

Gaussian Processes with Errors in Variables: Theory and Computation

Shuang Zhou

SZHOU98@ASU.EDU

*School of Mathematical and Statistical Sciences
Arizona State University
Tempe, AZ 85287-1804, USA*

Debdeep Pati

DEBDEEP@STAT.TAMU.EDU

*Department of Statistics
Texas A&M University
College Station, TX 77843-3143, USA*

Tianying Wang

TIANYINGW@TSINGHUA.EDU.CN

*Center for Statistical Science
Department of Industrial Engineering
Tsinghua University
Beijing 100084, China*

Yun Yang

YY84@ILLINOIS.EDU

*Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820-3633, USA*

Raymond J. Carroll

CARROLL@STAT.TAMU.EDU

*Department of Statistics
Texas A&M University
College Station, TX 77843-3143, USA
School of Mathematical and Physical Sciences
University of Technology Sydney
Ultimo NSW 2007, Australia*

Editor: Garvesh Raskutti

Abstract

Covariate measurement error in nonparametric regression is a common problem in nutritional epidemiology and geostatistics, and other fields. Over the last two decades, this problem has received substantial attention in the frequentist literature. Bayesian approaches for handling measurement error have only been explored recently and are surprisingly successful, although there still is a lack of a proper theoretical justification regarding the asymptotic performance of the estimators. By specifying a Gaussian process prior on the regression function and a Dirichlet process Gaussian mixture prior on the unknown distribution of the unobserved covariates, we show that the posterior distribution of the regression function and the unknown covariate density attain optimal rates of contraction adaptively over a range of Hölder classes, up to logarithmic terms. We also develop a novel surrogate prior for approximating the Gaussian process prior that leads to efficient computation and preserves the covariance structure, thereby facilitating easy prior elicitation. We demon-

strate the empirical performance of our approach and compare it with competitors in a wide range of simulation experiments and a real data example.

Keywords: Approximated Gaussian processes, measurement error model, nonparametric Bayes, smoothing and nonparametric regression, supersmooth errors

1. Introduction

The general formulation of a deconvolution problem assumes that the observations are the true underlying variables contaminated with measurement error. In an errors-in-variables regression problem, responses Y_i 's are observed corresponding to evaluations of an unknown regression function f_0 on noise-contaminated covariates W_i 's as

$$\begin{aligned} Y_i &= f_0(X_i) + \epsilon_i, & \epsilon_i &\stackrel{i.i.d.}{\sim} N(0, \sigma^2), \\ W_i &= X_i + u_i, & u_i &\stackrel{i.i.d.}{\sim} g_u, & X_i &\stackrel{i.i.d.}{\sim} p_0, \quad i = 1, \dots, n, \end{aligned} \tag{1}$$

where X_i 's are the unknown true covariates and we denote by p_0 the marginal distribution of the true covariate, and we write ‘‘i.i.d.’’ short for ‘‘identically and independently distributed’’. In model (1), we consider the centered Gaussian error ϵ_i with unknown standard deviation σ , and denote by g_u the known measurement error distribution. The goal is to recover the true regression function f_0 and the true density function p_0 .

From a frequentist perspective, there is a rich literature addressing these problems. Historically, the density deconvolution problem was first addressed in Carroll and Hall (1988); Fan (1991); Stefanski and Carroll (1990), where it was noted that the fundamental difficulty in recovering the true density lies in the nature of the distribution of the measurement errors, and a class of deconvolution kernel density estimators was proposed. In a nonparametric regression setting Fan and Truong (1993) developed a globally consistent deconvolution kernel type estimator. Later on, Ioannides and Alevizos (1997) generalized the estimator while Delaigle and Meister (2007) extended the theory to the heteroscedastic case. Refer to a review article (Delaigle, 2014) for a detailed discussion on kernel-based deconvolution estimators. Other methods such as deconvolution estimators based on Fourier-techniques, local linear and polynomial estimators are also popular, see Carroll et al. (1996, 1999); Cook and Stefanski (1994); Delaigle and Hall (2008); Delaigle et al. (2006, 2009); Du et al. (2011); Stefanski and Cook (1995).

It is well known that the optimal rate of convergence of deconvolution estimators can be quite slow compared to the classical minimax rate for estimating smooth densities or functions. The rate of convergence is controlled by the tail behavior of the characteristic function of the measurement error density; faster decaying rate of the characteristic function leads to a slower convergence rate and vice versa. In particular, the optimal rate is only of the logarithmic order when the measurement error distribution is a ‘‘supersmooth’’ distribution, whose characteristic function decays exponentially in the tails. This includes the Gaussian and the Cauchy densities. This slow rate of convergence renders estimation practically infeasible unless the measurement error variance is allowed to be sufficiently small (Carroll et al., 1999; Delaigle, 2008; Fan, 1992) with respect to the sample size. In particular, it has been shown in Delaigle (2008); Fan (1992) that the optimal rate of convergence in the ‘‘supersmooth’’ case is improved to $n^{-\beta/(2\beta+1)}$ for estimating a function in a Hölder class with regularity level β if the error standard deviation of a Gaussian error

density decreases to zero at the rate of $n^{-1/(2\beta+1)}$. This requirement on the error standard deviation can be easily satisfied by generating replicates $n^{1/(2\beta+1)}$ times per data point. In many applications, such as nutritional epidemiology, it is customary to collect multiple recalls of dietary intake from the respondents which serve as the replicated proxies and can boost the rate of convergence.

Another critical point regarding the performance of classical deconvolution estimators is the choice of an appropriate kernel and associated bandwidth. Many effective bandwidth selection procedures have been developed for practical purposes, refer to Delaigle and Gijbels (2004a,b); Delaigle and Hall (2008). In absence of the knowledge of the true regularity level, data-driven bandwidth selection procedures using the Lepski's method are employed (Comte and Lacour, 2013; Kappus and Mabon, 2014) with deconvolution kernel estimators, obtaining adaptivity with respect to the smoothness of the underlying function or density. Other types of the adaptive deconvolution estimator have been proposed, for instance, the ridge deconvolution estimator (Hall and Meister, 2007) and the thresholding wavelet deconvolution estimator (Fan and Koo, 2002).

On the other hand, Bayesian procedures are naturally suited for general nonparametric regression tasks because of their ability to adapt to the unknown smoothness and to allow quantifications of uncertainty. For classical density estimation problems with no measurement error, Bayesian nonparametric techniques including Dirichlet process Gaussian mixture model (Escobar and West, 1995; Ferguson, 1973; Lo, 1984) have demonstrated success in various applications, where the unknown density is modeled as a mixture of normals with a Dirichlet process prior on the mixing distribution. For the errors-in-variables regression estimation problem, Berry et al. (2002) were the first to develop a fully Bayesian procedure for the nonparametric regression problem using smoothing splines and P-splines. Variants of spline-based models are developed in Bayesian framework to approximate the density function and/or variance function in the heteroscedastic case (Sarkar et al., 2014; Staudenmayer et al., 2008). More recently, Cervone and Pillai (2015) developed a Bayesian analysis for Gaussian processes with location errors using hybrid Monte-Carlo techniques.

Bayesian approaches have been demonstrated to be very successful numerically, however, there is a clear dearth of theoretical results justifying these approaches. Few existing results for deconvolution density estimation are available recently in the Bayesian literature such as Gao and van der Vaart (2016); Donnet et al. (2018); Rousseau and Scricciolo (2021). To the best of our knowledge, a formal theoretical justification for the use of Bayesian procedures in the errors-in-variables regression problem is missing. As the main contribution of this paper, we propose a fully Bayesian framework for the errors-in-variables regression using a Gaussian process prior, and develop a new theoretical framework for studying its frequentist properties including consistency and the quantification of posterior convergence rates. As mentioned earlier, the optimal rate in the errors-in-variables problem with Gaussian error distribution has been proved to be extremely slow, rendering inference infeasible in applications. However, allowing the error variance to decrease to zero with sample size at an appropriate rate plays a very important role in improving the rate of convergence (Carroll et al., 1999; Fan, 1992). In this paper, we reexamine this situation from a Bayesian perspective assuming that the measurement error standard deviation decays at the order of $n^{-1/(2\beta+1)}$ where β is the smoothness of the true covariate density. However, we intend

to maintain adaptivity with respect to the smoothness level of the true function and the true covariate density.

As the main contribution, we show that in the errors-in-variables regression problem, when the Gaussian error variance decreases to zero at a certain rate, under appropriate regularity conditions on the true marginal density and regression function, the posterior distribution obtained from a suitably chosen hierarchical Gaussian process model with a Dirichlet process Gaussian mixture prior on the marginal density of the covariates converges to the ground truth at their respective minimax optimal rates, adaptively over a range of Hölder classes. By viewing density deconvolution as an inverse problem (Knapik et al., 2011; Ray, 2013), we follow the general recipe in Theorem 3.1 of Ray (2013) as sufficient conditions for posterior convergence in our setting. However, the work of Knapik et al. (2011) is restricted to conjugate priors, Ray (2013) considers only periodic function deconvolution using wavelets, and substantial technical hurdles remain. To address these challenges, we exploit the concentration properties of deconvolution kernel estimators to construct test functions with exponentially small type-I and type-II error bounds for the testing problem

$$H_0 : p = p_0, \quad \text{vs} \quad H_A : p \in \{p : d(p, p_0) > \xi_n\}. \quad (2)$$

Ray (2013) used concentration properties of thresholded wavelet based estimators based on standard results on concentration of Gaussian priors. However, analogous results for kernel density estimators suited to density deconvolution problems are lacking. One of our key technical contributions is to develop sharp concentration inequalities of the deconvolution kernel estimators to construct tests in (2).

On the computational side, although Bayesian spline models are quite successful in practice, the choice of knots as well as the number of basis functions are critical to obtain good empirical performance. This stimulates the development of other Bayesian approaches for modeling the unknown function of interest such as Gaussian process priors. Gaussian processes are routinely used for function estimation in a Bayesian context. However, their use in the context of measurement error in nonparametric regression models is limited, since the unobserved values of covariates are involved in the prior covariance matrix of Gaussian process and is no longer conditionally independent given the data. To alleviate this issue in errors-in-variables regression problem, we develop an approximation to the Gaussian process as a prior for the unknown regression function. The Gaussian process surrogate is computationally efficient as it avoids repeated computation of the matrix inversion. In addition to the appealing property of preserving the covariance kernel, we also show that the resulting surrogate process converges weakly to the original Gaussian process. This hints at the fact that the good properties of the original posterior distribution will be subsequently inherited by the surrogate posterior. For implementation, in addition to standard hyperparameters of a Gaussian process that control the smoothness of the sample paths, the Gaussian process surrogate contains a truncation parameter. Our result on the accuracy of such an approximation suggests that inference on the regression function is robust to the choice of the truncation parameter as long as it is chosen to be appropriately large. Hence the approximation retains all the potential advantages of a Gaussian process.

1.1 Review on Nonparametric Regression with Errors in Variables

Consider the regression model with errors in variables defined in Equation (1), where $\{(Y_i, W_i), i = 1, \dots, n\}$ are independent and identical draws from the joint unknown distribution. Recall that Y_i 's denote the observed responses and W_i 's are contaminated covariates. It is well known that in absence of any replicated proxy per data-point, the optimal rate for a ‘‘supersmooth’’ error distribution is only of the logarithmic order, rendering the estimators to be highly inefficient for practical purposes (Fan and Truong, 1993). In cases where the error distribution remains unknown, it can be estimated from the repeated observations or extra validation data (Hall and Ma, 2007; Johannes, 2009; Neumann, 2007). For the regular deconvolution kernel estimator, the deconvolution kernel function is constructed based on a suitable kernel function $K(\cdot)$ and the empirical estimator of the Fourier transform of the marginal density p of covariates. One can derive the deconvolution kernel density estimator (Fan and Truong, 1993) for both the marginal density p and the regression function f by

$$\widehat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^n K_n\{(x - W_i)/h\}, \quad (3)$$

$$\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K_n\{(x - W_i)/h\} Y_i / \widehat{p}_n(x), \quad (4)$$

$$K_n(x) = \frac{1}{2\pi} \int e^{-itx} \frac{\phi_K(t)}{\phi_u(t/h)} dt. \quad (5)$$

$K_n(\cdot)$ is the deconvolution kernel function, $\phi_K(\cdot)$ and $\phi_u(\cdot)$ are the Fourier transforms of the kernel function $K(\cdot)$ and the density of measurement error $g_u(\cdot)$, respectively. Usually $\phi_K(\cdot)$ is assumed to be compactly supported to ensure that the deconvolution kernel $K_n(\cdot)$ is well defined. Also, to achieve the rate optimality one requires that kernel function $K(\cdot)$ is a k th-order kernel function where k represents the regularity level of the true density function. However, in practice such deconvolution kernels typically do not admit closed-form expressions, and the estimation could suffer from extra errors due to numerical integrations.

1.2 Bayesian Nonparametric Regression with Errors in Variables

In this article, we focus on the normal distribution $N(0, \delta^2)$ with an unknown variance δ^2 as the measurement error distribution. We consider the following generic Bayesian hierarchical model for the nonparametric regression with errors in variables:

$$\begin{aligned} Y_i &= f(X_i) + \epsilon_i, & \epsilon_i &\sim N(0, \sigma^2), \\ W_i &= X_i + u_i, & u_i &\sim N(0, \delta^2), & X_i &\sim p, & i &= 1, \dots, n, \\ f &\sim \Pi_f, & p &\sim \Pi_p, & \sigma^2 &\sim \Pi_{\sigma^2}, & \delta^2 &\sim \Pi_{\delta^2}. \end{aligned} \quad (6)$$

We assume that Y_i is conditionally independent of W_i given X_i , for $i = 1, \dots, n$. In the Bayesian framework, we obtain the posterior distribution of unknown parameters $\theta = (f, p, \delta)$ given the observed values $D_n = \{(Y_i, W_i), i = 1, \dots, n\}$ via Bayes’ rule:

$$P(\theta \mid D_n) = \frac{P(D_n \mid \theta) P(\theta)}{P(D_n)}.$$

This posterior distribution $P(\theta | D_n)$ can then be used to conduct statistical inference on marginal density p and regression function f , such as constructing point estimators and their associated credible intervals or bands. Variants of the model defined in Equation (6) are used in the context of Bayesian methods in errors-in-variables regression problem (Berry et al., 2002; Sarkar et al., 2014). Although for practical purposes we assume a prior distribution on δ^2 , in the theoretical investigation, to obtain the minimax-optimal convergence rate results, we assume δ^2 to be known and let δ^2 decrease to 0 at a certain rate depending on n . For practical purpose we assign an objective prior on σ^2 , the details of which can be found in Appendix F. Whereas, for theoretical investigation we assume $\sigma = 1$ to simplify the analysis. Extension to general σ is straightforward.

By assigning proper priors on f and p , we show that the estimation of f and p can be made adaptive, which means the prior does not demand any knowledge on the smoothness of the true regression function, and yet a nearly optimal rate of posterior contraction can be achieved as if the smoothness is known. Different from the deconvolution kernel estimator, a Bayesian method does not require explicitly constructing a deconvolution kernel function $K_n(\cdot)$, but the existence of such kernel is used for constructing the test function aforementioned in the introduction. The details of choosing specific priors for f and p are discussed in the following section. We start describing the Gaussian process prior for f which requires specifying a covariance kernel analogous to the kernel $K(\cdot)$.

1.3 Prior Specifications

We consider a Gaussian process prior (Rasmussen and Williams, 2006) as the prior Π_f for f , which is a distribution over a space of functions such that the joint distribution of any finite evaluations of the random function is multivariate Gaussian. A Gaussian process is completely defined by a mean function $m(x) = E\{f(x)\}$ and a covariance kernel function $c(x, x') = \text{cov}\{f(x), f(x')\}$ for any $x, x' \in \mathbb{R}$. Therefore, any finite collection of random observation points $\{y_1(x_1), \dots, y_N(x_N)\}$ at locations x_1, \dots, x_N has a joint Gaussian distribution given by

$$\{y_1(x_1), \dots, y_N(x_N)\} \sim N(m, \tau^2 \Sigma),$$

where $m = \{m(x_1), \dots, m(x_N)\}$ and Σ is the $N \times N$ covariance matrix with the (i, j) th element $\Sigma_{ij} = c(x_i, x_j)$. The mean function reflects the expected center of the realization, and the covariance kernel function reflects its fluctuation and local dependence. The hyperparameter τ attached to the covariance kernel function controls the fluctuation magnitude. We use the notation $f(\cdot) \sim \text{GP}(m(\cdot), c(\cdot, \cdot))$ to denote that function f follows a Gaussian process with mean function m and covariance kernel function c . For the regular Gaussian process regression in the noised case with noise level σ , the predictive formula (Rasmussen and Williams, 2006) is

$$\begin{aligned} f(X^*) | X, Y, X^* &\sim N(\bar{f}^*, \text{cov}\{f(X^*)\}), \\ \bar{f}^* &= c(X^*, X)\{c(X, X) + \sigma^2 I\}^{-1}Y, \\ \text{cov}\{f(X^*)\} &= c(X^*, X^*) - c(X^*, X)\{c(X, X) + \sigma^2 I\}^{-1}c(X, X^*), \end{aligned} \tag{7}$$

where X, Y are the given data, X^* is a new data point, $f(X^*)$ is the prediction at X^* and $c(X^*, X)$ denotes the covariance matrix between X^* and X . The posterior is a multivariate

normal involved with the original data and the new data point. Refer to Rasmussen and Williams (2006) for a detailed explanation of a Gaussian process. Choice of the covariance kernel c is crucial to obtain a desirable functional estimation. A squared exponential covariance or more generally, a Matérn covariance kernel are commonly used in practice. Also, the covariance kernel is often associated with hyperparameters which control the smoothness of the sample paths (Adler, 1990). We shall discuss specific choices in Section 2.2.

