Anti-Proportional Bandwidth Selection for Smoothing (Non-)Sparse Functional Data with Covariate Adjustments

Dominik Liebl
Institute for Financial Economics and Statistics, University of Bonn

June 24, 2016

Abstract

This paper deals with the nonparametric estimation of the conditional mean and covariance function of a stationary time series of weakly dependent random functions with covariate-adjustments. As in the context of sparse functional data, it is assumed that only the noisy discretization points of a random function are observable. Estimation is done using multivariate local linear estimators. By means of a double asymptotic we consider all cases from sparsely to densely sampled discretization points per function and therefore take into account the vague cases typically found in applications. We show that the first- and second-order variance terms of the estimators can switch places depending on the asymptotic scenario. This has a surprising effect on the solution of the involved multiple bandwidth selection problem. By contrast to classical bandwidth results, it turns out that an *anti*-proportional bandwidth choice can become optimal. Ignoring these results and using the classical bandwidth expressions instead can lead to a large and diverging estimation error. Our research is motivated by the problem of estimating and testing the differences in the electricity prices before and after Germany's abrupt nuclear phaseout after the nuclear disaster in Fukushima Daiichi, Japan, in mid-March 2011.

Keywords: functional data, local linear estimation, multiple bandwidth selection, time series analysis, electricity spot prices, nuclear power phaseout

1 Introduction

This research is motivated by the problem of modeling hourly electricity prices $Y_{it} \in \mathbb{R}$ using the covariables hourly electricity demand $X_{it} \in \mathbb{R}$ and daily mean air temperature $Z_t \in \mathbb{R}$. We approach this problem from a functional data perspective using the qualitative assumption that there are underlying unobserved daily random price functions of electricity demand $P_t(.,z) \in L^2([a(z), b(z)])$ that are affected by the covariate $Z_t = z$, such that

$$Y_{it} = P_t(X_{it}, Z_t) + \epsilon_{it}, \tag{1}$$

where $i \in \{1, \ldots, n\}$ indexes the hour, $t \in \{1, \ldots, T\}$ indexes the day, and ϵ_{it} is an independent statistical error term. This functional perspective is theoretically well underpinned by the so-called merit order model—an economic model for electricity spot prices (see Liebl, 2013, and Section 5 below).

As an exemplary data example we investigate the conditional mean prices before and after Germany's abrupt reaction to the nuclear meltdown in Fukushima Daiichi, Japan: the shutdown of 40% of its nuclear power plants on March 15, 2011. This substantial loss of cheap (in terms of marginal costs) nuclear power raised concerns about increases in electricity prices and subsequent problems for industry and households; however, empirical studies do not report any clear price effects (see, e.g., Nestle, 2012). This is a surprising finding that we want to reconsider in our application in Section 5.

The analyzed data is shown in Figure 1, where the left plot shows the period one year before Germany's nuclear phaseout (from March 15, 2010 to March 14, 2011) and the right plot shows the period one year after (from March 15, 2011 to March 14, 2012). Interestingly, the daily mean air temperature Z_t (color-code) affects the general shape as well as the domains of the price-demand functions $P_t(., Z_t)$. Note that pre-smoothing is only used to visualize the underlying functions; we do *not* use pre-smoothed functional data in our theory, but rather work with the raw discretization points Y_{it}, X_{it}, and Z_t plotted as gray-filled circle points. This allows us to analyze data with only a small or moderate number n of discretization points per function, for which pre-smoothing is impossible or questionable as, e.g., in our application, where $n = 24$.

The first research on covariate-adjusted functional data is found in Cardot (2007), who

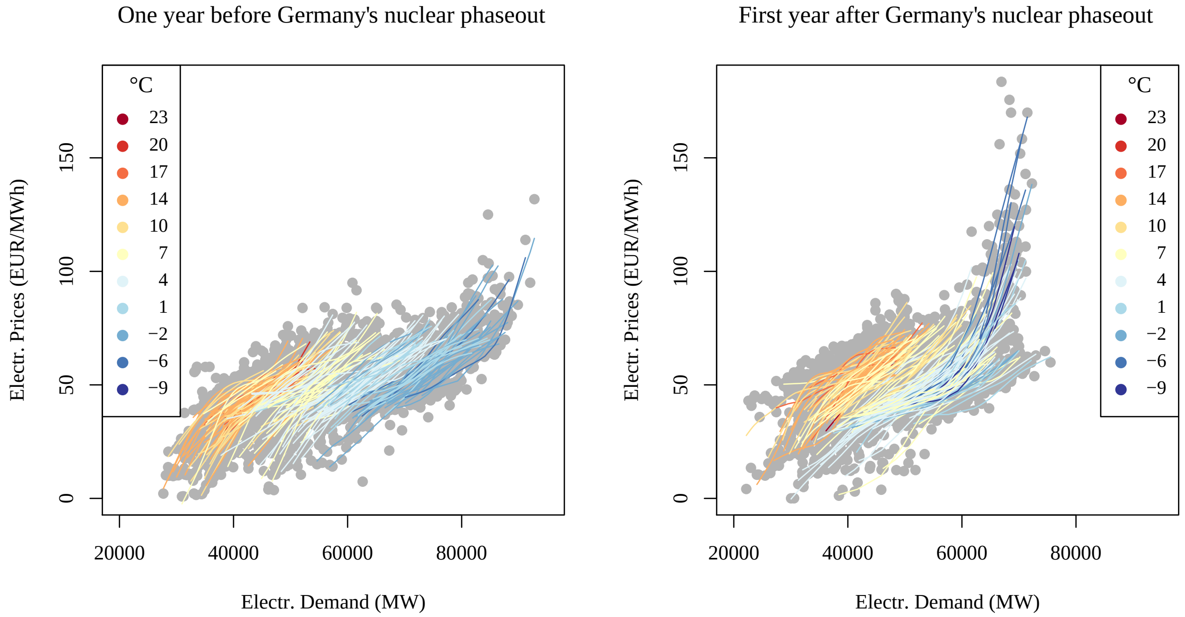

Figure 1: Scatter plot of hourly electricity spot prices Y_{it} against hourly electricity demand values X_{it} (gray filled circles) and the corresponding daily price functions (colored lines), where colors refer to the daily mean air temperature values Z_t.

considers the case of fully observed random functions. To date the only other article that is concerned with covariate-adjusted functional data is that of Jiang and Wang (2010). Like Jiang and Wang (2010) we derive the bias and variance expressions for the local linear estimators of the bivariate mean $\mu(x,z)=\mathbb{E}(P_t(x,z))$ and the trivariate covariance function $\gamma(x_1,x_2,z)=\text{Cov}(P_t(x_1,z),P_t(x_2,z))$ from the scalar observations Y_{it}, X_{it}, and Z_t. This is quite an involved task, therefore we follow Jiang and Wang (2010) and do not explicitly consider functional principal component analysis (FPCA). Consistency of the eigenvalues, eigenfunctions, and principal component (pc-)scores is, however, generally implied by our consistency results for the mean and covariance function estimators. Readers with a particular interest in FPCA are referred to existing results such as in Bosq (2000), Kneip and Utikal (2001), Yao et al. (2005), and Hall and Hosseini-Nasab (2006). More general introductions to functional data analysis can be found in the textbooks of Ramsay and Silverman (2005), Ferraty and Vieu (2006), and Horváth and Kokoszka (2012).

We extend the pointwise consistency results of Jiang and Wang (2010) by considering the local linear estimators for the bivariate mean and the trivariate covariance function under a double asymptotic. This asymptotic contains all cases from (very) sparsely to (very) densely sampled discretization points per function and therefore takes into account the somewhat intermediate case in our application. Furthermore, we consider the practically relevant case of an autocorrelated time series of (latent) random functions P_t. This time series context translates the so-called "within-function" covariances (Li and Hsing, 2010) into additional "between-functions" covariances, which together lead to rather specific covariance components that are unique for smoothing functional data. Besides pointwise consistency, Jiang and Wang (2010) also derive asymptotic normality and uniform consistency of the local linear estimators. These results are not of use for the specific focus of this paper, but can be derived using similar arguments.

The particular emphasis of this paper lies in the derivation of asymptotically optimal bandwidth expressions for smoothing in X- and Z-direction. By means of our double asymptotic, we show that the asymptotic first- and second-order variance terms of the local linear estimators can switch places depending on how fast n diverges with T. As long as n remains constant (i.e., the case of sparse functional data) or diverges only very slowly with T, namely $n \sim T^{\theta}$ with $0 \leq \theta < 1/5$, the multiple bandwidth selection problem follows the classical bias-variance tradeoff. Though, already for a diverging $n \sim T^{\theta}$ with $\theta > 1/5$, the first- and second-order variance terms switch places and thereby annul the classical bias-variance tradeoff with respect to smoothing in X-direction. Then, a rather unconventional multiple bandwidth choice becomes optimal, where the single X- and Z-directional bandwidths are anti-proportional to each other for given values of T and n. It turns out that these new bandwidth expressions essentially lead to an under-smoothing of the functional data components for the sake of better estimates of the mean and covariance function. Within the context of pre-smoothing functional data, this under-smoothing strategy is described by Benko et al. (2009). Our results reveal the optimality of the under-smoothing strategy for the context of pooling (non-) sparse functional data. In fact, our new bandwidth expressions allow the classical multivariate local linear estimators to *inherently* under-smooth the functional data components.

We show that the multivariate local linear estimators of μ and γ can achieve the fast convergence rates of univariate nonparametric estimators. The latter is assumed in Jiang and Wang (2010), though, our asymptotic results show that it is necessary to switch to the above-mentioned unconventional bandwidths in order to achieve these univariate convergence rates. Beyond this, we demonstrate that a naive usage of the classical bandwidth expressions, may even lead to arbitrarily large estimation errors.

Our theoretical results are of direct practical use in our real data application. There the size of the data set makes it practically impossible to use cross-validation to solve the multiple bandwidth selection problem – particularly for estimating the covariance function. Our bandwidth results, when combined with one of the well-known rule of thumb procedures, are therefore very useful in practice. Furthermore, our asymptotic results are also of direct use for constructing confidence bands that allow us to test for differences in the mean electricity spot prices one year before and one year after Germany's nuclear phaseout.

The rest of this paper is structured as follows: The next section introduces our statistical models. Section 3 presents our main results. Section 4 contains a small simulation study. Our real data study can be found Section 5 and Section 6 concludes. All proofs and further detailed discussions can be found in the supplemental paper.

2 Model and assumptions

We consider the following statistical model for the noisy discretization points $(Y_{it}, X_{it}) \in \mathbb{R}^2$ of a latent centered weakly stationary time series of random functions $P_t^c(.,Z_t) \in L^2[a(Z_t), b(Z_t)]$ adjusted by the covariate $Z_t \in \mathbb{R}$:

$$Y_{it} = \mu(X_{it}, Z_t) + P_t^c(X_{it}, Z_t) + \epsilon_{it}, \tag{2}$$

with $\mu(X_{it}, Z_t) = \mathbb{E}(P_t(X_{it}, Z_t)|\mathbf{X}, \mathbf{Z})$, $i \in \{1, \dots, n\}$, and $t \in \{1, \dots, T\}$, where n refers to the number of discretization points per function and T to the total number of functions, $\mathbf{X}$ and $\mathbf{Z}$ denote the data vectors with $\mathbf{X} = (X_{11}, X_{21}, \dots, X_{nT})^\top$ and $\mathbf{Z} = (Z_1, \dots, Z_T)^\top$, and ϵ_{it} is a classical iid error term with mean zero and finite variance $\mathbb{V}(\epsilon_{it}) = \sigma_\epsilon^2$.

The mean model in Eq. (2) directly implies the following covariance model:

$$C_{ijt} = \gamma(X_{it}, X_{jt}, Z_t) + \tilde{P}_t^c(X_{it}, X_{jt}, Z_t) + \varepsilon_{ijt} \tag{3}$$

with $i \neq j \in \{1, \dots, n\}$ and $t \in \{1, \dots, T\}$, where the "raw covariances" C_{ijt}, the centered stationary functional time series term $\tilde{P}_t^c$, and the covariance function γ are defined as

$$\begin{aligned} C_{ijt} &= (Y_{it} - \mu(X_{it}, Z_t))(Y_{jt} - \mu(X_{jt}, Z_t)), \\ \tilde{P}_t^c(X_{it}, X_{jt}, Z_t) &= P_t^c(X_{it}, Z_t)\, P_t^c(X_{jt}, Z_t) - \gamma(X_{it}, X_{jt}, Z_t), \text{ and} \\ \gamma(X_{it}, X_{jt}, Z_t) &= \mathbb{E}(P_t^c(X_{it}, Z_t)\, P_t^c(X_{jt}, Z_t)|\mathbf{X}, \mathbf{Z}). \end{aligned} \tag{4}$$

The scalar error term ε_{ijt} with $\varepsilon_{ijt} = P_t^c(X_{it}, Z_t)\epsilon_{jt} + P_t^c(X_{jt}, Z_t)\epsilon_{it} + \epsilon_{it}\epsilon_{jt}$ is uncorrelated, but by construction heteroscedastic with $\mathbb{V}(\varepsilon_{ijt}) = \sigma_\varepsilon^2(x_1, x_2, z)$, where $\sigma_\varepsilon^2(x_1, x_2, z) = \gamma(x_1, x_1, z)\, \sigma_\epsilon^2 + \gamma(x_2, x_2, z)\sigma_\epsilon^2 + \sigma_\epsilon^4$. Note that $\mathbb{E}(\varepsilon_{ijt}) \neq 0$ for all $i = j$, therefore all raw covariance points C_{ijt} with $i = j$ have to be excluded (see also Yao et al., 2005). Correspondingly, the number of raw covariance points for a time point t is given by $N = n^2 - n$, which makes it necessary that $n \geq 2$. In order to improve the readability of our results we assume that $n = n_t$ for all t, although minor modifications would allow for different n_ts; see also our remarks to assumption A-AS in Section 2.1.

The statistical properties of the local linear estimator for the regression function $\mu(X_{it}, Z_t) = \mathbb{E}(Y_{it}|\mathbf{X}, \mathbf{Z})$ depend on the covariance structure of the error term, where here additionally to the standard error term ϵ_{it} the random function P_t^c in Eq. (2) acts like a nonstandard error term. The functional nature of this error term leads to "within-function" covariances

$$\mathrm{Cov}(Y_{it}, Y_{jt}|\mathbf{X}, \mathbf{Z}) = \mathbb{E}(P_t^c(X_{it}, Z_t)P_t^c(X_{jt}, Z_t)|\mathbf{X}, \mathbf{Z}) = \gamma(X_{it}, X_{jt}, Z_t)$$

for all $i \neq j$. Our time series context translates the latter into additional "between-functions" covariances, such that in summary

$$\begin{aligned} \mathrm{Cov}(Y_{it}, Y_{js}|\mathbf{X}, \mathbf{Z}) \;=\; & (\gamma(X_{it}, X_{js}, Z_t) + \sigma_\epsilon^2)\, \mathbb{1}_{[(i,t)=(j,s)]} && \big\}\text{ classical ("scalar-like")} \\ & +\gamma(X_{it}, X_{js}, Z_t)\mathbb{1}_{[i \neq j \text{ AND } t=s]} && \big\}\text{ within-fct} \\ & +\gamma_{|t-s|}((X_{it}, Z_t), (X_{js}, Z_s))\mathbb{1}_{[t \neq s]}, && \big\}\text{ between-fct} \end{aligned} \tag{5}$$

where $\gamma_{|t-s|}((x_1, z), (x_2, z)) = \mathbb{E}(P_t^c(x_1, z)P_s^c(x_2, z))$ and $\mathbb{1}_{[.]}$ denotes the indicator function. The equivalent structure is found for the composed error structure in Eq. (3), though, the within-function covariance component becomes more involved, i.e., $\mathrm{Cov}(C_{ijt}, C_{kls}|\mathbf{X}, \mathbf{Z}) =$

$$\begin{aligned}
&\big(\tilde{\gamma}((X_{it}, X_{jt}), (X_{ks}, X_{ls}), Z_t) + \sigma_\varepsilon^2(X_{it}, X_{jt}, Z_t)\big)\, \mathbb{1}_{[(i,j,t)=(k,l,s)]} && \big\}\ \text{classical (``scalar-like'')} \\
&+ (\tilde{\gamma}((X_{it}, X_{jt}), (X_{ks}, X_{ls}), Z_t))\, \mathbb{1}_{[\ (i,j)\neq(k,l)\ \ \text{AND } t=s]} && \big\}\ \text{within-fct (1. part)} \\
&\qquad\qquad +\gamma(X_{it}, X_{ks}, Z_t)\sigma_\epsilon^2\, \mathbb{1}_{[i\neq k \text{ AND } j=l \text{ AND } t=s]} && \big\}\ \text{within-fct (2. part)} \qquad (6) \\
&\qquad\qquad +\gamma(X_{jt}, X_{ls}, Z_t)\sigma_\epsilon^2\, \mathbb{1}_{[i=k \text{ AND } j\neq l \text{ AND } t=s]} && \big\}\ \text{within-fct (3. part)} \\
&+\tilde{\gamma}_{|t-s|}((X_{it}, X_{jt}, Z_t), (X_{ks}, X_{ls}, Z_s))\mathbb{1}_{[t\neq s]}, && \big\}\ \text{between-fct}
\end{aligned}$$

where $\tilde{\gamma}((x_1, x_2), (x_1', x_2'), z) = \mathbb{E}(\tilde{P}_t^c(x_1, x_2, z)\tilde{P}_t^c(x_1', x_2', z))$ and correspondingly $\tilde{\gamma}_u((x_1, x_2, z), (x_1', x_2', z')) = \mathbb{E}(\tilde{P}_t^c(x_1, x_2, z)\tilde{P}_{t+u}^c(x_1', x_2', z'))$.

