Inference for Functional Data with Covariate Adjustments: From Sparse to Dense and Everything In-Between

Dominik Liebl *

Institute for Financial Economics and Statistics, University of Bonn

September 8, 2017

Abstract

We consider inference for the mean and covariance functions of covariate adjusted functional data using Local Linear Kernel (LLK) estimators. By means of a double asymptotic, we differentiate between "sparse" and "dense" covariate adjusted functional data – depending on the relative order of m (discretization points per function) and n (number of functions). Our simulation results demonstrate that the existing asymptotic normality results can lead to severely misleading inferences in finite samples. We explain this phenomenon based on our theoretical results and propose finite-sample corrections which provide practically useful approximations for inference in "sparse" and "dense" data scenarios as well as any hybrid data scenario in-between.

Keywords: functional data analysis, local linear kernel estimation, asymptotic normality, multiple bandwidth selection, finite-sample correction

*I want to thank my colleague Alois Kneip (University of Bonn) and Piotr Kokoszka (Colorado State University) for fruitful discussions and valuable comments which helped to improve this research work. Additionally, I want to thank Irène Gijbels (KU Leuven) and Jane-Ling Wang (UC Davis) for valuable comments on my talk at the ERCIM-2015 where I presented on earlier version of this manuscript.

1 Introduction

This work considers the case of independently identically distributed (iid) covariate adjusted functional data $X_i(., Z_i) \in L^2([0,1])$ with random covariate $Z_i \in [0,1] \subset \mathbb{R}$. As typically done in the literature on sparse functional data, we take into account measurement and discretization errors. That is, the observed data points $(Y_{ij}, U_{ij}, Z_i) \in \mathbb{R}^3$ are considered as error-prone random discretization points of an underlying unobserved random function $X_i(., Z_i)$ which is adjusted by an observed covariate Z_i,

$$Y_{ij} = X_i(U_{ij}, Z_i) + \epsilon_{ij}, \quad j = 1, \ldots, m, \quad i = 1, \ldots, n, \tag{1}$$

where $\epsilon_{ij} \in \mathbb{R}$ is an iid error term with mean zero and $\mathbb{V}(\epsilon_{ij}) = \sigma_\epsilon^2 < \infty$, independent from X_i, U_{ij}, and Z_i.

We derive inferential results for the LLK estimators of the mean function $\mu(u,z) = \mathbb{E}(X_i(u,z))$ and the covariance function $\gamma(u_1, u_2, z) = \mathrm{Cov}(X_i(u_1, z), X_i(u_2, z))$. So far the only other existing asymptotic normality results in this context are those of Jiang and Wang (2010), who consider the case of *sparse* covariate adjusted functional data where m is bounded while $n \to \infty$. However, as shown in our simulation study, the asymptotic variance expressions derived in Jiang and Wang (2010) tend to severely underestimate the actual variances in finite samples.

We are able to explain this finding based on our asymptotic normality results – our main contributions. The "finite-m asymptotic" considered by Jiang and Wang (2010) neglects an additional (functional-data-specific) variance term which is typically not negligible in practice. In contrast to Jiang and Wang (2010), we consider all cases from sparse to dense functional data depending on the relative order of m and n. This approach is related to the work of Zhang and Wang (2016), who, however, consider classical functional data *without* covariate adjustments[1].

Additionally, we derive the explicit optimal multiple bandwidth expressions for the case of sparse and dense covariate adjusted functional data. For dense functional data, this leads to rather unconventional bandwidth expressions with different convergence rates for the bandwidths in U- and Z-direction. Effectively, this imposes a necessary under-smoothing in U-direction, which guarantees that the U-related bias and variance components become negligible in comparison to the Z-related bias and variance components.

Our third contribution is concerned with finite-samples. The differentiation between sparse and dense functional data is based on pure theoretical considerations. In practice, however, it is usually impossible to differentiate between these two asymptotic data sce-

[1]Our results are based on the authors's PhD thesis (Liebl, 2013) and were developed independently from the work of Zhang and Wang (2016).

narios. Therefore, we contribute finite-sample corrections that allow for robust inferences with sparse and dense functional data as well as with any hybrid data scenario in-between.

The literature on covariate adjusted functional was initiated by the work of Cardot (2007), who considers functional principal component analysis for dense functional data, but does not provide inferential results. Jiang and Wang (2010) focus on the case of sparse functional data and their asymptotic normality results serve as a benchmark for our results. Li et al. (2015) consider a copula-based model and Zhang and Wei (2015) propose an iterative algorithm for computing functional principal components, though neither provides inferential results for the covariate adjusted mean and covariance functions. For the case *without* covariate adjustments there are several papers considering inference. Zhang and Chen (2007) and Hall and Van Keilegom (2007) consider inference in the pre-smoothing context for dense functional data. Benko et al. (2009) develop bootstrap procedures for the case of dense functional data. Cao et al. (2012) derive simultaneous confidence bands in the case of dense functional data. Gromenko and Kokoszka (2012) consider a L^2 based test statistic and address computational issues in finite samples and Horváth et al. (2013) focus on the case of dependent functional data within the same framework. Although related, the case without covariate adjustments is fundamentally different from our case, since the presence of a covariate affects the involved bandwidth selection problem in a nontrivial manner. Readers with a general interest in functional data analysis are referred to the textbooks of Ramsay and Silverman (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012), Hsing and Eubank (2015), and Kokoszka and Matthew (2017).

The rest of this paper is structured as following. The next section introduces the considered regression models and LLK estimators. Section 3 presents our assumptions and asymptotic results. Our simulation study is in Section 4. Section 5 introduces rule-of-thumb approximations to our theoretical bandwidth expressions and practical plug-in estimates for the unknown bias and variance components. All proofs can be found in the appendix.

2 Nonparametric regression models and estimators

Let X_i^c denote the centered random function $X_i^c(U_{ij}, Z_i) = X_i(U_{ij}, Z_i) - \mathbb{E}(X_i(U_{ij}, Z_i)|\mathbf{U}, \mathbf{Z})$, where $\mathbf{U} = (U_{11}, \dots, U_{nm})^\top$ and $\mathbf{Z} = (Z_1, \dots, Z_n)^\top$. Model (1) can be written as a nonparametric regression model with the bivariate mean function $\mu(U_{ij}, Z_i) = \mathbb{E}(X_i(U_{ij}, Z_i)|\mathbf{U}, \mathbf{Z})$ as its regression function,

$$Y_{ij} = \mu(U_{ij}, Z_i) + X_i^c(U_{ij}, Z_i) + \epsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, m, \tag{2}$$

where $X_i^c(.,z) \in L^2[0,1]$ is an iid centered random function, $U_{ij} \in [0,1]$ and $Z_i \in [0,1]$ are iid random predictors, and $\epsilon_{it} \in \mathbb{R}$ is an iid random error term independent from X_i^c, U_{ij} and Z_i. Note that Model (2) has a rather unusual composed error term $X_i^c(U_{ij}, Z_i) + \epsilon_{ij}$ consisting of a function- and a scalar-valued component. This structure of the error term leads the additional functional-data-specific variance term.

Likewise to Model (2) we can define the following nonparametric regression model with the trivariate covariance function $\gamma(U_{ij}, U_{ik}, Z_i) = \mathrm{Cov}(X_i(U_{ij}, Z_i), X_i(U_{ik}, Z_i)|\mathbf{U}, \mathbf{Z})$ as its regression function:

$$C_{ijk} = \gamma(U_{ij}, U_{ik}, Z_i) + \tilde{X}_i^c(U_{ij}, U_{ik}, Z_i) + \varepsilon_{ijk}, \quad i = 1, \ldots, n, \quad j \neq k \in \{1, \ldots, m\}, \quad (3)$$

where the "raw-covariances" C_{ijk}, the centered random function $\tilde{X}_i^c$, and the scalar-valued error term ε_{ijk} are defined as

$$\begin{aligned} C_{ijk} &= (Y_{ij} - \mu(U_{ij}, Z_i))(Y_{ik} - \mu(U_{ik}, Z_i)), \quad (4) \\ \tilde{X}_i^c(U_{ij}, U_{ik}, Z_i) &= X_i^c(U_{ij}, Z_i)\, X_i^c(U_{ik}, Z_i) - \gamma(U_{ij}, U_{ik}, Z_i), \text{ and} \\ \varepsilon_{ijk} &= X_i^c(U_{ij}, Z_i)\epsilon_{ik} + X_i^c(U_{ik}, Z_i)\epsilon_{ij} + \epsilon_{ij}\epsilon_{ik}. \end{aligned}$$

In contrast to ϵ_{ij}, the scalar error term ε_{ijk} is heteroscedastic with $\mathbb{V}(\varepsilon_{ijk}) = \sigma_\varepsilon^2(u_1, u_2, z)$, where $\sigma_\varepsilon^2(u_1, u_2, z) = \gamma(u_1, u_1, z)\,\sigma_\epsilon^2 + \gamma(u_2, u_2, z)\sigma_\epsilon^2 + \sigma_\epsilon^4$. Note that $\mathbb{E}(\varepsilon_{ijk}) \neq 0$ for all $j = k$, therefore all raw covariance points C_{ijk} with $j = k$ need to be excluded (see also Yao et al., 2005). Correspondingly, the number of raw covariance points for each i is $M = m^2 - m$, which makes it necessary that $m \geq 2$. As in Model (2), the error term of Model (3), $\tilde{X}_i^c(U_{ij}, U_{ik}, Z_i) + \varepsilon_{ijk}$, consists of a function- and a scalar-valued component.

We estimate the mean function $\mu(u, z)$ using the LLK estimator $\hat{\mu}(u, z; h_{\mu,U}, h_{\mu,Z})$ defined as the following locally weighted least squares estimator (see, e.g., Ruppert and Wand, 1994):

$$\hat{\mu}(u, z; h_{\mu,U}, h_{\mu,Z}) = e_1^\top \left([\mathbf{1}, \mathbf{U}_u, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,uz}[\mathbf{1}, \mathbf{U}_u, \mathbf{Z}_z]\right)^{-1} [\mathbf{1}, \mathbf{U}_u, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,uz}\mathbf{Y}, \quad (5)$$

where the vector $e_1 = (1, 0, 0)^\top$ selects the estimated intercept parameter and $[\mathbf{1}, \mathbf{U}_u, \mathbf{Z}_z]$ is a partitioned $nm \times 3$ dimensional data matrix with typical rows $(1, U_{ij} - u, Z_i - z)$. The $nm \times nm$ dimensional diagonal weighting matrix $\mathbf{W}_{\mu,uz}$ holds the bivariate multiplicative kernel weights $K_{\mu,h_{\mu,U},h_{\mu,Z}}(U_{ij} - u, Z_i - z) = h_{\mu,U}^{-1}\,\kappa(h_{\mu,U}^{-1}(U_{ij} - u))\, h_{\mu,Z}^{-1}\,\kappa(h_{\mu,Z}^{-1}(Z_i - z))$, where κ is a usual second-order kernel such as, e.g., the Epanechnikov or the Gaussian kernel. The usual kernel constants are denoted by $\nu_2(K_\mu) = (\nu_2(\kappa))^2$, with $\nu_2(\kappa) = \int u^2\kappa(u)du$, and $R(K_\mu) = R(\kappa)^2$, with $R(\kappa) = \int \kappa(u)^2 du$. All vectors and matrices are filled in correspondence with the nm dimensional vector $\mathbf{Y} = (Y_{11}, Y_{12}, \ldots, Y_{n,m-1}, Y_{n,m})^\top$.

The LLK estimator for the covariance function $\gamma(u_1, u_2, z)$ is defined correspondingly as $\hat{\gamma}(u_1, u_2, z; h_{\gamma,U}, h_{\gamma,Z}) =$

$$= e_1^\top \left([\mathbf{1}, \mathbf{U}_{u_1}, \mathbf{U}_{u_2}, \mathbf{Z}_z]^\top \mathbf{W}_{\gamma,u_1u_2z}[\mathbf{1}, \mathbf{U}_{u_1}, \mathbf{U}_{u_2}, \mathbf{Z}_z]\right)^{-1} [\mathbf{1}, \mathbf{U}_{u_1}, \mathbf{U}_{u_2}, \mathbf{Z}_z]^\top \mathbf{W}_{\gamma,u_1u_2z}\mathbf{C}, \quad (6)$$

where $e_1 = (1, 0, 0, 0)^\top$ and $[\mathbf{1}, \mathbf{U}_{u_1}, \mathbf{U}_{u_2}, \mathbf{Z}_z]$ is a $nM \times 4$ dimensional data matrix with typical rows $(1, U_{ij} - u_1, U_{ik} - u_2, Z_i - z)$. The $nM \times nM$ dimensional diagonal weighting matrix $\mathbf{W}_{\gamma,u_1u_2z}$ holds the trivariate multiplicative kernel weights $K_{\gamma,h_{\gamma,U},h_{\gamma,Z}}(U_{ij} - u_1, U_{ik} - u_2, Z_i - z) = h_{\gamma,U}^{-1}\, \kappa(h_{\gamma,U}^{-1}(U_{ij} - u_1))\; h_{\gamma,U}^{-1}\, \kappa(h_{\gamma,U}^{-1}(U_{ik} - u_2))\; h_{\gamma,Z}^{-1}\, \kappa(h_{\gamma,Z}^{-1}(Z_i - z))$, where κ is as defined above, with kernel constants are $\nu_2(K_\gamma) = (\nu_2(\kappa))^3$ and $R(K_\gamma) = R(\kappa)^3$. All vectors and matrices are filled in correspondence with the nM dimensional vector $\mathbf{C} = (C_{112}, C_{113}, \ldots, C_{n,m,m-2}, C_{n,m,m-1})^\top$.

Remark We follow the simplification as used in our main references Jiang and Wang (2010) and Zhang and Wang (2016) and use C_{ijk} as the raw-covariances instead of the empirical $\hat{C}_{ijk} = (Y_{ij} - \hat{\mu}(U_{ij}, Z_i; h_{\mu,U}, h_{\mu,Z}))(Y_{ik} - \hat{\mu}(U_{ik}, Z_i; h_{\mu,U}, h_{\mu,Z}))$.

3 Theoretical results

Before we present our asymptotic results, we list our additional assumptions which are equivalent to those in Ruppert and Wand (1994) with some straight forward adjustments to our functional data context.