It might appear on the surface that one can assume a parametric distribution for the unknown X if the interest is solely on recovering the unknown function f . However as we will show in the simulation studies and also observed in Sarkar et al. (2014), a parametric distribution on X is not capable of recovering the unknown infinite dimensional parameters f . As a flexible prior distribution on the density p , we propose to use a Dirichlet process Gaussian mixture prior defined by

$$X \sim g(\cdot), \quad g(\cdot) = \int \phi_{\sqrt{\tau}}(\cdot - \mu) G(d\mu, d\tau), \quad G \sim \text{DP}(\alpha G_0). \quad (8)$$

Here $\phi_{\sqrt{\tau}}(\cdot - \mu)$ denotes the normal density function with mean μ and variance τ . $\text{DP}(\alpha G_0)$ denotes a Dirichlet process prior (Ferguson, 1973) with G_0 as the base probability measure on $\mathbb{R} \times \mathbb{R}^+$ and $\alpha > 0$ is a precision parameter. Given a probability space \mathcal{P} , for any $P \in \mathcal{P}$ we define the measure space (\mathcal{X}, Ω, P) with Ω denoting the Borel sets of \mathcal{X} , a Dirichlet process satisfies that for any finite and measurable partition B_1, \dots, B_k on \mathcal{X} , $\{P(B_1), \dots, P(B_k)\} \sim \text{Dir}\{\alpha G_0(B_1), \dots, \alpha G_0(B_k)\}$, where $\text{Dir}\{a_1, \dots, a_k\}$ denotes the Dirichlet distribution with parameters a_1, \dots, a_k . A Dirichlet process Gaussian mixture prior is known to be a highly flexible nonparametric prior on the space of densities having a common support as the base measure G_0 (Escobar and West, 1995). It has thus become a very popular Bayesian density estimation method which received considerable attention over the last two decades both from computational (Kalli et al., 2011; Neal, 2000) and theoretical perspectives (Ghosal and van der Vaart, 2007; Kruijer et al., 2010; Shen et al., 2013). Recently, Dirichlet process mixture models have also been commonly used for studying the posterior consistency and contraction rate for Bayesian deconvolution problem under various settings (Gao and van der Vaart, 2016; Su et al., 2020; Rousseau and Scricciolo, 2021). In the next section, we shall discuss in detail that applying a Gaussian process prior to recover the true regression combined with modeling the covariate density with a finite approximation of the Dirichlet process Gaussian mixture prior, we can correct for the bias due to the measurement error.

2. Posterior Contraction Properties

In this section, we study the frequentist large sample properties of the proposed Bayesian errors-in-variables model. We begin with a description of notations used throughout the rest of the paper in Section 2.1, then state assumptions on the true functions and priors in Section 2.2. Section 2.3 contains our main result on the posterior contraction rate.

2.1 Notation and Preliminaries

Let $\lfloor x \rfloor$ denote the greatest integer that is strictly less than or equal to x for all $x \in \mathbb{R}$. We define the L_1 norm as $\|f\|_1 = \int |f(x)| dx$, and define the supremum norm as $\|f\|_\infty =$

$\sup_{x \in S} |f(x)|$, where S is the domain of function f . We say a sequence of measures P_n converges weakly to a measure P , denoted by $P_n \rightsquigarrow P$ if $\int \phi dP_n \rightarrow \int \phi dP$, for all bounded continuous function ϕ . Denote by $\mathcal{C}[0, 1]$ the space of continuous functions defined on $[0, 1]$ and denote by $\mathcal{C}^\beta[0, 1]$ the Hölder space of β -smooth functions $f : [0, 1] \rightarrow \mathbb{R}$ satisfying

$$|f(x+y)^{[\beta]} - f(x)^{[\beta]}| \leq L|y|^{\beta-[\beta]}, \quad (x, y) \in [0, 1],$$

for some constant $L > 0$. For any probability measure F on \mathbb{R} , let $p_{F, \sigma}(x) = \int \phi_\sigma(x-z)dF(z)$ be the location mixture of normals induced by F . For any finite positive measure α write $\bar{\alpha} = \alpha/\alpha(\mathbb{R})$, where $\alpha(\mathbb{R})$ denotes a measure on \mathbb{R} . Let $\text{DP}(\alpha)$ denote the Dirichlet process with the base measure α . We denote the prior distribution by $\Pi(\cdot)$ and the posterior distribution by $\Pi_n(\cdot | D_n)$. For two positive sequences a_n, b_n , we write $a_n \asymp b_n$ if a_n/b_n can be bounded from below and above by finite constants. In addition, we use “ \lesssim ” (“ \gtrsim ”) to indicate inequalities up to finite universal constants.

2.2 Assumptions

Assumption 1 *The regression function $f_0 \in \mathcal{C}^\beta[0, 1]$ with $\beta > 1/2$. We also assume $\|f_0\|_\infty < A_0$ for some large enough constant A_0 .*

We assume that β is unknown while fitting the model and our optimal convergence rate results are adaptive for any choice of $\beta > 1/2$. This is achieved easily in a Bayesian paradigm through a suitable prior on the smoothness parameter of the Gaussian process. The finite upper bound assumption is common to achieve the adaptivity in the errors-in-variables problem, similar assumptions can be found in Chesneau (2010); Chichignoud et al. (2017). In practice, we can obtain a reasonable upper bound as a multiple of averaged responses from additional validation data sets (Yang and Dunson, 2016). The lower bound on the smoothness is also a common assumption in the random design regression problem, refer to Baraud (2002); Birgé (1979); Brown et al. (2002) for further discussion on this topic.

Assumption 2 *The marginal density p_0 of the unobserved covariates X is in $\mathcal{C}^{\beta'}[0, 1]$ for some $\beta' \geq \beta$, where β is defined in Assumption 1. Also, we assume there exists a finite constant $B > 0$ such that $\inf_{x \in [0, 1]} p_0(x) \geq B^{-1}$.*

Smoothness assumptions and the lower bound assumption on the marginal density ensure a better control of the numerator and the denominator of the deconvolution kernel estimator defined in Equation (4) separately. Analogous smoothness assumptions can be found in Fan and Truong (1993), that the regression function and marginal density are assumed to have the same smoothness level. Refer also to Delaigle and Meister (2007) where $f_0 p_0$ and p_0 are assumed to have same regularity level.

The assumption $\beta' > \beta$ in Assumption 2 requires discussion. From model (1), the deconvolution density estimation problem for p_0 can be reduced to a random design regression function estimation problem for f_0 by conditioning on a density p in the parameter space. Hence the overall convergence rate will be determined by the slowest contraction rates for estimating p_0 and f_0 . Although our theory is derived for compactly supported p_0 , it can be extended to the unbounded support case under desirable tail conditions (Kruijer et al., 2010) on p_0 .

In the Bayesian errors-in-variables model defined in Equation (6), we assign a centered and rescaled Gaussian process prior on f , denoted by $\text{GP}(0, c; A)$, associated with the squared exponential covariance kernel $c(x, x'; A) = \exp\{-A^2\|x - x'\|^2\}$ with the rescaled random variable A satisfying the following Assumption 3. This choice is motivated by the fact that a properly scaled squared exponential covariance kernel is known to lead to the optimal rate of posterior convergence (van der Vaart and van Zanten, 2007, 2009). In addition, we consider a Dirichlet process Gaussian mixture prior on the marginal density p defined as $p_{F, \tilde{\sigma}}$, with $F \sim \text{DP}(\alpha)$ and $\tilde{\sigma} \sim G$, where G satisfies Assumption 4 below.

Assumption 3 *We assume the rescaled parameter A possesses a density m satisfying for sufficiently large $a > 0$,*

$$C_1 a^p \exp(-D_1 a \log^q a) \leq m(a) \leq C_2 a^p \exp(-D_2 a \log^q a),$$

for constants $C_1, C_2, D_1, D_2 > 0$ and $p, q \geq 0$. We assume a conditional Gaussian process prior on the sets of all functions $\mathcal{A} = \{f \in \mathcal{C}[0, 1] : \|f\|_\infty < A_0\}$, for the same constant A_0 in Assumption 1.

Assumption 3 includes the gamma density as a special case when $q = 0$. A similar assumption appears in van der Vaart and van Zanten (2009). We restrict the Gaussian prior over the set \mathcal{A} based on Assumption 1.

Assumption 4 *The Dirichlet process Gaussian mixture prior on the marginal density $p(x)$ defined by $p_{F, \tilde{\sigma}}$ with $F \sim \text{DP}(\alpha)$ and $\tilde{\sigma} \sim G$, satisfy the following conditions:*

$$\begin{aligned} 1 - \bar{\alpha}[-x, x] &\leq \exp(-b_1 x^{\tau_1}) \quad \text{for all sufficiently large } x > 0, \\ G(\tilde{\sigma}^{-2} \geq x) &\leq c_1 \exp(-b_2 x^{\tau_2}) \quad \text{for all sufficiently large } x > 0, \\ G(\tilde{\sigma}^{-2} < x) &\leq c_2 x^{\tau_3} \quad \text{for all sufficiently small } x > 0, \\ G(s < \tilde{\sigma}^{-2} < s(1+t)) &\leq c_3 s^{c_4} t^{c_5} \exp(-b_3 s^{1/2}) \quad \text{for } s > 0 \text{ and } t \in (0, 1), \end{aligned}$$

for positive constants $\tau_1, \tau_2, \tau_3, b_1, b_2, b_3, c_1, \dots, c_5$.

The inverse-gamma density on $\tilde{\sigma}$ satisfies the above assumptions, whereas the inverse-gamma density on $\tilde{\sigma}^2$ does not. This is a fairly standard assumption in the Bayesian asymptotics literature on the Dirichlet process mixture of Gaussians, for similar assumptions refer to the posterior convergence analysis for density estimation in Shen et al. (2013).

2.3 Main Theorem on Posterior Contraction

For the model defined in Equation (6), we first define the marginal likelihood of random pairs $\{(Y_i, W_i), i = 1, \dots, n\}$ as $g_{f,p}(y, w) = (2\pi\delta)^{-1} \int \phi_1\{y - f(x)\} \phi_\delta(w - x) p(x) dx$ and denote its distribution measure by $G_{f,p}$. Recall that we assume the noise level of the random error $\sigma = 1$ to simplify the calculation. Based on the Baye's rule, the posterior distribution given n pairs of observations denoted by $\{Y_{1:n}, W_{1:n}\}$ can be written as

$$\Pi_n\{(f, p) \in B \mid Y_{1:n}, W_{1:n}\} = \frac{\int_B \prod_{j=1}^n g_{f,p}(Y_j, W_j) d\Pi(f) d\Pi(p)}{\int_{\mathcal{P}} \prod_{j=1}^n g_{f,p}(Y_j, W_j) d\Pi(f) d\Pi(p)}, \quad (9)$$

where B is any measurable subset of $\mathcal{P} = \{(f, p) : f \in \mathcal{C}[0, 1]; p : [0, 1] \rightarrow \mathbb{R}, \int p(x) dx = 1\}$.

Theorem 1 *Suppose f_0 and p_0 satisfy Assumptions 1 and 2, respectively, and the prior Π on (f, p) satisfies Assumptions 3 and 4. Then for some sufficiently large constant $M > 0$ and the standard deviation δ_n of normal measurement error,*

$$\begin{aligned} \Pi_n\{(f, p) : \|f - f_0\|_1 < M \max(\epsilon_n, \delta_n^\beta), \|p - p_0\|_1 < M \max(\epsilon'_n, \delta_n^{\beta'}) \mid Y_{1:n}, W_{1:n}\} \\ \rightarrow 1 \text{ almost surely in } G_{f_0, p_0}, \text{ as } n \rightarrow \infty, \end{aligned}$$

where $\epsilon_n = n^{-\beta/(2\beta+1)}(\log n)^t$, $\epsilon'_n = n^{-\beta'/(2\beta'+1)}(\log n)^t$ with $t = \max\{(2 \vee q)\beta/(2\beta + 1), t'\}$, where $t' > (\gamma + 1/\beta')/(2 + 1/\beta')$ for some $\gamma > 2$. When $\delta_n \lesssim \epsilon_n^{1/\beta}$, the convergence rate for recovering f_0 is a multiple of ϵ_n .

It has been known that fixing $\delta_n \equiv 1$ leads to a logarithmic minimax error rate for errors-in-variables regression estimation. We remark that Theorem 1 does not yield the optimal rate in this case, as the current method to deliver posterior contraction rate for nonparametric models is sharp only up to logarithmic terms. However, when $\delta_n \lesssim n^{-1/(2\beta+1)}$, Theorem 1 shows that optimal rates for regression and density estimation under the EIV setting is the same under the regular nonparametric setting without measurement errors, respectively.

The proof of Theorem 1 can be found in Appendix B. Existing contraction rate results in the frequentist deconvolution literature (Fan and Truong, 1993) require the knowledge of the smoothness of both the true covariate density and the regression function to achieve the optimal convergence rate for the regression function. Theorem 1, on the other hand, achieves minimax optimal rate of posterior convergence adaptively over all smoothness levels (β', β) for β, β' defined in Assumptions 1 and 2, given the knowledge of decaying rate of the error standard deviation δ_n (or the number of replications). To understand the implication of the posterior convergence rate of f in Theorem 1 let us focus on the case where $\beta = 1$. Since $\{f(X) - f_0(X)\} \asymp \{f(W) - f_0(W)\} + \{f'(W) + f'_0(W)\}(X - W)$, the convergence rate for estimating f is limited by how fast the marginal density of X can be recovered from observations W . This intuitively justifies the rate $\max(\epsilon_n, \epsilon'_n, \delta_n^\beta)$ in EIV model. We remark that the rate results in Theorem 1 also hold for f, p even if $\beta > \beta'$. In that case, the posterior of p always attains the near-minimax rate in recovering the true density, whilst the best obtainable posterior rate for recovering f_0 is limited to ϵ'_n , which is slower than ϵ_n .

Analyzing the posterior distribution following the seminal work (Ghosal et al., 2000) requires upper-bounding the numerator of the posterior defined in Equation (9) over some set B of interest and lower-bounding the marginal likelihood. In our proof, the numerator can be bounded above by constructing a sequence of test functions that is used to test the true model against models outside a small neighborhood of the truth under proper metric. As a key technical contribution, we obtain sharp bounds for Type I and Type II errors of the constructed tests by developing large deviation bounds for the deconvolution density estimator, which generalizes some results in Pati et al. (2015) for random design regression to errors-in-variables problem. To bound the marginal likelihood from below, it requires the priors assigned on the regression and covariate density assigning enough mass around the truth. A component-wise Gaussian prior on the covariate cannot concentrate enough over a small neighborhood of the true locations, simply because the concentration of n -dimensional standard Gaussian vector cannot exploit the smoothness of the density and hence cannot assign enough mass within a small neighborhood around the true density.

On the other hand, a mixture of normals prior allows borrowing of information, naturally exploits the smoothness and provides adequate concentration. A similar treatment to the covariate density can also be found in recent Bayesian deconvolution literature (Gao and van der Vaart, 2016; Donnet et al., 2018; Rousseau and Scricciolo, 2021).

3. Posterior Computation

In order to sample from the posterior distribution of (f, p, δ, σ) , we employ a Gibbs sampler and sample from each of the parameters given the others. Posterior sampling methods for Bayesian density estimation using Dirichlet process Gaussian mixture prior is popular, refer to the Pólya urn sampler (Escobar and West, 1995; MacEachern and Müller, 1998) and blocked Gibbs sampler with stick-breaking representation (Ishwaran and James, 2001). In this article, we use the finite approximation of the Dirichlet process Gaussian mixture prior with the stick-breaking representation. The major bottleneck of the computation stems from sampling the Gaussian process term f which requires a) inversion of $n \times n$ matrices depending on the latent covariates and b) sampling from the conditional distribution of the true covariates, which is intractable. Task a) makes the algorithm computationally inefficient and unstable specifically for the errors-in-variables regression problem, since it requires evaluating the inverse of the covariance matrix repeatedly along with the updates of covariates. To bypass $O(n^3)$ computation steps associated with inverting an unstructured $n \times n$ covariance matrix, numerous powerful techniques have been proposed in the last decade; fixed rank kriging (Banerjee et al., 2008; Finley et al., 2009), covariance tapering (Furrer et al., 2006; Kaufman et al., 2008), composite likelihood methods (Guan, 2006; Heagerty and Lele, 1998). In using these techniques, often the original covariance kernel itself is not preserved, which means the covariance function of the approximate process differs from the covariance function of the original process. More recently, Stroud et al. (2017) and Guinness and Fuentes (2017) derived a fast algorithm of sampling from stationary Gaussian processes on the large-scale lattice data, using the circulant embedding technique proposed in Wood and Chan (1994). Such techniques typically require the assumption of equally spaced covariates. In the absence of equally spaced design, the idea is to define a larger lattice and consider the prediction as missing data imputation (Guinness and Fuentes, 2017; Stroud et al., 2017). However, it is not straightforward to translate these ideas to the errors-in-variables regression problem as the true covariates are contaminated and the true marginal distribution remains unknown. Instead, we propose using a lower dimensional mapping to approximate the Gaussian process based on the random Fourier basis proposed by Rahimi and Recht (2008a). And the random mapping to the Fourier domain preserves the covariance kernel associated with the original Gaussian process. This also avoids computing the inverse of covariance matrix by introducing moderate numbers of parameters associated with the Fourier basis. Moreover, this is suitable in applications where practitioners have a pre-conceived notion of using a particular covariance function and we require the approximated covariance to accurately reflect that prior opinion. The lower dimensional mapping is chosen to approximate the original Gaussian process arbitrarily well; refer to Theorem 2. We describe the approximate Gaussian process in the following Section 3.1.

3.1 An Approximation of the Gaussian Process

The low-rank projection of a stationary covariance kernel on a random feature space is a popular approach to scale up kernel-based regression methods (Rahimi and Recht, 2008a). Theoretical properties of the random Fourier feature projection have been extensively studied in the last decade, mostly in terms of the approximation accuracy of the covariance kernel function (Sutherland and Schneider, 2015; Sriperumbudur and Szabó, 2015), properties of the induced reproducing kernel Hilbert space (Rahimi and Recht, 2008b,c; Bach, 2017), and the expected risk bounds of an approximated kernel ridge regression estimator based on random Fourier features and their variants (Avron et al., 2017; Li et al., 2019; Zhang et al., 2019; Yang et al., 2021). For a detailed and categorized summary of existing results, one may refer to a recent work (Liu et al., 2021). In this section, we develop a low-rank random Fourier basis projection as an approximate of a stationary zero-mean Gaussian process $\text{GP}(0, c)$, which can be represented as a Bayesian linear model. Such a representation has been considered in Wilson et al. (2020) where they used a random feature projection to approximate the original GP, and further approximated the obtained posterior distribution to speed up posterior computation. In our case we study the exact posterior distribution resulting from the approximated GP prior.