In order to formally consider the autocovariances of the random functions we impose a weak dependency assumption formalized by a means of uniformly geometrically bounded autocovariance functions, i.e., that $\sup_{x_1,x_2,z_1,z_2} |\gamma_u((x_1, z_1), (x_2, z_2))| \leq c_\gamma r^u$ with $0 < c_\gamma < \infty$ and $0 < r < 1$. By construction it follows then that the autocovariance function $\tilde{\gamma}_u$ is also uniformly geometrically bounded, i.e., that $\sup_{x_1,x_1',x_2,x_2',z_1,z_2} |\tilde{\gamma}_u((x_1, x_2, z), (x_1', x_2', z'))| \leq c_{\tilde{\gamma}} r^u$ with $0 < c_{\tilde{\gamma}} < \infty$ and $0 < r < 1$. These weak dependency assumptions include the important case of functional AR processes as discussed, e.g., in Bosq (2000). The conceptual simplicity of geometrically bounded autocovariances is used so as not to distract from the already non trivial covariance structure. More general weak dependency concepts, such as strong mixing or the L^p-m approximability of Hörmann and Kokoszka (2010) could be used as well without changing our basic results.

For simplicity, the conditional random variables $X_{it}|Z_t$ and $X_{it}, X_{jt}|Z_t$ are assumed to be iid as $X|Z$ and $X_1, X_2|Z$ with joint pdfs f_{XZ} and f_{XXZ} and marginal pdfs f_X and f_Z, where all pdfs are assumed to be bounded away from zero over their compact supports S_{XZ}, S_{XXZ}, S_X, and S_Z and equal zero everywhere else. As usually in nonparametric estimation, all pdfs are assumed to be continuously differentiable for all points within their supports. Furthermore, all second-order derivatives of μ are assumed to be continuous for all points within its compact support S_{XZ} and the (auto-)covariance functions γ and γ_u are continuously differentiable for all points within their supports S_{XXZ} and S_{XZ}. (The equivalent assumptions apply to γ, $\tilde{\gamma}$ and $\tilde{\gamma}_u$.) Further classical, yet more detailed assumptions are listed in Section 2.1.

2.1 Assumptions

A-AS (Asymptotic Scenario): $Tn \to \infty$, where $n = n(T) \geq 2$ such that $n(T) \sim T^{\theta}$ with $0 \leq \theta < \infty$. Hereby, "$n(T) \sim T^{\theta}$" denotes that the two sequences $n(T)$ and T^{θ} are asymptotically equivalent, i.e., that $0 < \lim_{T\to\infty}(n(T)/T^{\theta}) < \infty$.

Remark on assumption A-AS: The number of functions T drives the asymptotics as $Tn \sim T^{(1+\theta)} \to \infty$. The case $\theta = 0$ implies that n is bounded and corresponds to the case of sparse functional data as considered by Yao et al. (2005). For $0 < \theta < \infty$ we are able to consider any scenario from sparsely to (very) densely sampled discretization points per function. We follow Hall et al. (2006) and consider deterministic sequences n, though minor modifications would allow us to deal with a random variable n. For instance, our results are directly applicable to the case in which $n_t = n_t(T)$ is a random variable defined as $n_t(T) = n(T) + Z_t(T)$, and where $n(T)$ is deterministic with $n(T) \sim T^{\theta}$ and $Z_t(T)$ is an independent (except for its dependency on T) random variable with realizations in some appropriate finite subset of $\mathbb{N}$, such that $n_t(T) \geq 2$ a.s. and $\mathbb{E}(n_t(T)) = n(T)$ for all T as $T \to \infty$.

A-RD (Random Design) The triple (Y_{it}, X_{it}, Z_t) has the same distribution as (Y, X, Z) with pdf f_{YXZ}. The conditional r.v. $X|Z$ is iid with pdf $f_{X|Z}(x) > 0$ for all $x \in S_{X|Z}$ and zero else, where $S_{X|Z} = [a(Z), b(Z)] \subset \mathbb{R}$, and $a(z) < b(z)$ for all $z \in [z_{\min}, z_{\max}]$. The marginal pdf of X is $f_X(x) > 0$ for all $x \in S_X$, where $S_X = [\min_z a(z), \max_z b(z)]$. The pdf of Z is $f_Z(z) > 0$ for all $z \in S_Z$ and zero else, where $S_Z = [z_{\min}, z_{\max}] \subset \mathbb{R}$. The joint pdf is then given by $f_{XZ}(x, z) = f_{X|Z}(x)f_Z(z) > 0$ for all $(x, z) \in S_{XZ}$, where $S_{XZ} = S_{X|Z} \times S_Z$. This implies that the conditional pdf of $(X, X)|Z$ is $f_{XX|Z}(x_1, x_2) > 0$ for all $(x_1, x_2) \in S^2_{X|Z}$ and zero else. The joint pdf is given by $f_{XXZ}(x_1, x_2, z) = f_{XX|Z}(x_1, x_2)f_Z(z) > 0$ for all $(x_1, x_2, z) \in S_{XXZ}$, where $S_{XXZ} = S^2_{X|Z} \times S_Z$. All pdf's in are continuously differentiable for all points within their supports.

A-SM (Smoothness) All second-order derivatives of μ (and γ) are continuous for all points within its support. The autocovariance functions γ (and $\tilde{\gamma}$) and γ_u (and $\tilde{\gamma}_u$) are continuously differentiable for all points within their supports.

A-BW (Bandwidths) $h_{\mu,X}, h_{\mu,Z} \to 0$ and $(Tn)^{-1}h_{\mu,X}, h_{\mu,Z} \to \infty$ as $Tn \to \infty$. $h_{\mu,X}, h_{\mu,Z} \to 0$ and $(Tn)^{-1}h_{\mu,X}^2, h_{\mu,Z} \to \infty$ as $Tn \to \infty$.

2.2 Local linear estimators

We estimate the mean function $\mu(x, z)$ using the local linear estimator $\hat{\mu}(x, z; h_{\mu,X}, h_{\mu,Z})$ defined as the following locally weighted least squares estimator (as in Ruppert and Wand, 1994):

$$\hat{\mu}(x, z; h_{\mu,X}, h_{\mu,Z}) = u_1^\top \left([\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]\right)^{-1} [\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz}\mathbf{Y}, \qquad (7)$$

where the vector $u_1 = (1, 0, 0)^\top$ selects the estimated intercept, the partitioned $nT \times 3$ dimensional data matrix $[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]$ has typical rows $(1, X_{it} - x, Z_t - z)$, the $nT \times nT$ dimensional diagonal weighting matrix $\mathbf{W}_{\mu,xz}$ holds the bivariate kernel weights $K_{\mu,h}(X_{it} - x, Z_t - z)$, and all vectors and matrices are filled in correspondence with the nT dimensional vector $\mathbf{Y} = (Y_{11}, Y_{21}, \dots, Y_{n-1,T}, Y_{n,T})^\top$. As bivariate kernel weights we use multiplicative kernel functions $K_{\mu,h}(u, v) = h_{\mu,X}^{-1}\,\kappa(h_{\mu,X}^{-1}(u)) \cdot h_{\mu,Z}^{-1}\,\kappa(h_{\mu,Z}^{-1}(v))$, where κ is a univariate, symmetric, probability density function (pdf) with compact support $\text{supp}(\kappa) = [-1, 1]$, such as, e.g., the univariate Epanechnikov kernel. The usual kernel constants are $\nu_2(K_\mu) = (\nu_2(\kappa))^2$, with $\nu_2(\kappa) = \int u^2\kappa(u)du$, and $R(K_\mu) = R(\kappa)^2$, with $R(\kappa) = \int \kappa(u)^2 du$.

The estimator for the covariance function $\gamma(x_1, x_2, z)$ is defined correspondingly as

$$\begin{aligned}&\hat{\gamma}(x_1, x_2, z; h_{\gamma,X} h_{\gamma,Z}) = \\ &= u_1^\top \left([\mathbf{1}, \mathbf{X}_{x_1}, \mathbf{X}_{x_2}, \mathbf{Z}_z]^\top \mathbf{W}_{\gamma,x_1x_2z}[\mathbf{1}, \mathbf{X}_{x_1}, \mathbf{X}_{x_2}, \mathbf{Z}_z]\right)^{-1} [\mathbf{1}, \mathbf{X}_{x_1}, \mathbf{X}_{x_2}, \mathbf{Z}_z]^\top \mathbf{W}_{\gamma,x_1x_2z}\hat{\mathbf{C}}, \qquad (8)\end{aligned}$$

where $u_1 = (1, 0, 0, 0)^\top$, $[\mathbf{1}, \mathbf{X}_{x_1}, \mathbf{X}_{x_2}, \mathbf{Z}_z]$ is a $NT \times 4$ dimensional data matrix with typical rows $(1, X_{it} - x_1, X_{jt} - x_2, Z_t - z)$, the $NT \times NT$ dimensional diagonal weighting matrix $\mathbf{W}_{\gamma,x_1x_2z}$ holds the trivariate kernel weights $K_{\mu,h}(X_{it} - x_1, X_{jt} - x_2, Z_t - z)$, and all vectors and matrices are filled in correspondence with the NT dimensional vector $\hat{\mathbf{C}} = (\hat{C}_{211}, \hat{C}_{311}, \dots, \hat{C}_{n-2,n,T}, \hat{C}_{n-1,n,T})^\top$, where the empirical raw covariances $\hat{C}_{ijt}$'s are defined as

$$\hat{C}_{ijt} = (Y_{it} - \hat{\mu}(X_{it}, Z_t))(Y_{jt} - \hat{\mu}(X_{jt}, Z_t)). \qquad (9)$$

As trivariate kernel weights $K_{\gamma,h}(X_{it}-x_1, X_{jt}-x_2, Z_t-z)$ we use multiplicative kernel functions $K_{\gamma,h}(u_1,u_2,v) = h_{\gamma,X}^{-1}\,\kappa(h_{\gamma,X}^{-1}(u_1)) \cdot h_{\gamma,X}^{-1}\,\kappa(h_{\gamma,X}^{-1}(u_2)) \cdot h_{\gamma,Z}^{-1}\,\kappa(h_{\mu,Z}^{-1}(v))$. The usual kernel constants are $\nu_2(K_\gamma) = (\nu_2(\kappa))^3$ and $R(K_\gamma) = R(\kappa)^3$.

It might be considered as restrictive to use only two bandwidths $h_{\gamma,X}$ and $h_{\gamma,Z}$ for the trivariate nonparametric estimator $\hat{\gamma}$ in Eq. (8). However, the use of equal bandwidths in both X-directions is not restrictive at all, since by construction (identical pairs at mirrored positions) the prediction points X_{it} and X_{jt} are uncorrelated and measured on the same scale.

3 Main Results

Below in the Theorems 3.1 and 3.2 we consider only interior points, as the consideration of boundary points does not add any new insights. It is well known that local linear estimators automatically correct for boundary effects (Fan and Gijbels (1996)).

Theorem 3.1 *Under our setup the conditional asymptotic bias and variance of the local linear estimator in Eq.* (7) *are given by*

(i) $\text{Bias}\,\{\hat{\mu}(x,z;h_{\mu,X},h_{\mu,Z})|\mathbf{X},\mathbf{Z}\} = \frac{1}{2}\,\nu_2(K_\mu)\,\left(h_{\mu,X}^2\,\mu^{(2,0)}(x,z) + h_{\mu,Z}^2\,\mu^{(0,2)}(x,z)\right)(1+o_p(1))$

(ii) $\mathbb{V}\,\{\hat{\mu}(x,z;h_{\mu,X},h_{\mu,Z})|\mathbf{X},\mathbf{Z}\} = (S_1^\mu(x,z) + S_2^\mu(x,z))\,(1+o_p(1))$ *with*

$S_1^\mu(x,z) = (Tn)^{-1}\left[h_{\mu,X}^{-1}h_{\mu,Z}^{-1}\,R(K_\mu)\dfrac{\gamma(x,x,z)+\sigma_\epsilon^2}{f_{XZ}(x,z)}\right]$ *and*

$S_2^\mu(x,z) = T^{-1}\left[h_{\mu,Z}^{-1}\,R(\kappa)\dfrac{\gamma(x,x,z)}{f_Z(z)} + c_r\right]$,

where $\mu^{(k,l)}(x,z) = (\partial^{k+l}/(\partial x^k \partial z^l))\mu(x,z)$ *and* $|c_r| \le \frac{2c_\gamma r}{1-r} < \infty$.

Theorem 3.2 *Under our setup the conditional asymptotic bias and variance of the local* *linear estimator in Eq.* (8) *are given by*

(i) $\operatorname{Bias}\{\hat{\gamma}(x_1, x_2, z; h_{\gamma,X}, h_{\gamma,Z})|\mathbf{X}, \mathbf{Z}\} =$

$$= \frac{1}{2}\nu_2(K_\gamma)\left(h_{\gamma,X}^2\left(\gamma^{(2,0,0)}(x_1,x_2,z) + \gamma^{(0,2,0)}(x_1,x_2,z)\right) + h_{\gamma,Z}^2\gamma^{(0,0,2)}(x_1,x_2,z)\right)(1+o_p(1))$$

(ii) $\mathbb{V}\{\hat{\gamma}(x_1, x_2, z; h_{\gamma,X}, h_{\gamma,Z})|\mathbf{X}, \mathbf{Z}\} = (S_1^\gamma(x_1,x_2,z) + S_2^\gamma(x_1,x_2,z))(1+o_p(1))$ *with*

$$S_1^\gamma(x_1,x_2,z) = (TN)^{-1}\left[h_{\gamma,X}^{-2}h_{\gamma,Z}^{-1}\,R(K_\gamma)\frac{\tilde{\gamma}((x_1,x_2),(x_1,x_2),z) + \sigma_\varepsilon^2(x_1,x_2,z)}{f_{XXZ}(x_1,x_2,z)}\right] \quad \textit{and}$$

$$S_2^\gamma(x_1,x_2,z) = T^{-1}\left[h_{\gamma,Z}^{-1}\,R(\kappa)\frac{\tilde{\gamma}((x_1,x_2),(x_1,x_2),z) + \left(\gamma(x_1,x_2,z)\sigma_\epsilon^2\right)\,\mathbb{1}_{[\theta=0]}}{f_Z(z)} + \tilde{c}_r\right],$$

where $\gamma^{(k,l,m)}(x_1,x_2,z) = (\partial^{k+l+m}/(\partial x_1^k\,\partial x_2^l\,\partial z^m))\gamma(x_1,x_2,z)$ *and* $|\tilde{c}_r| \leq \frac{2c_{\tilde{\gamma}}r}{1-r} < \infty$*.*

The proofs of Theorems 3.1 and 3.2 can be found in Section 1.1 of the supplemental paper.