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

A-RD (Random Design) The triple $(Y_{ij}, U_{ij}, Z_i) \in \mathbb{R} \times [0, 1]^2$ has the same distribution as (Y, U, Z) with pdf f_{YUZ}, where $f_{YUZ}(y, u, z) > 0$ for all $(y, u, z) \in \mathbb{R} \times [0, 1]^2$ and zero else. Equivalently, $(C_{ijk}, U_{ij}, U_{ik}, Z_i) \in \mathbb{R} \times [0, 1]^3$ has the same distribution as (C, U, U', Z) with pdf f_{CUUZ}, where $f_{CUUZ}(c, u, u', z) > 0$ for all $(c, u, u', z) \in \mathbb{R} \times [0, 1]^3$ and zero else.

A-SK (Smoothness & Kernel) The pdfs f_{YUZ} and f_{CUUZ} and their marginals are continuously differentiable. All second-order derivatives of μ and γ are continuous. The multiplicative kernel functions K_μ and K_γ are products of second-order kernel functions κ.

A-MO (Moments) $\mathbb{E}((X_i(u, z))^4) < \infty$ for all (u, z) and $\mathbb{E}(\epsilon_{ij}^2) < \infty$.

A-BW (Bandwidths) $h_{\mu,U}, h_{\mu,Z} \to 0$ and $(nm)h_{\mu,U}h_{\mu,Z} \to \infty$ as $nm \to \infty$. $h_{\mu,U}, h_{\mu,Z} \to 0$ and $(nM)h_{\mu,U}^2 h_{\mu,Z} \to \infty$ as $nM \to \infty$.

Remark Assumption A-AS is a simplified version of the asymptotic setup of Zhang and Wang (2016). The case $\theta = 0$ implies that m is bounded, which corresponds to the finite-m asymptotic as considered by Jiang and Wang (2010). For $0 < \theta < \infty$ we can consider all further scenarios from sparse to dense functional data.

The following two Theorems 3.1 and 3.2 build the basis of our theoretical results.

Theorem 3.1 (Bias and Variance of $\hat{\mu}$) *Let (u, z) be an interior point of $[0,1]^2$. Under our setup the conditional asymptotic bias and variance of the LLK estimator in Eq. (5) are then given by*

$$(i)\ \mathrm{Bias}\left\{\hat{\mu}(u, z; h_{\mu,U}, h_{\mu,Z})|\mathbf{U}, \mathbf{Z}\right\} = B_\mu(u, z)\,(1 + o_p(1))\ \ \textit{with}$$

$$B_\mu(u, z) = \frac{1}{2}\,\nu_2(K_\mu)\,\left(h^2_{\mu,U}\,\mu^{(2,0)}(u, z) + h^2_{\mu,Z}\,\mu^{(0,2)}(u, z)\right),$$

where $\mu^{(k,l)}(u, z) = (\partial^{k+l}/(\partial u^k \partial z^l))\mu(u, z)$.

$$(ii)\ \mathbb{V}\left\{\hat{\mu}(u, z; h_{\mu,U}, h_{\mu,Z})|\mathbf{U}, \mathbf{Z}\right\} = \left(V^I_\mu(u, z) + V^{II}_\mu(u, z)\right)(1 + o_p(1))\ \ \textit{with}$$

$$V^I_\mu(u, z) = (nm)^{-1}\left[h^{-1}_{\mu,U} h^{-1}_{\mu,Z}\, R(K_\mu)\frac{\gamma(u, u, z) + \sigma^2_\epsilon}{f_{UZ}(u, z)}\right]\ \ \textit{and}$$

$$V^{II}_\mu(u, z) = n^{-1}\left[\left(\frac{m-1}{m}\right) h^{-1}_{\mu,Z}\, R(\kappa)\frac{\gamma(u, u, z)}{f_Z(z)}\right].$$

Theorem 3.2 (Bias and Variance of $\hat{\gamma}$) *Let (u_1, u_2, z) be an interior point of $[0,1]^3$. Under our setup the conditional asymptotic bias and variance of the LLK estimator in Eq. (6) are then given by*

$$(i)\ \mathrm{Bias}\left\{\hat{\gamma}(u_1, u_2, z; h_{\gamma,U}, h_{\gamma,Z})|\mathbf{U}, \mathbf{Z}\right\} = B_\gamma(u_1, u_2, z)\,(1 + o_p(1))\ \ \textit{with}$$

$$B_\gamma(u_1, u_2, z) = \frac{1}{2}\nu_2(K_\gamma)\left(h^2_{\gamma,U}\left(\gamma^{(2,0,0)}(u_1, u_2, z) + \gamma^{(0,2,0)}(u_1, u_2, z)\right) + h^2_{\gamma,Z}\gamma^{(0,0,2)}(u_1, u_2, z)\right),$$

where $\gamma^{(k,l,m)}(u_1, u_2, z) = (\partial^{k+l+m}/(\partial u_1^k\, \partial u_2^l\, \partial z^m))\gamma(u_1, u_2, z)$.

$$(ii)\ \mathbb{V}\left\{\hat{\gamma}(u_1, u_2, z; h_{\gamma,U}, h_{\gamma,Z})|\mathbf{U}, \mathbf{Z}\right\} = \left(V^I_\gamma(u_1, u_2, z) + V^{II}_\gamma(u_1, u_2, z)\right)(1 + o_p(1))\ \textit{with}$$

$$V^I_\gamma(u_1, u_2, z) = (nM)^{-1}\left[h^{-2}_{\gamma,U} h^{-1}_{\gamma,Z}\, R(K_\gamma)\frac{\tilde{\gamma}((u_1, u_2), (u_1, u_2), z) + \sigma^2_\varepsilon(u_1, u_2, z)}{f_{UUZ}(u_1, u_2, z)}\right]\ \textit{and}$$

$$V^{II}_\gamma(u_1, u_2, z) = n^{-1}\left[\left(\frac{M-1}{M}\right) h^{-1}_{\gamma,Z}\, R(\kappa)\frac{\tilde{\gamma}((u_1, u_2), (u_1, u_2), z)}{f_Z(z)}\right],$$

$$\begin{aligned} \textit{where}\quad \tilde{\gamma}((u_1, u_2), (u_1, u_2), z) &= \mathrm{Cov}(\tilde{X}^c_i(u_1, u_2, z), \tilde{X}^c_i(u_1, u_2, z))\quad \textit{and} \\ \sigma^2_\varepsilon(u_1, u_2, z) &= \gamma(u_1, u_1, z)\,\sigma^2_\epsilon + \gamma(u_2, u_2, z)\sigma^2_\epsilon + \sigma^4_\epsilon. \end{aligned}$$

The bias expressions in Theorems 3.1 and 3.2 correspond to the classical bias results (see, e.g., Ruppert and Wand, 1994). The first variance terms $V^I_\mu(u, z)$ and $V^I_\gamma(u_1, u_2, z)$ are equivalent to those in Theorems 3.2 and 3.4 of Jiang and Wang (2010) who consider the LLK estimators $\hat{\mu}(u,z)$ and $\hat{\gamma}(u,z)$ under the finite-m asymptotic. The second (functional-data-specific) variance terms, $V_\mu^{II}(u,z)$ and $V_\gamma^{II}(u_1,u_2,z)$, are negligible under such a finite-m asymptotic, but generally not negligible when considering a double asymptotic with $m \to \infty$ and $n \to \infty$.

Whether the first variance terms, V_μ^I and V_γ^I, or the second, functional-data-specific variance terms, V_μ^{II} and V_γ^{II}, are the leading variance terms depends on the bandwidth choices and on the relative order of m and n, i.e., on the value of θ in $m \asymp n^\theta$. In order to determine the decisive θ value we postulate optimal bandwidth choices determined from minimizing the usual Asymptotic Mean Integrated Squared Error (AMISE) criteria,

$$\mathrm{AMISE}_{\hat{\mu}} = \int \left([\mathrm{Bias}\{\hat{\mu}(u,z)|\mathbf{U},\mathbf{Z}\}]^2 + \mathbb{V}\{\hat{\mu}(u,z)|\mathbf{U},\mathbf{Z}\} \right) f_{UZ}(u,z)\, d(u,z) \quad \text{and}$$

$$\mathrm{AMISE}_{\hat{\gamma}} = \int \left([\mathrm{Bias}\{\hat{\gamma}(u_1,u_2,z)|\mathbf{U},\mathbf{Z}\}]^2 + \mathbb{V}\{\hat{\gamma}(u_1,u_2,z)|\mathbf{U},\mathbf{Z}\} \right) f_{UUZ}(u_1,u_2,z)\, d(u_1,u_2,z).$$

In anticipation of some of our results: Under AMISE optimal bandwidth choices, the discriminating θ-threshold is given by $\theta = 1/5$. That is, if $m/n^{1/5} \to 0$ and $\sqrt{M}/n^{1/5} \to 0$, the first variance terms V_μ^I and V_γ^I are the leading variance terms. This asymptotic scenario comprises situations where m and $\sqrt{M}$ are eventually *very* small in comparison to n. For simplicity, we refer to this asymptotic scenario as "sparse" covariate adjusted functional data, even though a more appropriate description would be "ultra-sparse". If, however, $m/n^{1/5} \to \infty$ and $\sqrt{M}/n^{1/5} \to \infty$, then the functional-data-specific variance terms V_μ^{II} and V_γ^{II} are the leading variance terms. In contrast to the "sparse" case, this asymptotic scenario comprises quite general situations where m and $\sqrt{M}$ may eventually be smaller than n, comparable to n, or larger than n. For simplicity, we refer to this asymptotic scenario as "dense" covariate adjusted functional data, even though a more appropriate description would be "sparse-to-dense".

Remark This differentiation between "sparse" and "dense" functional data is inspired by the work of Zhang and Wang (2016), who consider classical functional data *without* covariate adjustments. Zhang and Wang (2016) differentiate between "non-dense" (there $m/n^{1/4} \to 0$ and $\sqrt{M}/n^{1/4} \to 0$), "dense" (there $m/n^{1/4} \to C$ and $\sqrt{M}/n^{1/4} \to C$), and "ultra-dense" (there $m/n^{1/4} \to \infty$ and $\sqrt{M}/n^{1/4} \to \infty$) functional data. We do not follow their notation of "dense" functional data (here $m/n^{1/5} \to C$ and $\sqrt{M}/n^{1/5} \to C$), since this leads to a stalemate with variance terms having equal orders of magnitude which makes it impossible to derive explicit optimal bandwidth expressions.

3.1 Sparse functional data

The explicit AMISE optimal bandwidth expressions for the case of leading first variance terms, V_μ^I and V_γ^I, can be found in the following two Theorems:

Theorem 3.3 (Sparse - optimal bandwidths for $\hat{\mu}$) *Let $m/n^{1/5} \to 0$ and (u,z) be an interior point of $[0,1]^2$. Under our setup the AMISE optimal bandwidths for the LLK estimator in Eq. (5) are then given by*

$$h_{\mu,U}^S = \left(\frac{R(K_\mu)\, Q_{\mu,1}\, \mathcal{I}_{\mu,ZZ}^{3/4}}{nm\ (\nu_2(K_\mu))^2 \left[\mathcal{I}_{\mu,UU}^{1/2}\, \mathcal{I}_{\mu,ZZ}^{1/2} + \mathcal{I}_{\mu,UZ}\right] \mathcal{I}_{\mu,UU}^{3/4}} \right)^{1/6} \tag{7}$$

$$h_{\mu,Z}^S = \left(\frac{\mathcal{I}_{\mu,UU}}{\mathcal{I}_{\mu,ZZ}} \right)^{1/4} h_{\mu,U}^S, \ \ \textit{where} \tag{8}$$

$Q_{\mu,1} = \int (\gamma(u,u,z) + \sigma_\epsilon^2)\ d(u,z)$, $\mathcal{I}_{\mu,UZ} = \int \mu^{(2,0)}(u,z)\mu^{(0,2)}(u,z)\, f_{UZ}(u,z)\, d(u,z)$, $\mathcal{I}_{\mu,UU} = \int \left(\mu^{(2,0)}(u,z)\right)^2 f_{UZ}(u,z)\, d(u,z)$, *and* $\mathcal{I}_{\mu,ZZ} = \int \left(\mu^{(0,2)}(u,z)\right)^2 f_{UZ}(u,z)\, d(u,z)$.

Theorem 3.4 (Sparse - optimal bandwidths for $\hat{\gamma}$) *Let $\sqrt{M}/n^{1/5} \to 0$ and (u_1,u_2,z) be an interior point of $[0,1]^3$. Under our setup the AMISE optimal bandwidths for the LLK estimator in Eq. (6) are then given by*

$$h_{\gamma,U}^S = \left(\frac{R(K_\gamma)\ Q_{\gamma,1}\ 4\sqrt{2}\, \mathcal{I}_{\gamma,ZZ}^{3/2}}{nM\ (\nu_2(K_\gamma))^2 \left(2\ (\nu_2(K_\gamma))^2\, \mathcal{I}_{\gamma,U_{(1)}Z} + C_{\mathcal{I}}\right) \left(C_{\mathcal{I}} - \mathcal{I}_{\gamma,U_{(1)}Z}\right)^{3/2}} \right)^{1/7} \tag{9}$$

$$h_{\gamma,Z}^S = \left(\frac{C_{\mathcal{I}} - \mathcal{I}_{\gamma,U_{(1)}Z}}{2\,\mathcal{I}_{\gamma,ZZ}} \right)^{1/2} h_{\gamma,U}^S, \tag{10}$$

$$\begin{aligned}
\textit{where}\quad C_{\mathcal{I}} &= (\mathcal{I}_{\gamma,U_{(1)}Z}^2 + 4\,(\mathcal{I}_{\gamma,U_{(1)}U_{(1)}} + \mathcal{I}_{\gamma,U_{(1)}U_{(2)}})\,\mathcal{I}_{\gamma,ZZ})^{1/2},\\
Q_{\gamma,1} &= \int (\tilde{\gamma}((u_1,u_2),(u_1,u_2),z) + \sigma_\varepsilon^2(u_1,u_2,z))\ d(u_1,u_2,z)\\
\mathcal{I}_{\gamma,U_{(1)}U_{(1)}} &= \int \left(\gamma^{(2,0,0)}(u_1,u_2,z)\right)^2 f_{UUZ}(u_1,u_2,z)\, d(u_1,u_2,z),\\
\mathcal{I}_{\gamma,U_{(1)}U_{(2)}} &= \int \left(\gamma^{(2,0,0)}(u_1,u_2,z)\gamma^{(0,2,0)}(u_1,u_2,z)\right) f_{UUZ}(u_1,u_2,z)\, d(u_1,u_2,z),\\
\mathcal{I}_{\gamma,U_{(1)}Z} &= \int \gamma^{(2,0,0)}(u_1,u_2,z)\gamma^{(0,0,2)}(u_1,u_2,z)\, f_{UUZ}(u_1,u_2,z)\, d(u_1,u_2,z), \quad \textit{and}\\
\mathcal{I}_{\gamma,ZZ} &= \int \left(\gamma^{(0,0,2)}(u_1,u_2,z)\right)^2 f_{UUZ}(u_1,u_2,z)\, d(u_1,u_2,z).
\end{aligned}$$

The bandwidth rates are well-known for bi- and trivariate nonparametric estimators and essentially equivalent results can be found, e.g., in Herrmann et al. (1995). The superscript "S" stands for "Sparse" covariate adjusted functional data.