Denote by the corresponding spectral density $\phi_c(\cdot)$ defined through $c(h) = \int e^{ihx} \phi_c(x) dx$. For a suitably chosen large integer N , we define

$$\tilde{f}_N(x) = (2/N)^{1/2} \sum_{j=1}^N a_j \cos(w_j x + s_j), \quad (10)$$

where $a_j \stackrel{i.i.d.}{\sim} \text{N}(0, 1)$, $w_j \stackrel{i.i.d.}{\sim} \phi_c$ and $s_j \stackrel{i.i.d.}{\sim} \text{Unif}[0, 2\pi]$ for $j = 1, \dots, N$. The random process $\tilde{f}_N(x)$ is an N -dimensional approximation to a GP such that its covariance function coincides with the kernel function of original GP. In addition, Theorem 2 shows that the approximate also converges to the original Gaussian process $\text{GP}(0, c)$ weakly.

Theorem 2 *Suppose f is the original Gaussian process $\text{GP}(0, c)$ and \tilde{f}_N is defined in Equation (10), we have*

$$\tilde{f}_N \rightsquigarrow f, \quad \text{as } N \rightarrow \infty.$$

Also, for any $x, y \in \mathbb{R}$,

$$E\{\tilde{f}_N(x)\} = 0; \quad \text{cov}\{\tilde{f}_N(x), \tilde{f}_N(y)\} = c(x, y).$$

The proof of Theorem 2 is deferred to Appendix C. The construction \tilde{f}_N is related to the random feature mapping in the Fourier domain (Rahimi and Recht, 2008a), used to project the kernel onto a lower-dimension space \mathbb{R}^N . It is straightforward to show that preservation of the covariance kernel associated with the original Gaussian process for the proposed process defined over the real area, due to the expression of Fourier features. However, the weak convergence result of \tilde{f}_N is non-trivial and the proof provides a framework to study the asymptotic property of random processes constructed based on Fourier projection.

Theorem 2 validates the usage of \tilde{f}_N to approximate a stationary GP in an asymptotic manner. Allowing \tilde{f}_N to be adaptive to the unknown smooth level of the true regression

function, we now assume $\{\omega_j\}$ are independently and identically generated from the spectral measure of a rescaled squared exponential kernel function. The rescaling parameter is unknown and endowed with the prior satisfying Assumption 3. Adopting the same notation of a rescaled mean-zero Gaussian process $f \sim \text{GP}(0, c^A)$ considered in Section 2 and the rescaling $A \sim g(\cdot)$ for some distribution g , analogously, we define a rescaled version of (10) as

$$\tilde{f}_{A,N}(x) = (2/N)^{1/2} \sum_{j=1}^N a_j \cos(w_j^A x + s_j),$$

where $\{a_j, s_j\}$ are same as in (10), and for any $a > 0$, $w_j^A | (A = a) \stackrel{i.i.d.}{\sim} \phi_c^a$ for $j = 1, \dots, N$, recall that $\phi_c^a(\lambda) = a^{-1} \phi_c(\lambda/a)$ denotes the spectral density of GP associated with a squared exponential kernel function indexed with the rescaling parameter a . It is straightforward to show that Theorem 2 holds with $A = a$ for any fixed $a > 0$. In addition, Theorem 3 below verifies that the posterior of $\tilde{f}_{A,N}$ converges towards the true regression curve f_0 at a near minimax rate in the EIV regression problem as well, given an appropriate number of the random Fourier features.

Theorem 3 *Suppose f_0 and p_0 satisfy Assumptions 1 and 2, respectively, and Assumptions 3 and 4 hold. Then for some fixed large constant $M > 0$ and for the number of features N satisfying $N \asymp n \epsilon_n^2 (\log n)^{\tilde{t}}$ for some constant $\tilde{t} > 0$ which is free of n, N , and recall the standard deviation δ_n of normal measurement error,*

$$\begin{aligned} \Pi_n \left\{ (\tilde{f}_{A,N}, p) : \|\tilde{f}_{A,N} - f_0\|_1 < M \max(\epsilon_n, \delta_n^\beta), \|p - p_0\|_1 < M \max(\epsilon_n', \delta_n^{\beta'}) \mid Y_{1:n}, W_{1:n} \right\} \\ \rightarrow 1 \text{ almost surely in } G_{f_0, p_0}, \text{ as } n \rightarrow \infty, \end{aligned}$$

where ϵ_n, ϵ_n' are same as in Theorem 1. Again, when $\delta_n \lesssim \epsilon_n^{1/\beta}$, the posterior contraction rate of $\tilde{f}_{A,N}$ is a multiple of ϵ_n .

Theorem 3 provides an asymptotic result on the posterior distribution of approximated GP, provided the rank of the random feature projection increases at a certain rate with the sample size. To the best of our knowledge, this is the first theoretical result on low-rank random feature projection of GPs in L_1 norm under a Bayesian framework. This result can be easily adapted to other regression/learning problems beyond the EIV context, such as nonparametric regression with random designs. The proof of Theorem 3 is deferred to Appendix, which follows a similar line of arguments as in the proof of Theorem 1. Theorem 3 delineates a specific increasing rate of the number of random features in order to attain the best rate. A minimum requirement on the number of random features has been determined in literature on kernel ridge regression (KRR) estimator with random features, which conveys the idea that larger is the number of the random features, the better is the approximation with random Fourier features to the original KRR estimator. However, when all $\{a_j, \omega_j, s_j\}$ in (10) are treated as random parameters, the number of random features N cannot increase too fast in order to retain a minimum prior concentration over a small KL-neighborhood of the truth, due to the concentration of measure phenomenon of high-dimensional Gaussian random vectors.

To implement $\tilde{f}_{A,N}$, it suffices to treat $\{a_j, w_j, s_j\}$ as unknown parameters endowed with suitable independent priors. More details of the posterior computation is deferred to Appendix F. We remark that the order of the the number of the features in Theorem 3 is primarily of theoretical interest as the smoothness of the function is unknown. Although we do not have adaptive results, in the empirical study, we find the Gaussian process surrogate performs almost as well as the original Gaussian process when N is chosen within the range $(n/8, n/2)$ for data sets of moderate sizes.

4. Numerical Results

In this section, we empirically illustrate applications of the proposed Gaussian process surrogate and its variants to Bayesian errors-in-variables model in the following synthetic examples. We consider a uniform marginal distribution $X \sim \text{Unif}[-3, 3]$ and the regression function: $f(x) = \sin(\pi x/2)/[1 + 2x^2\{\text{sign}(x) + 1\}]$. We consider three choices of sample size $n \in \{100, 250, 500\}$, and consider additive normal regression errors independently and identically drawn from $N(0, \sigma^2)$ with a fixed noise level $\sigma = 0.2$. We confine ourself to the centered normal distribution $N(0, \delta^2)$ for the measurement error with a sequence of gradually increasing variances $\{\delta^2\}$, for the purpose of checking empirical performance of proposed methods in the presence of measurement errors of varying degrees. Specifically, for $n = 100, 250$, we consider $\delta^2 \in \{0.01, 0.2, 0.4, 0.6, 0.8, 1\}$; for $n = 500$, we consider $\delta^2 \in \{0.001, 0.005, 0.01, 0.1, 0.5, 1\}$. Under each setting, we compare the following methods:

1. GPEV_a: Approximated Gaussian process model described in Section 3.1 with a Dirichlet process Gaussian mixture prior on the marginal density.
2. GPEV_f: Full scale Gaussian process model using the predictive formula in Equation (7), with a Dirichlet process Gaussian mixture prior on the marginal density.
3. GPEV_n: Approximated Gaussian process model described in Section 3.1 with a univariate normal prior on the covariate component-wise.
4. GP: Full scale Gaussian process model that ignores the measurement error.
5. decon: Deconvolution kernel method from <https://github.com/TimothyHyndman/deconvolve>.

To implement GPEV_a and GPEV_n, we consider the following combinations of the sample size n and the number of Fourier basis functions N : $(n, N) \in \{(100, 40), (250, 60), (500, 80)\}$. We remark that the values of N are chosen based on preliminary numerical experiments. We only present the numerical results for $n = 100, 500$ in this section, the result for $n = 250$ is similar and thus deferred to Appendix G. For Bayesian approaches, we ran the Gibbs sampler with 2,000 iterations and discarded the first 1,500 iterations as a burn-in. The derivation of a full conditional and detail on hyperparameter choices can be found in Appendix F. The investigation on the mixing behavior of the Gibbs sampler for estimated marginal and regression functions as well as other diagnostic checks are deferred to Appendix G. For the Bayesian methods, the posterior mean denoted by \hat{f} , is our estimator of the unknown regression function f and its pointwise 95% credible interval is obtained by

constructing $U(x)$ and $L(x)$ such that

$$\Pi_n\{f(x) \in [L(x), U(x)] \mid D_n\} = 0.95.$$

We also consider simultaneous credible bands centered at the posterior mean \hat{f} with level $\gamma \in (0, 1)$,

$$\text{CB}_n(\gamma) = \left\{f : \|f - \hat{f}\|_\infty \leq r\right\},$$

where the half length r is chosen so that posterior probability of f falling into the credible band is γ ,

$$\Pi_n\{f \in \text{CB}_n(\gamma) \mid D_n\} = \gamma.$$

A detailed construction of $\text{CB}_n(\gamma)$ can be found in Appendix F.

		δ^2					
n	Method	0.01	0.2	0.4	0.6	0.8	1
100	GPEV _a	0.58 (0.43)	1.82 (1.23)	3.89 (4.00)	4.64 (3.81)	5.88 (5.55)	6.31 (5.02)
	GPEV _f	0.55 (0.41)	1.85 (1.22)	3.20 (2.83)	4.24 (3.19)	5.54 (5.21)	5.82 (4.96)
	GPEV _n	0.60 (0.45)	4.82 (2.35)	10.98 (4.91)	15.29 (6.07)	19.26 (7.78)	20.98 (9.36)
	GP	3.29 (0.31)	6.11 (1.39)	9.35 (2.34)	12.00 (2.99)	14.80 (3.54)	16.81 (3.94)
	decon	1.18 (1.00)	5.07 (2.46)	10.46 (3.79)	14.72 (4.08)	18.25 (3.96)	20.59 (3.52)
		δ^2					
n	Method	0.001	0.005	0.01	0.1	0.5	1
500	GPEV _a	0.11 (0.04)	0.12 (0.04)	0.13 (0.04)	0.37 (0.21)	1.69 (1.20)	3.35 (3.27)
	GPEV _f	0.10 (0.04)	0.11 (0.04)	0.12 (0.04)	0.35 (0.19)	1.59 (1.04)	3.94 (5.76)
	GPEV _n	0.11 (0.04)	0.12 (0.05)	0.14 (0.05)	1.51 (0.44)	12.09 (2.02)	20.38 (4.12)
	GP	1.78 (0.08)	1.80 (0.09)	1.80 (0.09)	2.57 (0.26)	8.45 (1.08)	14.37 (1.60)
	decon	0.35 (0.21)	0.38 (0.26)	0.38 (0.26)	1.14 (0.46)	9.48 (1.55)	18.03 (1.53)

Table 1: Averaged mean squared errors (AMSE) defined as $\mathbb{E}[K^{-1} \sum_{k=1}^K \{\hat{f}(t_k) - f(t_k)\}^2]$ ($\hat{f}(\cdot)$ denotes the estimator of f , $\mathbb{E}(\cdot)$ denotes taking average over replicates) on an evenly spaced test grid $\{t_1, \dots, t_K\}$ of size $K = 100$ over $[-3, 3]$ with standard errors ($\times 10^2$) over 50 replicated data sets of sizes $n = 100, 500$.

Table 1 summarizes out-of-sample prediction results for all methods in terms of the averaged mean squared errors (AMSE) given different values of δ^2 . The results show GPEV_f performs the best among compared methods. However, we observe that the performance of GPEV_a is very close to that of GPEV_f for all combinations of n and δ^2 . This observation suggests that the approximation error of the proposed GP surrogate to the original GP is almost negligible in out-of-sample prediction despite the magnitude of the measurement error. We now investigate the performance of considered models in detail against the noise level of measurement errors. When δ^2 is small, all considered methods are robust to the measurement error except the GP model, implying that ignoring measurement errors could compromise the estimation significantly even though covariates are mildly contaminated.

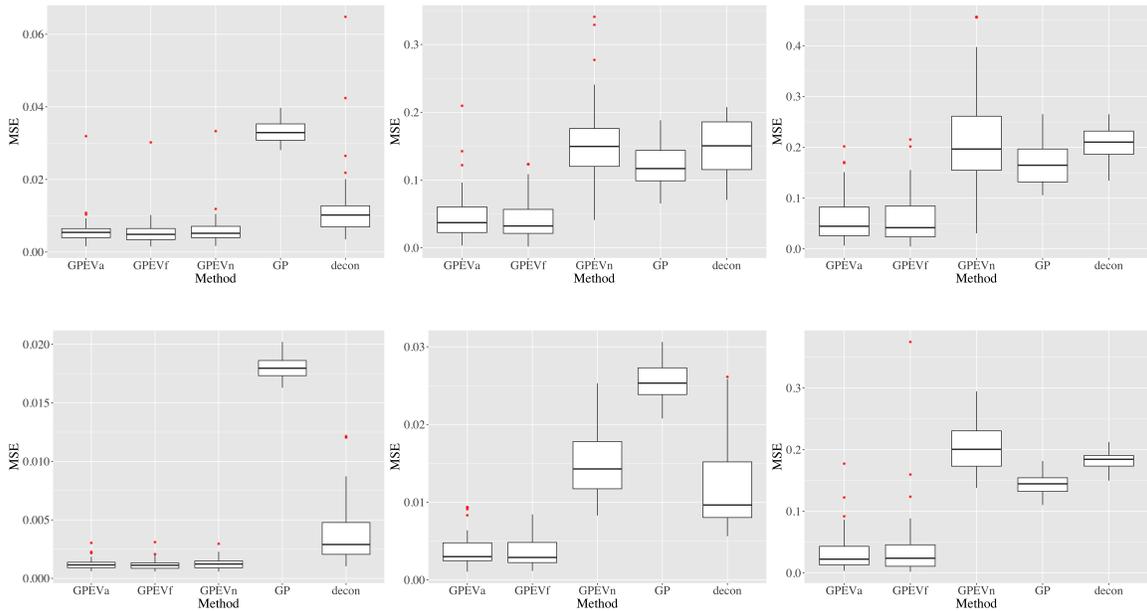


Figure 1: Boxplots of mean squared values for $f(x)$ when $n = 100$ (the top row) and $n = 500$ (the bottom row) over 50 replicated data sets. For $n = 100$, set $\delta^2 = 0.01$ (left panel), $\delta^2 = 0.6$ (middle panel) and $\delta^2 = 1$ (right panel); for $n = 500$, set $\delta^2 = 0.005$ (left panel), $\delta^2 = 0.1$ (middle panel) and $\delta^2 = 1$ (right panel). In each panel the methods displayed from left to right are GPEV_a , GPEV_f , GPEV_n , GP and decon.

As δ^2 increases, the AMSEs for GPEV_a and GPEV_f increase only by a relatively small margin, whereas other methods have suffered a drastic increase in AMSE values. For instance, the AMSE values obtained by GP and decon models are three times greater than those by GPEV_a and GPEV_f when $\delta^2 \geq 0.6$. The robustness of GPEV-based models to large measurement errors empirically justifies our theoretical claim that a DPMM prior is necessary for recovering the covariate density and thus allows the regression recovery to be robust to measurement errors.

Similar results can be also observed from the boxplots of mean squared error (MSE) values in Figure 1. The increasing amount of MSEs for all methods along with δ^2 is due to that the true covariate density p_0 turns harder to recover when the errors in covariates become more disturbing. On the other hand, this implies that increasing the number of replicates can improve the performance significantly of the Bayesian estimator in practice. Beyond the investigation on MSEs, the model fitting result in Figure 2 graphically displays the prediction performance of compared methods over various values of δ^2 . In particular, one can observe that when δ^2 increases the performance of decon and GP deteriorates fast and both fail to recover the curvature of the true function. On the contrary, even when $\delta^2 = 1$, the posterior mean of GPEV_a aligns with the true curve closely and its 95% pointwise credible interval contains the whole true function. A wider credible interval for larger values of δ^2 is expected due to an increasing amount of uncertainty in retrieving the covariate density. Overall, the GPEV-based models stand out among other competitors in terms of prediction.

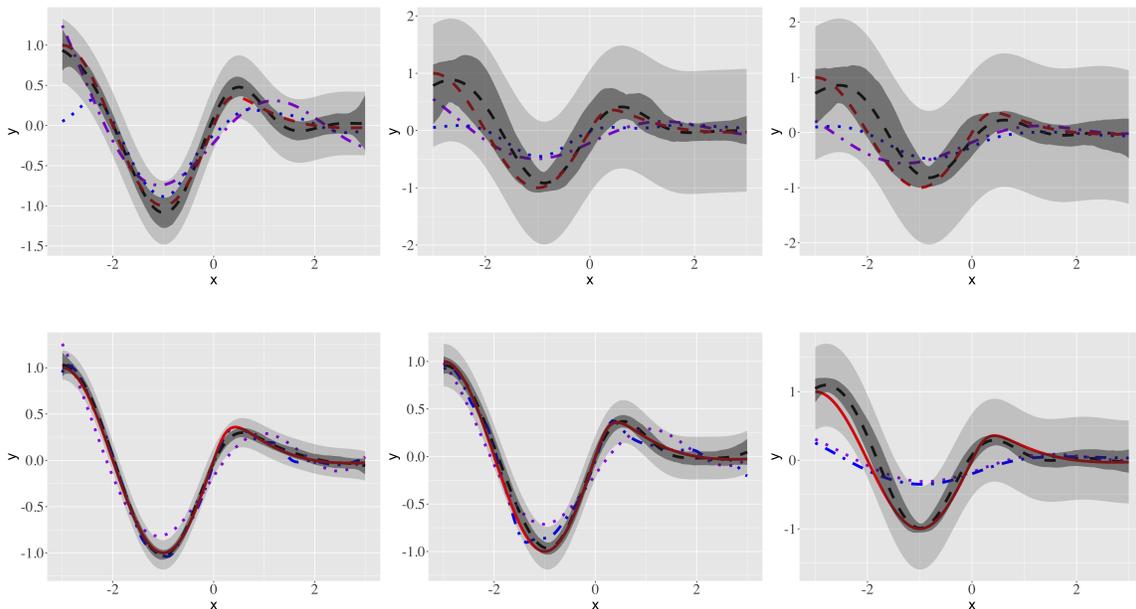


Figure 2: Out-of-sample predictions of $f(x)$ with $n = 100$ (first row) and $n = 500$ (second row). For $n = 100$, set $\delta^2 = 0.01$ (left panel), $\delta^2 = 0.6$ (middle panel) and $\delta^2 = 1$ (right panel). For $n = 500$, set $\delta^2 = 0.005$ (left panel), $\delta^2 = 0.1$ (middle panel) and $\delta^2 = 1$ (right panel). In each plot, the red solid line is the true function, the black dashed line stands for the prediction based on GPEV_a , the blue dot-dashed line based on decon , the purple dotted line based on GP . The darker and the lighter shades are the pointwise and simultaneous 95% credible intervals obtained with GPEV_a .