The bias expressions correspond to the classical bias results (see, e.g., Ruppert and Wand, 1994). Also the first summands $S_1^\mu(x,z)$ and $S_1^\gamma(x_1,x_2,z)$ of the variance results are essentially classical as they capture the classical variance effect of the scalar error terms, namely σ_ϵ^2 and $\sigma_\varepsilon^2(x_1,x_2,z)$, and the "scalar-like" variance parts of the functional error terms; see Eqs. (5) and (6).

The second summands $S_2^\mu(x,z)$ and $S_2^\gamma(x_1,x_2,z)$ in the variance expressions are more interesting as they are induced by the within- and between-function covariances of the functional error terms. The first terms within the square brackets quantify the within-function variance effects and the second summands (i.e., c_r and $\tilde{c}_r$) quantify the between-function variance effects. As Eq. (6) demonstrates, the within-function variance effects are more involved for the case of estimating the covariance function. The 2. and 3. parts of the within-function variance effects in Eq. (6) cause the additional $\left(\gamma(x_1,x_2,z)\sigma_\epsilon^2\right)\,\mathbb{1}_{[\theta=0]}$ term.

The pure cross-sectional case is included, since for $r \to 0$ in our weak dependency assumptions we have that $\lim_{r\to 0} c_r = \lim_{r\to 0}\tilde{c}_r = 0$. Note that the between-function variance effects, i.e., c_r and $\tilde{c}_r$ are only shown to improve understanding. However, they are actually of a negligible order of magnitude as under our classical bandwidth assumptions A-BW $T^{-1}c_r = o(T^{-1}h_{\mu,Z}^{-1})$ and $T^{-1}\tilde{c}_r = o(T^{-1}h_{\gamma,Z}^{-1})$. The reason for this is that the data are localized in the domain of the X and Z variables and not in the time domain. The resulting decorrelation effect is often referred to as the "whitening window" property (Fan and Yao, 2003, Ch. 5.3). In the case without covariate-adjustments, though, the between-function variance effect will not be negligible.

The most striking result is that we cannot readily identify which of the two variance terms S_1 and S_2 in each of the Theorems 3.1 and 3.2 the dominating ones are. The answer depends on how fast n diverges with T, i.e., on the value of θ in $n \sim T^{\theta}$; see assumption A-AS. In order to determine the decisive θ values we conduct a case-by-case study and initially postulate that either the first or the second variance summands are dominating. This allows us to derive the casewise optimal bandwidth sequences with respect to the conditional asymptotic mean integrated squared error (AMISE) criterion. Once we know the casewise AMISE-optimal bandwidths, we can determine the decisive θ values.

In correspondence with the literature (see, e.g., Fan and Gijbels, 1996) we use the following weighted AMISE criteria:

$$\begin{aligned}\text{AMISE}_{\hat{\mu}}&= \int_{S_{XZ}} \left([\text{Bias}\left\{\hat{\mu}(x,z)|\mathbf{X},\mathbf{Z}\right\}]^2 + \mathbb{V}\left\{\hat{\mu}(x,z)|\mathbf{X},\mathbf{Z}\right\}\right) w_{\mu}\, d(x,z)\\ \text{AMISE}_{\hat{\gamma}}&= \int_{S_{XXZ}} \left([\text{Bias}\left\{\hat{\gamma}(x_1,x_2,z)|\mathbf{X},\mathbf{Z}\right\}]^2 + \mathbb{V}\left\{\hat{\gamma}(x_1,x_2,z)|\mathbf{X},\mathbf{Z}\right\}\right) w_{\gamma}\, d(x_1,x_2,z)\end{aligned}$$

with weight functions $w_{\mu} = w_{\mu}(x,z)$ and $w_{\gamma} = w_{\gamma}(x_1,x_2,z)$ defined as $w_{\mu}(x,z) = f_{XZ}(x,z)$ and $w_{\gamma}(x_1,x_2,z) = f_{XXZ}(x_1,x_2,z)$. Let us differentiate between the following two AMISE-cases:

AMISE. I: The first (integrated and weighted) variance summands $\int S_{\mu,1}w_{\mu}$ and $\int S_{\gamma,1}w_{\gamma}$ are strictly dominating, where $S_{\mu,1}$ and $S_{\gamma,1}$ are defined in Theorems 3.1 and 3.2.

AMISE. II: The second (integrated and weighted) variance summands $\int S_{\mu,2}w_{\mu}$ and $\int S_{\gamma,2}w_{\gamma}$ are strictly dominating, where $S_{\mu,2}$ and $S_{\gamma,2}$ are defined in Theorems 3.1 and 3.2.

In anticipation of some of our results: Under AMISE-optimal bandwidth choices, the discriminating θ-threshold is given by $\theta = 1/5$. That is, the AMISE. I scenario applies to "low" θ-values of $0 \leq \theta < 1/5$ (representing relatively small values of n) and the AMISE. II scenario applies to "great" θ-values of $1/5 < \theta < \infty$ (representing relatively large values of n). The threshold $\theta = 1/5$, however, leads to a stalemate for which the competing asymptotic variance summands are of equal orders of magnitudes which makes it impossible to derive explicit AMISE-optimal bandwidth expressions.

3.1 AMISE. I optimal bandwidth selection

The AMISE. I hypothesis, i.e., that the first variance summands are strictly dominating can be formalized by the following inequalities:

$$T^{-1}\,h_{\mu,Z}^{-1}=o\left(T^{-(1+\theta)}\,h_{\mu,X}^{-1}\,h_{\mu,Z}^{-1}\right)\quad\text{and}\quad T^{-1}\,h_{\gamma,Z}^{-1}=o\left(T^{-(1+2\theta)}\,h_{\gamma,X}^{-2}\,h_{\gamma,Z}^{-1}\right),$$

where we used that by assumption A-AS $Tn\sim T^{1+\theta}$ and $TN\sim T^{1+2\theta}$.

Particularly for the covariance estimator, the derivation of the multiple AMISE. I optimal bandwidths is a bit tedious, but nevertheless follows the usual steps for minimizing the classical bias-variance tradeoff. Therefore, the AMISE. I optimal bandwidth sequences are of the well-known orders of magnitude for two and three dimensional nonparametric estimators: For the estimator $\hat{\mu}(x,z,h_X,h_Z)$ we have $h_{\mu,X,\text{AMISE.I}}\sim h_{\mu,Z,\text{AMISE.I}}\sim(Tn)^{1/6}$ and for $\hat{\gamma}(x_1,x_2z,h_X,h_Z)$ we have $h_{\gamma,X,\text{AMISE.I}}\sim h_{\gamma,Z,\text{AMISE.I}}\sim(TN)^{1/7}$. The explicit AMISE. I optimal bandwidth expressions can be found in Eqs. (34), (35), (36), and (37) of the supplemental paper.

Plugging these AMISE. I optimal bandwidth rates into the above strict order relations shows that the latter hold for all θ with $0\leq\theta<1/5$. However, if, e.g., $\theta=1/5$ both variance summands will be of the same order of magnitude, i.e.,

$$T^{-1}\,h_{\mu,Z,\text{AMISE.I}}^{-1}\sim\left(T^{-(1+\theta)}\,h_{\mu,X,\text{AMISE.I}}^{-1}\,h_{\mu,Z,\text{AMISE.I}}^{-1}\right)\quad\text{and}\tag{10}$$

$$T^{-1}\,h_{\gamma,Z,\text{AMISE.I}}^{-1}\sim\left(T^{-(1+2\theta)}\,h_{\gamma,X,\text{AMISE.I}}^{-2}\,h_{\gamma,Z,\text{AMISE.I}}^{-1}\right),\tag{11}$$

such that the AMISE. I hypothesis is false and the AMISE. I optimal bandwidths in Eqs. (34), (35), (36), and (37) are no longer the asymptotically optimal bandwidths that minimize the AMISE. Similar arguments apply to the case $\theta>1/5$. Interestingly, the same θ-threshold value of 1/5 applies to both estimators $\hat{\mu}$ and $\hat{\gamma}$.

3.2 AMISE. II optimal bandwidth selection

Under the hypothesis that the second variance summands are dominating, it is possible to archive convergence rates of univariate nonparametric estimators. In the following this is explained with a focus on the estimator $\hat{\mu}$. The equivalent arguments apply to the estimator $\hat{\gamma}$ and therefore are shifted to the supplemental paper.

By contrast to the AMISE. I case, it is impossible to determine the optimal bandwidths in X- and Z-direction by using only the asymptotic first order terms. The problem is that the second variance term S_2^μ does not depend on the X-directional bandwidth. This (partially) annuls the classical bias-variance tradeoff for the X-direction and it becomes tempting to lower the X-directional bias by choosing a (very) small X-bandwidth, e.g., $h_{\mu,X} \approx 0$. Unfortunately, too small a X-bandwidth recovers the dominance of the (integrated and weighted) first variance summand $\int S_1^\mu w_\mu$ over the (integrated and weighted) second variance summand $\int S_1^\mu w_\mu$, since the former is scaled by the reciprocal value $h_{\mu,X}^{-1}$; see Eq. (12). Therefore, in order to derive the expression for the optimal X-bandwidth under the AMISE. II scenario, we need to consider both competing variance summands. That is, the appropriate AMISE. II expression of the two-dimensional local linear estimator $\hat{\mu}$ needs to include both variance terms, such that

$$\begin{aligned} \text{AMISE. II}_{\hat{\mu}}\,(h_{\mu,X}, h_{\mu,Z}) = {} & \underbrace{\overbrace{(Tn)^{-1}\, h_{\mu,X}^{-1}\, h_{\mu,Z}^{-1}\, R(K_\mu)\, Q_{\mu,1}}^{\int S_1^\mu w_\mu}}_{\text{2nd Order}} + \underbrace{\overbrace{T^{-1}\, h_{\mu,Z}^{-1}\, R(\kappa)\, Q_{\mu,2}}^{\int S_2^\mu w_\mu}}_{\text{1st Order}} + \\ & + \frac{1}{4}\,(\nu_2(K_\mu))^2 \left[\underbrace{h_{\mu,X}^4\, \mathcal{I}_{\mu,XX}}_{\text{3rd Order}} + \underbrace{2\, h_{\mu,X}^2\, h_{\mu,Z}^2\, \mathcal{I}_{\mu,XZ}}_{\text{2nd Order}} + \underbrace{h_{\mu,Z}^4\, \mathcal{I}_{\mu,ZZ}}_{\text{1st Order}} \right], \end{aligned} \tag{12}$$

where $\mathcal{I}_{\mu,XX} = \int_{S_{XZ}} (\mu^{(2,0)}(x,z))^2\, w_\mu\, d(x,z)$, $\mathcal{I}_{\mu,ZZ} = \int_{S_{XZ}} (\mu^{(0,2)}(x,z))^2\, w_\mu\, d(x,z)$, $\mathcal{I}_{\mu,XZ} = \int_{S_{XZ}} \mu^{(2,0)}(x,z)\mu^{(0,2)}(x,z)\, w_\mu\, d(x,z)$, $Q_{\mu,1} = \int_{S_{XZ}} \left(\gamma(x,x,z) + \sigma_\epsilon^2\right)\, d(x,z)$, and $Q_{\mu,2} = \int_{S_{XZ}} \gamma(x,x,z)\, f_X(x)\, d(x,z)$. The corresponding AMISE. II$_\gamma$ expression is found in Eq. (38) of the supplemental paper.

Note that it is impossible to derive explicit AMISE optimal X- and Z-bandwidth expressions through minimizing Eqs. (12) and (38) simultaneously for both bandwidths. However, originating from the AMISE. II expressions in Eqs. (12) and (38) we can find asymptotically optimal X- and Z-bandwidth expressions by optimizing, first, with respect to the first order terms in order to get an optimal Z-bandwidth. Second, by optimizing with respect to the second order terms in order to get an optimal X-bandwidth for a given optimal Z-bandwidth.

To do so we need to determine assumptions on the X-bandwidth that first allow us to profit from the (partial) annulment of the X-directional bias-variance tradeoff through a bias reduction in X-direction. Second, they must assure that the AMISE. II scenario is maintained through only a sufficiently small increase of the first (integrated and weighted) variance term $\int S_1^\mu w_\mu$ such that the first variance term remains asymptotically negligible in comparison to the (integrated and weighted) second variance term $\int S_2^\mu w_\mu$. The first requirement is achieved if the X-bandwidth is of a smaller order of magnitude than the Z-bandwidth, i.e., if $h_{\mu,X} = o(h_{\mu,Z})$. This restriction makes those bias components that depend on $h_{\mu,X}$ asymptotically negligible, since it implies that $h_{\mu,X}^2\, h_{\mu,Z}^2 = o(h_{\mu,Z}^4)$ and therefore that $h_{\mu,X}^4 = o(h_{\mu,X}^2\, h_{\mu,Z}^2)$. The latter two strict inequalities lead to the order relations between the third, fourth, and fifth AMISE. II term as indicated in Eq. (12). The second requirement is achieved if the X-bandwidth does not converge to zero too fast, namely if $(nh_{\mu,X})^{-1} = o(1)$, which implies the order relation between the first two AMISE. II terms indicated in Eq. (12).

Let us initially assume that is possible to find an X-bandwidth that fulfills both the above requirements, namely $h_{\mu,X} = o(h_{\mu,Z})$ and $(nh_{\mu,X})^{-1} = o(1)$. With such an X-bandwidth we can make use of the order relations indicated in Eq. (12) (and in Eq. (38) for the covariance function). That is, instead of directly minimizing the AMISE. II expressions in Eqs. (12) and (38) over both X- and Z-bandwidths, we can minimize the following simpler, but asymptotically equivalent, first order approximations to the MISE of the estimators $\hat{\mu}$ and $\hat{\gamma}$, which depend only on the Z-bandwidth:

$$\text{AMISE. II}_{\hat{\mu}}^{\text{1st Order}}\,(h_{\mu,Z}) = T^{-1}\, h_{\mu,Z}^{-1}\, R(\kappa)\, Q_{\mu,2} + \frac{1}{4}\,\left(\nu_2(K_\mu)\right)^2\, h_{\mu,Z}^4\, \mathcal{I}_{\mu,ZZ}$$
$$\text{AMISE. II}_{\hat{\gamma}}^{\text{1st Order}}\,(h_{\gamma,Z}) = T^{-1}\, h_{\gamma,Z}^{-1}\, R(\kappa)\, Q_{\gamma,2} + \frac{1}{4}\,(\nu_2(K_\gamma))^2\, h_{\gamma,Z}^4\, \mathcal{I}_{\gamma,ZZ}.$$

The above equations are minimized by the following Z-bandwidths:

$$\tilde{h}_{\mu,Z,\text{AMISE.II}} = \left(\frac{R(\kappa)\, Q_{\mu,2}}{T\,\left(\nu_2(K_\mu)\right)^2\, \mathcal{I}_{\mu,ZZ}}\right)^{1/5} \quad \text{and} \tag{13}$$

$$\tilde{h}_{\gamma,Z,\text{AMISE.II}} = \left(\frac{R(\kappa)\, Q_{\gamma,2}}{T\,\left(\nu_2(K_\gamma)\right)^2\, \mathcal{I}_{\gamma,ZZ}}\right)^{1/5}, \tag{14}$$

where $Q_{\gamma,2} = \int_{S_{XXZ}} \tilde{\gamma}((x_1,x_2),(x_1,x_2),z)\, f_{XX}(x_1,x_2)\, d(x_1,x_2,z)$ and
$\mathcal{I}_{\gamma,ZZ} = \int_{S_{XXZ}} (\gamma^{(0,0,2)}(x_1,x_2,z))^2\, w_\gamma\, d(x_1,x_2,z)$.