The following Corollaries 3.1 and 3.2 contain our asymptotic normality results for the estimators $\hat{\mu}$ and $\hat{\gamma}$ for sparse functional data.

Corollary 3.1 (Sparse - asymptotic normality of $\hat{\mu}$) *Let $m/n^{1/5} \to 0$, let (u, z) be an interior point of $[0, 1]^2$, and assume optimal bandwidth choices. Under our setup the LLK estimator in Eq.* (5) *is then asymptotically normal.*
(a) Without finite sample correction:

$$\left(\frac{\hat{\mu}(u, z; h^S_{\mu,U}, h^S_{\mu,Z}) - B_\mu(u, z; h^S_{\mu,U}, h^S_{\mu,Z}) - \mu(u, z)}{\sqrt{V^I_\mu(u, z; h^S_{\mu,U}, h^S_{\mu,Z})}} \right) \overset{a}{\sim} N(0, 1)$$

(b) With finite sample correction:

$$\left(\frac{\hat{\mu}(u, z; h^S_{\mu,U}, h^S_{\mu,Z}) - B_\mu(u, z; h^S_{\mu,U}, h^S_{\mu,Z}) - \mu(u, z)}{\sqrt{V^I_\mu(u, z; h^S_{\mu,U}, h^S_{\mu,Z}) + V^{II}_\mu(u, z; h^S_{\mu,Z})}} \right) \overset{a}{\sim} N(0, 1)$$

Corollary 3.2 (Sparse - asymptotic normality of $\hat{\gamma}$) *Let $\sqrt{M}/n^{1/5} \to 0$, let (u_1, u_2, z) be an interior point of $[0, 1]^3$, and assume optimal bandwidth choices. Under our setup the LLK estimator in Eq.* (6) *is then asymptotically normal.*
(a) Without finite sample correction:

$$\left(\frac{\hat{\gamma}(u_1, u_2, z; h^S_{\gamma,U}, h^S_{\gamma,Z}) - B_\gamma(u_1, u_2, z; h^S_{\gamma,U}, h^S_{\gamma,Z}) - \gamma(u_1, u_2, z)}{\sqrt{V^I_\gamma(u_1, u_2, z; h^S_{\gamma,U}, h^S_{\gamma,Z})}} \right) \overset{a}{\sim} N(0, 1)$$

(b) With finite sample correction:

$$\left(\frac{\hat{\gamma}(u_1, u_2, z; h^S_{\gamma,U}, h^S_{\gamma,Z}) - B_\gamma(u_1, u_2, z; h^S_{\gamma,U}, h^S_{\gamma,Z}) - \gamma(u_1, u_2, z)}{\sqrt{V^I_\gamma(u_1, u_2, z; h^S_{\gamma,U}, h^S_{\gamma,Z}) + V^{II}_\gamma(u_1, u_2, z; h^S_{\gamma,Z})}} \right) \overset{a}{\sim} N(0, 1)$$

The above corollaries imply that the standard optimal convergence rates for bivariate ($nm^{-1/3}$) and trivariate ($nM^{-2/7}$) LLK estimators are attained. Corollaries 3.1 *(a)* and 3.2 *(a)* are essentially equivalent to Theorems 3.2 and 3.4 of Jiang and Wang (2010) who, however, consider the LLK estimators under the finite-m asymptotic. In contrast, we show that these results hold for all $n \to \infty$ and $m, M \to \infty$ with $m/n^{1/5} \to 0$ and $\sqrt{M}/n^{1/5} \to 0$ respectively. Corollaries 3.1 *(b)* and 3.2 *(b)* contain our finite-sample corrections that allow for robust inferences; see our simulation study in Section 4.

3.2 Dense functional data

If the second variance summands, V^{II}_μ and V^{II}_γ, are the leading variance terms, it is possible to achieve fast *uni*variate convergence rates for the bi- and trivariate estimators $\hat{\mu}(u, z)$ and

$\hat{\gamma}(u_1, u_2, z)$. By contrast to the preceding section, however, it is impossible to determine the optimal bandwidths by using only the leading variance terms V_μ^{II} and V_γ^{II} respectively. The trick is to determine the bandwidth expressions in a hierarchical manner: The optimal Z-bandwidths $h^D_{\mu,Z}$ and $h^D_{\gamma,Z}$ must be derived by optimizing with respect to the leading (i.e., Z-related) bias and variance terms. Given the optimal Z-bandwidths, the optimal U-bandwidths $h^D_{\mu,U}$ and $h^D_{\gamma,U}$ can be determined by optimizing the subsequent lower-order bias and variance terms. This leads to the following optimal bandwidth expressions, where the superscript "D" suggests that we are considering the case of "Dense" covariate adjusted functional data.

Theorem 3.5 (Dense - optimal bandwidths for $\hat{\mu}$) *Let $m/n^{1/5} \to \infty$ and (u, z) be an interior point of $[0, 1]^2$. Under our setup the AMISE optimal bandwidths for the LLK estimator in Eq. (5) are then given by*

$$h^D_{\mu,Z} = \left(\frac{R(\kappa)\, Q_{\mu,2}}{n\ (\nu_2(K_\mu))^2\ \mathcal{I}_{\mu,ZZ}}\right)^{1/5} \quad \textit{and} \tag{11}$$

$$h^D_{\mu,U} = \left(\frac{R(K_\mu)\, Q_{\mu,1}}{nm\ (\nu_2(K_\mu))^2\ \mathcal{I}_{\mu,UZ}}\right)^{1/3} \left(h^D_{\mu,Z}\right)^{-1}, \quad \textit{where} \tag{12}$$

$Q_{\mu,1} = \int (\gamma(u,u,z) + \sigma_\epsilon^2)\, d(u,z)$, $\mathcal{I}_{\mu,UZ} = \int \mu^{(2,0)}(u,z)\mu^{(0,2)}(u,z)\, f_{UZ}(u,z)\, d(u,z)$,
$Q_{\mu,2} = \int \gamma(u,u,z)\, f_U(u)\, d(u,z)$, *and* $\mathcal{I}_{\mu,ZZ} = \int (\mu^{(0,2)}(u,z))^2\, f_{UZ}(u,z)\, d(u,z)$.

Theorem 3.6 (Dense - optimal bandwidths for $\hat{\gamma}$) *Let $\sqrt{M}/n^{1/5} \to \infty$ and (u_1, u_2, z) be an interior point of $[0, 1]^3$. Under our setup the AMISE optimal bandwidths for the LLK estimator in Eq. (6) are then given by*

$$h^D_{\gamma,Z} = \left(\frac{R(\kappa)\, Q_{\gamma,2}}{n\ (\nu_2(K_\gamma))^2\ \mathcal{I}_{\gamma,ZZ}}\right)^{1/5} \quad \textit{and} \tag{13}$$

$$h^D_{\gamma,U} = \left(\frac{R(K_\gamma)\, Q_{\gamma,1}}{nM\ (\nu_2(K_\gamma))^2\ \mathcal{I}_{\gamma,U_{(1)}Z}}\right)^{1/4} \left(h^D_{\gamma,Z}\right)^{-3/4}, \tag{14}$$

$$\begin{aligned}
\textit{where} \quad Q_{\gamma,1} &= \int (\tilde{\gamma}((u_1,u_2),(u_1,u_2),z) + \sigma_\varepsilon^2(u_1,u_2,z))\, d(u_1,u_2,z),\\
Q_{\gamma,2} &= \int \tilde{\gamma}((u_1,u_2),(u_1,u_2),z)\, f_{UU}(u_1,u_2)\, d(u_1,u_2,z),\\
\mathcal{I}_{\gamma,ZZ} &= \int (\gamma^{(0,0,2)}(u_1,u_2,z))^2\, f_{UUZ}(u_1,u_2,z)\, d(u_1,u_2,z), \quad \textit{and}\\
\mathcal{I}_{\gamma,U_{(1)}Z} &= \int \gamma^{(2,0,0)}(u_1,u_2,z)\gamma^{(0,0,2)}(u_1,u_2,z)\, f_{UUZ}(u_1,u_2,z)\, d(u_1,u_2,z).
\end{aligned}$$

Note that the AMISE optimal bandwidths $h^D_{\mu,U}$ and $h^D_{\mu,Z}$ in Eqs. (11) and (12) and $h^D_{\gamma,U}$ and $h^D_{\gamma,Z}$ in Eqs. (13) and (14) are in a sense *anti*-proportional to each other. A larger

Z-bandwidth implies a smaller U-bandwidth, and vice versa, for given n and m. This is contrary to the classical multiple bandwidth results where the single bandwidths are directly proportional to each other.

To explain this finding, observe that a larger Z-bandwidth implies that more functions $X_i(., Z_i)$ are used for computing (local) averages. However, taking averages over an increased amount of data reduces variance so that we can afford some further increase in variance by using *under*smoothing bandwidths in U-direction (as implemented by Eqs. (12) and (14)). All in all, this (partial) undersmoothing strategy leads to better estimation performance. A related result can be found in Benko et al. (2009), who, however, consider the simpler context without covariate adjustments.

The following corollaries contain our asymptotic normality result for the estimators $\hat{\mu}$ and $\hat{\gamma}$ in the case of dense functional data:

Corollary 3.3 (Dense - asymptotic normality of $\hat{\mu}$) *Let $m/n^{1/5} \to \infty$, let (u, z) be an interior point of $[0,1]^2$, and assume optimal bandwidth choices. Under our setup the LLK estimator in Eq. (5) is then asymptotically normal.*
(a) Without finite sample correction:

$$\left(\frac{\hat{\mu}(u, z; h^D_{\mu,U}, h^D_{\mu,Z}) - B^D_\mu(u, z; h^D_{\mu,Z}) - \mu(u, z)}{\sqrt{V^{II}_\mu(u, z; h^D_{\mu,Z})}}\right) \overset{a}{\sim} N(0,1)$$

(b) With finite sample correction:

$$\left(\frac{\hat{\mu}(u, z; h^D_{\mu,U}, h^D_{\mu,Z}) - B^D_\mu(u, z; h^D_{\mu,Z}) - \mu(u, z)}{\sqrt{V^{I}_\mu(u, z; h^D_{\mu,U}, h^D_{\mu,Z}) + V^{II}_\mu(u, z; h^D_{\mu,Z})}}\right) \overset{a}{\sim} N(0,1)$$

where $B^D_\mu(u, z; h^D_{\mu,Z}) = \frac{1}{2}\nu_2(K_\mu)(h^D_{\mu,Z})^2\mu^{(0,2)}(u, z)$.

Corollary 3.4 (Dense - asymptotic normality of $\hat{\gamma}$) *Let $\sqrt{M}/n^{1/5} \to \infty$, let (u_1, u_2, z) be an interior point of $[0,1]^3$, and assume optimal bandwidth choices. Under our setup the LLK estimator in Eq. (6) is then asymptotically normal.*
(a) Without finite sample correction:

$$\left(\frac{\hat{\gamma}(u_1, u_2, z; h^D_{\gamma,U}, h^D_{\gamma,Z}) - B^D_\gamma(u_1, u_2, z; h^D_{\gamma,Z}) - \gamma(u_1, u_2, z)}{\sqrt{V^{II}_\gamma(u_1, u_2, z; h^D_{\gamma,Z})}}\right) \overset{a}{\sim} N(0,1)$$

(b) With finite sample correction:

$$\left(\frac{\hat{\gamma}(u_1, u_2, z; h^D_{\gamma,U}, h^D_{\gamma,Z}) - B^D_\gamma(u_1, u_2, z; h^D_{\gamma,Z}) - \gamma(u_1, u_2, z)}{\sqrt{V^{I}_\gamma(u_1, u_2, z; h^D_{\gamma,U} + h^D_{\gamma,Z}) + V^{II}_\gamma(u_1, u_2, z; h^D_{\gamma,Z})}}\right) \overset{a}{\sim} N(0,1),$$

where $B^D_\gamma(u_1, u_2, z; h^D_{\gamma,Z}) = \frac{1}{2}\nu_2(K_\gamma)(h^D_{\mu,Z})^2\gamma^{(0,0,2)}(u_1, u_2, z)$.

The above corollaries imply that the optimal convergence rate ($n^{-2/5}$) of *uni*variate LLK estimators is attained, although, we are considering bi- and trivariate estimators $\hat{\mu}(u,z)$ and $\hat{\gamma}(u_1,u_2,z)$. Indeed, if $m/n^{1/5}\to\infty$ and $\sqrt{M}/n^{1/5}\to\infty$, the LLK estimators $\hat{\mu}(u,z)$ and $\hat{\gamma}(u_1,u_2,z)$ behave like LLK estimators for univariate regression functions with asymptotically negligible U-related bias and variance components and with Z_i as their only covariate. That is, LLK estimators behave as if the sample of n functions $\{X_1(.,Z_1),\dots,X_n(.,Z_n)\}$ were *fully* observed such that smoothing needs to be done only in Z-direction.

This is essentially equivalent to the results in Corollaries 3.2 and 3.5 of Zhang and Wang (2016), but with the additional complexity of considering covariate adjustments. Their LLK estimators behave as if they were classical parametric moment estimators applied to a sample of n fully observed random functions without covariate adjustments.