A careful inspection of our theory implies that placing a component-wise normal prior on the covariate results in a slower posterior contraction rate in recovering both the true covariate density and the true regression function. This is supported by the empirical observation that much larger AMSE values are obtained by GPEV_n when δ^2 becomes large. Additional investigation regarding the recovery of covariate can be found in Figure 8 of Appendix G. By comparing the posterior density function of covariates based on GPEV_a and GPEV_n , one can see that a component-wise normal prior is unable to identify the true covariate from the contaminated observations. In Figure 8 in Appendix G, we display a few examples of the posterior marginal density function obtained by GPEV_a and GPEV_n , when $n = 500$ and $\delta^2 = 0.001, 0.1, 0.5$, respectively. Recall that the true covariate distribution is $\text{Unif}[-3, 3]$ in our simulation setting. When $\delta^2 = 0.001$, both GPEV_a and GPEV_n recover the true underlying density quite well, indicating that a DPMM prior on the marginal density has a similar performance with independent normal priors on the locations. When $\delta^2 = 0.1, 0.5$, the performance of both methods deteriorate dramatically in estimating the marginal density function in Figure 8, which is expected since the best obtainable rate of convergence becomes slower with large δ^2 . However, as δ^2 increases, one can still notice an improvement in estimating the covariate density using GPEV_a . The posterior density function of the covariate obtained from GPEV_n is more erratic, suggesting that assigning

independent normal priors on locations results in a poor recovery when the measurement errors are more significant. In addition, we compared these two methods in terms of the averaged mean squared error in recovering true locations in all cases of sample sizes in Table 3 in Appendix G, which tells a similar story regarding the performance of GPEV_a and GPEV_n . In each iteration of the Gibbs sampler, we update the covariate values and update the rest of parameters upon those, a better performance of recovering the true locations leads to a better result in updating other parameters, which partially explains the outperformance of GPEV_a in estimating the regression curve.

In addition to a comparable performance in prediction, GPEV_a is more computationally efficient than GPEV_f . GPEV_a avoids repeated computation of the inverse of covariance matrix associated with a full GP, at a price of updating hyperparameters of a relatively moderate size (a fraction of sample size) related to Fourier basis functions. This is particularly beneficial for the errors-in-variables problem as covariates are treated as unknown parameters and both covariates and the covariance matrices are updated in each iteration. Also, for GP models, the choice of covariance kernel and treatment to the associated hyperparameters play an important role in the mixing of the Markov chains (Murray and Adams, 2010). To implement GPEV_a , we consider a squared exponential covariance kernel associated with a bandwidth parameter, denoted by λ , which is treated as an unknown parameter. The conjugate form of its spectral density induces a closed-form conditional of the bandwidth parameter λ based on the Fourier basis representation. Figure 7 in Appendix G shows the trace plots of posterior samples of the bandwidth parameter λ , where one can observe that the mixing of the chain based on GPEV_a is much better than that based on GPEV_f . We also remark that auto-correlation of the Markov chains obtained from GPEV_a and GPEV_f are similar, which is from the boxplots of the effective sample sizes (ESS) of estimated function values based on GPEV_a and GPEV_f over training data points in Figure 9 of Appendix G. To gauge the computational efficiency of GPEV-based methods, we report that the computation time of GPEV_a , GPEV_n , GPEV_f for a single Markov chain iteration when $n = 500$ are 0.025, 0.022, 0.197 second separately, on an 8-Core Intel Core i9 computer with 32 GB RAM. It is evident that implementing the proposed GP surrogate improves the computation speed substantially and the improvement becomes more pronounced as the sample size increases. In conclusion, GPEV_a stands out as a more robust and computationally efficient method for tackling the errors-in-variables regression problem.

5. A Case Study

We re-analyzed the real data set studied in Berry et al. (2002) using the proposed GPEV method. As described in Berry et al. (2002), the data set was collected from a randomized study where the actual content is not allowed to be disclosed. Basically, the data contains a treatment group and a control group. In each group we have the surrogate measurement W evaluated at baseline, and the observed response Y evaluated at the end of study. We know smaller values of W and Y indicate a worse case in the study. As discussed in Berry et al. (2002), the quantity of interest is the change from the baseline $\Delta(X) = f(X) - X$. We assume a normal zero-mean measurement error with two choices of variance, 1) a fixed variance $\delta^2 = 0.35$, adopting the estimated value from the study; and 2) an unknown variance δ^2 which will be treated as an unknown parameter in the model. To implement

the GPEV_a model, we choose $N = 60$ based on the simulation results, and consider an exponential prior $\exp(\lambda_0)$ with hyperparameter $\lambda_0 = 1.5$ on the bandwidth parameter λ associated with the squared exponential kernel. Besides, we treat the response error variance σ^2 as unknown and we consider an objective prior for σ^2 , namely, $\Pi(\sigma^2) \propto 1/\sigma^2$, allowing the data to update the parameter. To update σ^2 , we refer to Step 7 of the Gibbs sampler in Appendix F. For both cases of δ^2 , we ran the Gibbs sampler with 1,500 iterations with the first 1,000 being discarded as a burn-in. We consider the posterior mean as our Bayesian estimator and report the 95% pointwise credible interval.

Figure 3 shows the prediction results of the changes by GPEV_a with $\delta^2 = 0.35$. We observe that for both the treatment and control groups, the change from the baseline increases first and then decreases as the true baseline score increases, which coincides with the results presented in Berry et al. (2002). In Figure 4, we compare the estimated changes by GPEV_a with fixed δ^2 and unknown δ^2 for both groups. We observe that for both treatment and control groups, an objective prior on δ^2 results in a similar prediction of $\Delta(X)$ as that with fixed δ^2 . Since the true regression and true covariates are unknown, we compute the mean squared error as $\text{MSE} = \frac{1}{n} \sum_{i=1}^n \{(\hat{f}(\hat{x}_i) - y_i)^2\}$, where $\{y_i\}$ are observed responses and $\{\hat{x}_i\}$ denote the posterior mean of covariates obtained by GPEV_a . The defined MSE value accounts for randomness in the responses and errors in estimating the regression function and the covariates. Although this MSE value does not directly reflect the accuracy of predicting the true function, it provides some insights when comparing the performance of various methods. For the treatment group, the MSE values for GPEV_a with $\delta^2 = 0.35$, GPEV_a with unknown δ^2 and the decon method are 4.35, 4.54, and 8.23 separately; and for the control group, the MSE for the three competitors are 3.87, 4.35 and 5.18, respectively. Theoretical results have shown that with relatively large δ^2 , all methods may obtain an extremely slow rate of convergence (Fan, 1991; Fan and Truong, 1993), which explains the large MSE values. However, MSE values for GPEV_a are smaller than those of decon for both control and treatment data, despite of knowledge of the measurement error variance, showing a superior performance to decon in the real data example. The diagnostic results show the mixing of Markov chains for $\{w_j, s_j, x_j\}$ are good in both scenarios for the GPEV_a model, refer to Figure 10 in Appendix G for more details on trace plots and density plots of posterior samples of selected subsets of $\{a_i, \omega_j, s_j\}$.

6. Discussion

The article revisits error-in-variables regression problem from a Bayesian framework and addresses two fundamental challenges. Theoretical guarantees on the convergence of the posterior are established for the first time in a Bayesian framework. More specifically, optimal rates of posterior convergence are obtained simultaneously for the regression function as well as the covariate density. From a computational perspective, we provide a new Gaussian process approximation which facilitates posterior sampling and avoids costly matrix operations associated with a standard Gaussian process framework.

In addition to showing weak convergence of the approximate Gaussian process to the original ones, we have also shown that when it is employed as a prior process, the resultant posterior maintains same contraction results as those of original GPs in recovering both the regression curve and covariate density function in EIV problem. As the procedure can be

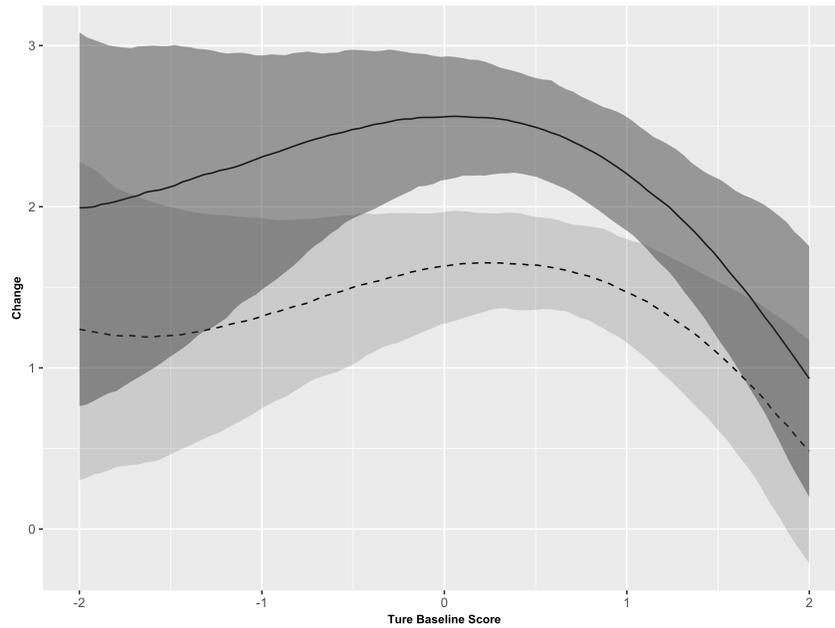


Figure 3: Estimate of $\Delta(X)$ on an evenly spaced test grid over $[-2, 2]$ with $\delta^2 = 0.35$. The solid line indicates the treatment group with the darker shade as its 95% pointwise credible interval and the dashed line indicates the control group with the lighter shade as its 95% pointwise credible interval.

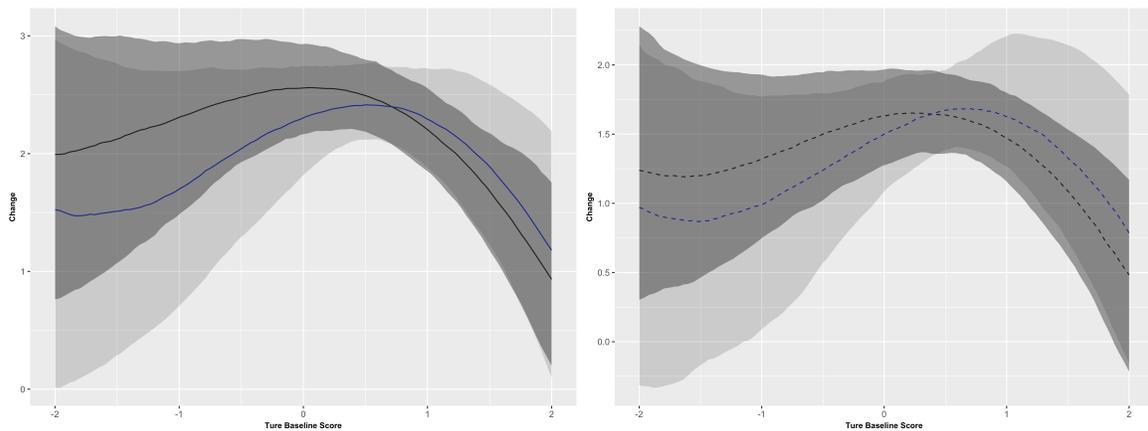


Figure 4: Estimate of $\Delta(X)$ on an evenly spaced test grid over $[-2, 2]$ with different choices of δ^2 . The left panel indicates the treatment group and the right panel indicates the control group. In both panels, the black lines stand for the posterior mean with $\delta^2 = 0.35$ with the dark shades as 95% pointwise credible intervals, the blue lines stand for the posterior mean with unknown δ^2 with the light shades as 95% pointwise credible intervals.

easily generalized to other nonparametric setting, this result implies some statistical guarantee of the predictive performance of projection technique with the random Fourier features under a Bayesian framework, which is a new addition to the theoretical investigations of the random Fourier features.

Acknowledgments

Pati's research was partially supported by NSF DMS (1613156, 1854731, 1916371) and Yang's research was partially supported by NSF DMS 1810831. The research of Wang and Carroll was supported by a grant from the National Cancer Institute (U01-CA057030).

Appendix A. Technical Results

Section A.1 introduces notations used throughout the rest of the paper and some background knowledge on the Gaussian process prior and its associated reproducing kernel Hilbert space. Section A.2 collects all auxiliary results used to prove Theorem 1.

A.1 Notations and Backgrounds

We first introduce some notations used in the proofs. Denote by E_X the marginal expectation with respect to random variable X ; denote $P_{X,Y}^{f,p}$ as the probability measure of random pair (X, Y) which has a joint density denoted by (f, p) . Let $*$ denote the convolution, say, for two functions f and g we define $f * g(\cdot) = \int f(\cdot - t)g(t) dt$. Denote the Kullback–Leibler divergence between functions f and g with respect to the Lebesgue measure μ by $KL(f, g) = \int f \log(f/g) d\mu$ and denote the second moment of the Kullback–Leibler divergence by $V(f, g) = \int f \{\log(f/g)\}^2 d\mu$. Define the ϵ -Kullback–Leibler neighborhood around f_0 as $B_{f_0}(\epsilon) = \{f : KL(f_0, f) \leq \epsilon^2, V(f_0, f) \leq \epsilon^2\}$. We also define the Hellinger distance between two densities f and g as $H(f, g) = \{\int (\sqrt{f} - \sqrt{g})^2 d\mu\}^{1/2}$. And define the L_2 -norm as $\|f\|_2 = \sqrt{\int f(x)^2 dx}$. Let $\mathbf{1}_C(\cdot)$ denote the indicator function on any set $C \subset \mathbb{R}$. For two sets A, B , we denote their Cartesian product by $A \otimes B$, the set contains all pairs (x, y) , where $x \in A$ and $y \in B$. For two positive sequences a_n, b_n , we write $a_n = O(b_n)$ if $\lim_{n \rightarrow \infty} (a_n/b_n) = c$ for some constant $c > 0$, and $a_n = o(b_n)$ if $\lim_{n \rightarrow \infty} (a_n/b_n) = 0$. At last, we define a k th order kernel function $K(\cdot)$ that satisfies,

$$\begin{aligned} \int K(u) du &= 1, & \int K^2(u) du &< \infty, & \int u^{[\beta]} K(u) du &\neq 0, \\ \int u^{i-1} K(u) du &= 0, & \text{for } i &= 1, \dots, [\beta] - 1, & \beta &\geq 2. \end{aligned} \tag{11}$$

Now we briefly recall the definition of the reproducing kernel Hilbert space of a Gaussian process prior; a detailed review can be found in van der Vaart and van Zanten (2008). A Borel measurable random element W with values in a separable Banach space denoted by $(\mathbb{B}, \|\cdot\|)$, for instance, the space of continuous functions $C[0, 1]$, is called Gaussian if the random variable b^*W is normally distributed for any element $b^* \in \mathbb{B}^*$, the dual space of \mathbb{B} . The reproducing kernel Hilbert space \mathbb{H} attached to a zero-mean Gaussian process W is defined as the completion of the linear space of functions $t \mapsto EW(t)H$ relative to the inner product

$$\langle E(W(\cdot)H_1); E(W(\cdot)H_2) \rangle_{\mathbb{H}} = E(H_1H_2),$$

where H, H_1 and H_2 are finite linear combinations of the form $\sum_i a_i W(s_i)$ with $a_i \in \mathbb{R}$ and s_i in the index set of W .

Let $W = (W_t : t \in \mathbb{R})$ be a Gaussian process associated with a squared exponential covariance kernel, which is

$$C(t, t') = e^{-(t-t')^2}.$$

The spectral measure m_w of W is absolutely continuous with respect to the Lebesgue measure λ on \mathbb{R} with the Radon-Nikodym derivative given by

$$\frac{dm_w}{d\lambda}(x) = \frac{1}{(2\pi)^{1/2}} e^{-x^2/4}.$$

Define a scaled Gaussian process $W^a = (W_{at} : t \in [0, 1])$, viewed as a map in $C[0, 1]$. Let \mathbb{H}^a denote the reproducing kernel Hilbert space of W^a , with the corresponding norm $\|\cdot\|_{\mathbb{H}^a}$. The unit balls in reproducing kernel Hilbert space and in the Banach space are denoted by \mathbb{H}_1^a and \mathbb{B}_1 , respectively.

Next we describe the construction of the sieve \mathcal{P}_n on the parameter space of (f, p) , the parameter space of p . For fixed constants $m, \underline{\sigma}, \bar{\sigma} > 0$ and integer $H \geq 1$. Let

$$\mathcal{F} = \left\{ p_{F, \tilde{\sigma}} = \phi_{\tilde{\sigma}} * F : F = \sum_{h=1}^{\infty} \pi_h \delta_{z_h}, z_h \in [-m, m], h \leq H, \sum_{h>H} \pi_h < \epsilon_n, \underline{\sigma} \leq \tilde{\sigma} < \bar{\sigma} \right\}.$$

Set $\mathcal{P}_n = \tilde{B}_n \otimes \mathcal{F}$, where $\tilde{B}_n = B_n \cap \mathcal{A}$ with $B_n = M_n \mathbb{H}_1^{a_n} + \epsilon_n \mathbb{B}_1$ and \mathcal{A} as in Assumption 3.

A.2 Auxiliary Results

In this section, we collect auxiliary results that are needed for the proofs of main theorems. The proofs of Lemmata 5-8 are deferred to Appendix E.