Though, we still need to find X-bandwidths that fulfill the postulated requirements. To do so we suggest plugging in the above optimal Z-bandwidths expressions of Eqs. (13)

and (14) into the AMISE. II expressions of Eqs. (12) and (38) and to minimize the (then classical) bias-variance tradeoff between the asymptotic second order terms, which leads to the following second order approximations of the AMISE. II optimal X-bandwidths:

$$\tilde{h}_{\mu,X,\text{AMISE.II}} = \left(\frac{R(K_\mu)\, Q_{\mu,1}}{Tn\ (\nu_2(K_\mu))^2\ \mathcal{I}_{\mu,XZ}}\right)^{1/3} \left(\tilde{h}_{\mu,Z,\text{AMISE.II}}\right)^{-1} \quad \text{and} \tag{15}$$

$$\tilde{h}_{\gamma,X,\text{AMISE.II}} = \left(\frac{R(K_\gamma)\, Q_{\gamma,1}}{TN\ (\nu_2(K_\gamma))^2\ \mathcal{I}_{\gamma,X_1Z}}\right)^{1/4} \left(\tilde{h}_{\gamma,Z,\text{AMISE.II}}\right)^{-3/4}, \tag{16}$$

where $\mathcal{I}_{\gamma,X_1Z} = \int_{S_{XXZ}} \gamma^{(2,0,0)}(x_1,x_2,z)\gamma^{(0,0,2)}(x_1,x_2,z)\, w_\gamma\, d(x_1,x_2,z)$.

In order to check whether the X-bandwidths in Eqs. (15) and (16) actually fulfill the two necessary requirements, we apply some rearrangements. Using that by assumption $n \sim T^\theta$ and that by construction $N \sim n^2$, leads to the following more transparent presentation of the bandwidth rates:

$$\tilde{h}_{\mu,Z,\text{AMISE.II}} \sim n^{-1/(5\,\theta)} \quad \text{and} \quad \tilde{h}_{\mu,X,\text{AMISE.II}} \sim n^{-\eta_\mu(\theta)} \quad \text{with} \quad \eta_\mu(\theta) = \frac{1}{3} + \frac{2}{15\,\theta} \tag{17}$$

$$\tilde{h}_{\gamma,Z,\text{AMISE.II}} \sim N^{-1/(10\,\theta)} \quad \text{and} \quad \tilde{h}_{\gamma,X,\text{AMISE.II}} \sim N^{-\eta_\gamma(\theta)} \quad \text{with} \quad \eta_\gamma(\theta) = \frac{1}{4} + \frac{1}{20\,\theta}. \tag{18}$$

With Eqs. (17) and (18) it is now easily verified that the first requirements, i.e., that $\tilde{h}_{\mu,X,\text{AMISE.II}} = o(\tilde{h}_{\mu,Z,\text{AMISE.II}})$ and $\tilde{h}_{\gamma,X,\text{AMISE.II}} = o(\tilde{h}_{\gamma,Z,\text{AMISE.II}})$ are fulfilled iff $\theta > 1/5$. Furthermore, the second requirements, i.e., that $(n\tilde{h}_{\mu,X,\text{AMISE.II}})^{-1} = o(1)$ and $(N\tilde{h}^2_{\gamma,X,\text{AMISE.II}})^{-1} = o(1)$, are also fulfilled iff $\theta > 1/5$.

Summing up, if $\theta > 1/5$, we can exploit the order relations as indicated in Eq. (12) (and in Eq. (38) for the covariance function) in order to derive explicit first and second order approximations of the AMISE. II optimal Z- and X-bandwidths as given in Eqs. (13), (14), (15), and (16). It follows as a simple corollary from the above results that the differences between our explicit yet approximatively AMISE. II optimal Z- and X-bandwidths and the algebraically infeasible truly AMISE. II optimal Z- and X-bandwidths are asymptotically negligible as $Tn \to \infty$ with $n \sim T^\theta$ and $\theta > 1/5$.

3.3 Discussion of the AMISE. II optimal bandwidth results

A surprising result is that the AMISE. II optimal X-bandwidths in Eqs. (15) and (16) are *anti*-proportional to the corresponding AMISE. II optimal Z bandwidths. This is completely contrary to the classical multiple bandwidth choices where the single directional bandwidths are directly proportional to each other; see, e.g., Eqs. (35) and (37) for the AMISE. I scenario with $\theta < 1/5$. This counter-intuitive result is a direct consequence of our particular data structure in Eq. (1). If n is large (i.e., $n \gg T$), we approach the situation in which one could reveal the random functions $P_1(., Z_1), \dots, P_T(., Z_T)$ through pre-smoothing. It is, however, a scientific fact that it is optimal to use under-smoothing bandwidths in the pre-smoothing step if the actual aim is to estimate the mean or the covariance functions (Benko et al., 2009). See also Wang et al. (2008), who essentially advocate the same argument, but in a slightly different context. The reason is that estimating the mean and the covariance functions involves taking averages over the pre-smoothed functions. Taking averages reduces variance which therefore opens the possibility of further increasing the variance in the pre-smoothing step through the use of under-smoothing bandwidths for the sake of an overall lower bias.

It is exactly this logic that is reflected by the anti-proportional bandwidth relations in Eqs. (15) and (16). If n is large enough (i.e., here $n \sim T^{\theta}$ with $\theta > 1/5$), the functional nature of the data becomes "visible" in the asymptotic first order variance terms. Then we essentially compute z-localized averages across the total functions $\hat{P}_t(., Z_t)$ with $Z_t \approx z$ as if a pre-smoothing step (in X-direction) had been conducted in advance to reveal the functions $\hat{P}_t(., Z_t)$. I.e., this is then essentially a *univariate* smoothing problem only in Z-direction, since the Z-bandwidth localizes the *total* functions and therefore equally applies to all evaluation points x of the random functions $\hat{P}_t(x, Z_t)$. This univariate nature is reflected by the AMISE. II optimal Z bandwidths in Eqs. (13), (14), which are, by their rates and constant components, univariate smoothing parameters. This inherent pre-smoothing of the functions in X-direction explains the anti-proportional bandwidth relations in Eqs. (15) and (16): The larger the Z-bandwidth, the more functional data units are used to compute a z-localized averages. Consequently, the choice of a smaller X-bandwidth which lowers the bias and increases the variance becomes optimal. If, however, n is small (i.e., here $n \sim T^{\theta}$ with $\theta < 1/5$) then, the functional nature of the data remains "hidden" and the effect of the classical scalar valued error term Eq. (1) is visible in the asymptotic first order terms. Under this situation the classical directly proportional bandwidth relations apply as derived in Eqs. (34), (35), (36), and (37).

A further consequence of the above considerations is that the convergence rates of the estimators for the covariate-adjusted mean and covariance functions are bounded from above by the univariate convergence rate of the Z-directional smoothing error. This is not surprising, since even if we could observe the random functions $P_t(.,Z_t)$ perfectly we are left with a Z-directional smoothing error. For this see also the following section.

3.4 Converging and diverging AMISE rates

The above considerations and derived bandwidths results directly lead to the AMISE rates for general sequences n and T with $n \sim T^\theta$, which we formalize in the following theorem:

Theorem 3.3 *Under our set up, the minimal conditional* AMISE *rates for the estimators in Eqs.* (7) *and* (8) *are given by*

$$\min_{h_{\mu,X},\,h_{\mu,Z}\in\mathbb{R}} \text{AMISE}_{\hat{\mu}}(h_{\mu,X},\,h_{\mu,Z}) = \begin{cases} O_p\left((Tn)^{-2/3}\right) & \text{if } 0 \leq \theta \leq 1/5 \\ O_p\left(T^{-4/5}\right) & \text{if } \theta > 1/5 \end{cases} \tag{19}$$

$$\min_{h_{\gamma,X},\,h_{\gamma,Z}\in\mathbb{R}} \text{AMISE}_{\hat{\gamma}}(h_{\gamma,X},\,h_{\gamma,Z}) = \begin{cases} O_p\left((TN)^{-4/7}\right) & \text{if } 0 \leq \theta \leq 1/5 \\ O_p\left(T^{-4/5}\right) & \text{if } \theta > 1/5 \end{cases}. \tag{20}$$

The proof of Theorem 3.3 for the cases $0 < \theta < 1/5$ and $\theta > 1/5$ follows directly from our above AMISE. I and AMISE. II optimal bandwidth results. For the case $\theta = 1/5$ we have no explicit AMISE-optimal bandwidth expression, though, the corresponding AMISE rate follows from the results in Eqs. (10) and (11).

This summarizes a striking result: For rather slowly diverging sequences $n \sim T^\theta$ with θ slightly greater than 1/5 the two- and three-dimensional local linear estimators $\hat{\mu}$ and $\hat{\gamma}$ obtain univariate nonparametric convergence rates in T. That is, the estimation problems become first order equivalent to the case in which the random functions $P_t(.,Z_z)$ are observed directly, such that smoothing has to be done only across the covariate Z_t. In order to achieve this, however, it is necessary to switch to the AMISE. II optimal bandwidths as given in Eqs. (13), (14), (15), and (16) with their non-standard convergence rates.

It is important to emphasize that Theorem 3.3 requires a correct choice of the AMISE. I and AMISE. II optimal bandwidth expressions. In fact, using the AMISE. I optimal bandwidths under the $\theta > 1/5$ scenario instead of the correct AMISE. II optimal bandwidths can result in diverging variance components and therefore in *diverging* AMISE rates. More precisely, note that the orders of magnitudes of the second (integrated and weighted) variance terms, $\int S_2^\mu w_\mu$ and $\int S_2^\gamma w_\gamma$, under AMISE. I optimal bandwidths are given by

$$\text{mean function:} \quad T^{-1}\, h^{-1}_{\mu,Z,\text{AMISE.I}} \sim T^{-1}\,(Tn)^{1/6} \sim T^{-5/6}\,T^{\theta/6} \quad \text{and}$$
$$\text{covariance function:} \quad T^{-1}\, h^{-1}_{\gamma,Z,\text{AMISE.I}} \sim T^{-1}\,(TN)^{1/7} \sim T^{-6/7}\,T^{2\theta/7},$$

where we use that $n \sim T^\theta$. The above rates $T^{(\theta-5)/6}$ and $T^{(2\theta-6)/7}$ are diverging sequences as $T \to \infty$ for $\theta > 5$ and $\theta > 3$, respectively.

4 Simulation

To compare the finite sample properties of the AMISE. I and AMISE. II optimal bandwidths we conduct a small simulation study. Let us define the following model specifications for random functions $P_t(x,z) = \mu(x,z) + P_t^c(x,z)$ with $\mu(x,z) = 0.5\sin(\pi\, x\, z/2)$, $P_t^c(x,z) = \xi_{t1}\,\psi_1(x,z) + \xi_{t2}\,\psi_2(x,z)$, $\psi_1(x,z) = \sin(\pi\, x\, z/2)$ and $\psi_2(x,z) = \sin(\pi\, x\, z)$, where the random coefficients $\xi_{t1} \in \mathbb{R}$ and $\xi_{t1} \in \mathbb{R}$ have mean zero and variances $\mathbb{V}(\xi_1) = \lambda_1$ and $\mathbb{V}(\xi_2) = \lambda_2$ further specified below. The prediction points X_{it} and covariate values Z_t are iid uniform pdfs on $[0,1]$. Finally, in order to generate the dependent variables $Y_{it} = P_t(X_{it}, Z_t) + \epsilon_{it}$ we sample the error term ϵ_{it} from a Gaussian with mean zero and variance $\sigma_\epsilon^2 = 2/3$. We investigate two different Data Generating Processes (DGPs), which differ with respect to the eigenvalues λ_1 and λ_2. That is, **DGP-1:** $\lambda_1 = \frac{1}{3}$ and $\lambda_2 = \frac{1}{4}$ and **DGP-2:** $\lambda_1 = 1$ and $\lambda_2 = \frac{1}{2}$. This way the signal-to-noise ratio in DGP-1 is lower than one and in DGP-2 greater than one. In order to design a simulation study of practical relevance, let us fix $T = 100$ and consider different sample sizes $n \in \{2, 25, 50\}$; for sample sizes greater than $n = 50$ practitioners usually start pre-smoothing their functional data.

The unknown quantities in our AMISE. I and AMISE. II expressions are approximated using preliminary global polynomial regressions. A detailed description on this simple rule-of-thumb procedure can be found in Section 2.3 of the supplemental paper. In order to investigate the different finite sample effects of the AMISE. I and AMISE. II optimal bandwidths, we approximate the respective AMISE values by means of a Monte Carlo

(MC) simulation based on $B = 500$ replications. Table 1 shows the values of the quotients $\text{MSE. I}_\mu / \text{MSE. II}_\mu$ and $\text{MSE. I}_\gamma / \text{MSE. II}_\gamma$, where

$$\text{MSE. I}_\mu = (BTn)^{-1} \sum_{b=1}^{B} \sum_{t=1}^{T} \sum_{i=1}^{n} (\mu(X_{it}, Z_t) - \hat{\mu}(X_{it}, Z_t; h_{\mu,X,\text{AMISE.I}}, h_{\mu,Z,\text{AMISE.I}}))^2$$

and likewise for MSE. II_μ, MSE. I_γ, and MSE. II_γ.

As predicted by our theory we do not observe any gain in using the AMISE. II optimal bandwidths in the case of the small values $n = 2$ and $N = 4$, which can be seen to represent the case of $\theta < 1/5$. In fact, it is surprising that the AMISE. II optimal bandwidths perform equally well as the AMISE. I optimal bandwidths, since ratios smaller than one would be consistent with our theory, too. Though, for the moderate values $n \in \{25, 50\}$ and $N = n^2 - n \in \{600, 2450\}$, which can be seen to represent the case of $\theta > 1/5$, we can see clear improvements when using the AMISE. II optimal bandwidths. In the case of estimating the mean function the MSE values based on the AMISE. I optimal bandwidths are 6% and 9% higher than those based on the AMISE. II optimal bandwidths. In the case of estimating the covariance function they are even 40% ($N = 600$) and 60% ($N = 2450$) higher than under the use of the AMISE. II optimal bandwidths.

	Mean Fct			Covariance Fct ($N = n^2 - n$)		
	$n = 2$	$n = 25$	$n = 50$	$N = 4$	$N = 600$	$N = 2450$
DGP-1	1.00	1.07	1.10	1.00	1.38	1.59
DGP-2	1.00	1.06	1.08	1.00	1.41	1.62

Table 1: Mean Squared Error ratios $\text{MSE. I}_\mu / \text{MSE. II}_\mu$ and $\text{MSE. I}_\gamma / \text{MSE. II}_\gamma$.

5 Application

In our real data study, we analyze electricity spot prices of the German power market traded at the European Energy Power Exchange (EPEX). The EPEX spot price is of fundamental importance as a benchmark and reference price for many other markets, such as over-the-counter and forward markets (Grimm et al. (2008), Ch. 1). The data for our analysis come from four different sources that are described in detail in the supplemental paper.

5.1 Preliminary descriptions and limitations

The German electricity market, like many others, provides purchase guarantees for renewable energy sources (RES). Therefore, the relevant variable for pricing is electricity demand minus electricity infeeds from RES. Correspondingly, in our application X_{it} refers to *residual* electricity demand defined as $X_{it} = \texttt{Elect.Demand}_{it} - \texttt{RES}_{it}$, where $\texttt{RES}_{it} = \texttt{Wind.Infeed}_{it} + \texttt{Solar.Infeed}_{it}$. The effect of further RES such as biomass is still negligible for the German electricity market.