4 Simulation

In order to assess the finite-sample properties of our asymptotic normality results, we consider the performance of the following pointwise confidence intervals:

Sparse - without finite-sample correction (Corollary 3.1 *(a)*):

$$\mathrm{CI}^S(u,z)=\hat{\mu}^S_{\mathrm{bc}}(u,z;h^S_{\mu,U},h^S_{\mu,Z})\pm z_{1-\alpha/2}\sqrt{V^I_\mu(u,z;h^S_{\gamma,U},h^S_{\gamma,Z})}$$

Sparse - with finite-sample correction (Corollary 3.1 *(b)*):

$$\mathrm{CI}^S_C(u,z)=\hat{\mu}^S_{\mathrm{bc}}(u,z;h^S_{\mu,U},h^S_{\mu,Z})\pm z_{1-\alpha/2}\sqrt{V^I_\mu(u,z;h^S_{\gamma,U},h^S_{\gamma,Z})+V^{II}_\mu(u,z;h^S_{\gamma,Z})}$$

Dense - without finite-sample correction (Corollary 3.3 *(a)*):

$$\mathrm{CI}^D(u,z)=\hat{\mu}^D_{\mathrm{bc}}(u,z;h^D_{\mu,U},h^D_{\mu,Z})\pm z_{1-\alpha/2}\sqrt{V^I_\mu(u,z;h^D_{\gamma,U},h^D_{\gamma,Z})}$$

Dense - with finite-sample correction (Corollary 3.3 *(b)*):

$$\mathrm{CI}^D_C(u,z)=\hat{\mu}^D_{\mathrm{bc}}(u,z;h^D_{\mu,U},h^D_{\mu,Z})\pm z_{1-\alpha/2}\sqrt{V^I_\mu(u,z;h^D_{\gamma,U},h^D_{\gamma,Z})+V^{II}_\mu(u,z;h^D_{\gamma,Z})},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution and

$$\begin{aligned}\hat{\mu}^S_{\mathrm{bc}}(u,z;h^S_{\mu,U},h^S_{\mu,Z})&=\hat{\mu}(u,z;h^S_{\mu,U},h^S_{\mu,Z})-B_\mu(u,z;h^S_{\mu,U},h^S_{\mu,Z})\\ \hat{\mu}^D_{\mathrm{bc}}(u,z;h^D_{\mu,U},h^D_{\mu,Z})&=\hat{\mu}(u,z;h^D_{\mu,U},h^D_{\mu,Z})-B^D_\mu(u,z;h^D_{\mu,Z})\end{aligned}\tag{15}$$

denote the bias-corrected mean estimates.

The above theoretical confidence intervals are infeasible as they depend on the unknown bandwidth, bias, and variance expressions. To achieve feasible confidence intervals we replace the unknown theoretical bandwidth parameters ($h^S_{\mu,U}$, $h^S_{\mu,Z}$, $h^D_{\mu,U}$, and $h^D_{\mu,Z}$) using simple but effective rule-of-thumb estimates ($\hat{h}^S_{\mu,U}$, $\hat{h}^S_{\mu,Z}$, $\hat{h}^D_{\mu,U}$, and $\hat{h}^D_{\mu,Z}$), based on our theoretical bandwidth expressions as described in Section 5.1. This allows us to asses the our explicit bandwidth expressions, though, (Generalized) Cross Validation (GCV or CV) could be used, too. The unknown bias (B_μ and B_μ^D) and variance (V_μ^I and V_μ^{II}) terms are estimated using LLK estimators ($\hat{B}_\mu$, $\hat{B}_\mu^D$, $\hat{V}_\mu^I$, and $\hat{V}_\mu^{II}$) as described in Section 5.2.

We focus only on confidence intervals for the mean function. Nonparametric confidence intervals for the covariance function are typically not used in practice as they involve the nonparametric estimation of the fourth-moment function $\tilde{\gamma}$ contained in the unknown variance terms V_γ^I and V_γ^{II}. The latter is complicated and typically leads to very unstable estimates due to an accumulation of preceding estimation errors. Our theoretical results on the LLK estimator $\hat{\gamma}$ are, nevertheless, of crucial importance for estimating the unknown variance expressions in the feasible versions of the above confidence intervals (see Section 5.2).

We simulate data from $Y_{ij} = \mu(U_{ij}, Z_i) + X_i^c(U_{ij}, Z_i) + \epsilon_{ij}$, where $U_{ij} \sim \text{Unif}(0,1)$, $Z_i \sim \text{Unif}(0,1)$, $\epsilon_{ij} \sim \text{N}(0,1)$, $X_i^c(u,z) = \xi_{i1}\psi_1(u,z) + \xi_{i2}\psi_2(u,z)$, $\xi_{i1} \sim \text{N}(0,3)$, $\xi_{i2} \sim \text{N}(0,2)$, $\gamma(u_1,u_2,z) = 2\psi_1(u_1,z)\psi_1(u_2,z) + \psi_2(u_1,z)\psi_2(u_2,z)$, $m \in \{5,10,15\}$, and $n = 100$. The following two Data Generating Processes (DGPs) are used:

$$
\begin{array}{llll}
\textbf{DGP 1} & \mu(u,z) = 5\sin(\pi uz/2), & \psi_1(u,z) = \sin(\pi uz), & \psi_2(u,z) = \sin(2\pi uz), \\
\textbf{DGP 2} & \mu(u,z) = 5\sin(\pi uz), & \psi_1(u,z) = \sin(2\pi uz), & \psi_2(u,z) = \sin(3\pi uz),
\end{array}
$$

where DGP-2 is the more complex process as its mean and basis functions have a higher curvature. For each DGP and each sample size combination we draw 1000 Monte-Carlo repetitions and compute the empirical coverage probabilities of the pointwise confidence intervals at $(u,z) = (0.5, 0.5)$. At this point the second order derivatives of the mean functions are non-zero, namely,

$$
\begin{array}{ll}
\text{DGP-1} & \mu^{(2,0)}(0.5,0.5) = \mu^{(0,2)}(0.5,0.5) \approx -1.2 \\
\text{DGP-2} & \mu^{(2,0)}(0.5,0.5) = \mu^{(0,2)}(0.5,0.5) \approx -8.7,
\end{array}
$$

which allows us to asses the performance of the empirical the bias corrections in Eq. (15). The data scenario with $m = 5$ and $n = 100$ represents our "sparse" data scenario. The two further scenarios with $m \in \{10, 15\}$ deviate from this "sparse" data scenario towards a more "dense" data scenario. The implementation is done in R (R Core Team, 2017) and the R-codes are available from the author.

The two top panels in Figure 1 show the empirical coverage probabilities of the feasible confidence intervals with plugged-in estimates $\hat{h}_{\mu,U}^S$, $\hat{h}_{\mu,Z}^S$, $\hat{h}_{\mu,U}^D$, $\hat{h}_{\mu,Z}^D$, $\hat{B}_\mu$, $\hat{B}_\mu^D$, $\hat{V}_\mu^I$, and $\hat{V}_\mu^{II}$ from Sections 5.1 and 5.2. The two bottom panels show the empirical coverage probabilities of the infeasible theoretical confidence intervals based on the theoretical bandwidth, bias, and variance expressions. The infeasible confidence intervals serve as validating bench-

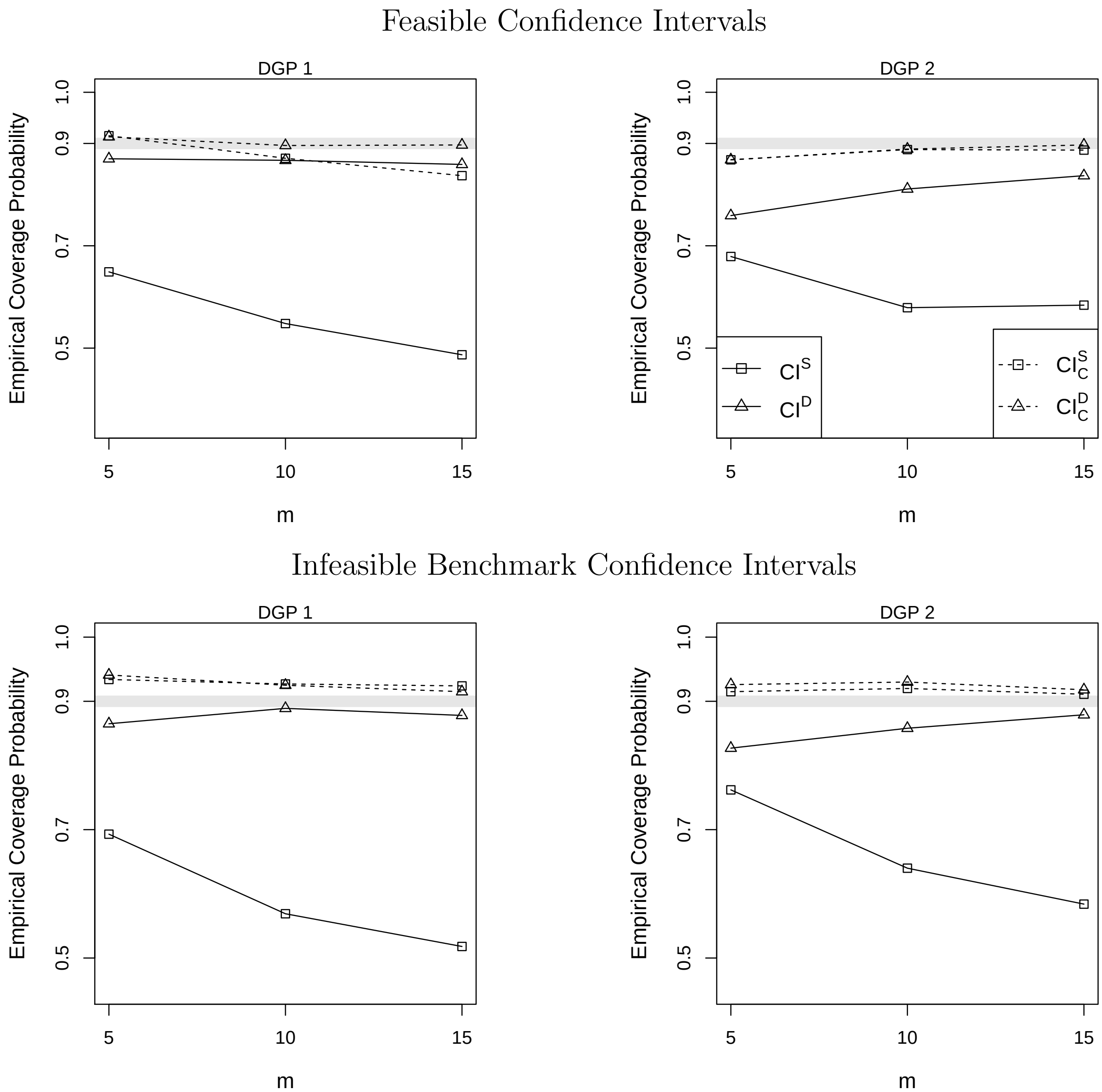

Figure 1: Empirical coverage probabilities of the feasible and infeasible confidence intervals.

marks, since they allow us to abstract from the additional estimation errors that are due to the plug-in estimates. We use $(1 - \alpha/2) = 0.9$ as our nominal coverage probability.

Let us first consider the feasible version of the "sparse" confidence interval CI^S. This is an interesting special case, since essentially the same confidence interval would be used by a practitioner who takes the asymptotic normality result in Theorem 3.2 of Jiang and Wang (2010) as a theoretical basis. This confidence interval, however, shows a very poor performance with far to small coverage probabilities. The problem occurs in our sparse data scenario with sample sizes $m = 5$ and $n = 100$ and – as expected – becomes worse as m increases.

By contrast, the feasible version of the "dense" confidence interval CI^D performs quite satisfactorily. Though, the best and most stable results are achieved by the feasible versions of the confidence intervals with finite-sample corrections, CI_C^S and CI_C^D, both showing an almost equally good performance. All of our results on the feasible confidence intervals are essentially equivalent to those for the infeasible theoretical benchmark confidence intervals shown in the two bottom panels in Figure 1. This comparison serves as a validation of our simulation results, since it shows that the results are not driven by bad and imprecise plug-in estimates.

The reason for the poor performance of the "sparse" confidence interval CI^S is shown in the first row of Table 1. The variance term $V_\mu^I = V_\mu^I(u,z)$ used to construct CI^S, severely underestimates the finite-sample variance $\widehat{\mathrm{Var}}(\hat{\mu}) = \widehat{\mathrm{Var}}(\hat{\mu}(u,z))$ of the LLK estimator $\hat{\mu}(u,z)$, where $\widehat{\mathrm{Var}}(\hat{\mu})$ is computed from the 1000 Monte Carlo replications. For $m = 5$, the first variance term, V_μ^I, is 3.7 and 2.5 times smaller than the finite-sample variance $\widehat{\mathrm{Var}}(\hat{\mu})$ and – as expected – the ratio becomes worse as m increases. This leads to too narrow confidence intervals CI^S and hence to too small coverage probabilities.

		DGP-1			DGP-2		
	m	5	10	15	5	10	15
Sparse	$\widehat{\mathrm{Var}}(\hat{\mu})/V_\mu^I$	3.7	4.5	7.2	2.5	3.9	5.2
	$\widehat{\mathrm{Var}}(\hat{\mu})/(V_\mu^I + V_\mu^{II})$	1.3	1.4	1.3	0.9	0.8	0.8
Dense	$\widehat{\mathrm{Var}}(\hat{\mu})/V_\mu^{II}$	1.7	1.3	1.6	1.4	1.2	1.1
	$\widehat{\mathrm{Var}}(\hat{\mu})/(V_\mu^I + V_\mu^{II})$	**1.1**	**1.0**	**1.3**	**0.9**	**0.9**	**0.9**

Table 1: Empirical versus theoretical variances. The figures in the upper panel are based on the sparse bandwidths, $\hat{h}_{\mu,U}^S$ and $\hat{h}_{\mu,Z}^S$, from Eq.s (16) and (17); those of the lower panel are based on the dense bandwidths, $\hat{h}_{\mu,U}^D$ and $\hat{h}_{\mu,Z}^D$, from Eq.s (20) and (21).

This observation is in line with the relatively small $\theta = 1/5$ threshold for differentiating between "sparse" (or better "ultra-sparse") and "dense" (or better "sparse-to-dense") functional data. The small $\theta = 1/5$ threshold implies that the functional-data-specific second variance term V_μ^{II} will be non-negligible in real data scenarios where m is relatively small (in comparison to n). Therefore, including the second variance terms $V_\mu^{II}(u,z)$ leads to strongly improved approximations of the finite-sample variances of $\hat{\mu}$; see the second and fourth row in Table 1. This explains the superior performances of the confidence intervals, CI_C^S and CI_C^D, incorporating our finite-sample corrections.

Summing up, the best and most robust inferential results are achieved using the confidence intervals with finite-sample corrections, CI_C^S and CI_C^D, with a slight advantage for the confidence interval CI_C^D incorporating "dense" (or "sparse-to-dense") bandwidths, which is due to the more general ("sparse-to-dense") applicability of the "dense" bandwidths.