Lemma 5 *Suppose Assumptions 1, 2, 3 and 4 hold, by taking $M_n = a_n = \epsilon_n^{-1/\beta}$, $H \lesssim n \epsilon_n^2$, $m^{\tau_1} \lesssim n$, $\underline{\sigma} \lesssim n^{-1/2\tau_2}$ and $\bar{\sigma}^{2\tau_3} \lesssim e^n$, we have $\Pi(\mathcal{P}_n^c) \leq e^{-n\epsilon_n^2}$ with $\epsilon_n = n^{-\beta/(2\beta+1)}(\log n)^t$, where $t = \max\{(2 \vee q)\beta/(2\beta+1), 1\}$.*

Lemma 6 *For model (1), \hat{p}_n and \hat{f}_n defined in Equations (3) and (4) and $f, p \in \mathcal{P}_n$ for any small constant $\epsilon_0 > 0$,*

$$P_{W,X}^p(\|\hat{p}_n - p\|_{\infty} > \epsilon_0) \leq e^{-C_1 n \epsilon_0 h_n^2}, \quad (12)$$

$$P_{W,X}^p(\|\hat{p}_n - p\|_1 > \epsilon_n) \leq e^{-n\epsilon_n^2}, \quad (13)$$

$$P_{Y,W,X}^{f,p}(\|\hat{f}_n \hat{p}_n - f p\|_1 > \epsilon_n) \leq e^{-n\epsilon_n^2}, \quad (14)$$

where $h_n \asymp \epsilon_n^{1/\beta}$, $\epsilon_n = n^{-\beta/(2\beta+1)}(\log n)^t$ with $t = \max\{(2 \vee q)\beta/(2\beta+1), 1\}$ and some constant $C_1 > 0$.

Lemma 7 *Suppose Assumptions 2, 3 and 4 hold, then $\Pi\{KL(p_0, \epsilon_n)\} \geq e^{-n\epsilon_n^2}$, where $\epsilon_n = n^{-\beta/(2\beta+1)}(\log n)^t$ with $t = \max\{(2 \vee q)\beta/(2\beta+1), 1\}$.*

Lemma 8 *Under the conditions in Theorem 1 and suppose Lemma 7 hold, for sufficiently large n ,*

$$\Pi\{B_{(f_0, p_0)}(\epsilon_n)\} \geq e^{-n\epsilon_n^2},$$

where $B_{(f_0, p_0)}(\epsilon_n)$ is defined in (16) and $\epsilon_n = n^{-\beta/(2\beta+1)}(\log n)^t$ with $t = \max\{(2 \vee q)\beta/(2\beta+1), 1\}$.

Lemma 9 (Theorem 7.3 in Bousquet 2003) Suppose \mathcal{G} is a countable set of functions $g : \mathcal{X} \rightarrow \mathbb{R}$ and assume all functions $g \in \mathcal{G}$ are measurable, squared-integrable and satisfy $E\{g(X_k)\} = 0$. Assume $\sup_{g \in \mathcal{G}} \text{ess sup } g$ is bounded and define $Z = \sup_{g \in \mathcal{G}} \sum_{k=1}^n g(X_k)$. Let $\sigma_{\mathcal{G}}$ be a positive real number such that $n\sigma_{\mathcal{G}}^2 \geq \sum_{k=1}^n \sup_{g \in \mathcal{G}} E\{g^2(X_k)\}$, then for all $t > 0$ with $\nu = n\sigma_{\mathcal{G}}^2 + 2E(Z)$, we have

$$P\left\{Z \geq E(Z) + (2t\nu)^{1/2} + \frac{t}{3}\right\} \leq e^{-t}.$$

Lemma 10 (Borell's inequality in Adler 1990) Let $\{f(x) : x \in [0, 1]\}$ be a centered Gaussian process and denote $\|f\|_{\infty} = \sup_{x \in [0, 1]} f(x)$ and $\sigma_f^2 = \sup_{x \in [0, 1]} E\{f^2(x)\}$. Then $E(\|f\|_{\infty}) < \infty$ and for any $t > 0$,

$$P(|\|f\|_{\infty} - E\|f\|_{\infty}| > t) \leq 2e^{-\frac{1}{2}t^2/\sigma_f^2}.$$

Appendix B. Proof of Theorem 1

In this section, we provide the proof of Theorem 1. Given ϵ_n, ϵ'_n in Theorem 1, define $U_n = \{f, p : \|f - f_0\|_1 < M\epsilon_n, \|p - p_0\|_1 < M\epsilon'_n\}$, our goal is to show $\Pi_n(U_n^c | Y_{1:n}, W_{1:n}) \rightarrow 0$ almost surely in G_{f_0, p_0} as $n \rightarrow \infty$. To that end, note that

$$\begin{aligned} & \Pi_n(U_n^c | Y_{1:n}, W_{1:n}) \\ & \leq \Pi_n(f, p : \|f - f_0\|_1 > M\epsilon_n | Y_{1:n}, W_{1:n}) + \Pi_n(p : \|p - p_0\|_1 > M\epsilon'_n | W_{1:n}) := S_1 + S_2. \end{aligned} \quad (15)$$

It suffices to estimate S_1 and S_2 in the preceding separately. We shall analyze term S_1 in detail and only provide a brief discussion about bounding term S_2 as it can be considered as an immediate application of existing results.

Bounding term S_1 in Equation (15). Define the ϵ_n -Kullback–Leibler neighborhood around (f_0, p_0) as

$$B_{f_0, p_0}(\epsilon_n) = \left\{ \int g_{f_0, p_0} \log \frac{g_{f_0, p_0}}{g_{f, p}} \leq \epsilon_n^2, \quad \int g_{f_0, p_0} \left(\log \frac{g_{f_0, p_0}}{g_{f, p}} \right)^2 \leq \epsilon_n^2 \right\}. \quad (16)$$

The following Theorem provides sufficient conditions showing $S_1 \rightarrow 0$ almost surely as $n \rightarrow \infty$. A sketch of the proof is provided in the following.

Theorem 4 (Contraction Theorem) Consider model (1) and under the conditions in Theorem 1, let $\mathcal{U}_n = \{\|f - f_0\|_1 > M\epsilon_n\}$. If there exist a sequence of $\epsilon_n \rightarrow 0$ and $n\epsilon_n^2 \rightarrow \infty$ and a sequence of sieve $\mathcal{P}_n \subset \mathcal{P}$, and a sequence of test functions $\phi_n = \mathbb{1}_{\{\|\hat{f}_n - f_0\|_1 > (M-1)\epsilon_n\}}$ satisfying the following conditions,

$$G_{f_0, p_0} \phi_n \leq e^{-(C+4)n\epsilon_n^2}, \quad \sup_{(f, p) \in \mathcal{P}_n \cap \mathcal{U}_n} G_{f, p} (1 - \phi_n) \leq e^{-(C+4)n\epsilon_n^2}, \quad (17)$$

$$\Pi\{B_{f_0, p_0}(\epsilon_n)\} \geq e^{-n\epsilon_n^2}, \quad (18)$$

$$\Pi(\mathcal{P}_n^c) \leq e^{-(C+4)n\epsilon_n^2}, \quad (19)$$

for some constant $C > 0$, then $\Pi_n(\mathcal{U}_n^c | Y_{1:n}, W_{1:n}) \rightarrow 0$ almost surely in G_{f_0, p_0} , for the constant M same as in Theorem 1.

Proof (Sketch) Define the set

$$C_n = \left\{ \int \frac{\prod_{j=1}^n g_{f,p}(Y_j, W_j)}{\prod_{j=1}^n g_{f_0,p_0}(Y_j, W_j)} d\Pi(f) d\Pi(p) \geq e^{-(C+3)n\epsilon_n^2} \Pi\{B_{f_0,p_0}(\epsilon_n)\} \right\}.$$

Under the conditions in Theorem 1, from Lemma 8.1 in Ghosal et al. (2000), it follows $G_{f_0,p_0}(C_n) \geq 1 - 1/(C'n\epsilon_n^2)$, for some constant $C' > 0$. Hence for any sequence of test functions $\{\phi_n\}$,

$$\begin{aligned} \Pi_n(\mathcal{U}_n^c | Y_{1:n}, W_{1:n}) &\leq G_{f_0,p_0}\phi_n + G_{f_0,p_0}\{(C_n^c) + G_{f_0,p_0}\Pi(\mathcal{P}_n^c | Y_{1:n}, W_{1:n})\mathbf{1}_{C_n}\} \\ &\quad + G_{f_0,p_0}\{\Pi(\mathcal{U}_n \cap \mathcal{P}_n | Y_{1:n}, W_{1:n})(1 - \phi_n)\mathbf{1}_{C_n}\}. \end{aligned}$$

We suppress the term ‘‘almost surely’’ in the following argument. According to Conditions (18) and (19), the third term in the above display goes to 0. From Conditions (17) and (18), the first and the fourth terms in the preceding go to 0. Then we have shown $S_1 \rightarrow 0$ as $n \rightarrow \infty$. \blacksquare

We now verify three conditions in Theorem 4 under the conditions in Theorem 1, based on the auxiliary results summarized in Appendix A.2. The main steps are

- Condition (18) of Theorem 4: Follows from Lemma 8 under the conditions of Theorem 1.
- Condition (19) of Theorem 4: Follows from Lemma 5 under the conditions of Theorem 1.
- Condition (17) of Theorem 4: For model (1), recall \hat{p}_n and \hat{f}_n defined in Equations (3) and (4), and $f, p \in \mathcal{P}_n$, it suffices to estimate $P_{Y,W,X}^{f_0,p_0}(\|\hat{f}_n - f_0\|_1 > \epsilon_n)$ and $P_{Y,W,X}^{f,p}(\|\hat{f}_n - f\|_1 > \epsilon_n)$. Following a similar line of argument in Meister (2009), for any marginal density p_0 satisfying Assumption 2, for any $p \in \mathcal{P}_n \cup p_0$ define $\Delta p = (\hat{p}_n - p)/p$, then for any $f \in \mathcal{P}_n \cup f_0$ we have

$$\|\hat{f}_n - f\|_1 \leq \frac{|\hat{f}_n \hat{p}_n - fp|}{|p|} \left(\frac{|\Delta p|}{|\Delta p + 1|} + 1 \right) + |f| \frac{|\Delta p|}{|\Delta p + 1|}.$$

By Assumption 2, p_0 is lower-bounded by some constant $B^{-1} > 0$. Then applying the Equation (12) in Lemma 6, for any constant $\epsilon_0 > 0$ we have $\|\hat{p}_n - p\|_\infty < \epsilon_0$ with probability at least $1 - e^{-n\epsilon_0 h_n^2}$. Define the set $\mathcal{A}_\epsilon = \{p : \|\hat{p}_n - p\|_\infty < \epsilon_0\}$. Thus for $p \in \mathcal{P}_n \cap \mathcal{A}_\epsilon$, $\|p - p_0\|_\infty \leq \|\hat{p}_n - p_0\|_\infty + \|\hat{p}_n - p\|_\infty \leq 2\epsilon_0$. Then $\|p\|_\infty \geq \|p_0\|_\infty - \|p - p_0\|_\infty \geq B_1$, for some constant $B_1 > 0$ by choosing $\epsilon_0 < B^{-1}/2$. Thus for $f \in \mathcal{P}_n \cup f_0$ and $p \in \mathcal{P}_n \cap \mathcal{A}_\epsilon$, we have

$$\|\hat{f}_n - f\|_1 \leq \frac{1}{B_1} \|\hat{f}_n \hat{p}_n - fp\|_1 \left(\left\| \frac{\Delta p}{\Delta p + 1} \right\|_\infty + 1 \right) + \|f\|_\infty \left\| \frac{\Delta p}{\Delta p + 1} \right\|_1. \quad (20)$$

Since $\|\Delta p\|_\infty \leq \epsilon_0/B_1$, choosing ϵ_0 such that $\epsilon_0/B_1 \leq 1/2$, then we have $\|\Delta p/(\Delta p + 1)\|_\infty \leq 1$ and $1/2 \leq \|\Delta p + 1\|_\infty \leq 3/2$ and therefore $1/\|\Delta p + 1\|_\infty \leq 2$. Thus we

have,

$$\left\| \frac{\Delta p}{\Delta p + 1} \right\|_1 \leq \frac{1}{\|\Delta p + 1\|_\infty \|p\|_\infty} \int_0^1 |\widehat{p}_n(x) - p(x)| dx \leq \frac{2}{B_1} \|\widehat{p}_n - p\|_1.$$

Similarly for $p = p_0 \in \mathcal{A}_\epsilon$, we bound $\|\Delta p / (\Delta p + 1)\|_1 \leq 2B \|\widehat{p}_n - p_0\|_1$. Combining the above results and the result in Equation (20), we obtain,

$$\begin{aligned} P(\|\widehat{f}_n - f\|_1 > \epsilon_n) &\leq P(\|\widehat{f}_n \cdot \widehat{p}_n - f \cdot p\|_1 > B_1 \epsilon_n / 4) \\ &\quad + P\{\|\widehat{p}_n - p\|_1 > B_1 \epsilon_n / (4\|f\|_\infty)\} \\ &\quad + P(\|\widehat{p}_n - p\|_\infty > \epsilon_0). \end{aligned}$$

Since we assume f_0 and $f \in \mathcal{P}_n$ are bounded, applying Lemma 6 verifies Condition (17).

Bounding term S_2 in Equation (15). To estimate S_2 , we apply an inversion inequality built upon a special kernel function (the *sinc* kernel) considered in Donnet et al. (2018), then apply the existing posterior contraction result for the direct density problem. Recall the Fourier transform of the error density $\widehat{\phi}_{\delta_n}(t) \asymp \delta_n e^{-\pi^2 \delta_n^2 t^2}$. Then with a careful inspection of the proof of Proposition 1 in Donnet et al. (2018), one can obtain the inversion inequality

$$\begin{aligned} \|p - p_0\|_2^2 &\lesssim \delta^{2\beta'} + \|\phi_{\delta_n} * p - \phi_{\delta_n} * p_0\|_1^2 \times \int_{|t| \leq 1/\delta} |\widehat{\phi}_{\delta_n}|^{-2} dt \\ &\lesssim \delta_n^{2\beta'} + \|\phi_{\delta_n} * p - \phi_{\delta_n} * p_0\|_1^2, \end{aligned} \tag{21}$$

where β' denotes the regularity level of p_0 . The last inequality in Equation (21) holds by choosing $\delta \asymp \delta_n$ and the fact that $\int_{|t| \leq 1/\delta} |\widehat{\phi}_{\delta_n}|^{-2} dt \lesssim (\delta_n/\delta) e^{2\pi^2 (\delta_n/\delta)^2} = O(1)$. Denote the observed density and the true density of W by $f_W = \phi_{\delta_n} * p$ and $f_{0W} = \phi_{\delta_n} * p_0$ separately. By Cauchy-Schwarz inequality, Equation (21) implies $\|p - p_0\|_1 \lesssim \max\{\delta_n^{\beta'}, \|f_W - f_{0W}\|_1\}$. Then under the Assumptions 2 and 4, for δ_n, ϵ'_n defined in Theorem 1, one can easily show $S_2 = o(1)$, by applying posterior contraction results for direct density estimation problem from the seminal work (Ghosal and van Der Vaart, 2001; Shen et al., 2013), which leads to the error rate $\epsilon'_n = n^{-\beta'/(2\beta'+1)} (\log n)^{t'}$ with $t' > (\gamma + 1/\beta')/(2 + 1/\beta')$ for some $\gamma > 2$ under Assumption 2.

Combining above results for terms S_1 and S_2 completes the proof of Theorem 1, and letting $\epsilon_n = n^{-\beta/(2\beta+1)} (\log n)^t$, $\epsilon'_n = n^{-\beta'/(2\beta'+1)} (\log n)^t$ with $t = \max\{(2 \vee q)\beta/(2\beta + 1), t'\}$ yields the desired rates in Theorem 1.

Appendix C. Proof of Theorem 2

In this section, we provide a proof of Theorem 2. In Part I, we shall first show the weak convergence of \widetilde{f}_N to the original Gaussian process; in Part II, we derive expressions of expectation and covariance of \widetilde{f}_N .

Part I. We now show \widetilde{f}_N weakly converges to the Gaussian process f . Based on Theorem 1.5.7 in van der Vaart and Wellner (1996), it suffices to show the marginal weak convergence and asymptotical tightness of \widetilde{f}_N .

First, we show the marginal weak convergence. For any finite sequence $\{x_1, \dots, x_k\}$ in $[0, 1]$ of size k where k is arbitrary positive integer, applying multivariate central limit theorem with the expectation and covariance of \tilde{f}_N derived in Part II, one can easily show that as $N \rightarrow \infty$,

$$\{\tilde{f}_N(x_1), \dots, \tilde{f}_N(x_k)\} \rightarrow N(0, c_{k,k}),$$

in distribution, where $c_{k,k} = (c_{ij})$ is a $k \times k$ covariance matrix with the (i, j) th element $c_{ij} = c(x_i, x_j)$.

Next, we show the asymptotic tightness of \tilde{f}_N . By definition, it suffices to verify the following three conditions. First, it is straightforward that $[0, 1]$ is totally bounded. Second, for any fixed $x_0 \in [0, 1]$, we shall show the tightness of $\tilde{f}_N(x_0)$. It is equivalent to show, by definition, for any $\epsilon > 0$, there exists a compact set K such that,

$$P\{\tilde{f}_N(x_0) \in K\} > 1 - \epsilon. \quad (22)$$

For any $x_0 \in [0, 1]$, we bound $\tilde{f}_N(x_0)$ from above as

$$|\tilde{f}_N(x_0)| \leq (2/N)^{1/2} \sum_{i=1}^N |a_i|,$$

where $a_j \stackrel{i.i.d.}{\sim} N(0, 1)$, $j = 1, \dots, N$. It is well-known that $|a_j|$ is a sub-gaussian random variable for $j = 1, \dots, N$. For any $t > 0$, we have

$$P\{|\tilde{f}_N(x_0)| \geq t\} \leq P\left\{ (2/N)^{1/2} \sum_{i=1}^N |a_i| \geq t \right\} \leq 2 \exp(-ct^2),$$

for some constant $c > 0$. For any $\epsilon > 0$, we choose $t = \{2 \log(1/\epsilon)\}^{1/2}$ and the set $K = \{|\tilde{f}_N(x_0)| \leq t\}$, then Equation (22) holds, thus we show the tightness of $\tilde{f}_N(x_0)$ for any $x_0 \in [0, 1]$.