At the EPEX, the 24 hourly electricity spot prices of a day t are determined in a separate auction and all these 24 auctions are settled simultaneously at 12 am on day $t-1$ (see, e.g., Grimm et al., 2008). This justifies our assumption of autocorrelations only across the indices t. From a mirco-economic perspective, prices settled at electricity exchanges can be described by the so-called merit order curve, i.e., the inverse supply curve. The plot in Figure 2 sketches the merit order curve of the German electricity market and is in line with Cludius et al. (2014); see also Ch. 4 of Burger et al. (2008). The interplay of the inverse demand curve (dashed line) with the merit order curve determines the electricity prices. In fact, our random functions P_t could be interpreted as empirical merit order curves, if one is willing to assume a completely price-inelastic demand. We do not want to be that restrictive and therefore interpret the random functions P_t more generally as price functions of demand and temperature.

There are clear limitations in our empirical results. The fundamental price drivers are (residual) electricity demand, temperate, and prices for uranium, lignite, coal, and gas. Our approach, however, only allows to control for the nonlinear effects of demand and temperature. We cannot disentangle shifts in the mean prices into their demand and supply side components. Furthermore, we cannot isolate the mean price effects that are due to Germany's nuclear phase out as this event was accompanied by an increase in Germany's resource prices which further increased the electricity prices. That is, we cannot identify the underlying causalities of changes in the conditional mean prices.

Nevertheless, only after controlling for the nonlinear effects of electricity demand we can reasonably compare electricity mean prices. Electricity prices are unit prices measured in "euros per 1megawatt·1hour" (EUR/MWh). But unit prices are only meaningful in

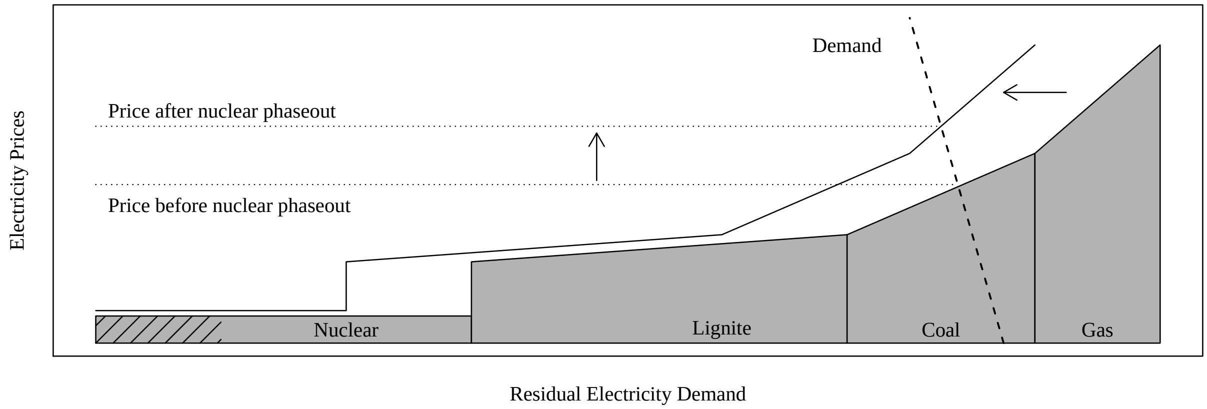
Demand
Price after nuclear phaseout
Price before nuclear phaseout
Electricity Prices
Nuclear
Lignite
Coal
Gas
Residual Electricity Demand

Figure 2: Sketch of the merit order curve and the theoretical price effect of a nuclear power phaseout. The dashed region signifies the proportion of phased out nuclear power plants.

relation with the total amount of demanded units. Clearly, a unit price of 40EUR/MWh for in total 6000MW is much cheaper than a unit price of 40EUR/MWh for in total 3000MW. Therefore, a reasonable comparison of electricity spot prices necessarily has to condition on electricity demand. Using our statistical model we can compute these conditional means which is not a trivial task due to the involved nonlinearities.

5.2 Empirical analysis

On March 15, 2011, just after the nuclear meltdown in Fukushima Daiichi, Japan, Germany decided to switch to a renewable energy economy and initiated this by an immediate and permanent shutdown of about 40% its nuclear power plants. This substantial loss of nuclear power with its low marginal production costs raised concerns about increases in electricity prices and subsequent problems for industry and households. However, empirical studies that build upon univariate time series analysis do not report any clear price effects (see, e.g., Nestle, 2012). A look at the univariate time series of Germany's hourly electricity spot prices, as shown in Figure 3, confirms this finding: Except for the very high prices at the end of the first year after Germany's (partial) nuclear phaseout, it is impossible to identify obvious mean shifts. Though, such a naive approach without conditioning on electricity demand does not allow for a reasonable comparison of the electricity mean prices; see our discussion in the latter section.

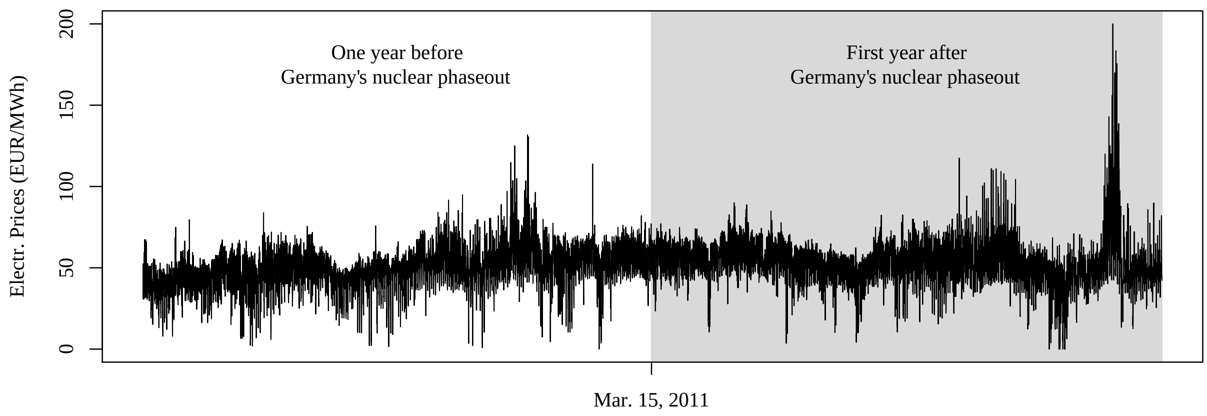

Figure 3: Time series of Germany's hourly electricity spot prices traded on the European Energy Exchange.

In order to estimate and to test the conditional mean prices one year before and one year after Germany's nuclear phaseout, we apply our local linear estimators in Eqs. (7) and (8) to the two data sets shown in Figure 1. The multiple bandwidth selection problem is solved by approximating the unknown quantities of our AMISE. I optimal bandwidth expressions in Eqs. (34), (35), (36), (37), and the AMISE. II optimal bandwidth expressions in Eqs. (15), (16), (13), (14) by fitting global polynomial pilot models of order four. Due to the relatively simple structured data, this rule-of-thumb (ROT) method works surprisingly well (c.f. the global polynomial fits in Figure 6 of the supplementary material). Note that these parametric pilot models need to include interaction terms, otherwise the required partial derivatives would degenerate. All necessary technical details of this straightforward ROT method can be found in Section 2.3 of the supplemental paper.

A practical problem of having two different bandwidth scenarios is that we have to decide which of the two AMISE cases is the better fitting scenario for our data at hand. Using our global polynomial approximations we can approach this question heuristically by comparing the approximated AMISE values (under consideration of both variance terms) when using either the AMISE. I optimal bandwidths or the AMISE. II optimal bandwidths. Figure 4 shows the ROT-approximated AMISE trajectories for the time period one year before Germany's nuclear phase-out. The corresponding plots for the second time period essentially bear the same information. According to these approximations, we choose the

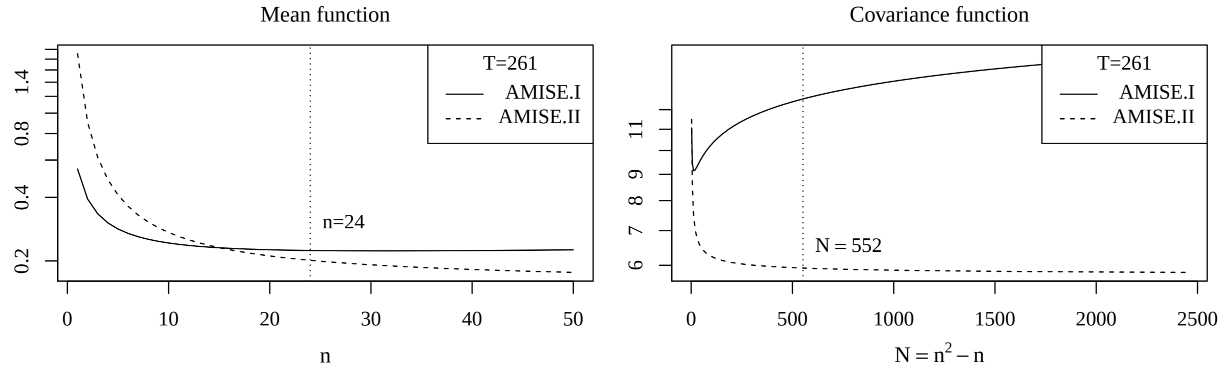

Figure 4: Approximated AMISE trajectories under optimal bandwidth choices y for the time period one year before Germany's nuclear phase-out. The corresponding plots for the second time period bear the same information.

AMISE. II optimal bandwidths for estimating the mean and the covariance function. Remember from our discussion in Section 3.4 that diverging AMISE trajectories as in the right plot of Figure 4 are not surprising under a false bandwidth choice.

Figure 5 shows the two estimated mean functions $\hat{\mu}(x, z; \hat{h}_{\mu,X,\text{AMISE.II}}, \hat{h}_{\mu,Z,\text{AMISE.II}})$ one year before and one year after Germany's (partial) nuclear phaseout, where $\hat{h}_{\mu,X,\text{AMISE.II}}$ and $\hat{h}_{\mu,Z,\text{AMISE.II}}$ denote the ROT-approximated AMISE. II optimal bandwidths. Here and below the dependency on the two time periods is suppressed. It is striking that the supports of the mean functions are relatively complicated objects. This explains our assumption that the price functions have z-dependent domains, namely $P_t(., z) \in L^2[a(z), b(z)]$. The boundary functions $a(z)$ and $b(z)$ of the kidney-shaped supports are estimated by the local linear estimation approach of Martins-Filho and Yao (2007).

In both periods the estimated mean prices range from about 30 to 100 EUR/MWh which is consistent with the ranges of the univariate time series in Figure 3. In the year after Germany's (partial) nuclear phaseout the *conditional* mean prices for fixed values of electricity demand and temperature are throughout higher than their corresponding counterparts one year before. These differences in the conditional means range from about +5 EUR/MWh for moderate factor values, which relate to the relatively flat middle part of the merit order curve, up to +42 EUR/MWh for very high values of (residual) demand at which the merit order curve is very steep.

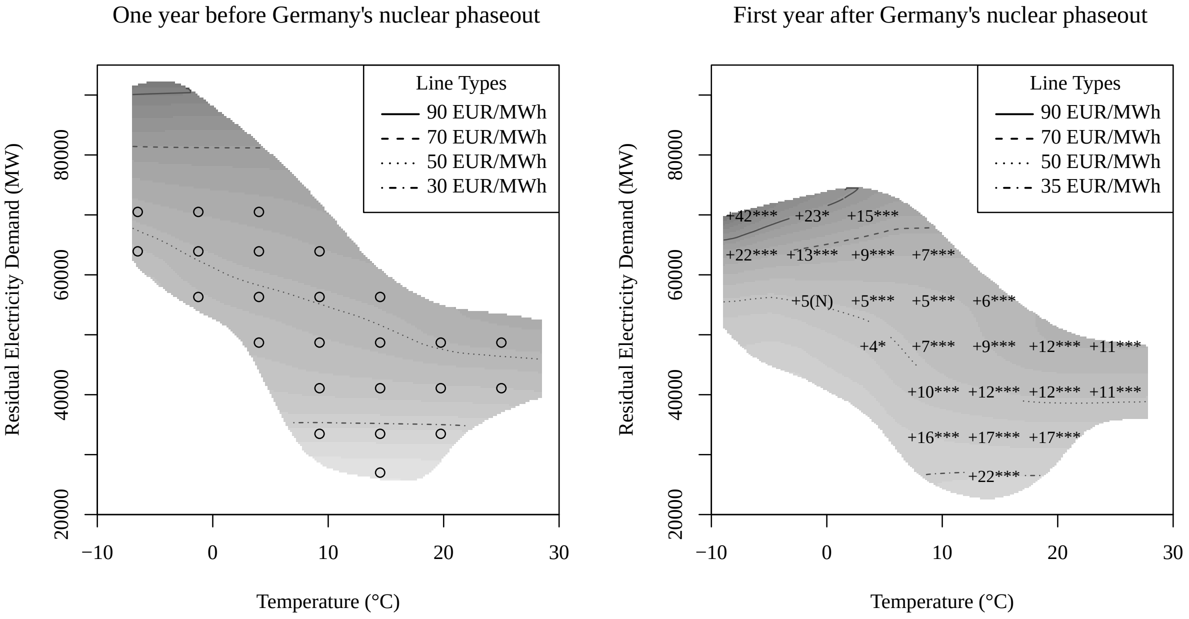

Figure 5: Comparison of the estimated mean functions one year before and after Germany's nuclear phaseout. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05.

In order to test the hypothesis that the estimated price shifts are equal to zero against the alternative that they are strictly greater than zero, we select a two-dimensional regular grid that covers the intersection of the supports of both mean functions. This grid of in total 24 grid points is shown by the circle points in the left plot of Figure 5. The test procedure is conducted using the following Bonferroni-type confidence bands (computed separately for each period):

$$\hat{\mu}(x_i, z_j; \hat{h}_{\mu,X,\text{AMISE.II}}, \hat{h}_{\mu,Z,\text{AMISE.II}}) \pm l_{B\alpha} \quad \text{with} \tag{21}$$

$$\begin{aligned} l_{B\alpha} = (Tn)^{-1} &\left[\hat{h}_{\mu,X}^{-1} \hat{h}_{\mu,Z}^{-1} \, R(K_\mu) \frac{\hat{\gamma}(x_i, x_i, z_j; \hat{h}_{\gamma,X,\text{AMISE.II}}, \hat{h}_{\gamma,Z,\text{AMISE.II}}) + \hat{\sigma}_\epsilon^2}{\hat{f}_{XZ}(x_i, z_j)} \right] z_{\alpha/24} + \\ + T^{-1} &\left[\hat{h}_{\gamma,Z}^{-1} \, R(\kappa) \frac{\hat{\gamma}(x_i, x_i, z_j; \hat{h}_{\gamma,X,\text{AMISE.II}}, \hat{h}_{\gamma,Z,\text{AMISE.II}})}{\hat{f}_Z(z_i)} \right] z_{\alpha/24}, \end{aligned} \tag{22}$$

where (x_i, z_j) label the 24 grid points and $z_{\alpha/24}$ is the $100(1-\alpha/24)$th percentile of the standard normal distribution with Bonferroni adjustment for the 24 tested grid points, and where for the Epanechnikov kernel we have that $R(K_\mu) = (3/5)^2$ and $R(\kappa) = (3/5)$.

These Bonferroni-type confidence bands are very useful in practice as they have an asymptotic coverage of at least $1 - \alpha$, even without explicitly accounting for the bias (Eubank and Speckman, 1993). In order to achieve this convenient result one cannot use classical asymptotic normal theory. Instead, Eubank and Speckman (1993) impose the restrictive assumption on the error term ε_{it} to posses 19 moments. However, this is not an issue here as the electricity spot prices Y_{it} are in anyway bounded from above by 3000 EUR/MWh and below by -500 EUR/MWh due the particular design of the auction market. As we are only concerned with positive shifts of the mean prices, we conduct a one-sided significance test. A shift in the mean at a grid point (x_i, z_j) is significant if the lower one-sided confidence "interval" of the second period is strictly greater than the upper one-sided confidence "interval" of the first period, i.e., if

$$\left(\hat{\mu}^{\text{2nd Period}}(x_i, z_j) - l_{B\alpha}^{\text{2nd Period}}\right) - \left(\hat{\mu}^{\text{1st Period}}(x_i, z_j) + l_{B\alpha}^{\text{1st Period}}\right) > 0. \tag{23}$$

As significance levels we choose $\alpha \in \{0.001, 0.01, 0.05\}$.