5 Bandwidth, bias and variance approximations

5.1 Rule-of-thumb bandwidth approximations

Our above bandwidth expressions are infeasible as they depend on the unknown quantities $\mathcal{I}_{\mu,UU}$, $\mathcal{I}_{\mu,UZ}$, $\mathcal{I}_{\mu,ZZ}$, $Q_{\mu,1}$, $Q_{\mu,2}$, $\mathcal{I}_{\gamma,U_{(1)}U_{(1)}}$, $\mathcal{I}_{\gamma,U_{(1)}Z}$, $\mathcal{I}_{\gamma,ZZ}$, $Q_{\gamma,1}$, and $Q_{\gamma,2}$. Following Fan and Gijbels (1996), we suggest approximating them using global polynomial regression models. In the following we list our rule-of-thumb approximations for the bandwidths in Eq.s (7)-(14):

Sparse - rule-of-thumb bandwidths for $\hat{\mu}$:

$$\hat{h}_{\mu,U}^S = \left(\frac{R(K_\mu)\, Q_{\mu,1}\, \hat{\mathcal{I}}^{3/4}_{\mu_{\text{poly}},ZZ}}{nm\ (\nu_2(K_\mu))^2\ \left[\hat{\mathcal{I}}^{1/2}_{\mu_{\text{poly}},UU}\, \hat{\mathcal{I}}^{1/2}_{\mu,ZZ} + \hat{\mathcal{I}}_{\mu_{\text{poly}},UZ}\right] \hat{\mathcal{I}}^{3/4}_{\mu_{\text{poly}},UU}} \right)^{1/6} \tag{16}$$

$$\hat{h}_{\mu,Z}^S = \left(\frac{\hat{\mathcal{I}}_{\mu_{\text{poly}},UU}}{\hat{\mathcal{I}}_{\mu_{\text{poly}},ZZ}} \right)^{1/4} \hat{h}_{\mu,U}^S \tag{17}$$

Sparse - rule-of-thumb bandwidths for $\hat{\gamma}$:

$$\hat{h}_{\gamma,U}^S = \left(\frac{R(K_\gamma)\ \hat{Q}_{\gamma_{\text{poly}},1}\ 4\,\sqrt{2}\,\hat{\mathcal{I}}^{3/2}_{\gamma_{\text{poly}},ZZ}}{nM\ (\nu_2(K_\gamma))^2\ \left(2\ (\nu_2(K_\gamma))^2\ \hat{\mathcal{I}}_{\gamma_{\text{poly}},U_{(1)}Z} + \hat{C}_{\mathcal{I}}\right)\ \left(\hat{C}_{\mathcal{I}} - \hat{\mathcal{I}}_{\gamma_{\text{poly}},U_{(1)}Z}\right)^{3/2}} \right)^{1/7} \tag{18}$$

$$\hat{h}_{\gamma,Z}^S = \left(\frac{\hat{C}_{\mathcal{I}} - \hat{\mathcal{I}}_{\gamma_{\text{poly}},U_{(1)}Z}}{2\,\hat{\mathcal{I}}_{\gamma,ZZ}} \right)^{1/2} \hat{h}_{\gamma,U}^S,\ \text{where} \tag{19}$$

$$\hat{C}_{\mathcal{I}} = (\hat{\mathcal{I}}^2_{\gamma_{\text{poly}},U_{(1)}Z} + 4\,(\hat{\mathcal{I}}_{\gamma_{\text{poly}},U_{(1)}U_{(1)}} + \hat{\mathcal{I}}_{\gamma_{\text{poly}},U_{(1)}U_{(2)}})\,\hat{\mathcal{I}}_{\gamma_{\text{poly}},ZZ})^{1/2}.$$

Dense - rule-of-thumb bandwidths for $\hat{\mu}$:

$$\hat{h}_{\mu,Z}^D = \left(\frac{R(\kappa)\,\hat{Q}_{\mu_{\text{poly}},2}}{n\ (\nu_2(K_\mu))^2\ \hat{\mathcal{I}}_{\mu_{\text{poly}},ZZ}} \right)^{1/5} \tag{20}$$

$$\hat{h}_{\mu,U}^D = \left(\frac{R(K_\mu)\,\hat{Q}_{\mu_{\text{poly}},1}}{nm\ (\nu_2(K_\mu))^2\ \hat{\mathcal{I}}_{\mu_{\text{poly}},UZ}} \right)^{1/3} \left(\hat{h}_{\mu,Z}^D\right)^{-1} \tag{21}$$

Dense - rule-of-thumb bandwidths for $\hat{\gamma}$:

$$\hat{h}^D_{\gamma,Z} = \left(\frac{R(\kappa)\,\hat{Q}_{\gamma_{\text{poly}},2}}{n\ (\nu_2(K_\gamma))^2\ \hat{\mathcal{I}}_{\gamma_{\text{poly}},ZZ}} \right)^{1/5} \tag{22}$$

$$\hat{h}^D_{\gamma,U} = \left(\frac{R(K_\gamma)\,\hat{Q}_{\gamma_{\text{poly}},1}}{nM\ (\nu_2(K_\gamma))^2\ \hat{\mathcal{I}}_{\gamma_{\text{poly}},U_{(1)}Z}} \right)^{1/4} \left(\hat{h}^D_{\gamma_{\text{poly}},Z}\right)^{-3/4} \tag{23}$$

The above rule-of-thumb bandwidth expressions are based on the following estimates for the "sparse" rule-of-thumb bandwidths:

$$\begin{aligned}
\hat{\mathcal{I}}_{\mu_{\text{poly}},UU} &= \int_{\text{supp}(f_{UZ})} (\hat{\mu}^{(2,0)}_{\text{poly}}(u,z))^2\,\hat{f}_{UZ}(u,z)\,d(u,z),\\
\hat{\mathcal{I}}_{\mu_{\text{poly}},UZ} &= \int_{\text{supp}(f_{UZ})} \hat{\mu}^{(2,0)}_{\text{poly}}(u,z)\hat{\mu}^{(0,2)}_{\text{poly}}(u,z)\,\hat{f}_{UZ}(u,z)\,d(u,z),\\
\hat{\mathcal{I}}_{\mu_{\text{poly}},ZZ} &= \int_{\text{supp}(f_{UZ})} (\hat{\mu}^{(0,2)}_{\text{poly}}(u,z))^2\,\hat{f}_{UZ}(u,z)\,d(u,z),\\
\hat{Q}_{\mu_{\text{poly}},1} &= \int_{\text{supp}(f_{UZ})} \hat{\gamma}^{\text{ND}}_{\text{poly}}(u,u,z)\,d(u,z),\\
\hat{Q}_{\mu_{\text{poly}},2} &= \int_{\text{supp}(f_{UZ})} \hat{\gamma}_{\text{poly}}(u,u,z)\,\hat{f}_U(u)\,d(u,z),
\end{aligned}$$

and for the "dense" rule-of-thumb bandwidths:

$$\begin{aligned}
\hat{\mathcal{I}}_{\gamma_{\text{poly}},U_{(1)}U_{(1)}} &= \int_{\text{supp}(f_{UZ})} (\hat{\gamma}^{(2,0,0)}_{\text{poly}}(u,u,z))^2\,\hat{f}_{UZ}(u,z)\,d(u,z),\\
\hat{\mathcal{I}}_{\gamma_{\text{poly}},U_{(1)}Z} &= \int_{\text{supp}(f_{UUZ})} \hat{\gamma}^{(2,0,0)}_{\text{poly}}(u,u,z)\hat{\gamma}^{(0,0,2)}_{\text{poly}}(u,u,z)\,\hat{f}_{UUZ}(u,u,z)\,d(u,u,z),\\
\hat{\mathcal{I}}_{\gamma_{\text{poly}},ZZ} &= \int_{\text{supp}(f_{UUZ})} (\hat{\gamma}^{(0,0,2)}_{\text{poly}}(u,u,z))^2\,\hat{f}_{UUZ}(u,u,z)\,d(u,u,z),\\
\hat{Q}_{\gamma_{\text{poly}},1} &= \int_{\text{supp}(f_{UUZ})} \hat{\tilde{\gamma}}^{\text{ND}}_{\text{poly}}((u_1,u_2),(u_1,u_2),z)\,d(u_1,u_2,z), \quad \text{and}\\
\hat{Q}_{\gamma_{\text{poly}},2} &= \int_{\text{supp}(f_{UUZ})} \hat{\tilde{\gamma}}_{\text{poly}}((u_1,u_2),(u_1,u_2),z)\ \hat{f}_{UU}(u_1,u_2)\,d(u_1,u_2,z).
\end{aligned}$$

The estimates $\hat{\mu}_{\text{poly}}$, $\hat{\gamma}^{\text{ND}}_{\text{poly}}$, $\hat{\gamma}_{\text{poly}}$, $\hat{\tilde{\gamma}}^{\text{ND}}_{\text{poly}}$, $\hat{\tilde{\gamma}}_{\text{poly}}$, $\hat{\mu}^{(2,0)}_{\text{poly}}$, $\hat{\mu}^{(0,2)}_{\text{poly}}$, $\hat{\gamma}^{(2,0,0)}_{\text{poly}}$, and $\hat{\gamma}^{(0,0,2)}_{\text{poly}}$ are the ordinary least squares estimates (and their derivatives) of the following polynomial regression models:

$\boldsymbol{\mu}_{\textbf{poly}}$**:** The model $\mu_{\text{poly}}(u,z)$ is fitted via regressing Y_{ij} on powers (each up to the fourth power) of U_{ij}, Z_i, and $U_{ij}\cdot Z_i$ for all $i\in\{1,\dots,n\}$ and $j\in\{1,\dots,m\}$, i.e., $Y_{ij} = \mu_{\text{poly}}(U_{ij},Z_i)+\text{error}_{ij}$, where $\mu_{\text{poly}}(U_{ij},Z_i) = \beta_0+\sum_{q=1}^4\left(\beta^U_q U^q_{ij}+\beta^Z_q Z^q_{ij}+\beta^{UZ}_q(U_{ij}Z_i)^q\right)$.

$\boldsymbol{\gamma}_{\textbf{poly}}$**:** The model $\gamma_{\text{poly}}(u_1,u_2,z)$ is fitted via regressing C^{poly}_{ijk} on powers (each up to the fourth power) of U_{ij}, U_{ik}, Z_i, $U_{ij}\cdot Z_i$, and $U_{ik}\cdot Z_i$ for all $i\in\{1,\dots,n\}$ and all $j,k\in\{1,\dots,m\}$ with $j\neq k$, i.e., $C^{\text{poly}}_{ijk} = \gamma_{\text{poly}}(U_{ij},U_{ik},Z_i)+\text{error}_{ijk}$, where $C^{\text{poly}}_{ijk} = (Y_{ij}-\mu_{\text{poly}}(U_{ij},Z_i))(Y_{jt}-\mu_{\text{poly}}(U_{ik},Z_i))$ and
$\gamma_{\text{poly}}(U_{ij},U_{ik},Z_i) = \beta_0+\sum_{q=1}^4\left(\beta^{U,1}_q U^q_{ij}+\beta^{U,2}_q U^q_{ik}+\beta^Z_q Z^q_i+\beta^{UZ,1}_q(U_{ij}Z_i)^q+\beta^{UZ,2}_q(U_{ik}Z_i)^q\right)$.

$\tilde{\boldsymbol{\gamma}}_{\textbf{poly}}$**:** The model $\tilde{\gamma}_{\text{poly}}((u_1,u_2),(u_3,u_4),z)$ is fitted via regressing $\mathbb{C}^{\text{poly}}_{ijk\ell m}$ on powers (each up to the fourth power) of U_{ij}, U_{ik}, $U_{i\ell}$, U_{im}, and Z_i for all $i\in\{1,\dots,n\}$ and all $j,k,\ell,m\in$

$\{1,\ldots,m\}$ such that ($j \neq \ell$ AND $k \neq m$), i.e., $\mathbb{C}^{\text{poly}}_{ijk\ell m} = \tilde{\gamma}_{\text{poly}}((U_{ij}, U_{ik}), (U_{i\ell}, U_{im}), Z_i) +$ error$_{ijk\ell m}$, where $\mathbb{C}^{\text{poly}}_{ijk\ell m} = (C^{\text{poly}}_{ijk} - \gamma_{\text{poly}}(U_{ij}, U_{ik}, Z_i))(C^{\text{poly}}_{i\ell m} - \gamma_{\text{poly}}(U_{i\ell}, U_{im}, Z_i))$ and $\tilde{\gamma}_{\text{poly}}((U_{ij}, U_{ik}), (U_{i\ell}, U_{im}), Z_i) = \beta_0 + \sum_{q=1}^{4} \big(\beta^{U,1}_q U^q_{ij} + \beta^{U,2}_q U^q_{ik} + \beta^{U,3}_q U^q_{i\ell} + \beta^{U,4}_q U^q_{im} + \beta^Z_q Z^q_i +$ $+\beta^{UZ,1}_q (U_{ij}Z_i)^q + \beta^{UZ,2}_q (U_{ik}Z_i)^q + \beta^{UZ,3}_q (U_{i\ell}Z_i)^q + \beta^{UZ,4}_q (U_{im}Z_i)^q\big)$.

$\boldsymbol{\gamma}^{\mathbf{S}}_{\mathbf{poly}}$**:** The model $\gamma^{\text{ND}}_{\text{poly}}(u,u,z)$ is fitted via regressing the noise-contaminated diagonal values C^{poly}_{ij} on powers (each up to the fourth power) of U_{ij}, and Z_i for all $i \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,m\}$, i.e., $C^{\text{poly}}_{ij} = \gamma^{\text{ND}}_{\text{poly}}(U_{ij}, U_{ij}, Z_i) + \text{error}_{ij}$, where $C^{\text{poly}}_{ij} = (Y_{ij} - \mu_{\text{poly}}(U_{ij}, Z_i))^2$ and $\gamma^{\text{ND}}_{\text{poly}}(U_{ij}, U_{ij}, Z_i) = \beta_0 + \sum_{q=1}^{4} \big(\beta^U_q U^q_{ij} + \beta^Z_q Z^q_i + \beta^{UZ}_q (U_{ij}Z_i)^q\big)$. The "ND" in $\gamma^{\text{ND}}_{\text{poly}}$ suggest that we are estimating the noise-contaminated diagonal values $\gamma(u,u,z) + \sigma_\epsilon$.