Third, we shall show \tilde{f}_N is asymptotically uniformly equicontinuous with respect to the Euclidean norm, which is defined as $d(x, y) = |x - y|$, for $x, y \in \mathbb{R}$. It suffices to show that for any $\epsilon, \eta > 0$, there exists some $\delta > 0$ such that

$$\limsup_{N \rightarrow \infty} P\left\{ \sup_{d(x,y) < \delta} |\tilde{f}_N(x) - \tilde{f}_N(y)| > \epsilon \right\} < \eta. \quad (23)$$

Without loss of the generality, we assume $0 \leq x \leq y \leq 1$. Then

$$\begin{aligned} \sup_{|x-y| \leq \delta} |\tilde{f}_N(x) - \tilde{f}_N(y)| &= \sup_{|x-y| \leq \delta} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N a_j \{\cos(w_j x + u_j) - \cos(w_j y + u_j)\} \right| \\ &\leq \sup_{\theta \in [0, 1]^N} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N a_j w_j \sin(w_j \theta_j + u_j) \right| \delta. \end{aligned}$$

The inequality in the preceding holds by applying the mean-value theorem, namely, there exists a sequence $\{\theta_1, \dots, \theta_N\}$ such that we have $\theta_j \in (x, y)$ satisfying $\cos(w_j x + u_j) -$

$\cos(w_j y + u_j) = -w_j \sin(w_j \theta_j + u_j)(x - y)$ for $j = 1, \dots, N$. To show Equation (23), it remains to show

$$\limsup_{N \rightarrow \infty} P \left\{ \sup_{\theta \in [0,1]^N} \left| \frac{1}{N} \sum_{j=1}^N a_j w_j \sin(w_j \theta_j + u_j) \right| > \epsilon/\delta \right\} < \eta.$$

For any fixed $\lambda > 0$, recall that $w_j \stackrel{i.i.d.}{\sim} N(0, 2/\lambda)$, for $j = 1, \dots, N$. Then $(\lambda/2) \sum_{j=1}^N w_j^2$ is a chi-square random variable with the degree of freedom N . Let $K = c\sqrt{2/\lambda}$ with some constant $c \in (1, 2)$, then by the sub-exponential tail bound of a chi-square random variable, we have

$$P \left(\frac{1}{N} \sum_{j=1}^N w_j^2 > K \right) \leq \exp(-NK^2/8). \quad (24)$$

Now define the set $\mathcal{A} = \{w \in \mathbb{R}^N : (1/N) \sum_{j=1}^N w_j^2 \leq K\}$ and define the truncated variable $\tilde{w} = w \mathbb{1}_{\mathcal{A}}(w)$ over the set \mathcal{A} , the density function of \tilde{w} follows as $\Pi_{\tilde{w}}(\cdot) = N(\cdot; 0, 2/\lambda) \mathbb{1}_{\mathcal{A}}(\cdot) / P(w \in \mathcal{A})$. For any fixed N ,

$$\begin{aligned} & P \left(\sup_{\theta \in [0,1]^N} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N a_j w_j \sin(w_j \theta_j + u_j) \right| > \epsilon/\delta \right) \\ & \leq P \left(\left\{ \sup_{\theta \in [0,1]^N} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N a_j w_j \sin(w_j \theta_j + u_j) \right| > \epsilon/\delta \right\} \cap \mathcal{A} \right) + P(\mathcal{A}^c). \end{aligned} \quad (25)$$

By Equation (24), we see that $\lim_{N \rightarrow \infty} P(\mathcal{A}^c) = 0$. Now we estimate the first term on the right hand side of Equation (25). First, we consider

$$\begin{aligned} & P \left(\left\{ \sup_{\theta \in [0,1]^N} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N a_j w_j \sin(w_j \theta_j + u_j) \right| > \epsilon/\delta \right\} \cap \mathcal{A} \right) \\ & = E_{\tilde{w}, u} \left\{ P_{a|\tilde{w}, u} \left(\sup_{\theta \in [0,1]^N} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N a_j \tilde{w}_j \sin(\tilde{w}_j \theta_j + u_j) \right| > \epsilon/\delta \mid \tilde{w}, u \right) \right\}. \end{aligned}$$

With fixed $\{\tilde{w}_j, u_j\}$, define the set of indexes $J_N = \{1 \leq j \leq N : a_j \tilde{w}_j \geq 0\}$, then

$$\sup_{\theta \in [0,1]^N} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N a_j \tilde{w}_j \sin(\tilde{w}_j \theta_j + u_j) \right| \leq \frac{1}{\sqrt{N}} \left| \sum_{j'=1}^{J_N} a_{j'} \tilde{w}_{j'} \right|. \quad (26)$$

Then

$$\begin{aligned} & P_{a|\tilde{w}, u} \left(\sup_{\theta \in [0,1]^N} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N a_j \tilde{w}_j \sin(\tilde{w}_j \theta_j + u_j) \right| > \epsilon/\delta \mid \tilde{w}, u \right) \\ & \leq P_{a|\tilde{w}, u} \left(\left| \sum_{j'=1}^{J_N} a_{j'} \tilde{w}_{j'} \right| > \epsilon\sqrt{N}/\delta \mid \tilde{w}, u \right) \\ & \leq 2 \exp \left(- \frac{c N \epsilon^2}{2\delta^2 \sum_{j'=1}^{J_N} \tilde{w}_{j'}^2} \right) \leq 2 \exp \left(- \frac{c \epsilon^2}{2K\delta^2} \right), \end{aligned}$$

where $c > 0$ is some constant. The first inequality in the preceding applies the bound in Equation (26); the second inequality holds by applying the general Hoeffding's inequality for independent Gaussian random variables; the third inequality is due to the fact that $J_N \leq N$ for any fixed N . Therefore we have

$$\limsup_{N \rightarrow \infty} P\left(\left\{\sup_{\theta \in [0,1]^N} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N a_j w_j \sin(w_j \theta_j + u_j) \right| > \epsilon/\delta\right\} \cap \mathcal{A}\right) \leq 2 \exp\left(-\frac{c\epsilon^2}{2K\delta^2}\right).$$

Combining the above result with the bound for $P(\mathcal{A}^c)$, we show that for any $\epsilon, \eta > 0$, Equation (23) holds by choosing $\delta = \sqrt{c\epsilon^2/\{K \log(1/\eta)\}}$. Therefore, we have verified that \tilde{f}_N is asymptotically uniformly equicontinuous with respect to Euclidean norm. Then we complete the proof of weak convergence of \tilde{f}_N to the original Gaussian process.

Part II. Now we compute the expectation and covariance of \tilde{f}_N . For any $x \in \mathbb{R}$,

$$\begin{aligned} & E\{\tilde{f}_N(x)\} \\ &= (2/N)^{-1/2} \sum_{j=1}^N \int \int \frac{1}{2\pi} \cos(w_j x + s_j) \phi_c(w_j) dw_j ds_j \\ &= (2/N)^{-1/2} \sum_{j=1}^N \int \frac{1}{2\pi} \left\{ \cos(w_j x) \int_{-\pi}^{\pi} \cos s_j ds_j - \sin(w_j x) \int_{-\pi}^{\pi} \sin s_j ds_j \right\} \phi_c(w_j) dw_j = 0. \end{aligned}$$

For any $x, y \in \mathbb{R}$,

$$\begin{aligned} & \text{cov}\{\tilde{f}_N(x), \tilde{f}_N(y)\} \\ &= (2/N) \sum_{j=1}^N \text{cov}\{\cos(w_j x + s_j), \cos(w_j y + s_j)\} = 2E_{w,s} \cos(xw + s)^2 \\ &= \frac{1}{2\pi} \int_w \int_{-\pi}^{\pi} [\cos\{(x+y)w + 2s\} + \cos\{(x-y)w\}] \phi_c(w) ds dw \\ &= \frac{1}{2\pi} \int_w \left(\int_{-\pi}^{\pi} [\cos\{(x+y)w\} \sin(2s) + \sin\{(x+y)w\} \cos(2s)] ds + \cos\{(x-y)w\} \right) \phi_c(w) dw \\ &= \frac{1}{2\pi} \int_w \cos\{(x-y)w\} \phi_c(w) dw = c(x, y). \end{aligned}$$

We now have completed the proof of Theorem 2.

Appendix D. Proof of Theorem 3

To prove Theorem 3, it suffices to prove Theorem 4 by verifying Conditions (17), (18), and (19). As the line of argument is similar to the proof of Theorem 1, below we only highlight different steps. We use bold letters $\mathbf{a}, \boldsymbol{\omega}, \mathbf{s}$ to denote the vector form of parameters $\{a_j\}, \{\omega_j\}, \{s_j\}$, respectively.

First we define the sequence of sieves for parameters $\{a_j, \omega_j, s_j\}_{j=1}^N$ in \tilde{f}_N as

$$\mathcal{D}_N = \{(a_j, \omega_j, s_j) : a_j \in [-\sqrt{n}\epsilon_n, \sqrt{n}\epsilon_n], \omega_j \in [-n\epsilon_n^2, n\epsilon_n^2], s_j \in [0, 2\pi], j = 1, \dots, N.\}, \quad (27)$$

for any fixed positive integer $N > 0$ and ϵ_n is defined in Theorem 1. Denote $\tilde{\mathcal{D}}_N = \mathcal{D}_N \cap \tilde{\mathcal{A}}_N$, where $\tilde{\mathcal{A}}_N = \{(a_i, \omega_j, s_j) : \|\tilde{f}_N\|_\infty \leq A_0, j = 1, \dots, N\}$ for the same constant A_0 defined in Assumption 3.

We now start from verifying Condition (18), which suffices to show Lemma 8 for \tilde{f}_N . Define the set $\tilde{B}_{(f_0, p_0)}(\epsilon_n)$ as the KL-neighborhood of \tilde{f}_N centered at (f_0, p_0) , by replacing f with \tilde{f}_N in the definition of $B_{(f_0, p_0)}(\epsilon_n)$ in (16). It suffices to lower bound

$$\Pi\{\tilde{B}_{(f_0, p_0)}(\epsilon_n)\} \geq \int_{r_0}^{r_1} \Pi\{\tilde{B}_{(f_0, p_0)}(\epsilon_n) \mid A = a\} g(a) da, \quad (28)$$

for arbitrary fixed constants $r_0, r_1 > 0$. Then, following a same argument in the proof of Lemma 8 leads to

$$\Pi\{\tilde{B}_{(f_0, p_0)}(\epsilon_n) \mid A = a\} \geq \Pi(\|\tilde{f}_{A, N} - f_0\|_\infty \leq \epsilon_n \mid A = a) \Pi(KL(p_0, p) \leq \epsilon_n^2).$$

It suffices to lower bound the first probability term on the right hand side of the preceding. For simplicity, we use the shorthand $\tilde{f}_{a, N}$ for $\tilde{f}_{A, N}$ conditioning on $A = a$. For any $a > 0$, recall an original Gaussian process $f \sim \text{GP}(0, c^a(\cdot, \cdot))$ and apply the triangular inequality that $\|\tilde{f}_{a, N} - f_0\|_\infty \leq \|\tilde{f}_{a, N} - f\|_\infty + \|f - f_0\|_\infty \leq \|\tilde{f}_{a, N}\|_\infty + \|f\|_\infty + \|f - f_0\|_\infty$. Then one obtains

$$\begin{aligned} \Pi(\|\tilde{f}_{a, N} - f_0\|_\infty \leq \epsilon_n) &\geq \Pi(\|\tilde{f}_{a, N}\|_\infty + \|f\|_\infty + \|f - f_0\|_\infty \leq \epsilon_n) \\ &\geq \Pi(\|\tilde{f}_{a, N}\|_\infty \leq \epsilon_n/3) \Pi(\|f\|_\infty \leq \epsilon_n/3) \Pi(\|f - f_0\|_\infty \leq \epsilon_n/3). \end{aligned} \quad (29)$$

In van der Vaart and van Zanten (2009), it is shown that $\Pi(\|f\|_\infty \leq \epsilon_n/3) \gtrsim e^{-c n \epsilon_n^2}$ and $\Pi(\|f - f_0\|_\infty \leq \epsilon_n/3) \gtrsim e^{-c' n \epsilon_n^2}$ for some constants $c, c' > 0$. Then to bound (29), it suffices to bound

$$\begin{aligned} \Pi(\|\tilde{f}_{a, N}\|_\infty \leq \epsilon_n/3) &= \Pi\left(\left\|\left(2/N\right)^{1/2} \sum_{j=1}^N a_j \cos(w_j x + s_j)\right\|_\infty \leq \epsilon_n/3\right) \\ &\geq \Pi\left(\left(2/N\right)^{1/2} \sum_{j=1}^N |a_j| \leq \epsilon_n/3\right) \geq \Pi(\|\mathbf{a}\|^2 \leq \epsilon_n^2/18), \end{aligned}$$

where $\|\mathbf{a}\|^2 = \sum_{j=1}^N a_j^2 \sim \chi_N^2$, which is a centered chi-square random variable with the degree of freedom N . Further, we have

$$\begin{aligned} \Pi(\|\mathbf{a}\|^2 \leq \epsilon_n^2/18) &= \int_0^{\epsilon_n^2/18} 2^{-N/2} \{\Gamma(N/2)\}^{-1} x^{N/2-1} \exp(-x/2) dx \\ &\geq (\epsilon_n^2/36)^{N/2} \sqrt{2\pi N} (N/2)^{-N/2} \exp(-\epsilon_n^2/36) \\ &\geq \{\epsilon_n/(C\sqrt{N})\}^N \asymp \exp\{-N \log(C\sqrt{N}/\epsilon_n)\} \gtrsim \exp(-c n \epsilon_n^2), \end{aligned}$$

by choosing N such that $N \log(\sqrt{N}/\epsilon_n) \lesssim n \epsilon_n^2$. Notice that the above bound holds uniformly for all $a > 0$. Also by Assumption 3, we can show for any fixed constants $0 < r_0 < r_1$

$$P(r_0 \leq A \leq r_1) \geq (r_1 - r_0) C_1 r_0^p \exp(-D_1 r_0 \log^q r_0). \quad (30)$$

Then, invoke the above results in (28) and (29), we have $\Pi\{\tilde{B}_{(f_0, p_0)}(\epsilon_n)\} \gtrsim e^{-c_1 n \epsilon_n^2} P(r_0 \leq A \leq r_1) \gtrsim e^{-c_2 n \epsilon_n^2}$ for some constants $c_2 > c_1 > 0$. Then, we have verified Condition (18).

Next, we verify Condition (19), it suffices to show the desired bound for $\Pi(\tilde{\mathcal{D}}_N^c)$. Note that $\Pi(\tilde{\mathcal{D}}_N^c) = \Pi(\mathcal{D}_N^c \mid \tilde{\mathcal{A}}_N) \leq \Pi(\mathcal{D}_N^c) / \Pi(\tilde{\mathcal{A}}_N)$. Similar to the proof of Theorem 1, it is easy to show that

$$\begin{aligned} \Pi(\tilde{\mathcal{A}}_N) &\geq \int_{r_0}^{r_1} \Pi(\|\tilde{f}_{a,N}\|_\infty \leq A_0) g(a) da \\ &\geq \Pi\left(\left(2/N\right)^{1/2} \sum_{j=1}^N |a_j| \leq A_0\right) P(r_0 < A < r_1) \geq c' \exp\{-N \log(C' \sqrt{N})\}, \end{aligned} \quad (31)$$

for some constants $c', C' > 0$. Again, the second inequality holds by the fact that $\Pi(\|\tilde{f}_{a,N}\|_\infty \leq A_0) \leq \Pi\{(2/N)^{1/2} \sum_{j=1}^N |a_j| \leq A_0\}$ universally for all $a > 0$, and the last inequality uses the result in (30). Next we have

$$\Pi(\mathcal{D}_N^c) \leq 2 \left[\Pi_{\mathbf{a},N}(\{\cap_{j=1}^N [-\sqrt{n}\epsilon_n, \sqrt{n}\epsilon_n]\}^c) + \Pi_{\boldsymbol{\omega},N}(\{\cap_{j=1}^N [-n\epsilon_n^2, n\epsilon_n^2]\}^c) \right]. \quad (32)$$

Recall $a_j \stackrel{i.i.d.}{\sim} N(0, 1)$ for $j = 1, \dots, N$. First, we can show that

$$\begin{aligned} \Pi_{\mathbf{a},N}(\{\cap_{j=1}^N [-\sqrt{n}\epsilon_n, \sqrt{n}\epsilon_n]\}^c) &= \Pi_{\mathbf{a},N}\left(\max_{1 \leq j \leq N} |a_j| \geq \sqrt{n}\epsilon_n\right) \\ &\leq \Pi_{\mathbf{a},N}\left(\max_{1 \leq j \leq N} |a_j| - \mathbb{E} \max_{1 \leq j \leq N} |a_j| \geq \sqrt{n}\epsilon_n/2\right) \leq \exp(-n\epsilon_n^2/8). \end{aligned} \quad (33)$$

The inequality holds by the known result that $\mathbb{E} \max_{1 \leq j \leq N} |a_j| \leq c \sigma_{\max}^2 \sqrt{2 \log N}$ for $\{a_j\}$ are independent centered Gaussian random variables with $\sigma_{\max}^2 = \max_j \{\text{var}(a_j)\} = 1$. Given the chosen N , it is obvious that $\mathbb{E} \max_{1 \leq j \leq N} |a_j| < n\epsilon_n^2/2$. The last inequality uses the tail bound for the maximum of independent Gaussian random variables.

For any fixed $a > 0$, denote $\omega_j \mid (A = a)$ by $\omega_j^a \stackrel{i.i.d.}{\sim} N(0, a^2)$ for $j = 1, \dots, N$. Similarly, we can show

$$\begin{aligned} \Pi_{\boldsymbol{\omega},N}(\{\cap_{j=1}^N [-n\epsilon_n^2, n\epsilon_n^2]\}^c \mid A = a) &= \Pi\left(\max_{1 \leq j \leq N} |\omega_j^a| \geq n\epsilon_n^2\right) \\ &\leq \Pi\left(\max_{1 \leq j \leq N} |\omega_j^a| - \mathbb{E} \max_{1 \leq j \leq N} |\omega_j^a| \geq n\epsilon_n^2/2\right) \leq \exp\{- (n\epsilon_n^2)^2 / 8a^2\}. \end{aligned} \quad (34)$$

The last inequality holds due to facts that $\mathbb{E} \max_{1 \leq j \leq N} |\omega_j| \leq c'a \sqrt{2 \log(2N)}$ for some constant $c' > 0$ and choosing $R_n \asymp \sqrt{n}\epsilon_n$, and an application of the concentration bound of the maximum of independent Gaussian random variables.