The chosen grid size of 24 grid points is found to be appropriate for balancing the variance inflation effect due to the Bonferroni adjustment and the need to capture enough information which allows to compare the mean function by means of a pointwise comparison. The above confidence bands can be seen as rather conservative. On the one hand the Bonferroni adjustment is known to be conservative. On the other hand our significance criterion in Eq. (23) is conservative, too. While a positive value necessarily implies a significant test result, a non-positive value does not necessarily imply a non-significant test result. Nevertheless, we interpret non-positive values in Eq. (23) as non-significant. The significant mean shifts at the chosen grid points are depicted in the right plot of Figure 5. The numerical values stand for the amount of the price shift measured in EUR/MWh. Insignificant mean shifts are denoted by the letter "N" and for the significant results we use the typical significance stars ranging from one to three stars. Except for one grid point all pointwise comparisons are significant - most of them at the significance level of $\alpha = 0.001$ (three stars).

In addition, Figure 5 gives empirical evidence for a further important issue: Germany managed to reduce its residual demand for electricity in the year after the nuclear phaseout, which is reflected by the lower support of the mean function for low temperatures values in the second period. This reduction was mainly due to a politically promoted higher amount of electricity infeeds from RES and obviously helped to avoid the occurrence of some very high electricity prices.

6 Conclusion

The theoretical side of our paper is concerned with the nonparametric estimation of the conditional mean and covariance function of a stationary time series of weakly dependent random functions with covariate-adjustments. Using a double asymptotic we investigate all cases from sparsely sampled to densely sampled functional data which takes into account the vague cases typically found in applications. A specific emphasis lies on the derivation of AMISE optimal bandwidths for the multivariate local linear estimators under this double asymptotic which leads to some non-classical bandwidth expressions. It turns out that these new bandwidth expressions allow the local linear estimators to *inherently* under-smooth the functional data for the sake of a better estimate of the mean or covariance function. The practical side of our paper applies our bias, variance, and bandwidth results in order to test for differences in conditional mean prices before and after Germany's abrupt nuclear phaseout in mid-March 2011.

SUPPLEMENTARY MATERIAL

Supplemental Paper: Proofs and further discussions.

References

Benko, M., W. Härdle, and A. Kneip (2009). Common functional principal components. *The Annals of Statistics 37*(1), 1–34.

Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*, Volume 149. Springer.

Burger, M., B. Graeber, and G. Schindlmayr (2008). *Managing Energy Risk: An Integrated View on Power and Other Energy Markets* (1. ed.). Wiley.

Cardot, H. (2007). Conditional functional principal components analysis. *Scandinavian Journal of Statistics 34*(2), 317–335.

Cludius, J., H. Hermann, F. C. Matthes, and V. Graichen (2014). The merit order effect of wind and photovoltaic electricity generation in germany 2008–2016: Estimation and distributional implications. *Energy Economics 44*, 302–313.

Eubank, R. and P. Speckman (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association 88*(424), 1287–1301.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and its Applications* (1. ed.), Volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.

Fan, J. and Q. Yao (2003). *Nonlinear Time Series* (1. ed.). Springer Series in Statistics. Springer.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice* (1. ed.). Springer Series in Statistics. Springer.

Grimm, V., A. Ockenfels, and G. Zoettl (2008). Strommarktdesign: Zur ausgestaltung der auktionsregeln an der eex. *Zeitschrift für Energiewirtschaft 32*(3), 147–161.

Hall, P. and M. Hosseini-Nasab (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 109–126.

Hall, P., H. G. Müller, and J. L. Wang (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics 34*(3), 1493–1517.

Hörmann, S. and P. Kokoszka (2010). Weakly dependent functional data. *The Annals of Statistics 38*(3), 1845–1884.

Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*, Volume 200. Springer.

Jiang, C.-R. and J.-L. Wang (2010). Covariate adjusted functional principal components analysis for longitudinal data. *The Annals of Statistics 38*(2), 1194–1226.

Kneip, A. and K. J. Utikal (2001). Inference for density families using functional principal component analysis. *Journal of the American Statistical Association 96*(454), 519–532.

Li, Y. and T. Hsing (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics 38*(6), 3321–3351.

Liebl, D. (2013). Modeling and forecasting electricity spot prices: a functional data perspective. *The Annals of Applied Statistics 7*(3), 1562–1592.

Martins-Filho, C. and F. Yao (2007). Nonparametric frontier estimation via local linear regression. *Journal of Econometrics 141*(1), 283–319.

Nestle, U. (2012). Does the use of nuclear power lead to lower electricity prices? an analysis of the debate in germany with an international perspective. *Energy Policy 41*(0), 152–160.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (2. ed.). Springer Series in Statistics. Springer.

Ruppert, D. and M. Wand (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics 22*(3), 1346–1370.

Wang, L., L. D. Brown, T. T. Cai, and L. Michael (2008). Effect of mean on variance function estimation in nonparametric regression. *The Annals of Statistics 36*(2), 646–664.

Yao, F., H. G. Müller, and J. L. Wang (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association 100*(470), 577–590.

Supplemental Paper for:

Anti-Proportional Bandwidth Selection for Smoothing (Non-)Sparse Functional Data with Covariate Adjustments

Dominik Liebl

Institute for Financial Economics and Statistics

University of Bonn

Adenauerallee 24-42

53113 Bonn, Germany

Outline:

Section 1 contains the proofs of the Theorems 3.1 and 3.2. Further discussions and supplementary results on the bandwidth results and the rule-of-thumb procedure used in the application are found in Section 2. The data sources are described in detail in Section 3.

1 Proofs

1.1 Proof of Theorem 3.1

Proof of Theorem 3.1, part *(i)*: Let x and z be interior points of $\text{supp}(f_{XZ})$ and define $\mathbf{H}_\mu = \text{diag}(h^2_{\mu,X}, h^2_{\mu,Z})$, $\mathbf{X} = (X_{11}, \dots, X_{nT})^\top$, and $\mathbf{Z} = (Z_1, \dots, Z_T)^\top$. Taylor-expansion of μ around (x, z), the conditional bias of the estimator $\hat{\mu}(x, z; \mathbf{H})$ can be written as

$$\begin{aligned}\mathbb{E}(\hat{\mu}(x, z; \mathbf{H}_\mu) - \mu(x, z)|\mathbf{X}, \mathbf{Z}) = \frac{1}{2} u_1^\top \left((Tn)^{-1}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]\right)^{-1} \times \\ \times (Tn)^{-1}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz} \left(\boldsymbol{\mathcal{Q}}_\mu(x, z) + \mathbf{R}_\mu(x, z)\right),\end{aligned} \tag{24}$$

where $\boldsymbol{\mathcal{Q}}_\mu(x, z)$ is a $Tn \times 1$ vector with typical elements

$$(X_{it} - x, Z_t - z)\boldsymbol{\mathcal{H}}_\mu(x, z)(X_{it} - x, Z_t - z)^\top \in \mathbb{R}$$

with $\boldsymbol{\mathcal{H}}_\mu(x,z)$ being the Hessian matrix of the regression function $\mu(x,z)$. The $Tn \times 1$ vector $\mathbf{R}_\mu(x,z)$ holds the remainder terms

$$o\left((X_{it}-x, Z_t-z)(X_{it}-x, Z_t-z)^\top\right) \in \mathbb{R}.$$

Next we derive asymptotic approximations for the 3×3 matrix $\left((Tn)^{-1}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]\right)^{-1}$ and the $3{\times}1$ matrix $(Tn)^{-1}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz}\boldsymbol{\mathcal{Q}}_\mu(x,z)$ of the right hand side of Eq. (24). Using standard procedures from kernel density estimation it is easy to derive that

$$(Tn)^{-1}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z] =$$

$$\begin{pmatrix} f_{XZ}(x,z) + o_p(1) & \nu_2(K_\mu)\mathbf{D}_{f_{XZ}}(x,z)^\top \mathbf{H}_\mu + o_p(\mathbf{1}^\top \mathbf{H}_\mu) \\ \nu_2(K_\mu)\mathbf{H}_\mu \mathbf{D}_{f_{XZ}}(x,z) + o_p(\mathbf{H}_\mu \mathbf{1}) & \nu_2(K_\mu)\mathbf{H}_\mu f_{XZ}(x,z) + o_p(\mathbf{H}_\mu) \end{pmatrix},$$

where $\mathbf{1} = (1,1)^\top$ and $\mathbf{D}_{f_{XZ}}(x,z)$ is the vector of first order partial derivatives (i.e., the gradient) of the pdf f_{XZ} at (x,z). Inversion of the above block matrix yields

$$\left((Tn)^{-1}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]\right)^{-1} = \tag{25}$$

$$\begin{pmatrix} (f_{XZ}(x,z))^{-1} + o_p(1) & -\mathbf{D}_{f_{XZ}}(x,z)^\top (f_{XZ}(x,z))^{-2} + o_p(\mathbf{1}^\top) \\ -\mathbf{D}_{f_{XZ}}(x,z)\,(f_{XZ}(x,z))^{-2} + o_p(\mathbf{1}) & (\nu_2(K_\mu)\mathbf{H}_\mu f_{XZ}(x,z))^{-1} + o_p(\mathbf{H}_\mu) \end{pmatrix}.$$

The 3×1 matrix $(Tn)^{-1}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz}\boldsymbol{\mathcal{Q}}_\mu(x,z)$ can be partitioned as following:

$$(Tn)^{-1}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]^\top \mathbf{W}_{xz}\boldsymbol{\mathcal{Q}}_\mu(x,z) \quad = \quad \begin{pmatrix} \texttt{upper element} \\ \texttt{lower bloc} \end{pmatrix},$$

where the 1×1 dimensional `upper element` can be approximated by a scalar variable equal to

$$\begin{aligned} &(Tn)^{-1}\sum_{it} K_{\mu,h}(X_{it}-x, Z_t-z)(X_{it}-x, Z_t-z)\boldsymbol{\mathcal{H}}_\mu(x,z)(X_{it}-x, Z_t-z)^\top \\ &= (\nu_2(\kappa))^2\, tr\left\{\mathbf{H}_\mu \boldsymbol{\mathcal{H}}_\mu(x,z)\right\} f_{XZ}(x,z) + o_p(tr(\mathbf{H}_\mu)) \end{aligned} \tag{26}$$

and the 2×1 dimensional `lower bloc` is equal to

$$\begin{aligned} &(Tn)^{-1}\sum_{it}\left\{K_{\mu,h}(X_{it}-x, Z_t-z)(X_{it}-x, Z_t-z)\boldsymbol{\mathcal{H}}_\mu(x,z)(X_{it}-x, Z_t-z)^\top\right\} \times \\ &\times (X_{it}-x, Z_t-z)^\top = O_p(\mathbf{H}_\mu^{3/2}\mathbf{1}). \end{aligned} \tag{27}$$

Plugging in the approximations of Eqs. (25)-(27) into the first summand of the conditional bias expression in Eq. (24) leads to the following expression

$$\begin{aligned}&\frac{1}{2}u_1^\top((Tn)^{-1}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,xz}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z])^{-1}(Tn)^{-1}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,xz}\boldsymbol{\mathcal{Q}}_\mu(x,z)=\\&=\frac{1}{2}\left(\nu_2(\kappa)\right)^2 tr\left\{\mathbf{H}_\mu\boldsymbol{\mathcal{H}}_\mu(x,z)\right\}+o_p(tr(\mathbf{H}_\mu)).\end{aligned}$$

Furthermore, it is easily seen that the second summand of the conditional bias expression in Eq. (24), which holds the remainder term, is given by

$$\frac{1}{2}u_1^\top((Tn)^{-1}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,xz}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z])^{-1}(Tn)^{-1}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,xz}\mathbf{R}_\mu(x,z)=o_p(tr(\mathbf{H}_\mu)).$$

Summation of the two latter expressions yields the asymptotic approximation of the conditional bias

$$\mathbb{E}(\hat{\mu}(x,z;\mathbf{H}_\mu)-\mu(x,z)|\mathbf{X},\mathbf{Z})=\frac{1}{2}\left(\nu_2(\kappa)\right)^2 tr\left\{\mathbf{H}_\mu\boldsymbol{\mathcal{H}}_\mu(x,z)\right\}+o_p(tr(\mathbf{H}_\mu)).$$

This is our bias statement of Theorem 3.1 part *(i)*.

Proof of Theorem 3.1, part *(ii)*: In the following we derive the conditional variance of the local linear estimator $\mathbb{V}(\hat{\mu}(x,z;\mathbf{H}_\mu)|\mathbf{X},\mathbf{Z})=$

$$\begin{aligned}=&u_1^\top([\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,xz}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z])^{-1}\,[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,xz}\,\mathrm{Cov}(\mathbf{Y}|\mathbf{X},\mathbf{Z})\,\mathbf{W}_{\mu,xz}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]\\&([\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,xz}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z])^{-1}u_1\\=&u_1^\top((Tn)^{-1}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,xz}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z])^{-1}\,((Tn)^{-2}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,xz}\,\mathrm{Cov}(\mathbf{Y}|\mathbf{X},\mathbf{Z})\,\mathbf{W}_{\mu,xz}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z])\\&((Tn)^{-1}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,xz}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z])^{-1}u_1,\end{aligned}\tag{28}$$

where $\mathrm{Cov}(\mathbf{Y}|\mathbf{X},\mathbf{Z})$ is the $Tn\times Tn$ matrix with typical elements

$$\mathrm{Cov}(Y_{it},Y_{js}|X_{it},X_{js},Z_t,Z_s)=\gamma_{|t-s|}((X_{it},Z_t),(X_{js},Z_s))+\sigma_\epsilon^2\mathbb{1}(i=j\text{ and }t=s)$$

with $\mathbb{1}(.)$ being the indicator function.

We begin with analyzing the 3×3 matrix

$$(Tn)^{-2}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,xz}\,\mathrm{Cov}(\mathbf{Y}|\mathbf{X},\mathbf{Z})\,\mathbf{W}_{\mu,xz}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]$$

using the following three Lemmas 1.1-1.3.

