cointegration type model. *Annals of Statistics 35,* 1 (February), 1–57.

Karlsen, H. and Tjøstheim, D. 1998. Nonparametric estimation in null–recurrent time series. *Working paper HU Berlin.*

Karlsen, H. and Tjøstheim, D. 2001. Nonparametric estimation in null–recurrent time series. *Annals of Statistics 29,* 2, 372–416.

Kwiatkowski, D., Phillips, P., Schmidt, P., and Shin, Y. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics 54,* 1-3, 159–178.

Lewbel, A. and Ng, S. 2005. Demand systems with nonstationary regressors. *Rev. of Econ. and Stat. 87,* 3, 479–494.

Linton, O. and Nielsen, J. P. 1995. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika 82*, 93–100.

Mammen, E. and Linton, O. 2005. Estimating semiparametric arch(∞) models by kernel smoothing methods,. *Econometrica 73,* 3, 771–836.

Mammen, E., Linton, O., and Nielsen, J. 1999. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics 27,* 5 (October), 1443–1490.

Mammen, E., Marron, J., Turlach, B. A., and Wand, M. 2001. A general projection framework for constrained smoothing. *Statistical Science 16,* 3, 232–248.

Mammen, E. and Nielsen, J. P. 2003. Generalised structured models. *Biometrika 90,* 3, 551–566.

Mammen, E. and Park, B. U. 2005. Bandwidth selection for smooth backfitting in additive models. *Ann. Statist. 33,* 3, 1260–1294.

Meese, R. and Rose, A. 1991. An Empirical Assessment of Non-Linearities in Models of Exchange Rate Determination. *The Review of Economic Studies 58,* 3, 603–619.

Meese, R. and Singleton, K. 1982. On Unit Roots and the Empirical Modeling of Exchange Rates. *The Journal of Finance 37,* 4, 1029–1035.

Meyn, S. and Tweedie, R. 1993. *Markov Chains and Stochastic Stability.* Springer.

Moloche, G. 2001. Kernel regression for nonstationary harris–recurrent processes. MIT working paper 2001.

Nielsen, J. P. and Sperlich, S. 2005. Smooth backfitting in practice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67,* 1, 43–61.

Nummelin, E. 1984. *General Irreducible Markov Chains and Non–negative Operators.* Cambridge University Press.

Opsomer, J. D. 2000. Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis 73*, 166 –179.

Opsomer, J. D. and Ruppert, D. 1997. Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics 25*, 186 – 211.

Park, J. Y. and Phillips, P. C. 1999. Asymptotics for nonlinear transformations of integrated time series. *Econometric Theory 15*, 269–298.

Park, J. Y. and Phillips, P. C. 2001. Nonlinear regressions with integrated time series. *Econometrica 69,* 1, 117–161.

Phillips, P. 1987. Time Series Regression with a Unit Root. *Econometrica 55,* 2, 277–301.

Phillips, P. C. and Park, J. Y. 1998. Nonstationary density estimation and kernel autoregression. Cowles Foundation Discussion Papers 1181, Cowles Foundation, Yale University. June.

Phillips, P. C. and Perron, P. 1988. Testing for a unit root in time series regression. *Biometrika 75,* 2, 335–346.

Resnick, S. and Greenwood, P. 1979. A bivariate stable characterization and domains of attraction. *J. Multivariate Anal 9*, 206–221.

Sato, K. 1999. *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press.

Schilling, R. 1998. Growth and h¨older conditions for the sample paths of feller processes. *Probability Theory and Related Fields*, 565 – 611.

Sun, Y. and Phillips, P. 2004. Understanding the Fisher equation. *Journal of Applied Econometrics 19,* 7, 869–886.

Tibshirani, R. and Hastie, T. 1990. *Generalized Additive Models*. Chapman and Hall, London.

Tjøstheim, D. and Auestad, B. 1994. Nonparameteric identification of nonlinear time series: Projections. *J.Amer.Stat.Assoc. 89,* 428, 1398–1409.

Tsay, R. 2002. *Analysis of financial time series*. Wiley New York.

Wang, Q. and Phillips, P. C. B. 2006. Asymptotic theory for local time density estimation and nonparametric cointegrating regression. Tech. Rep. 1594, Cowles Foundation Discussion Paper. December.

Yakowitz, S. 1989. Nonparametric density and regression estimation for markov sequences without mixing assumptions. *Journal of of Multivariate Analysis 30*, 359–372.

Yu, K., Park, B. U., and Mammen, E. 2007. Smooth backfitting in generalized additive models. *Annals of Statistics*.

Eidesstattliche Erklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig angefertigt und mich keiner anderen als der in ihr angegebenen Hilfsmittel bedient zu haben. Insbesondere sind sämtliche Zitate aus anderen Quellen als solche gekennzeichnet und mit Quellenangaben versehen.

Mannheim, 28. März 2008

Melanie Schienle

Lebenslauf

Persönliche Daten

Melanie Schienle
geboren am 10. Mai 1979 in Offenburg

Ausbildung

10/2003-03/2008	Promotion in Volkswirtschaftslehre an der Universität Mannheim, Mitgleid des Graduiertenkollegs VWL und CDSE
09/2003	Abschluß als Diplom-Mathematikerin
08/2000–08/2001	Auslandssemester an University of Toronto, Kanada
09/1998–09/2003	Studium der Mathematik an der Universität Karlsruhe
06/1998	Abitur
08/1990–06/1998	Grimmelshausen Gymnasium in Offenburg