$\tilde{\boldsymbol{\gamma}}^{\mathbf{S}}_{\mathbf{poly}}$**:** The model $\tilde{\gamma}^{\text{ND}}_{\text{poly}}(u_1, u_2, z)$ is fitted via regressing the noise-contaminated diagonal values $\mathbb{C}^{\text{poly}}_{ijk}$ on powers (each up to the fourth power) of U_{ij}, U_{ik}, and Z_i for all $i \in \{1,\ldots,n\}$, and $j,k \in \{1,\ldots,m\}$, i.e., $\mathbb{C}^{\text{poly}}_{ijk} = \tilde{\gamma}^{\text{ND}}_{\text{poly}}(U_{ij}, U_{ik}, Z_i) + \text{error}_{ijk}$, where $\mathbb{C}^{\text{poly}}_{ijk} = (C^{\text{poly}}_{ijk} - \gamma_{\text{poly}}(U_{ij}, U_{ik}, Z_i))^2$ and $\tilde{\gamma}^{\text{ND}}_{\text{poly}}(U_{ij}, U_{ik}, Z_i) = \beta_0 + \sum_{q=1}^{4} \big(\beta^{U,1}_q U^q_{ij} + \beta^{U,2}_q U^q_{ik} + \beta^Z_q Z^q_i\big)$. The "ND" in $\tilde{\gamma}^{\text{ND}}_{\text{poly}}$ suggest that we are estimating the noise-contaminated diagonal values $\tilde{\gamma}((u_1,u_2),(u_1,u_2),z) + \sigma_\varepsilon(u_1,u_2,z)$.

Estimates of the densities f_{UZ} and f_{UUZ} are computed as kernel density estimates using Gaussian kernels and bandwidth determined by cross-validation.

Remark It is important to specify the models $\boldsymbol{\mu}_{\text{poly}}$ and $\boldsymbol{\gamma}_{\text{poly}}$ using interaction terms, since otherwise their partial derivatives $\hat{\mu}^{(2,0)}_{\text{poly}}$, $\hat{\mu}^{(0,2)}_{\text{poly}}$, $\hat{\gamma}^{(2,0,0)}_{\text{poly}}$, and $\hat{\gamma}^{(0,0,2)}_{\text{poly}}$ would degenerate.

5.2 Bias and variance estimates

Following Härdle and Bowman (1988), we approximate the unknown second derivatives $\mu^{(2,0)}$ and $\mu^{(0,2)}$ in B_μ and B^D_μ using local polynomial estimators. That is, we approximate $B_\mu(u,z;h_{\mu,U},h_{\mu,Z})$ and $B^D_\mu(u,z;h_{\mu,Z})$ by

$$\begin{aligned}
\hat{B}_\mu(u,z;h_{\mu,U},h_{\mu,Z}) &= \tfrac{\nu_2(K_\mu)}{2}\Big(h^2_{\mu,U}\hat{\mu}^{(2,0)}\big(u,z;g_{\mu,U},g_{\mu,Z}\big) + h^2_{\mu,Z}\hat{\mu}^{(0,2)}\big(u,z;g_{\mu,U},g_{\mu,Z}\big)\Big)\\
\text{and}\quad \hat{B}^D_\mu(u,z;h_{\mu,Z}) &= \tfrac{\nu_2(K_\mu)}{2} h^2_{\mu,Z}\hat{\mu}^{(0,2)}\big(u,z;g_{\mu,U},g_{\mu,Z}\big)
\end{aligned}$$

where $\hat{\mu}^{(2,0)}$ and $\hat{\mu}^{(0,2)}$ are local polynomial (order 3) kernel estimators of $\mu^{(2,0)}$ and $\mu^{(0,2)}$:

$$\begin{aligned}
\hat{\mu}^{(2,0)}(u,z;g_{\mu,U},g_{\mu,Z}) &= 2!\, e_3^\top \left([\mathbf{1},\mathbf{U}^{1:3}_u,\mathbf{Z}^{1:3}_z]^\top \mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}^{1:3}_u,\mathbf{Z}^{1:3}_z]\right)^{-1}[\mathbf{1},\mathbf{U}^{1:3}_u,\mathbf{Z}^{1:3}_z]^\top \mathbf{W}_{\mu,uz}\mathbf{Y}\\
\hat{\mu}^{(0,2)}(u,z;g_{\mu,U},g_{\mu,Z}) &= 2!\, e_6^\top \left([\mathbf{1},\mathbf{U}^{1:3}_u,\mathbf{Z}^{1:3}_z]^\top \mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}^{1:3}_u,\mathbf{Z}^{1:3}_z]\right)^{-1}[\mathbf{1},\mathbf{U}^{1:3}_u,\mathbf{Z}^{1:3}_z]^\top \mathbf{W}_{\mu,uz}\mathbf{Y},
\end{aligned}$$

with $e_3^\top = (0,0,1,0,0,0,0)$, $e_6^\top = (0,0,0,0,0,1,0)$, $\mathbf{U}^{1:3}_u = [\mathbf{U}_u, \mathbf{U}^2_u, \mathbf{U}^3_u]$, $\mathbf{Z}^{1:3}_z = [\mathbf{Z}_z, \mathbf{Z}^2_z, \mathbf{Z}^3_z]$, and $\mathbf{W}_{\mu,uz} = \text{diag}(\ldots, g^{-1}_{\mu,U}\,\kappa(g^{-1}_{\mu,U}(U_{ij}-u))\; g^{-1}_{\mu,Z}\,\kappa(g^{-1}_{\mu,Z}(Z_i - z)),\ldots)$.

For estimating the bandwidths $g_{\mu,U}$ and $g_{\mu,Z}$ we use bivariate GCV based on second-order differences. We follow the procedure of Charnigo and Srinivasan (2015), but use a GCV-penalty instead of their proposed (asymptotically equivalent) C_p-penalty.

For approximating the unknown noise-contaminated diagonal of the covariance function, $\gamma(u,u,z)+\sigma^2_\epsilon$, contained in V^I_μ, we propose to use a LLK estimator. That is, we approximate V^I_μ by

$$\hat{V}^I_\mu(u,z;h_{\mu,U},h_{\mu,Z},h_{\gamma,U},h_{\gamma,Z}) = (nm)^{-1}\left[(h_{\mu,U}h_{\mu,Z})^{-1}R(K_\mu)\frac{\hat{\gamma}^{\text{ND}}(u,u,z;h_{\gamma,U},h_{\gamma,Z})}{\hat{f}_{UZ}(u,z)}\right]$$

where the estimator of the Noisy Diagonal (ND), $\hat{\gamma}^{\text{ND}}(u,u,z;h_{\gamma,U},h_{\gamma,Z}) \approx \{\gamma(u,u,z)+\sigma^2_\epsilon\}$, is defined as the following LLK estimator:

$$\hat{\gamma}^{\text{ND}}(u,u,z;h_{\gamma,U},h_{\gamma,Z}) = e_1^\top\left([\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top\mathbf{W}_{\gamma,uz}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]\right)^{-1}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top\mathbf{W}_{\gamma,uz}\hat{\mathbf{C}},$$

where $\hat{\mathbf{C}} = (\hat{C}_{111},\dots,C_{ijj}\dots,\hat{C}_{nmm})^\top$ consists only of the "diagonal" raw-covariances, i.e., $\hat{C}_{ijj} = (Y_{ij}-\hat{\mu}(U_{ij},Z_i;h_{\mu,U},h_{\mu,Z}))^2$. Note that $\hat{\gamma}^{\text{ND}}$ is equivalent to the LLK estimator "$\hat{V}$" in Jiang and Wang (2010).

Finally, for estimating the unknown quantity $\gamma(u,u,z)$ in V^{II}_μ we can use our LLK estimator $\hat{\gamma}(u,u,z;h_{\gamma,U},h_{\gamma,Z})$ as defined in (6). That is we estimate V^{II}_μ by

$$\hat{V}^{II}_\mu(u,z;h_{\mu,Z},h_{\gamma,U},h_{\gamma,Z}) = n^{-1}\left[\left(\frac{m-1}{m}\right)(h_{\mu,Z})^{-1}R(\kappa)\frac{\hat{\gamma}(u,u,z;h_{\gamma,U},h_{\gamma,Z})}{\hat{f}_Z(z)}\right].$$

The bandwidths for the LLK estimators $\hat{\gamma}$ and $\hat{\gamma}^{\text{ND}}$ are selected according to our Rule-of-thumb bandwidth approximations in Eq.s (18), (19), (22), and (23).

A Proofs

A.1 Proof of Theorem 3.1

Proof of Theorem 3.1, part *(i)*: For simplicity, consider a second-order kernel function κ with compact support such as the Epanechnikov kernel. This is, of course, without loss of generality, but allows for a more compact proof. Define $\mathbf{H}_\mu = \text{diag}(h^2_{\mu,U},h^2_{\mu,Z})$, $\mathbf{U} = (U_{11},\dots,U_{nm})^\top$, and $\mathbf{Z} = (Z_1,\dots,Z_n)^\top$. Using a Taylor-expansion of μ around (u,z), the conditional bias of the estimator $\hat{\mu}(u,z;\mathbf{H})$ can be written as

$$\begin{aligned}\mathbb{E}(\hat{\mu}(u,z;\mathbf{H}_\mu)-\mu(u,z)|\mathbf{U},\mathbf{Z}) = {} & \frac{1}{2}e_1^\top\left((nm)^{-1}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]\right)^{-1}\times \\ & \times(nm)^{-1}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,uz}\left(\boldsymbol{\mathcal{Q}}_\mu(u,z)+\mathbf{R}_\mu(u,z)\right),\end{aligned} \tag{24}$$

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

$$(U_{ij}-u, Z_i-z)\boldsymbol{\mathcal{H}}_\mu(u,z)(U_{ij}-u, Z_i-z)^\top \in \mathbb{R}$$

with $\boldsymbol{\mathcal{H}}_\mu(u,z)$ being the Hessian matrix of the regression function $\mu(u,z)$. The $nm \times 1$ vector $\mathbf{R}_\mu(u,z)$ holds the remainder terms as in Ruppert and Wand (1994).

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

$$(nm)^{-1}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top \mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z] =$$

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

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

$$\left((nm)^{-1}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top \mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]\right)^{-1} = \tag{25}$$

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

The 3×1 matrix $(nm)^{-1}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top \mathbf{W}_{\mu,uz}\boldsymbol{\mathcal{Q}}_\mu(u,z)$ can be partitioned as following:

$$(nm)^{-1}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top \mathbf{W}_{\mu,uz}\boldsymbol{\mathcal{Q}}_\mu(u,z) = \begin{pmatrix} \texttt{upper element} \\ \texttt{lower bloc} \end{pmatrix},$$

where the 1×1 dimensional `upper element` can be approximated by

$$(nm)^{-1}\sum_{it} K_{\mu,h}(U_{ij}-u,Z_i-z)(U_{ij}-u,Z_i-z)\boldsymbol{\mathcal{H}}_\mu(u,z)(U_{ij}-u,Z_i-z)^\top \tag{26}$$
$$=(\nu_2(\kappa))^2\,tr\left\{\mathbf{H}_\mu\boldsymbol{\mathcal{H}}_\mu(u,z)\right\}f_{UZ}(u,z)+o_p(tr(\mathbf{H}_\mu))$$

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

$$(nm)^{-1}\sum_{it}\left\{K_{\mu,h}(U_{ij}-u,Z_i-z)(U_{ij}-u,Z_i-z)\boldsymbol{\mathcal{H}}_\mu(u,z)(U_{ij}-u,Z_i-z)^\top\right\}\times \tag{27}$$
$$\times\,(U_{ij}-u,Z_i-z)^\top = O_p(\mathbf{H}_\mu^{3/2}\mathbf{1}).$$

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

$$\frac{1}{2}e_1^\top((nm)^{-1}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top \mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z])^{-1}(nm)^{-1}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top \mathbf{W}_{\mu,uz}\boldsymbol{\mathcal{Q}}_\mu(u,z)=$$
$$=\frac{1}{2}\,(\nu_2(\kappa))^2\,tr\left\{\mathbf{H}_\mu\boldsymbol{\mathcal{H}}_\mu(u,z)\right\}+o_p(tr(\mathbf{H}_\mu)).$$

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

$$\frac{1}{2}e_1^\top((nm)^{-1}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z])^{-1}(nm)^{-1}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,uz}\mathbf{R}_\mu(u,z)=o_p(tr(\mathbf{H}_\mu)).$$

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

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

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 $\mathbb{V}(\hat{\mu}(u,z;\mathbf{H}_\mu)|\mathbf{U},\mathbf{Z})=$

$$\begin{aligned}
=&e_1^\top([\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z])^{-1}\,[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,uz}\,\mathrm{Cov}(\mathbf{Y}|\mathbf{U},\mathbf{Z})\,\mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]\times\\
&\times([\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z])^{-1}u_1\\
=&e_1^\top((nm)^{-1}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z])^{-1}\times\\
&\times((nm)^{-2}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,uz}\,\mathrm{Cov}(\mathbf{Y}|\mathbf{U},\mathbf{Z})\,\mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z])\times\\
&\times((nm)^{-1}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z])^{-1}u_1,
\end{aligned}\tag{28}$$

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

$$\mathrm{Cov}(Y_{ij},Y_{\ell k}|U_{ij},U_{\ell k},Z_i,Z_\ell)=\gamma_{|i-\ell|}((U_{ij},Z_i),(U_{\ell k},Z_\ell))+\sigma_\epsilon^2\mathbb{1}(i=\ell\text{ and }j=k)$$

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

We begin with analyzing the 3×3 matrix

$$(nm)^{-2}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,uz}\,\mathrm{Cov}(\mathbf{Y}|\mathbf{U},\mathbf{Z})\,\mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]$$

using the following three Lemmas A.1-A.3.