Lemma 1.1 *The upper-left scalar (block) of the matrix* $(Tn)^{-2}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz} \mathrm{Cov}(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) \mathbf{W}_{\mu,xz} [\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]$ *is given by*

$$\begin{aligned}
& (Tn)^{-2} \mathbf{1}^\top \mathbf{W}_{\mu,xz} \mathrm{Cov}(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) \mathbf{W}_{\mu,xz} \mathbf{1} \\
= \;& (Tn)^{-1} f_{XZ}(x,z) |\mathbf{H}_\mu|^{-1/2} R(K_\mu) \left(\gamma(x,x,z) + \sigma_\epsilon^2\right) (1 + O_p(tr(\mathbf{H}_\mu^{1/2}))) \\
+ \;& T^{-1} (f_{XZ}(x,z))^2 \left[\left(\frac{n-1}{n} \right) h_{\mu,Z}^{-1} R(\kappa) \frac{\gamma(x,x,z)}{f_Z(z)} + c_r \right] (1 + O_p(tr(H^{1/2}))) \\
= \;& O_p((Tn)^{-1} |\mathbf{H}_\mu|^{-1/2}) + O_p(T^{-1} h_{\mu,Z}^{-1}).
\end{aligned}$$

Lemma 1.2 *The* 1×2 *dimensional upper-right block of the matrix* $(Tn)^{-2}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz} \mathrm{Cov}(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) \mathbf{W}_{\mu,xz} [\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]$ *is given by*

$$\begin{aligned}
& (Tn)^{-2} \mathbf{1}^\top \mathbf{W}_{\mu,xz} \mathrm{Cov}(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) \mathbf{W}_{\mu,xz} \begin{pmatrix} (X_{11} - x, Z_1 - z) \\ \vdots \\ (X_{nT} - x, Z_T - z) \end{pmatrix} \\
= \;& (Tn)^{-1} f_{XZ}(x,z) |\mathbf{H}_\mu|^{-1/2} (\mathbf{1}^\top \mathbf{H}_\mu^{1/2}) R(K_\mu) \left(\gamma(x,x,z) + \sigma_\epsilon^2\right) (1 + O_p(tr(\mathbf{H}_\mu^{1/2}))) \\
+ \;& T^{-1} (f_{XZ}(x,z))^2 (\mathbf{1}^\top \mathbf{H}_\mu^{1/2}) \left[\left(\frac{n-1}{n} \right) h_{\mu,Z}^{-1} R(\kappa) \frac{\gamma(x,x,z)}{f_Z(z)} + c_r \right] (1 + O_p(tr(\mathbf{H}_\mu^{1/2}))) \\
= \;& O_p((Tn)^{-1} |\mathbf{H}_\mu|^{-1/2} (\mathbf{1}^\top \mathbf{H}_\mu^{1/2})) + O_p(T^{-1} (\mathbf{1}^\top \mathbf{H}_\mu^{1/2}) h_{\mu,Z}^{-1}).
\end{aligned}$$

The 2×1 *dimensional lower-left block of the matrix* $(Tn)^{-2}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz} \mathrm{Cov}(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) \mathbf{W}_{\mu,xz} [\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]$ *is simply the transposed version of this result.*

Lemma 1.3 *The* 2×2 *lower-right block of the matrix* $(Tn)^{-2}[\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz} \mathrm{Cov}(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) \mathbf{W}_{\mu,xz} [\mathbf{1}, \mathbf{X}_x, \mathbf{Z}_z]$ *is given by*

$$\begin{aligned}
& (Tn)^{-2} \Big(((X_{11} - x), (Z_1 - z))^\top, \ldots, ((X_{nT} - x), (Z_T - z)^\top) \Big) \times \\
& \times \mathbf{W}_{\mu,xz} \mathrm{Cov}(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) \mathbf{W}_{\mu,xz} \begin{pmatrix} (X_{11} - x, Z_1 - z) \\ \vdots \\ (X_{nT} - x, Z_T - z) \end{pmatrix} \\
= \;& (Tn)^{-1} f_{XZ}(x,z) |\mathbf{H}_\mu|^{-1/2} \mathbf{H}_\mu R(K_\mu) \left(\gamma(x,x,z) + \sigma_\epsilon^2\right) (1 + O_p(tr(\mathbf{H}_\mu^{1/2}))) \\
+ \;& T^{-1} (f_{XZ}(x,z))^2 \mathbf{H}_\mu \left[\left(\frac{n-1}{n} \right) h_{\mu,Z}^{-1} R(\kappa) \frac{\gamma(x,x,z)}{f_Z(z)} + c_r \right] (1 + O_p(tr(\mathbf{H}_\mu^{1/2}))) \\
= \;& O_p((Tn)^{-1} |\mathbf{H}_\mu|^{-1/2} \mathbf{H}_\mu) + O_p(T^{-1} \mathbf{H}_\mu h_{\mu,Z}^{-1}).
\end{aligned}$$

Using the approximations for the bloc-elements of the matrix $(Tn)^{-2}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz}\mathrm{Cov}(\mathbf{Y}|\mathbf{X},\mathbf{Z})\mathbf{W}_{\mu,xz}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]$, given by the Lemmas 1.1-1.3, and the approximation for the matrix $\left((Tn)^{-1}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]^\top \mathbf{W}_{\mu,xz}[\mathbf{1},\mathbf{X}_x,\mathbf{Z}_z]\right)^{-1}$, given in (25), we can approximate the conditional variance of the bivariate local linear estimator, given in (28). Some tedious yet straightforward matrix algebra leads to $\mathbb{V}(\hat{\mu}(x,z;\mathbf{H}_\mu)|\mathbf{X},\mathbf{Z}) =$

$$(Tn)^{-1}|\mathbf{H}_\mu|^{-1/2}\left\{\frac{R(K_\mu)\left(\gamma(x,x,z)+\sigma_\epsilon^2\right)}{f_{XZ}(x,z)}\right\}(1+o_p(1))$$
$$+T^{-1}\left[\left(\frac{n-1}{n}\right)h_{\mu,Z}^{-1}R(\kappa)\frac{\gamma(x,x,z)}{f_Z(z)}+c_r\right](1+o_p(1)),$$

which is asymptotically equivalent to our variance statement of Theorem 3.1 part *(ii)*.

Next we proof Lemma 1.1; the proofs of Lemmas 1.2 and 1.3 can be done correspondingly. To show Lemma 1.1 it will be convenient to split the sum such that $(Tn)^{-2}\mathbf{1}^\top\mathbf{W}_{\mu,xz}\mathrm{Cov}(\mathbf{Y}|\mathbf{X},\mathbf{Z})\mathbf{W}_{\mu,xz}\mathbf{1} = s_1+s_2+s_3$. Using standard procedures from kernel density estimation leads to

$$\begin{aligned}
s_1 &= (Tn)^{-2}\sum_{it}(K_{\mu,h}(X_{it}-x,Z_t-z))^2\,\mathbb{V}(Y_{it}|\mathbf{X},\mathbf{Z}) && (29)\\
&= (Tn)^{-1}|\mathbf{H}_\mu|^{-1/2}f_{XZ}(x,z)R(K_\mu)\left(\gamma(x,x,z)+\sigma_\epsilon^2\right)+O((Tn)^{-1}|\mathbf{H}_\mu|^{-1/2}\,tr(\mathbf{H}_\mu^{1/2})),\\
s_2 &= (Tn)^{-2}\sum_{ij}\sum_{\substack{ts\\t\neq s}}K_{\mu,h}(X_{it}-x,Z_t-z)\,\mathrm{Cov}(Y_{it},Y_{js}|\mathbf{X},\mathbf{Z})\,K_{\mu,h}(X_{js}-x,Z_s-z) && (30)\\
&= T^{-1}(f_{XZ}(x,z))^2c_r+O_p(T^{-1}tr(\mathbf{H}_\mu^{1/2}))\\
s_3 &= (Tn)^{-2}\sum_{\substack{ij\\i\neq j}}\sum_{t}h_{\mu,X}^{-1}\kappa(h_{\mu,X}^{-1}(X_{it}-x))(h_{\mu,Z}^{-1}\kappa(h_{\mu,Z}^{-1}(Z_t-z)))^2\,\mathrm{Cov}(Y_{it},Y_{jt}|\mathbf{X},\mathbf{Z})\times && (31)\\
&\times h_{\mu,X}^{-1}\kappa(h_{\mu,X}^{-1}(X_{jt}-x))\\
&= T^{-1}(f_{XZ}(x,z))^2\left[\left(\frac{n-1}{n}\right)h_{\mu,Z}^{-1}R(\kappa)\frac{\gamma(x,x,z)}{f_Z(z)}\right]+O_p(T^{-1}tr(\mathbf{H}_\mu^{1/2})),
\end{aligned}$$

where $|c_r|\leq\frac{2c_\gamma r}{1-r}<\infty$. Summing up (29)-(30) leads to the result in Lemma 1.1. Lemmas 1.2 and 1.3 differ from Lemma 1.1 only with respect to the additional factors $\mathbf{1}^\top\mathbf{H}_\mu^{1/2}$ and $\mathbf{H}_\mu$. These come in due to the usual substitution step for the additional data parts $(X_{it}-x,Z_t-z)$.

1.2 Proof of Theorem 3.2

The proof of Theorem 3.2 follows the same arguments as in the proof of Theorem 3.1. The only additional issue we need to consider, is that we do not directly observe the dependent variables, namely, the true theoretical raw covariances C_{ijt} as defined in Eq. (4), but only their empirical versions $\hat{C}_{ijt}$ as defined in Eq. (9). In the following we show that this additional estimation error is asymptotically negligible in comparison to the approximation error of $\hat{\gamma}$ derived under the usage of the theoretical raw covariates C_{ijt} as given in Eq. (20) of Theorem 3.3.

Observe that we can expand $\hat{C}_{ijt}$ as following:

$$\begin{aligned}\hat{C}_{ijt} = C_{ijt} &+ (Y_{it} - \mu(X_{it}, Z_t))(\mu(X_{jt}, Z_t) - \hat{\mu}(X_{jt}, Z_t))\\ &+ (Y_{jt} - \mu(X_{jt}, Z_t))(\mu(X_{it}, Z_t) - \hat{\mu}(X_{it}, Z_t))\\ &+ (\mu(X_{it}, Z_t) - \hat{\mu}(X_{it}, Z_t))(\mu(X_{jt}, Z_t) - \hat{\mu}(X_{jt}, Z_t)).\end{aligned}$$

Further, it follows from Eq. (19) in Theorem 3.3 that

$$\left(\hat{\mu}(X_{it}, Z_t; h_{\mu,X,\mathrm{opt}}, h_{\mu,Z,\mathrm{opt}}) - \mu(X_{it}, Z_t)\right)\big|\mathbf{X}, \mathbf{Z} = \begin{cases} O_p\left((Tn)^{-2/3}\right) & \text{if } 0 \leq \theta \leq 1/5 \\ O_p\left(T^{-4/5}\right) & \text{if } \theta > 1/5, \end{cases}$$

for all i and t, where $h_{\mu,X,\mathrm{opt}}$ and $h_{\mu,Z,\mathrm{opt}}$ denote the θ-specific AMISE optimal bandwidth choices as defined in Eqs. (13), (15), (34), and (35). Therefore, under our setup we have that

$$\left(\hat{C}_{ijt} - C_{ijt}\right)\big|\mathbf{X}, \mathbf{Z} = \begin{cases} O_p\left((Tn)^{-2/3}\right) & \text{if } 0 \leq \theta \leq 1/5 \\ O_p\left(T^{-4/5}\right) & \text{if } \theta > 1/5, \end{cases}$$

which is of an order of magnitude smaller than the approximation error of $\hat{\gamma}$ derived under the usage of the theoretical raw covariates C_{ijt} as given in Eq. (20) of Theorem 3.3.

2 Further discussions

2.1 AMISE. I optimal bandwidth selection

The following expressions are derived under the hypothesis that the first variance terms in parts *(ii)* of Theorems 3.1 and 3.2 are dominating. The AMISE. I expression for the local linear estimator $\hat{\mu}$ is given by

$$\begin{aligned}\text{AMISE.}\,\mathrm{I}_{\hat{\mu}}\,(h_{\mu,X},h_{\mu,Z}) = (Tn)^{-1}\,h_{\mu,X}^{-1}\,h_{\mu,Z}^{-1}\,R(K_\mu)\,Q_{\mu,1}+ \\ +\frac{1}{4}\,(\nu_2(K_\mu))^2\,\left[h_{\mu,X}^4\,\mathcal{I}_{\mu,XX}+2\,h_{\mu,X}^2\,h_{\mu,Z}^2\,\mathcal{I}_{\mu,XZ}+h_{\mu,Z}^4\,\mathcal{I}_{\mu,ZZ}\right],\end{aligned} \qquad (32)$$

$$\begin{aligned} \text{where}\qquad Q_{\mu,1} &= \textstyle\int_{S_{XZ}}(\gamma(x,x,z)+\sigma_\epsilon^2)\;d(x,z),\\ \mathcal{I}_{\mu,XX} &= \textstyle\int_{S_{XZ}}\left(\mu^{(2,0)}(x,z)\right)^2\;w_\mu\,d(x,z),\\ \mathcal{I}_{\mu,ZZ} &= \textstyle\int_{S_{XZ}}\left(\mu^{(0,2)}(x,z)\right)^2\;w_\mu\,d(x,z),\quad\text{and}\\ \mathcal{I}_{\mu,XZ} &= \textstyle\int_{S_{XZ}}\mu^{(2,0)}(x,z)\mu^{(0,2)}(x,z)\,w_\mu\,d(x,z).\end{aligned}$$

This is a known expression for the AMISE of a two-dimensional local linear estimator with a diagonal bandwidth matrix (see, e.g., Herrmann et al. (1995)[1]) and can be derived using the formulas in Wand and Jones (1994)[2].

The AMISE. I expression for the local linear estimator $\hat{\gamma}$ is less common as it is based on two bandwidths, which leads to

$$\begin{aligned}\text{AMISE.}\,\mathrm{I}_{\hat{\gamma}}\,(h_{\gamma,X},h_{\gamma,Z}) = (TN)^{-1}\,h_{\gamma,X}^{-2}\,h_{\gamma,Z}^{-1}\,R(K_\gamma)\,Q_{\gamma,1}+ \\ +\frac{1}{4}\,(\nu_2(K_\gamma))^2\,\left[2\,h_{\gamma,X}^4\,(\mathcal{I}_{\gamma,X_1X_1}+\mathcal{I}_{\gamma,X_1X_2})+4\,h_{\gamma,X}^2\,h_{\gamma,Z}^2\,\mathcal{I}_{\gamma,X_1Z}+h_{\gamma,Z}^4\,\mathcal{I}_{\gamma,ZZ}\right],\end{aligned} \qquad (33)$$

where

$$\begin{aligned} Q_{\gamma,1} &= \textstyle\int_{S_{XXZ}}(\tilde{\gamma}((x_1,x_2),(x_1,x_2),z)+\sigma_\varepsilon^2(x_1,x_2,z))\;d(x_1,x_2,z)\\ \mathcal{I}_{\gamma,X_1X_1} &= \textstyle\int_{S_{XXZ}}\left(\gamma^{(2,0,0)}(x_1,x_2,z)\right)^2\;w_\gamma\,d(x_1,x_2,z),\\ \mathcal{I}_{\gamma,X_1X_2} &= \textstyle\int_{S_{XXZ}}\left(\gamma^{(2,0,0)}(x_1,x_2,z)\gamma^{(0,2,0)}(x_1,x_2,z)\right)\;w_\gamma\,d(x_1,x_2,z),\\ \mathcal{I}_{\gamma,X_1Z} &= \textstyle\int_{S_{XXZ}}\gamma^{(2,0,0)}(x_1,x_2,z)\gamma^{(0,0,2)}(x_1,x_2,z)\,w_\gamma\,d(x_1,x_2,z),\quad\text{and}\\ \mathcal{I}_{\gamma,ZZ} &= \textstyle\int_{S_{XXZ}}\left(\gamma^{(0,0,2)}(x_1,x_2,z)\right)^2\;w_\gamma\,d(x_1,x_2,z).\end{aligned}$$

Equation (33) can be derived again using the formulas in Wand and Jones (1994)[3] with the further simplifications that $\mathcal{I}_{\gamma,X_1X_1}=\mathcal{I}_{\gamma,X_2X_2}$, $\mathcal{I}_{\gamma,X_1X_2}=\mathcal{I}_{\gamma,X_2X_1}$, and $\mathcal{I}_{\gamma,X_1Z}=\mathcal{I}_{\gamma,X_2Z}$

[1] Herrmann, E., J. Engel, M. Wand, and T. Gasser (1995). A bandwidth selector for bi- variate kernel regression. Journal of the Royal Statistical Society. Series B (Methodological) 57 (1), 171180.

[2] Wand, M. and M. Jones (1994). Multivariate plug-in bandwidth selection. Computational Statistics 9 (2), 97116.

[3] Wand, M. and M. Jones (1994). Multivariate plug-in bandwidth selection. Computational Statistics 9 (2), 97116.

due to the symmetry of the covariance function, where the expressions $\mathcal{I}_{\gamma,X_2X_2}$, $\mathcal{I}_{\gamma,X_2X_1}$, and $\mathcal{I}_{\gamma,X_2Z}$ are defined in correspondence with those above.