Then we have for some $R_n > 0$ that depends on n ,

$$\Pi_{\boldsymbol{\omega},N}(\{\cap_{j=1}^N [-n\epsilon_n^2, n\epsilon_n^2]\}^c) \leq \int_0^{R_n} \Pi_{\boldsymbol{\omega}^a,N}(\{\cap_{j=1}^N [-n\epsilon_n^2, n\epsilon_n^2]\}^c) g(a) da + P(A > R_n).$$

Based on the final bound in (34), the second term on the right hand of the preceding can be upper bounded by

$$\int_0^{R_n} \exp\{- (n\epsilon_n^2)^2 / 8a^2\} g(a) da \leq \exp\{- (n\epsilon_n^2)^2 / 8R_n^2\} \asymp \exp(-c'n\epsilon_n^2),$$

for some constant $c' > 0$. Choosing $R_n \asymp \sqrt{n}\epsilon_n$ leads to the final bound in the preceding. And we have $P(A > R_n) \lesssim e^{-\tilde{c}n\epsilon_n^2}$ for some constant $\tilde{c} > 0$ based on Lemma 4.9 of van der Vaart and van Zanten (2009). Then invoking these results in (32) combined with the result in (33) leads to the desired result that $\Pi(\tilde{\mathcal{D}}_N^c) \lesssim e^{-c_3 n\epsilon_n^2}$ for some constant $c_3 > 0$, by choosing N such that $N \log N \lesssim n\epsilon_n^2$.

At last, we verify Condition (17). First, we estimate the entropy of the sieves. For arbitrary two parameter vectors $\theta = (\mathbf{a}, \boldsymbol{\omega}, \mathbf{s})$ and $\theta' = (\mathbf{a}', \boldsymbol{\omega}', \mathbf{s}') \in \tilde{\mathcal{D}}_N$ and $\theta \neq \theta'$, denote $\tilde{f}_N, \tilde{f}'_N$ associated with θ, θ' , respectively. Then

$$\begin{aligned} \|\tilde{f}_N - \tilde{f}'_N\|_\infty &= \sqrt{\frac{2}{N}} \left\| \sum_{j=1}^N a_j \cos(\omega_j x + s_j) - \sum_{j=1}^N a'_j \cos(\omega'_j x + s'_j) \right\|_\infty \\ &\leq \sqrt{\frac{2}{N}} \left[\left\| \sum_{j=1}^N (a_j - a'_j) \cos(\omega_j x + s_j) \right\|_\infty \right. \\ &\quad \left. + \left\| \sum_{j=1}^N a'_j \{ \cos(\omega_j x + s_j) - \cos(\omega'_j x + s'_j) \} \right\|_\infty \right] \\ &\leq \sqrt{\frac{2}{N}} \{ \|\mathbf{a} - \mathbf{a}'\|_1 + \sqrt{n}\epsilon_n (\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_1 + \|\mathbf{s} - \mathbf{s}'\|_1) \}. \end{aligned}$$

We now consider the partition S_n of length ϵ_n/\sqrt{N} on the interval $[-\sqrt{n}\epsilon_n, \sqrt{n}\epsilon_n]$ for each of $\{a_j\}$ for all j and the partition O_n of length $1/\sqrt{nN}$ on the interval $[-n\epsilon_n^2, n\epsilon_n^2]$ for $\{\omega_j\}$ and the partition M_n of length $1/\sqrt{nN}$ on the interval $[0, 2\pi]$ for $\{s_j\}$. For any $\tilde{f}_N \in \mathcal{D}_N$, we can always find $\{a'_j, \omega'_j, s'_j\}$ with $a'_j \in S_n$, $\omega'_j \in O_n$ and $s'_j \in M_n$ for $j = 1, \dots, N$, such that $\tilde{f}'_N(x) = \sqrt{2/N} \sum_{j=1}^N a'_j \cos(\omega'_j x + s'_j)$ satisfies $\|\tilde{f}_N - \tilde{f}'_N\|_\infty \leq \epsilon_n$. By definition, it is obvious that $N(\epsilon_n, \tilde{\mathcal{D}}_N, \|\cdot\|_\infty) \leq N(\epsilon_n, \mathcal{D}_N, \|\cdot\|_\infty)$. Then, it suffices to bound

$$\begin{aligned} N(\epsilon_n, \mathcal{D}_N, \|\cdot\|_\infty) &\leq [2\sqrt{n}\epsilon_n/(\epsilon_n/\sqrt{N}) + 1]^N \times [2n\epsilon_n^2/(\sqrt{nN})^{-1} + 1]^N \times (4\pi\sqrt{nN})^N \\ &\lesssim (n^{3/2}N^{1/2}\epsilon_n^2)^N. \end{aligned}$$

It is easy to see that $\log N(\epsilon_n, \mathcal{D}_N, \|\cdot\|_\infty) \lesssim N \log n \lesssim n\epsilon_n^2$ by choosing N such that $N \log n \lesssim n\epsilon_n^2$. Therefore we have verified the entropy condition.

Based on a same argument of verifying Condition (17) in the proof of Theorem 1, to complete the proof, it suffices to verify that Lemma 6 holds for all $\tilde{f}_N \in \tilde{\mathcal{D}}_N$. For any $\tilde{f}_N \in \tilde{\mathcal{D}}_N$, it is easy to show that Proposition 12 holds since $\|\tilde{f}_N\|_\infty \leq A_0$ and \tilde{f}_N is infinitely differentiable, which are key points to verify equations (28) and (29). Then we can verify equations (12), (13), and (14) and complete the proof of Lemma 6. Putting all pieces together, we have shown Theorem 4 for \tilde{f}_N , leading to the desired result in Theorem 3.

Appendix E. Proof of Auxiliary Results

In this section, we provide the proof of Lemmata 5, 6, 7 and 8 consecutively.

E.1 Proof of Lemma 5

Based on the definition of sieves \mathcal{P}_n , one has $\mathcal{P}_n^c = (B_n^c \otimes \mathcal{F}) \cup (B_n \otimes \mathcal{F}^c) \cup (B_n^c \otimes \mathcal{F}^c)$ and $\Pi(\mathcal{P}_n^c) \leq 2\{\Pi(B_n^c) + \Pi(\mathcal{F}^c)\}$. We first bound $\Pi(\mathcal{F}^c)$. Under Assumptions 3 and 4,

$$\begin{aligned} \Pi(\mathcal{F}^c) &\leq H\bar{\alpha}([-m, m]^c) + P(\tilde{\sigma} \notin [\underline{\sigma}, \bar{\sigma}]) + P\left(\sum_{h>H} \pi_h > \epsilon\right) \\ &\leq He^{-b_1 m^{\tau_1}} + c_2 \bar{\sigma}^{-2\tau_3} + c_1 e^{-b_2 \underline{\sigma}^{-2\tau_2}} + \left(\frac{e|\alpha|}{H} \log \frac{1}{\epsilon}\right)^H. \end{aligned}$$

Choosing $m^{\tau_1} \lesssim n$, $\underline{\sigma} \lesssim n^{-1/2\tau_2}$ and $\bar{\sigma}^{2\tau_3} \lesssim e^n$ with $\epsilon = \epsilon_n$ for ϵ_n defined in Theorem 1, the first three terms on the right hand side of second line in the preceding can be bounded by a multiple of e^{-n} , and by taking $H \lesssim n\epsilon_n^2$ the last term in the same line can be bounded from above by,

$$\left(\frac{e|\alpha|}{H} \log \frac{1}{\epsilon}\right)^H \lesssim e^{-H \log(H \log n)} \lesssim e^{-\frac{1}{2\alpha+1} n^{1/(2\alpha+1)} (\log n)^{2t+1}} \lesssim e^{-c_4 n^{1/(2\alpha+1)} (\log n)^{2t}}.$$

Thus $\Pi(\mathcal{F}^c) \lesssim e^{-c_4 n \epsilon_n^2}$ for every $c_4 > 0$.

Now we bound $\Pi(\tilde{B}_n^c)$. By definition, $\Pi(\tilde{B}_n^c) = \Pi(B_n^c | \mathcal{A}) \leq \Pi(B_n^c)/P(\mathcal{A})$, with \mathcal{A} defined in Assumption 3. Based on the facts that $E(\|f\|_\infty) < \infty$ and $\sigma_f^2 = \sup_{x \in [0,1]} E\{f(x)\}^2 < \infty$, applying Borell's inequality in Lemma 10, we have $P(\mathcal{A}) = P(\|f\|_\infty < A_0) \geq 1 - e^{-A_0^2/2\sigma_f^2} \geq a_0$, for some $A_0 > 0$ and $a_0 \in (0, 1)$. Thus $\Pi(\tilde{B}_n^c) \lesssim \Pi(B_n^c) \lesssim e^{-n\epsilon_n^2}$ with $M_n^2 \lesssim n\epsilon_n^2$ and $a_n^2 \lesssim n\epsilon_n^2$. More details can be found in the proof of Theorem 3.1 in van der Vaart and van Zanten (2009).

E.2 Proof of Lemma 6

To prove Lemma 6, we will prove the inequality in Equation (14) in detail and only mention the key elements in the proof of results in Equations (12) and (13) since they all follow the similar line of argument. The key elements of the proof are applications of Talagrand's inequality stated in Lemma 9, bounded L_1 -norm of the deconvolution kernel K_n and tight bounds on the bias terms of deconvolution estimators based on the construction in Equations (3) and (4). The last two results are stated in the following Proposition 11 and Proposition 12.

Proposition 11 *For any kernel function K satisfying conditions in Equation (11) and K_n defined in Equation (5), we have $\|K_n\|_1 < C_1$, for some constant $C_1 > 0$.*

Proof There exists a symmetric and integrable kernel function K such that Equation (11) hold and the Fourier transform $\phi_K(t) = \mathbb{1}_{[-1,1]}/(2\pi)$, which is symmetric, real-valued, bounded infinitely smooth function with a compact support. We remark that one example of kernels that satisfy the above conditions is the *sinc* kernel. For any fixed positive constant a , $\int |K_n(s)| ds = \int_{|s| \leq a} |K_n(s)| ds + \int_{|s| > a} |K_n(s)| ds$. We have

$$|K_n(s)| \leq \int |e^{-its}| \frac{|\phi_K(t)|}{|\phi_\delta(t/h_n)|} dt \leq \int_{-1}^1 \frac{|\phi_K(t)|}{|\phi_\delta(t/h_n)|} dt \lesssim \exp(\delta_n^2/2h_n^2),$$

thus $\int_{|s| \leq a} |K_n(s)| ds \lesssim \exp(\delta_n^2/2h_n^2) = O(1)$. For $|s| > a$, by Cauchy-Schwarz inequality,

$$\int_{|s| > a} |K_n(s)| ds \leq \left(\int_{|s| > a} \frac{1}{s^4} ds \right)^{1/2} \left\{ \int_{|s| > a} s^4 K_n(s)^2 ds \right\}^{1/2}.$$

By Parseval's theorem, $\int \{s^2 K_n(s)\}^2 ds = \int \{g''(t)\}^2 dt$ with

$$g(t) = \phi_K(t)/\phi_\delta(t/h_n) = \frac{1}{2\pi} e^{-t^2 \delta^2 / (2h_n^2)} \mathbf{1}_{[-1,1]}.$$

Since $g''(t)$ is the Fourier transform of $(is)^2 K_n(s)$, also $g(t), g'(t), g''(t)$ are continuous and therefore bounded on $[-1, 1]$. Thus $\int \{s^2 K_n(s)\}^2 ds$ is bounded and so is $\int_{|s| > a} 1/s^4 ds$, which yields the result that $\int |K_n(s)| ds$ is bounded. \blacksquare

The following Proposition provides tight bounds on the bias terms of \widehat{p}_n and $\widehat{f}_n \widehat{p}_n$ separately.

Proposition 12 For \widehat{p}_n and \widehat{f}_n defined in Equation (3) and Equation (4) and for any $f, p \in \mathcal{P}_n$ we have

$$\|E_{W,X}(\widehat{p}_n) - p\|_1 \lesssim \epsilon_n, \quad \text{and} \quad \|E_{Y,W,X}(\widehat{f}_n \widehat{p}_n) - fp\|_1 \lesssim \epsilon_n,$$

with ϵ_n defined in Theorem 1.

Proof By Fourier inversion theorem, it is easy to show that $E_{W,X}(\widehat{p}_n) = K_{h_n} * p(x)$ and $E_{Y,W,X}(\widehat{f}_n \widehat{p}_n) = K_{h_n} * (fp)$ with $K_{h_n} = K(\cdot/h_n)/h_n$. First for any $p = \phi_{\tilde{\sigma}} * F$, by Cauchy-Schwarz inequality we have $\|K_{h_n} * p - p\|_1 \leq \|K_{h_n} * p - p\|_2$. Recall the Fourier transform of the kernel function K is denoted by $\phi_K(t)$, applying Parseval's theorem again,

$$\begin{aligned} \|K_{h_n} * p - p\|_2^2 &= \int |2\pi\phi_K(h_n t) - 1|^2 |\widehat{p}(t)|^2 dt = \int_{|t| > 1/h_n} |\widehat{F}(t)|^2 |\widehat{\phi}_{\tilde{\sigma}}(t)|^2 dt \\ &\leq \int_{|t| > 1/h_n} |\widehat{\phi}_{\tilde{\sigma}}(t)|^2 dt \leq (h_n/\underline{\sigma}^2) e^{-(\underline{\sigma}/h_n)^2/2} \\ &\lesssim h_n^{-1} (\log n)^{-t_3} e^{-K^2(\log n)^{2t_3/2}} \lesssim \epsilon_n^2, \end{aligned}$$

for all $\tilde{\sigma} \geq \underline{\sigma}$. Let $h_n \asymp \epsilon_n^{1/\beta}$ with ϵ_n defined in Theorem 1 and by Lemma 5 we have $\underline{\sigma} \lesssim n^{-1/(2\tau_2)}$, where τ_2 is chosen such that $\underline{\sigma} = Kh_n(\log n)^{t_3}$ for some constants K, t_3 satisfying $K^2/2 > 1$ and $t_3 > 1/2$.

Now we bound the bias term of $\widehat{f}_n \widehat{p}_n$. By triangle inequality,

$$\|K_{h_n} * (fp) - fp\|_1 \leq \|K_{h_n} * (fp) - pK_{h_n} * f\|_1 + \|pK_{h_n} * f - fp\|_1. \quad (35)$$

By Cauchy-Schwarz inequality, the first term of the right hand side of Equation (35) can be bounded as

$$\begin{aligned} \|K_{h_n} * (fp) - pK_{h_n} * f\|_1 &= \int \int |K_{h_n}(x-y)\{p(y) - p(x)\}f(y) dy| dx \\ &\leq \|K_{h_n} * p - p\|_2 \|f\|_2 \lesssim \|K_{h_n} * p - p\|_2, \end{aligned} \quad (36)$$

since $\|f\|_2 \leq \|f\|_\infty \leq A_0$ under Assumption 3. The second term on the right hand side of Equation (35) can be bounded

$$\|pK_{h_n} * f - fp\|_1 \leq \|p\|_1 \|K_{h_n} * f - f\|_\infty = \|K_{h_n} * f - f\|_\infty \lesssim \epsilon_n. \quad (37)$$

The last inequality in the preceding holds based on the properties of higher order kernel as in Lemma 4.3 of van der Vaart and van Zanten (2009). \blacksquare

Proof of Equation (14). Now we are ready to prove the inequality in Equation (14). By triangle inequality,

$$\begin{aligned} \|\widehat{f}_n \widehat{p}_n - fp\|_1 &\leq \|\widehat{f}_n \widehat{p}_n - E_{Y,W|X}(\widehat{f}_n \widehat{p}_n)\|_1 + \|E_{Y,W|X}(\widehat{f}_n \widehat{p}_n) - E_{Y,W,X}(\widehat{f}_n \widehat{p}_n)\|_1 \\ &\quad + \|E_{Y,W,X}(\widehat{f}_n \widehat{p}_n) - f \cdot p\|_1 := I_{1,n} + I_{2,n} + I_{3,n}. \end{aligned} \quad (38)$$

First we estimate $P(I_{1,n} > \epsilon_n/2)$ for $I_{1,n}$ in Equation (38). By definition,

$$\begin{aligned} &\widehat{f}_n \widehat{p}_n - E_{Y,W|X}(\widehat{f}_n \widehat{p}_n) \\ &= \frac{1}{2\pi n h_n} \sum_{j=1}^n \int e^{-\frac{itx}{h_n}} \left\{ e^{itW_j/h_n} Y_j - E_{W|X}(e^{itW_j/h_n}) E_{Y|X}(Y_j) \right\} \frac{\phi_K(t)}{\phi_u(t/h_n)} dt \\ &= \frac{1}{2\pi n h_n} \sum_{j=1}^n \int e^{-\frac{it(x-W_j)}{h_n}} \frac{\phi_K(t)}{\phi_u(t/h_n)} dt \{Y_j - E_{Y|X}(Y_j)\} \\ &\quad + \frac{1}{2\pi n h_n} \sum_{j=1}^n \int e^{-\frac{itx}{h_n}} \left\{ e^{itW_j/h_n} - E_{W|X}(e^{itW_j/h_n}) \right\} \frac{\phi_K(t)}{\phi_u(t/h_n)} dt E_{Y|X}(Y_j) \\ &:= T_{1,n} + T_{2,n}. \end{aligned} \quad (39)$$