Lemma A.1 *The upper-left scalar (block) of the matrix* $(nm)^{-2}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]^\top\mathbf{W}_{\mu,uz}\mathrm{Cov}(\mathbf{Y}|\mathbf{U},\mathbf{Z})\mathbf{W}_{\mu,uz}[\mathbf{1},\mathbf{U}_u,\mathbf{Z}_z]$ *is given by*

$$\begin{aligned}
&(nm)^{-2}\mathbf{1}^\top\mathbf{W}_{\mu,uz}\mathrm{Cov}(\mathbf{Y}|\mathbf{U},\mathbf{Z})\mathbf{W}_{\mu,uz}\mathbf{1}\\
=\;&(nm)^{-1}f_{UZ}(u,z)|\mathbf{H}_\mu|^{-1/2}R(K_\mu)\left(\gamma(u,u,z)+\sigma_\epsilon^2\right)(1+O_p(tr(\mathbf{H}_\mu^{1/2})))\\
+\;&n^{-1}(f_{UZ}(u,z))^2\left[\left(\frac{m-1}{m}\right)h_{\mu,Z}^{-1}R(\kappa)\frac{\gamma(u,u,z)}{f_Z(z)}\right](1+O_p(tr(H^{1/2})))\\
=\;&O_p((nm)^{-1}|\mathbf{H}_\mu|^{-1/2})+O_p(n^{-1}h_{\mu,Z}^{-1}).
\end{aligned}$$

Lemma A.2 *The* 1×2 *dimensional upper-right block of the matrix* $(nm)^{-2}[\mathbf{1}, \mathbf{U}_u, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,uz}\mathrm{Cov}(\mathbf{Y}|\mathbf{U},\mathbf{Z})\mathbf{W}_{\mu,uz}[\mathbf{1}, \mathbf{U}_u, \mathbf{Z}_z]$ *is given by*

$$\begin{aligned}
& (nm)^{-2}\mathbf{1}^\top \mathbf{W}_{\mu,uz}\mathrm{Cov}(\mathbf{Y}|\mathbf{U},\mathbf{Z})\mathbf{W}_{\mu,uz}\begin{pmatrix}(U_{11}-u, Z_1-z)\\ \vdots \\ (U_{nm}-u, Z_n-z)\end{pmatrix}\\
= \;& (nm)^{-1} f_{UZ}(u,z)|\mathbf{H}_\mu|^{-1/2}(\mathbf{1}^\top\mathbf{H}_\mu^{1/2})R(K_\mu)\left(\gamma(u,u,z)+\sigma_\epsilon^2\right)(1+O_p(tr(\mathbf{H}_\mu^{1/2})))\\
+ \;& n^{-1}(f_{UZ}(u,z))^2(\mathbf{1}^\top\mathbf{H}_\mu^{1/2})\left[\left(\frac{m-1}{m}\right)h_{\mu,Z}^{-1}R(\kappa)\frac{\gamma(u,u,z)}{f_Z(z)}\right](1+O_p(tr(\mathbf{H}_\mu^{1/2})))\\
= \;& O_p((nm)^{-1}|\mathbf{H}_\mu|^{-1/2}(\mathbf{1}^\top\mathbf{H}_\mu^{1/2})) + O_p(n^{-1}(\mathbf{1}^\top\mathbf{H}_\mu^{1/2})h_{\mu,Z}^{-1}).
\end{aligned}$$

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

Lemma A.3 *The* 2×2 *lower-right block of the matrix* $(nm)^{-2}[\mathbf{1}, \mathbf{U}_u, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,uz}\mathrm{Cov}(\mathbf{Y}|\mathbf{U},\mathbf{Z})\mathbf{W}_{\mu,uz}[\mathbf{1}, \mathbf{U}_u, \mathbf{Z}_z]$ *is given by*

$$\begin{aligned}
& (nm)^{-2}\left(((U_{11}-u),(Z_1-z))^\top,\ldots,((U_{nm}-u),(Z_n-z)^\top)\right)\times\\
& \times\mathbf{W}_{\mu,uz}\mathrm{Cov}(\mathbf{Y}|\mathbf{U},\mathbf{Z})\mathbf{W}_{\mu,uz}\begin{pmatrix}(U_{11}-u, Z_1-z)\\ \vdots \\ (U_{nm}-u, Z_n-z)\end{pmatrix}\\
= \;& (nm)^{-1} f_{UZ}(u,z)|\mathbf{H}_\mu|^{-1/2}\mathbf{H}_\mu R(K_\mu)\left(\gamma(u,u,z)+\sigma_\epsilon^2\right)(1+O_p(tr(\mathbf{H}_\mu^{1/2})))\\
+ \;& n^{-1}(f_{UZ}(u,z))^2\mathbf{H}_\mu\left[\left(\frac{m-1}{m}\right)h_{\mu,Z}^{-1}R(\kappa)\frac{\gamma(u,u,z)}{f_Z(z)}\right](1+O_p(tr(\mathbf{H}_\mu^{1/2})))\\
= \;& O_p((nm)^{-1}|\mathbf{H}_\mu|^{-1/2}\mathbf{H}_\mu) + O_p(n^{-1}\mathbf{H}_\mu h_{\mu,Z}^{-1}).
\end{aligned}$$

Using the approximations for the bloc-elements of the matrix $(nm)^{-2}[\mathbf{1}, \mathbf{U}_u, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,uz}\mathrm{Cov}(\mathbf{Y}|\mathbf{U},\mathbf{Z})\mathbf{W}_{\mu,uz}[\mathbf{1}, \mathbf{U}_u, \mathbf{Z}_z]$, given by the Lemmas A.1-A.3, and the approximation for the matrix $\left((nm)^{-1}[\mathbf{1}, \mathbf{U}_u, \mathbf{Z}_z]^\top \mathbf{W}_{\mu,uz}[\mathbf{1}, \mathbf{U}_u, \mathbf{Z}_z]\right)^{-1}$, 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 $\mathbb{V}(\hat{\mu}(u,z;\mathbf{H}_\mu)|\mathbf{U},\mathbf{Z}) =$

$$\begin{aligned}
& (nm)^{-1}|\mathbf{H}_\mu|^{-1/2}\left\{\frac{R(K_\mu)\left(\gamma(u,u,z)+\sigma_\epsilon^2\right)}{f_{UZ}(u,z)}\right\}(1+o_p(1))\\
& + n^{-1}\left[\left(\frac{m-1}{m}\right)h_{\mu,Z}^{-1}R(\kappa)\frac{\gamma(u,u,z)}{f_Z(z)}\right](1+o_p(1)),
\end{aligned}$$

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

Next we prove Lemma A.1; the proofs of Lemmas A.2 and A.3 are equivalent. To show Lemma A.1 it will be convenient to split the sum such that $(nm)^{-2}\mathbf{1}^\top\mathbf{W}_{\mu,uz}\mathrm{Cov}(\mathbf{Y}|\mathbf{U},\mathbf{Z})\mathbf{W}_{\mu,uz}\mathbf{1} = s_1 + s_2$. Using standard procedures from kernel density estimation leads to

$$\begin{aligned} s_1 &= (nm)^{-2}\sum_{ij}(K_{\mu,h}(U_{ij}-u, Z_i-z))^2\,\mathbb{V}(Y_{ij}|\mathbf{U},\mathbf{Z}) \qquad (29)\\ &= (nm)^{-1}|\mathbf{H}_\mu|^{-1/2} f_{UZ}(u,z)R(K_\mu)\left(\gamma(u,u,z)+\sigma_\epsilon^2\right) + O((nm)^{-1}|\mathbf{H}_\mu|^{-1/2}\,tr(\mathbf{H}_\mu^{1/2})),\\ s_2 &= (nm)^{-2}\sum_{\substack{jk\\ j\neq k}}\sum_i h_{\mu,U}^{-1}\kappa(h_{\mu,U}^{-1}(U_{ij}-u))(h_{\mu,Z}^{-1}\kappa(h_{\mu,Z}^{-1}(Z_i-z)))^2\,\mathrm{Cov}(Y_{ij},Y_{ik}|\mathbf{U},\mathbf{Z})\times \qquad (30)\\ &\times h_{\mu,U}^{-1}\kappa(h_{\mu,U}^{-1}(U_{ik}-x))\\ &= n^{-1}(f_{UZ}(u,z))^2\left[\left(\frac{m-1}{m}\right)h_{\mu,Z}^{-1}R(\kappa)\frac{\gamma(u,u,z)}{f_Z(z)}\right] + O_p(n^{-1}tr(\mathbf{H}_\mu^{1/2})), \end{aligned}$$

Summing up (29)-(30) leads to the result in Lemma A.1. Lemmas A.2 and A.3 differ from Lemma A.1 only with respect to the additional factors $\mathbf{1}^\top\mathbf{H}_\mu^{1/2}$ and $\mathbf{H}_\mu$ which occur due to the usual substitution step for the additional data parts $(U_{ij}-u, Z_i-z)$.

A.2 Proof of Theorem 3.2

The proof of Theorem 3.2 follows exactly the same arguments as in the proof of Theorem 3.1 and therefore is omitted.

A.3 Proofs of the results in Section 3.1 (Sparse functional data)

A.3.1 Proof of Theorem 3.3

The AMISE function (i.e., the AMISE function with leading V_μ^I variance term) for the local linear estimator $\hat{\mu}$ is given by

$$\begin{aligned} &\mathrm{AMISE}_{\hat{\mu}}\,(h_{\mu,U},h_{\mu,Z}) = (nm)^{-1}\,h_{\mu,U}^{-1}\,h_{\mu,Z}^{-1}\,R(K_\mu)\,Q_{\mu,1}+ \qquad (31)\\ &+\frac{1}{4}\,(\nu_2(K_\mu))^2\,\left[h_{\mu,U}^4\,\mathcal{I}_{\mu,UU}+2\,h_{\mu,U}^2\,h_{\mu,Z}^2\,\mathcal{I}_{\mu,UZ}+h_{\mu,Z}^4\,\mathcal{I}_{\mu,ZZ}\right], \end{aligned}$$

$$\begin{aligned} \text{where}\qquad Q_{\mu,1} &= \int\left(\gamma(u,u,z)+\sigma_\epsilon^2\right)\,d(u,z),\\ \mathcal{I}_{\mu,UU} &= \int\left(\mu^{(2,0)}(u,z)\right)^2\,f_{UZ}(u,z)\,d(u,z),\\ \mathcal{I}_{\mu,ZZ} &= \int\left(\mu^{(0,2)}(u,z)\right)^2\,f_{UZ}(u,z)\,d(u,z),\quad\text{and}\\ \mathcal{I}_{\mu,UZ} &= \int\mu^{(2,0)}(u,z)\mu^{(0,2)}(u,z)\,f_{UZ}(u,z)\,d(u,z). \end{aligned}$$

This is a known expression for the AMISE function of a two-dimensional local linear estimator with a diagonal bandwidth matrix (see, e.g., Herrmann et al., 1995) and follows from the formulas in Wand and Jones (1994). Minimizing the above AMISE function with respect to $h_{\mu,U}$ and $h_{\mu,Z}$ leads to the optimal bandwidth expressions in Theorem 3.3 which correspond to the results in Herrmann et al. (1995).

It follows directly from Theorem 3.1 that the first variance summand V_μ^I is the leading variance term if the following order relation holds:

$$n^{-1}\,h_{\mu,Z}^{-1} = o\left(n^{-(1+\theta)}\,h_{\mu,U}^{-1}\,h_{\mu,Z}^{-1}\right), \tag{32}$$

where we used that by Assumption A-AS $nm \asymp n^{1+\theta}$. Plugging the AMISE optimal bandwidth rates of Theorem 3.3 into the order relation of Eq. (32) leads to the corresponding θ values of $0 \leq \theta < 1/5$ which describe the case we consider here as "sparse" functional data.

A.3.2 Proof of Theorem 3.4

The corresponding AMISE function (i.e., the AMISE function with leading V_γ^I variance term) for the local linear estimator $\hat{\gamma}$ is given by

$$\begin{aligned} \mathrm{AMISE}_{\hat{\gamma}}\left(h_{\gamma,U},h_{\gamma,Z}\right) &= (nM)^{-1}\,h_{\gamma,U}^{-2}\,h_{\gamma,Z}^{-1}\,R(K_\gamma)\,Q_{\gamma,1}+ \\ &+\frac{1}{4}\left(\nu_2(K_\gamma)\right)^2\left[2\,h_{\gamma,U}^4\left(\mathcal{I}_{\gamma,U_{(1)}U_{(2)}}+\mathcal{I}_{\gamma,U_{(1)}U_{(2)}}\right)+4\,h_{\gamma,U}^2\,h_{\gamma,Z}^2\,\mathcal{I}_{\gamma,U_{(1)}Z}+h_{\gamma,Z}^4\,\mathcal{I}_{\gamma,ZZ}\right], \end{aligned} \tag{33}$$

$$\begin{aligned} \text{where}\quad Q_{\gamma,1} &= \int\left(\tilde{\gamma}((u_1,u_2),(u_1,u_2),z)+\sigma_\varepsilon^2(u_1,u_2,z)\right)\,d(u_1,u_2,z) \\ \mathcal{I}_{\gamma,U_{(1)}U_{(1)}} &= \int\left(\gamma^{(2,0,0)}(u_1,u_2,z)\right)^2\,f_{UUZ}(u_1,u_2,z)\,d(u_1,u_2,z), \\ \mathcal{I}_{\gamma,U_{(1)}U_{(2)}} &= \int\left(\gamma^{(2,0,0)}(u_1,u_2,z)\gamma^{(0,2,0)}(u_1,u_2,z)\right)\,f_{UUZ}(u_1,u_2,z)\,d(u_1,u_2,z), \\ \mathcal{I}_{\gamma,U_{(1)}Z} &= \int\gamma^{(2,0,0)}(u_1,u_2,z)\gamma^{(0,0,2)}(u_1,u_2,z)\,f_{UUZ}(u_1,u_2,z)\,d(u_1,u_2,z),\quad\text{and} \\ \mathcal{I}_{\gamma,ZZ} &= \int\left(\gamma^{(0,0,2)}(u_1,u_2,z)\right)^2\,f_{UUZ}(u_1,u_2,z)\,d(u_1,u_2,z). \end{aligned}$$

Equation (33) again follows from the formulas in Wand and Jones (1994) and additionally by using the following equalities:
$\mathcal{I}_{\gamma,U_{(1)}U_{(1)}} = \mathcal{I}_{\gamma,U_{(2)}U_{(2)}}$, $\mathcal{I}_{\gamma,U_{(1)}U_{(2)}} = \mathcal{I}_{\gamma,U_{(2)}U_{(1)}}$, and $\mathcal{I}_{\gamma,U_{(1)}Z} = \mathcal{I}_{\gamma,U_{(2)}Z}$ due to the symmetry of the covariance function, where the expressions $\mathcal{I}_{\gamma,U_{(2)}U_{(2)}}$, $\mathcal{I}_{\gamma,U_{(2)}U_{(1)}}$, and $\mathcal{I}_{\gamma,U_{(2)}Z}$ are defined equivalently to their above defined counterparts.

Minimizing the AMISE function above with respect to $h_{\gamma,U}$ and $h_{\gamma,Z}$ leads to the optimal bandwidth expressions in Theorem 3.4. This is much more cumbersome than for the case of the mean function μ, but can easily done using, e.g., a computer algebra system.