The AMISE. I expressions (32) and (33) allow us to analytically derive the AMISE. I optimal bandwidth pairs. For the mean function we have

$$h_{\mu,X,\text{AMISE.I}} = \left(\frac{R(K_\mu)\, Q_{\mu,1}\, \mathcal{I}_{\mu,ZZ}^{3/4}}{Tn\ (\nu_2(K_\mu))^2 \left[\mathcal{I}_{\mu,XX}^{1/2}\, \mathcal{I}_{\mu,ZZ}^{1/2} + \mathcal{I}_{\mu,XZ} \right] \mathcal{I}_{\mu,XX}^{3/4}} \right)^{1/6} \tag{34}$$

$$h_{\mu,Z,\text{AMISE.I}} = \left(\frac{\mathcal{I}_{\mu,XX}}{\mathcal{I}_{\mu,ZZ}} \right)^{1/4} h_{\mu,X,\text{AMISE.I}} \tag{35}$$

which corresponds to the result in Herrmann et al. (1995)[4]. The AMISE. I optimal bandwidths for the covariance function are given by

$$h_{\gamma,X,\text{AMISE.I}} = \left(\frac{R(K_\gamma)\ Q_{\gamma,1}\ 4\ \sqrt{2}\, \mathcal{I}_{\gamma,ZZ}^{3/2}}{TN\ (\nu_2(K_\gamma))^2 \left(2\ (\nu_2(K_\gamma))^2\ \mathcal{I}_{\gamma,X_1Z} + C_{\mathcal{I}} \right)\ (C_{\mathcal{I}} - \mathcal{I}_{\gamma,X_1Z})^{3/2}} \right)^{1/7} \tag{36}$$

$$h_{\gamma,Z,\text{AMISE.I}} = \left(\frac{C_{\mathcal{I}} - \mathcal{I}_{\gamma,X_1Z}}{2\, \mathcal{I}_{\gamma,ZZ}} \right)^{1/2} h_{\gamma,X,\text{AMISE.I}}, \tag{37}$$

where $C_{\mathcal{I}} = \left(\mathcal{I}_{\gamma,X_1Z}^2 + 4\, (\mathcal{I}_{\gamma,X_1X_1} + \mathcal{I}_{\gamma,X_1X_2})\, \mathcal{I}_{\gamma,ZZ} \right)^{1/2}$. Obviously, the expressions in (36) and (37) are much less readable than those in (34) and (35). This is the burden of a two times higher dimension when estimating γ instead of μ.

2.2 AMISE. II optimal bandwidth selection

The AMISE. II expression of the three-dimensional local linear estimator $\hat{\gamma}$ is given by

$$\begin{aligned} \text{AMISE. II}_{\hat{\gamma}}\,(h_{\gamma,X}, h_{\gamma,Z}) &= \overbrace{(TN)^{-1}\, h_{\gamma,X}^{-2}\, h_{\gamma,Z}^{-1}\, R(K_\gamma)\, Q_{\gamma,1}}^{\text{2nd Order}} + \overbrace{T^{-1}\, h_{\gamma,Z}^{-1}\, R(\kappa)\, Q_{\gamma,2}}^{\text{1st Order}} + \qquad (38) \\ &+ \frac{1}{4}\, (\nu_2(K_\gamma))^2 \left[\underbrace{2\, h_{\gamma,X}^4\, (\mathcal{I}_{\gamma,X_1X_1} + \mathcal{I}_{X_1X_2})}_{\text{3rd Order}} + \underbrace{4\, h_{\gamma,X}^2\, h_{\gamma,Z}^2\, \mathcal{I}_{\gamma,X_1Z}}_{\text{2nd Order}} + \underbrace{h_{\gamma,Z}^4\, \mathcal{I}_{\gamma,ZZ}}_{\text{1st Order}} \right], \end{aligned}$$

where $Q_{\gamma,2} = \int_{S_{XXZ}} \tilde{\gamma}((x_1, x_2), (x_1, x_2), z)\, f_{XX}(x_1, x_2)\, d(x_1, x_2, z)$ and all other quantities are defined in the preceding section below of Eq. (33).

[4] Herrmann, E., J. Engel, M. Wand, and T. Gasser (1995). A bandwidth selector for bi- variate kernel regression. Journal of the Royal Statistical Society. Series B (Methodological) 57 (1), 171180.

The lowest possible AMISE value under the AMISE. II scenario can be achieved if there exists a X-bandwidth which, first, allows us to profit from the (partial) annulment of the classical bias-variance tradeoff, but, second, assures that the AMISE. II scenario remains maintained. The first requirement is achieved if the X-bandwidth is of a smaller order of magnitude than the Z-bandwidth, i.e., if $h_{\gamma,X} = o(h_{\gamma,Z})$. This restriction makes those bias components that depend on $h_{\gamma,X}$ asymptotically negligible, since it implies that $h^2_{\gamma,X}\, h^2_{\gamma,Z} = o(h^4_{\gamma,Z})$ and therefore that $h^4_{\gamma,X} = o(h^2_{\gamma,X}\, h^2_{\gamma,Z})$. The latter leads to the order relations between the third, fourth, and fifth AMISE. II term as indicated in Eq. (38). The second requirement is achieved if the X-bandwidth does not converge to zero too fast, namely if $(Nh^2_{\gamma,X})^{-1} = o(1)$, which implies the order relation between the first two AMISE. II terms as indicated in Eq. (38).

2.3 Global polynomial fits

We suggest approximating the unknown quantities of the AMISE. I optimal bandwidth expressions in Eqs. (34), (35), (36), and (37) and those of the AMISE. II optimal bandwidth expressions in Eqs. (15), (16), (13), and (14) using five different global polynomial models of order four referred to as: μ_{poly}, γ_{poly}, $\tilde{\gamma}_{\text{poly}}$, $[\gamma(x,x,z)+\sigma^2_\epsilon]_{\text{poly}}$, and $[\tilde{\gamma}((x_1,x_2),(x_1,x_2),z)+\sigma^2_\varepsilon(x_1,x_2,z)]_{\text{poly}}$. Given estimates of these polynomial models, allows us to approximate the unknown quantities $\mathcal{I}_{\mu,XX}$,

$\mathcal{I}_{\mu,XZ}$, $\mathcal{I}_{\mu,ZZ}$, $Q_{\mu,1}$, $Q_{\mu,2}$, $\mathcal{I}_{\gamma,X_1X_1}$, $\mathcal{I}_{\gamma,X_1Z}$, $\mathcal{I}_{\gamma,ZZ}$, $Q_{\gamma,1}$, and $Q_{\gamma,2}$ by the empirical versions of

$$\begin{aligned}
\mathcal{I}_{\mu_{\text{poly}},XX} &= \int_{S_{XZ}} (\mu_{\text{poly}}^{(2,0)}(x,z))^2\, w_\mu\, d(x,z),\\
\mathcal{I}_{\mu_{\text{poly}},XZ} &= \int_{S_{XZ}} \mu_{\text{poly}}^{(2,0)}(x,z)\mu_{\text{poly}}^{(0,2)}(x,z)\, w_\mu\, d(x,z),\\
\mathcal{I}_{\mu_{\text{poly}},ZZ} &= \int_{S_{XZ}} (\mu_{\text{poly}}^{(0,2)}(x,z))^2\, w_\mu\, d(x,z),\\
Q_{\mu_{\text{poly}},1} &= \int_{S_{XZ}} [\gamma(x,x,z)+\sigma_\epsilon^2]_{\text{poly}}\, d(x,z),\\
Q_{\mu_{\text{poly}},2} &= \int_{S_{XZ}} \gamma_{\text{poly}}(x,x,z)\, f_X(x)\, d(x,z),\\
\mathcal{I}_{\gamma_{\text{poly}},X_1X_1} &= \int_{S_{XZ}} (\gamma_{\text{poly}}^{(2,0,0)}(x,x,z))^2\, w_\mu\, d(x,z),\\
\mathcal{I}_{\gamma_{\text{poly}},X_1Z} &= \int_{S_{XXZ}} \gamma_{\text{poly}}^{(2,0,0)}(x,x,z)\gamma_{\text{poly}}^{(0,0,2)}(x,x,z)\, w_\gamma\, d(x,x,z),\\
\mathcal{I}_{\gamma_{\text{poly}},ZZ} &= \int_{S_{XXZ}} (\gamma_{\text{poly}}^{(0,0,2)}(x,x,z))^2\, w_\gamma\, d(x,x,z),\\
Q_{\gamma_{\text{poly}},1} &= \int_{S_{XXZ}} [\tilde{\gamma}((x_1,x_2),(x_1,x_2),z)+\sigma_\varepsilon^2]_{\text{poly}}\, d(x_1,x_2,z),\\
Q_{\gamma_{\text{poly}},2} &= \int_{S_{XXZ}} \tilde{\gamma}_{\text{poly}}((x_1,x_2),(x_1,x_2),z)\; f_{XX}(x_1,x_2)\, d(x_1,x_2,z).
\end{aligned}$$

Remember that the weight functions $w_\mu = w_\mu(x,z)$ and $w_\gamma = w_\gamma(x_1,x_2,z)$ are defined as $w_\mu(x,z) = f_{XZ}(x,z)$ and $w_\gamma(x_1,x_2,z) = f_{XXZ}(x_1,x_2,z)$. Estimates for these densities are constructed from kernel density estimates of the single pdfs f_X and f_Z, where we use the Epanechnikov kernel and the bandwidth selection procedure of Sheather and Jones (1991)[5]. Note that it is necessary to estimate the models μ_{poly} and γ_{poly} with interactions, since otherwise their partial derivatives would degenerate – all necessary further details are found in the following list:

- The model μ_{poly} is fitted via regressing Y_{it} on powers (each up to the fourth power) of X_{it}, Z_t, and $X_{it}\cdot Z_t$ for all i and t.

- The model γ_{poly} is fitted via regressing $C_{ijt}^{\text{poly}} = (Y_{it}-\mu_{\text{poly}}(X_{it},Z_t))(Y_{jt}-\mu_{\text{poly}}(X_{jt},Z_t))$ on powers (each up to the fourth power) of X_{it}, X_{jt}, Z_t, $X_{it}\cdot Z_t$, and $X_{jt}\cdot Z_t$ for all t and all i, j with $i\neq j$.

- The model $\tilde{\gamma}_{\text{poly}}$ is fitted via regressing $\mathbb{C}_{(ij),(kl),t}^{\text{poly}} = (C_{ijt}^{\text{poly}} - \gamma_{\text{poly}}(X_{it},X_{jt},Z_t))(C_{klt}^{\text{poly}} - \gamma_{\text{poly}}(X_{kt},X_{lt},Z_t))$ on powers (each up to the fourth power) of X_{it}, X_{jt}, X_{kt}, X_{lt}, and Z_t

[5] Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society. Series B (Methodological) 53 (3), 683690.

for all t and all i, j, k, l with $(i, j) \neq (k, l)$.

- The model $[\gamma(x, x, z) + \sigma_\epsilon^2]_{\text{poly}}$ is fitted via regressing the noise contaminated diagonal values C_{iit}^{poly} on powers (each up to the fourth power) of X_{it}, and Z_t for all i and t.
- The model $[\tilde{\gamma}((x_1, x_2), (x_1, x_2), z) + \sigma_\varepsilon^2(x_1, x_2, z)]_{\text{poly}}$ is fitted via regressing the noise contaminated diagonal values $\mathbb{C}_{(ij),(ij),t}^{\text{poly}}$ on powers (each up to the fourth power) of X_{it}, X_{jt}, and Z_t for all i, j, and t.

For the computation of the Bonferroni type confidence intervals we also need to approximate the error variance σ_ϵ^2. This is done via the empirical version of:

$$\sigma_{\text{poly},\epsilon}^2 = \frac{1}{\int_{S_{XZ}} 1 d(x, z)} \int_{S_{XZ}} \left([\gamma(x, x, z) + \sigma_\epsilon^2]_{\text{poly}} - \gamma_{\text{poly}}(x, x, z)\right) d(x, z). \qquad (39)$$

Of course, it is always possible (and often necessary) to use more complex polynomial models that include more interaction terms and higher order polynomials. Though, due to the relatively simple structured data this rule-of-thumb method works very well. The global polynomial fits of the mean functions are shown in Figure 6. Their general shapes are plausible and we can expect them to be very useful in approximating the above unknown quantities, though these pilot estimates are not perfect substitutes for the final local linear estimates. Particularly, the temperature effects in the second time period show some implausible wave shapes, which do not show up in the nonparametric fit shown in Figure 5.

3 Data Sources

The data for our analysis come from four different sources. Hourly spot prices of the German electricity market are provided by the European Energy Power Exchange (EPEX) (`www.epexspot.com`), hourly values of Germany's gross electricity demand and electricity exchanges with other countries are provided by the European Network of Transmission System Operators for Electricity (`www.entsoe.eu`), German wind and solar power infeed data are provided by the transparency platform of the European energy exchange (`www.eex-transparency.com`), and German air temperature data are available from the German Weather Service (`www.dwd.de`).

The data dimensions are given by $n = 24$, $N = 552$, $T = 261$, and $T = 262$, where the latter two numbers are the number of working days one year before and one year after Germany's nuclear phaseout. Very few (0.2%) of the data tuples (Y_{it}, X_{it}, Z_t) with prices $Y_{it} > 200$ EUR/MWh are

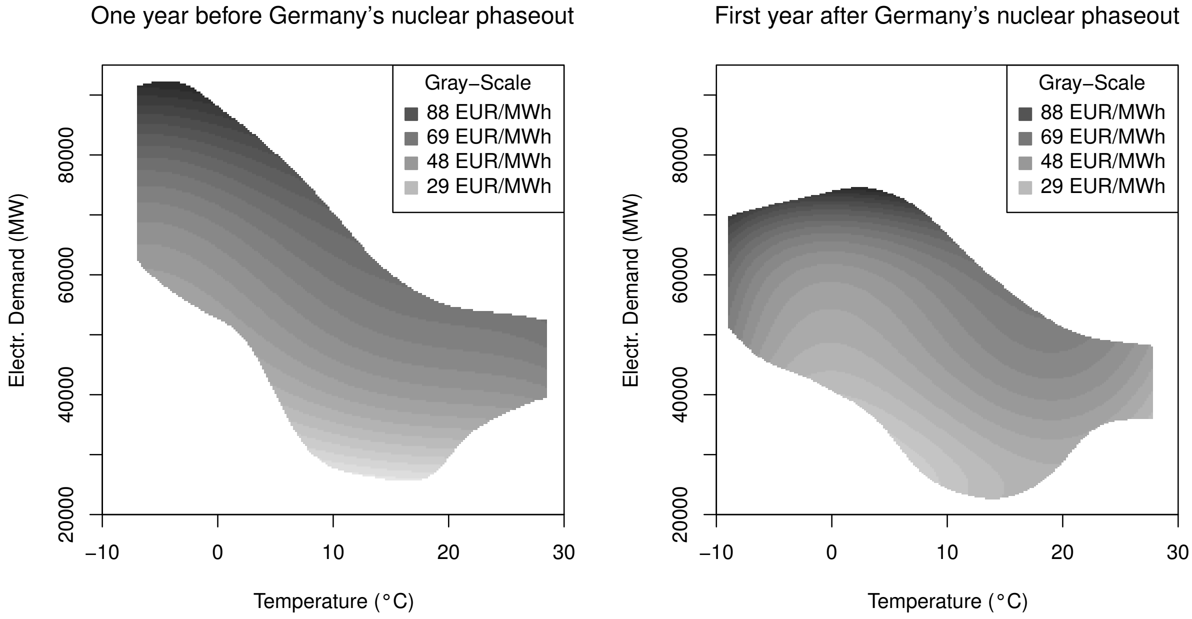

Figure 6: Approximated mean functions using global polynomial regression models.

considered as outliers and removed. Such extreme prices cannot be explained by the merit order model, since the marginal costs of electricity production usually do not exceed the value of about 200 EUR/MWh. Prices above this threshold are often referred to as "price spikes" and have to be modeled using different approaches (Ch. 4 in Burger (2008)[6]).

[6] Burger, M., B. Graeber, and G. Schindlmayr (2008). Managing Energy Risk: An Integrated View on Power and Other Energy Markets (1. ed.). Wiley.