It follows directly from the Theorem 3.2 that the first variance term V_γ^I is the leading variance term if the following order relation holds:

$$n^{-1}\,h_{\gamma,Z}^{-1} = o\left(n^{-(1+2\theta)}\,h_{\gamma,U}^{-2}\,h_{\gamma,Z}^{-1}\right), \tag{34}$$

where we used that by Assumption A-AS $nM \asymp n^{1+2\theta}$. Plugging the AMISE optimal bandwidth rates of Theorem 3.4 into the order relation of Eq. (34) leads to the corresponding θ

values of $0 \leq \theta < 1/5$ which describe the case considered here as "sparse" functional data. Observe that the same θ-threshold value of $1/5$ applies to both estimators $\hat{\mu}$ and $\hat{\gamma}$.

A.3.3 Proofs of Corollaries 3.1 and 3.2

Corollaries 3.1 and 3.2 follow directly from Theorems 3.1, 3.2, 3.3, and 3.4 and from applying a standard central limit theorem for iid data.

A.4 Proofs of the results in Section 3.2 (Dense functional data)

A.4.1 Proof of Theorem 3.5

The AMISE function of $\hat{\mu}$ including both variance terms V_μ^I and V_μ^{II} is given by

$$\mathrm{AMISE}_{\hat{\mu}}\left(h_{\mu,U}, h_{\mu,Z}\right) = \underbrace{\overbrace{(nm)^{-1}\, h_{\mu,U}^{-1}\, h_{\mu,Z}^{-1}\, R(K_\mu)\, Q_{\mu,1}}^{\int V_\mu^I(u,z) f_{UZ}(u,z) d(u,z)}}_{\text{2nd Order}} + \underbrace{\overbrace{n^{-1}\, h_{\mu,Z}^{-1}\, R(\kappa)\, Q_{\mu,2}}^{\int V_\mu^{II}(u,z) f_{UZ}(u,z) d(u,z)}}_{\text{1st Order}} + \tag{35}$$

$$+\,\frac{1}{4}\,(\nu_2(K_\mu))^2 \left[\underbrace{h_{\mu,U}^4\, \mathcal{I}_{\mu,UU}}_{\text{3rd Order}} + \underbrace{2\, h_{\mu,U}^2\, h_{\mu,Z}^2\, \mathcal{I}_{\mu,UZ}}_{\text{2nd Order}} + \underbrace{h_{\mu,Z}^4\, \mathcal{I}_{\mu,ZZ}}_{\text{1st Order}}\right],$$

$$\begin{array}{rrcl}
\text{where} & \mathcal{I}_{\mu,UU} & = & \int(\mu^{(2,0)}(u,z))^2\, f_{UZ}(u,z)\, d(u,z), \\
& \mathcal{I}_{\mu,ZZ} & = & \int(\mu^{(0,2)}(u,z))^2\, f_{UZ}(u,z)\, d(u,z), \\
& \mathcal{I}_{\mu,UZ} & = & \int \mu^{(2,0)}(u,z)\mu^{(0,2)}(u,z)\, f_{UZ}(u,z)\, d(u,z), \\
& Q_{\mu,1} & = & \int \left(\gamma(u,u,z) + \sigma_\epsilon^2\right)\, d(u,z), \quad \text{and} \\
& Q_{\mu,2} & = & \int \gamma(u,u,z)\, f_U(u)\, d(u,z).
\end{array}$$

Note that it is impossible to derive explicit AMISE optimal U- and Z-bandwidth expressions through minimizing Eq. (35) simultaneously for both bandwidths. If the second variance term V_μ^{II} is the leading variance term, the lowest possible AMISE value can be achieved if there exists a U-bandwidth which, first, allows us to profit from the (partial) annulment of the U-related bias-variance trade-off, but, second, assures that the second variance term V_γ^{II} remains the leading variance term.

The first requirement is achieved if the U-bandwidth is of a smaller order of magnitude than the Z-bandwidth, i.e., if $h_{\mu,U} = o(h_{\mu,Z})$. This restriction makes those bias components that depend on $h_{\mu,U}$ asymptotically negligible, since it implies that $h_{\mu,U}^2\, h_{\mu,Z}^2 = o(h_{\mu,Z}^4)$ and therefore that $h_{\mu,U}^4 = o(h_{\mu,U}^2\, h_{\mu,Z}^2)$. The latter two strict inequalities lead to the order relations between the three bias terms as indicated in Eq. (35). The second requirement is achieved if the U-bandwidth does not converge to zero too fast, namely if $mh_{\mu,U} \to \infty$, which implies the order relation between the two variance terms as indicated in Eq. (35).

Let us initially assume that it is possible to find an U-bandwidth that fulfills both the above requirements, namely $h_{\mu,U} = o(h_{\mu,Z})$ and $nh_{\mu,U} \to \infty$. With such an U-bandwidth we can make use of the order relations indicated in Eq. (35). That is, instead of minimizing the AMISE function in Eq. (35) over both bandwidths, we can minimize the following simpler and asymptotically equivalent AMISE function, which depends only on the Z-bandwidth:

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

The above equation is minimized by the following Z-bandwidth:

$$h_{\mu,Z}^D = \left(\frac{R(\kappa)\, Q_{\mu,2}}{n\, \left(\nu_2(K_\mu)\right)^2\, \mathcal{I}_{\mu,ZZ}}\right)^{1/5},$$

which is that of Eq. (11) in Theorem 3.5.

We still need to find U-bandwidth that fulfills the postulated requirements. To do so we suggest plugging the above optimal Z-bandwidth into the AMISE function in Eq. (35) and minimizing the (then classical) bias-variance trade-off between the asymptotic second order terms, which leads to the following expression for the U-bandwidths:

$$h_{\mu,U}^D = \left(\frac{R(K_\mu)\, Q_{\mu,1}}{nm\, \left(\nu_2(K_\mu)\right)^2\, \mathcal{I}_{\mu,UZ}}\right)^{1/3} \left(h_{\mu,Z}^D\right)^{-1},$$

which is that of Eq. (12) in Theorem 3.5.

In order to check whether this U-bandwidth actually fulfills the two necessary requirements, we apply some rearrangements. Using that by Assumption AS $m \asymp n^\theta$, leads to the following more transparent presentation of the bandwidth rates:

$$h_{\mu,Z}^D \asymp m^{-1/(5\,\theta)} \quad \text{and} \quad h_{\mu,U}^D \asymp m^{-\eta_\mu(\theta)} \quad \text{with} \quad \eta_\mu(\theta) = \frac{1}{3} + \frac{2}{15\,\theta} \tag{36}$$

With Eq. (36) it is easily verified that the necessary requirements ($h_{\mu,U,\text{AMISE}} = o(h_{\mu,Z}^D)$ and $mh_{\mu,U}^D \to \infty$) are fulfilled iff $\theta > 1/5$.

A.4.2 Proof of Theorem 3.6

The AMISE expression of $\hat{\gamma}$ including both variance terms V_γ^I and V_γ^{II} is given by

$$\begin{aligned}\text{AMISE}_{\hat{\gamma}}\,(h_{\gamma,U}, h_{\gamma,Z}) = {} & \overbrace{(nM)^{-1}\, h_{\gamma,U}^{-2}\, h_{\gamma,Z}^{-1}\, R(K_\gamma)\, Q_{\gamma,1}}^{\text{2nd Order}} + \overbrace{n^{-1}\, h_{\gamma,Z}^{-1}\, R(\kappa)\, Q_{\gamma,2}}^{\text{1st Order}} + \\ & + \frac{1}{4}\left(\nu_2(K_\gamma)\right)^2 \left[\underbrace{2\, h_{\gamma,U}^4\, (\mathcal{I}_{\gamma,U_{(1)}U_{(1)}} + \mathcal{I}_{U_{(1)}U_{(2)}})}_{\text{3rd Order}} + \underbrace{4\, h_{\gamma,U}^2\, h_{\gamma,Z}^2\, \mathcal{I}_{\gamma,U_{(1)}Z}}_{\text{2nd Order}} + \underbrace{h_{\gamma,Z}^4\, \mathcal{I}_{\gamma,ZZ}}_{\text{1st Order}}\right],\end{aligned} \tag{37}$$

$$
\begin{aligned}
\text{where} \quad \mathcal{I}_{\gamma,U_{(1)}U_{(1)}} &= \int \left(\gamma^{(2,0,0)}(u_1,u_2,z)\right)^2 \, f_{UUZ}(u_1,u_2,z)\, d(u_1,u_2,z),\\
\mathcal{I}_{\gamma,U_{(1)}U_{(2)}} &= \int \left(\gamma^{(2,0,0)}(u_1,u_2,z)\gamma^{(0,2,0)}(u_1,u_2,z)\right) \, f_{UUZ}(u_1,u_2,z)\, d(u_1,u_2,z),\\
\mathcal{I}_{\gamma,U_{(1)}Z} &= \int \gamma^{(2,0,0)}(u_1,u_2,z)\gamma^{(0,0,2)}(u_1,u_2,z)\, f_{UUZ}(u_1,u_2,z)\, d(u_1,u_2,z),\\
\mathcal{I}_{\gamma,ZZ} &= \int \left(\gamma^{(0,0,2)}(u_1,u_2,z)\right)^2 \, f_{UUZ}(u_1,u_2,z)\, d(u_1,u_2,z),\\
Q_{\gamma,1} &= \int \left(\tilde{\gamma}((u_1,u_2),(u_1,u_2),z) + \sigma^2_{\varepsilon}(u_1,u_2,z)\right)\, d(u_1,u_2,z), \quad \text{and}\\
Q_{\gamma,2} &= \int \tilde{\gamma}((u_1,u_2),(u_1,u_2),z)\, f_{UU}(u_1,u_2)\, d(u_1,u_2,z)
\end{aligned}
$$

By the same reasoning as in the preceding section, we initially determine requirements on the U-bandwidth that maintain the order relation between the two variance terms as indicated in Eq. (37). The first requirement is that $h_{\gamma,U} = o(h_{\gamma,Z})$. This restriction makes those bias components that depend on $h_{\gamma,U}$ asymptotically negligible, since it implies that $h^2_{\gamma,U}\, h^2_{\gamma,Z} = o(h^4_{\gamma,Z})$ and therefore that $h^4_{\gamma,U} = o(h^2_{\gamma,U}\, h^2_{\gamma,Z})$. The latter leads to the order relations between the three bias terms as indicated in Eq. (37). The second requirement is that the U-bandwidth does not converge to zero too fast, namely that $Mh^2_{\gamma,U} \to \infty$, which implies the order relation between the first two variance terms as indicated in Eq. (37).

Under these requirements on the U-bandwidths, we can minimize the following simpler and asymptotically equivalent AMISE function, which depends only on the Z-bandwidth:

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

The above equation is minimized by the following Z-bandwidth

$$
h^D_{\gamma,Z} = \left(\frac{R(\kappa)\, Q_{\gamma,2}}{n\, \left(\nu_2(K_\gamma)\right)^2\, \mathcal{I}_{\gamma,ZZ}}\right)^{1/5},
$$

which is that of Eq. (13) in Theorem 3.6.

Parallel to the preceding section, we determine the U-bandwidth by plugging the above optimal Z-bandwidth into the AMISE function in Eq. (37) and by minimizing the (then classical) bias-variance trade-off between the asymptotic second order terms, which leads to the following expression for the U-bandwidths:

$$
h^D_{\gamma,U} = \left(\frac{R(K_\gamma)\, Q_{\gamma,1}}{nM\, \left(\nu_2(K_\gamma)\right)^2\, \mathcal{I}_{\gamma,U_{(1)}Z}}\right)^{1/4} \left(h^D_{\gamma,Z}\right)^{-3/4},
$$

which is that of Eq. (14) in Theorem 3.6.

In order to check whether this U-bandwidth actually fulfills the two necessary requirements, we apply some rearrangements. Using that by Assumption AS $m \asymp n^\theta$ and that by construction $M \asymp m^2$, leads to the following more transparent presentation of the bandwidth rates:

$$
h^D_{\gamma,Z} \asymp M^{-1/(10\,\theta)} \quad \text{and} \quad h^D_{\gamma,U} \asymp M^{-\eta_\gamma(\theta)} \quad \text{with} \quad \eta_\gamma(\theta) = \frac{1}{4} + \frac{1}{20\,\theta}. \tag{38}
$$

With Eq. (38) it is easily verified that the necessary requirements, i.e., that $h_{\gamma,U,\mathrm{AMISE}} = o(h_{\gamma,Z,\mathrm{AMISE}})$ and $Mh^2_{\gamma,U,\mathrm{AMISE}} \to \infty$, are fulfilled iff $\theta > 1/5$.

A.4.3 Proofs of Corollaries 3.3 and 3.4

Corollaries 3.3 and 3.4 follow directly from Theorems 3.1 and 3.2 and from applying a standard central limit theorem for iid data.

References

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

Cao, G., L. Yang, and D. Todem (2012). Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics 24*(2), 359–377.

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

Charnigo, R. and C. Srinivasan (2015). A multivariate generalized cp and surface estimation. *Biostatistics 16*(2), 311–325.

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.

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

Gromenko, O. and P. Kokoszka (2012). Testing the equality of mean functions of ionospheric critical frequency curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*(5), 715–731.

Hall, P. and I. Van Keilegom (2007). Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica 17*(4), 1511–1531.

Härdle, W. and A. W. Bowman (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *Journal of the American Statistical Association 83*(401), 102–110.

Herrmann, E., J. Engel, M. Wand, and T. Gasser (1995). A bandwidth selector for bivariate kernel regression. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), 171–180.

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

Horváth, L., P. Kokoszka, and R. Reeder (2013). Estimation of the mean of functional time series and a two-sample problem. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(1), 103–122.

Hsing, T. and R. Eubank (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators.* John Wiley & Sons.

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.

Kokoszka, P. and R. Matthew (2017). *Introduction to Functional Data Analysis* (1. ed.). Texts in Statistical Science. Chapman & Hall/CRC.

Li, M., A.-M. Staicu, and H. D. Bondell (2015). Incorporating covariates in skewed functional data models. *Biostatistics 16*(3), 413–426.

Liebl, D. (2013). *Contributions to Functional Data Analysis with Applications to Modeling Time Series and Panel Data.* Dissertation. University of Cologne.

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing.

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.

Wand, M. and M. Jones (1994). Multivariate plug-in bandwidth selection. *Computational Statistics 9*(2), 97–116.

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.

Zhang, J.-T. and J. Chen (2007). Statistical inferences for functional data. *The Annals of Statistics 35*(3), 1052–1079.

Zhang, W. and Y. Wei (2015). Regression based principal component analysis for sparse functional data with applications to screening growth paths. *The Annals of Applied Statistics 9*(2), 597–620.

Zhang, X. and J.-L. Wang (2016). From sparse to dense functional data and beyond. *The Annals of Statistics 44*(5), 2281–2321.