First, we estimate $P(\|T_{2,n}\|_1 > \epsilon_n/2)$ with $T_{2,n}$ defined in Equation (39). By Hahn-Banach Theorem, there exists a bounded linear functional T such that $T(h) = \int T_{2,n}(x)h(x)dx$ for all $h \in L_\infty[0, 1]$, namely, for all $h(x)$ such that $\sup_{x \in [0,1]} |h(x)| < \infty$. And $\|T_{2,n}\|_1 = \|T\|_{\mathcal{F}_1}$ where $\|T\|_{\mathcal{F}_1} = \sup_{h \in \mathcal{F}_1} |T(h)|$ and \mathcal{F}_1 is a countable and dense subset of $L_\infty[0, 1]$. Thus we have

$$\mathcal{K} = \left\{ k(u, v) : (u, v) \mapsto \frac{1}{h_n} \int_0^1 \left[K_n \left(\frac{x-u}{h_n} \right) - E_{W|X} \left\{ K_n \left(\frac{x-W}{h_n} \right) \right\} \right] f(v)h(x)dx, \right. \\ \left. \text{for all } h \in \mathcal{F}_1 \right\},$$

and $\|nT_{2,n}\|_1 = \sup_{k \in \mathcal{K}} |\sum_{j=1}^n k(W_j, X_j)|$. To apply Lemma 9, we need to estimate the following quantities, $\sup_{k \in \mathcal{K}} \|k(u, v)\|_\infty$, $\sigma_{\mathcal{K}}^2 = E_{W|X} \{ \sup k^2(W, X) \}$ and $E \{ \sup_{k \in \mathcal{K}} k(W, X) \}$. Based on the Assumptions 3 and 4 we have $\|f\|_\infty \leq C_0$ and $\|h\|_\infty \leq 1$, then for any $k \in \mathcal{K}$,

$$|k(u, v)| \leq \frac{C_2}{h_n} \left[\int_0^1 \left| K_n \left(\frac{x-u}{h_n} \right) \right| dx + \int_0^1 \left| E_{W|X} \left\{ K_n \left(\frac{x-W}{h_n} \right) \right\} \right| dx \right],$$

for some constant $C_2 > 0$. For any u , by change of variables $s = (x - u)/h_n$, for any fixed positive constant a , one has $\int_0^1 |K_n\{(x - u)/h_n\}/h_n| dx \leq \int |K_n(s)| ds \leq C'$ for some constant C' . The second inequality holds by Proposition 11. Given $W | X \sim N(X, \delta_n^2)$,

$$\begin{aligned} E_{W|X} \left\{ K_n \left(\frac{x - W}{h_n} \right) \right\} &= \frac{1}{2\pi} \int E_{W|X} \left(e^{-it\{(x - X - (W - X))/h_n\}} \right) \frac{\phi_K(t)}{\phi_u(t/h_n)} dt \\ &= \frac{1}{2\pi} \int e^{-it(x - X/h_n)} \phi_K(t) dt = K\{(x - X)/h_n\}. \end{aligned}$$

Again by change of variables $r = (x - X)/h_n$, we have $\int_0^1 E_{W|X} [K\{(x - W)/h_n\}/h_n] dx = \int |K(r)| dr = 1$. There exists a constant K_1 such that $\|k\|_\infty \leq K_1$ for any $k \in \mathcal{K}$, then $\sup_{k \in \mathcal{K}} \|k\|_\infty \lesssim \max\{1, \exp(\delta_n^2/2h_n^2)\}$. Next we estimate the term $\sigma_{\mathcal{K}}^2$. For any $k \in \mathcal{K}$ and $W | X \sim N(X, \delta_n^2)$,

$$\begin{aligned} k(W, X)^2 &= \frac{1}{h_n^2} \left(\int_0^1 \left[K_n \left(\frac{x - u}{h_n} \right) - E_{W|X} \left\{ K_n \left(\frac{x - W}{h_n} \right) \right\} \right] f(X) h(x) dx \right)^2 \\ &\lesssim \frac{1}{h_n^2} \left\{ \int_0^1 K_n \left(\frac{x - u}{h_n} \right) dx \right\}^2 + \frac{1}{h_n^2} \left\{ \int_0^1 E_{W|X} K_n \left(\frac{x - W}{h_n} \right) dx \right\}^2 \\ &\lesssim \max\{1, \exp(\delta_n^2/h_n^2)\}. \end{aligned}$$

Therefore $\sup_{k \in \mathcal{K}} E_{W|X} \{k(W, X)^2\} \lesssim \max\{1, \exp(\delta_n^2/h_n^2)\}$.

Finally, we move to bound $E_{W|X}(\sup_{k \in \mathcal{K}} |\sum_{j=1}^n k(W_j, X_j)|)$. By Cauchy-Schwarz inequality,

$$\begin{aligned} E_{W|X} \left(\sup_{k \in \mathcal{K}} \left| \sum_{j=1}^n k(W_j, X_j) \right| \right) &\leq \left[E_{W|X} \left\{ \sup_{k \in \mathcal{K}} \left| \sum_{j=1}^n k(W_j, X_j) \right|^2 \right\} \right]^{1/2} \\ &\lesssim \left(\frac{1}{h_n^2} \sum_{j=1}^n E_{W|X} \left[\int \left| K_n \left(\frac{x - W_j}{h_n} \right) - E_{W|X} \left\{ K_n \left(\frac{x - W_j}{h_n} \right) \right\} \right|^2 dx \right] \right)^{1/2} \\ &\lesssim n^{1/2} \max\{1, \exp(\delta_n^2/2h_n^2)\}. \end{aligned}$$

To apply the Lemma 9, we choose $\delta_n \asymp h_n$ and same ϵ_n in Theorem 1, we have $\exp(\delta_n^2/2h_n^2) = O(1)$. By choosing $t = n\epsilon_n^2$, we have $n^{1/2} + \{2(n + n^{1/2})n\epsilon_n^2\}^{1/2} + n\epsilon_n^2/3 \lesssim n\epsilon_n$.

We now discuss bounding the probability $P(\|T_{1,n}\|_1 > \epsilon_n/2)$ with $T_{1,n}$ defined in Equation (39). Recall that

$$nT_{1,n} = \sum_{j=1}^n K_n\{(x - W_j)/h_n\}(Y_j - E_{Y|X}Y_j)/h_n = \sum_{j=1}^n K_n\{(x - W_j)/h_n\}\tilde{Y}_j/h_n,$$

with $\tilde{Y}_j \sim N(0, 1)$ i.i.d. for $j = 1, \dots, n$, given $Y_j | X_j \sim N(f(X_j), 1)$ for $j = 1, \dots, n$. Again by Hahn-Banach theorem, there exists a countable and dense subset $\mathcal{T} \in L_\infty[0, 1]$ and a class of bounded linear functionals on $L_\infty[0, 1]$,

$$\mathcal{Q} = \left\{ q = \sum_{j=1}^n \tilde{q}(u_j), \tilde{q}(u) = \int_0^1 \sum_{j=1}^n K_n \left(\frac{x - u}{h_n} \right) (Y_j - E_{Y|X}Y_j) t(x) dx, t \in \mathcal{T} \right\},$$

and $\|nT_{1,n}\|_1 = \sup_{q \in \mathcal{Q}} \|q\|_\infty$.

We now proceed to estimate $\sigma_{\mathcal{Q}}^2 = \sup_{q \in \mathcal{Q}} E_{Y|X} \{ \sum_{j=1}^n \tilde{q}(W_j) \}^2$ and $E_{Y|X}(\sup_{q \in \mathcal{Q}} \|q\|_\infty)$ in order to apply Lemma 10. We first estimate $\sigma_{\mathcal{Q}}^2$. Again, by change of variables and the fact $\|t\|_\infty \leq 1$ we have

$$\begin{aligned} E_{Y|X} \left\{ \sum_{j=1}^n \tilde{q}(W_j) \right\}^2 &= \frac{1}{h_n^2} \sum_{j=1}^n \left\{ \int_0^1 K_n \left(\frac{x - W_j}{h_n} \right) t(x) dx \right\}^2 \\ &\leq \frac{1}{h_n^2} \sum_{j=1}^n \left\{ \int_0^1 K_n \left(\frac{x - W_j}{h_n} \right) dx \right\}^2 \\ &\leq \sum_{j=1}^n \left(\int |K_n(u)| du \right)^2 \lesssim n \max\{1, \exp(\delta_n^2/h_n^2)\}. \end{aligned}$$

Next we estimate $E_{Y|X}(\sup_{q \in \mathcal{Q}} \|q\|_\infty)$, using the generalized Minkowski inequality, we obtain

$$\begin{aligned} E_{Y|X} \sup_{q \in \mathcal{Q}} \|q\|_\infty &= E_{Y|X}(\|nT_{1,n}\|_1) \leq \{E_{Y|X}(\|nT_{1,n}\|_1^2)\}^{1/2} \\ &\leq \|[E_{Y|X}\{(nT_{1,n})^2\}]^{1/2}\|_1 \\ &= \int \left\{ \frac{1}{h_n^2} \sum_{j=1}^n K_n \left(\frac{x - W_j}{h_n} \right)^2 \right\}^{1/2} dx. \end{aligned}$$

The last equation in the preceding holds because Y_j 's are independent. By Jensen's inequality and change of variables it can be bound by $\sum_{j=1}^n \int K_n \{(x - W_j)/h_n\}^2 dx\}^{1/2}/h_n = n^{1/2} \{\int K_n(u)^2 du\}^{1/2}/h_n$. Fixed any constant $a' > 0$, one has

$$\int K_n(u)^2 du \leq \int_{|u| > a'} (u^4/a'^4) K_n(u)^2 du + \int_{|u| \leq a'} K_n(u)^2 du.$$

It has been shown in the proof of Proposition 11 that $\int u^4 K_n(u)^2 du \lesssim \exp(\delta_n^2/h_n^2)$, and it is easy to see that $\int K_n(u)^2 du \lesssim \max\{1, \exp(\delta_n^2/h_n^2)\}$. Thus we have $E_{Y|X}(\sup_{q \in \mathcal{Q}} \|q\|_\infty) \lesssim n^{1/2} \max\{1, \exp(\delta_n^2/h_n^2)\}/\sqrt{h_n}$. Then applying Borell's inequality in Lemma 10 by choosing $x = n\epsilon_n$, $\delta_n \asymp h_n \asymp \epsilon_n^{1/\beta}$, where ϵ_n is defined in Theorem 1, we have shown that $P(\|T_{1,n}\|_1 > \epsilon_n/2) < e^{-n\epsilon_n^2/8}$.

We now estimate the probability $P(I_{2,n} > \epsilon_n/2)$, recall that $I_{2,n}$ is defined in Equation (38). By definition, $I_{2,n} = E_{Y,W|X}(\hat{f}_n \hat{p}_n) - E_{Y,W,X}(\hat{f}_n \hat{p}_n)$, then with simple calculation one can show that $E_{Y,W|X}(\hat{f}_n \hat{p}_n) = \sum_{j=1}^n K\{(x - X_j)/h_n\} f(X_j)/(nh_n)$. Similarly, by Hahn-Banach theorem, there exists a countable and dense set $\mathcal{H}_1 \in L_\infty[0, 1]$ such that we can construct a class of bounded linear functionals

$$\mathcal{L} = \left\{ l(u) : u \mapsto \int \left[K \left(\frac{x - u}{h_n} \right) f(u) - E_X \left\{ K \left(\frac{x - X}{h_n} \right) f(X) \right\} \right] h_1(x) dx, \quad h_1 \in \mathcal{H}_1 \right\},$$

and we have $\|nI_{2,n}\|_1 = \sup_{l \in \mathcal{L}} \|\sum_{j=1}^n l(X_j)\|_\infty$. To apply the Talagrand's inequality, we first bound $\sup_{l \in \mathcal{L}} \|l(u)\|_\infty \leq [\int |K\{(x - X_j)/h_n\}/h_n| dx] \|f\|_\infty$. Since $\int |K(u)| du \leq$

K_3 for some constant $K_3 > 0$, by change of variables and Assumption 3 one can show $\sup_{l \in \mathcal{L}} \|l(u)\|_\infty \leq K_4$, for some constant $K_4 > 0$.

Second, we bound $\sup_{l \in \mathcal{L}} E_X \{l(X)\}^2$. For any $l \in \mathcal{L}$,

$$\begin{aligned} E_X \{l(X)\}^2 &\leq 2E_X \left(\left\{ \int \left| K \left(\frac{x-X}{h_n} \right) \right| dx \right\}^2 + 2 \left[\int E_X \left\{ K \left(\frac{x-X}{h_n} \right) \right\} dx \right]^2 \right) \|f\|_\infty^2 / h_n^2 \\ &\leq K_5, \end{aligned}$$

for some constant $K_5 > 0$. Thus we show that $\sup_{l \in \mathcal{L}} E_X \{l(X)^2\} \leq K_5$.

At last, we have

$$\begin{aligned} &E_X \sup_{l \in \mathcal{L}} \left| \sum_{j=1}^n l(X_j) \right| \\ &\leq \left\{ E_X \left(\sup_{l \in \mathcal{L}} \left| \sum_{j=1}^n l(X_j) \right| \right)^2 \right\}^{1/2} \\ &\leq \frac{1}{h_n} \left(2n \left[E_X \left\{ \int K \left(\frac{x-X}{h_n} \right) dx \right\}^2 + \left\{ \int E_X K \left(\frac{x-X}{h_n} \right) dx \right\}^2 \right] \right)^{1/2} \|f\|_\infty \\ &\lesssim (n/h_n)^{1/2}. \end{aligned}$$

Choosing $h_n \asymp \epsilon_n^{1/\beta}$ with ϵ_n defined in the Theorem 1, then applying Talagrand's inequality yields the result $P(I_{2,n} > \epsilon_n/2) \leq e^{-n\epsilon_n^2/8}$.

Finally, for $I_{3,n}$ defined in Equation (38), it is easy to see $I_{3,n} \leq \epsilon_n$ by Proposition 12. Combining the results of $I_{1,n}, I_{2,n}$ and $I_{3,n}$, we prove the inequality in Equation (14).

Proof of Equation (13). Inequality in Equation (13) can be obtained directly from Equation (14), as it can be seen as a special case of Equation (14) by letting the regression function $f(x) \equiv c$ for some constant $c > 0$.

Proof of Equation (12). The proof of inequality (12) follows a same line of arguments in the proof of Equation (14) and we omit some details. Let $P_{1,n} = \hat{p}_n - E_{W|X}(\hat{p}_n)$, $P_{2,n} = E_{W|X}(\hat{p}_n) - E_{W,X}(\hat{p}_n)$ and $P_{3,n} = E_{W,X}(\hat{p}_n) - p$. First, we estimate $P(\|P_{1,n}\|_\infty > \epsilon_0/2)$. The difference is that we consider the empirical process directly in $\|\cdot\|_\infty$. Since the function $K_n(x)$ is continuous and bounded on $[0, 1]$, by the separability of $C[0, 1]$, there exists a countable and dense set T over $[0, 1]$ and consider the class,

$$\mathcal{M} = \left\{ m_x(u) : u \mapsto \int e^{-itx/h_n} \left\{ e^{itu/h_n} - E_{W|X} \left(e^{itW/h_n} \right) \right\} \frac{\phi_K(t)}{\phi_u(t/h_n)} dt, x \in T \right\},$$

then $\|nP_{1,n}\|_\infty = \sup_{x \in T} \left| \sum_{j=1}^n m_x(W_j) \right|$. Also we can show

$$\begin{aligned} \sup_{x \in T} \|m_x\|_\infty &\lesssim h_n^{-1} \exp(\delta_n^2/2h_n^2), \\ \sup_{x \in T} E_{W|X} [m_x(W)]^2 &\lesssim h_n^{-2} \exp(\delta_n^2/h_n^2), \\ E_{W|X} \sup_{x \in T} \left| \sum_{j=1}^n m_x(W_j) \right| &\lesssim n^{1/2} h_n^{-1} \exp(\delta_n^2/h_n^2). \end{aligned}$$

Therefore choosing $\delta_n = o(h_n)$ and $h_n = o(\epsilon_n)$ with same ϵ_n in Theorem 1. For any $\epsilon_0 > 0$, take $t = \epsilon_0 n h_n^2$, one has

$$\begin{aligned} & n^{1/2} h_n^{-1} \exp(\delta_n^2/2h_n^2) \\ & + \{2n h_n^{-2} \exp(\delta_n^2/h_n^2) + 4n^{1/2} h_n^{-1} \exp(\delta^2/2h_n^2)\}^{1/2} (n\epsilon_0 h_n^2)^{1/2} + \epsilon_0 n h_n^2 \\ & < n\epsilon_0. \end{aligned}$$

By applying Lemma 9, one can show $P(\|\hat{p}_n - E_{W|X}(\hat{p}_n)\|_\infty > \epsilon_0) \leq e^{-\epsilon_0 n h_n^2}$. Similarly, for $P_{2,n}$ one can write $P_{2,n} = E_{W|X}(\hat{p}_n) - E_{W,X}(\hat{p}_n) = \sum_{j=1}^n \tilde{g}_x(X_j)/(n h_n)$, where $\tilde{g}_x(u) = K\{(x-u)/h_n\} - E_X[K\{(x-X)/h_n\}]$ for any $x \in T$. Construct the class $\mathcal{G} = \{\tilde{g}_x, x \in T\}$ with the countable and dense set T over $[0, 1]$, with same calculation by choosing $t = n\epsilon_0 h_n^2$, $\delta_n = o(h_n)$ and $h_n = o(\epsilon_n)$, another application of Talagrand's inequality shows $P(P_{2,n} > \epsilon_0) \leq e^{-\epsilon_0 n h_n^2}$. Combining the above results for $P_{1,n}$, $P_{2,n}$ and applying Proposition 12 to $P_{3,n}$ completes the proof of Equation (12).

E.3 Proof of Lemma 7

The Kullback–Leibler neighborhood around f_0 has been studied extensively in Bayesian literature. We give a brief argument mentioning the difference in our case, refer to Shen et al. (2013) for extended proof. Under the Assumption 2, p_0 is compactly supported and lower-bounded. From Theorem 3 in Shen et al. (2013), there exists a density function h_σ supported on $[-a_0, a_0]$ satisfying $H(p_0, \phi_\sigma * h_\sigma) \lesssim \sigma^\beta$, for some constant $a_0 > 0$. Fix $\sigma^\beta = \tilde{\epsilon}_n \{\log(1/\tilde{\epsilon}_n)\}^{-1}$ and find $b' > \max(1, 1/(2\beta))$ such that $\tilde{\epsilon}_n^{b'} \{\log(1/\tilde{\epsilon}_n)\}^{5/4} \leq \tilde{\epsilon}_n$. By Lemma 2 of Ghosal and van der Vaart (2007) there is a discrete probability measure $F' = \sum_{j=1}^N p_j \delta_{z_j}$ with at most $N \leq D\sigma^{-1} \{\log(1/\sigma)\}^{-1}$ support points on $[-a_0, a_0]$, and F' satisfies $H(\phi_\sigma * h_\sigma, \phi_\sigma * F') \leq \tilde{\epsilon}_n^{b'} \{\log(1/\tilde{\epsilon}_n)\}^{1/4}$. We construct the partition $\{U_1, \dots, U_M\}$ in the flavor of $c\sigma \tilde{\epsilon}_n^{b'} \leq \alpha(U_j) \leq 1$ for $j = 1, \dots, M$, where $M \lesssim \tilde{\epsilon}_n^{1/\beta} \{\log(1/\tilde{\epsilon}_n)\}^{1+1/\beta}$. Further denote the set S_F of probability measure F with $\sum_{j=1}^M |F(U_j) - p_j| \leq 2\tilde{\epsilon}_n^{2b'}$ and $\min_{1 \leq j \leq M} F(U_j) \geq \tilde{\epsilon}_n^{Ab'}/2$ for sufficiently large n . Then $\Pi(S_F) \gtrsim \exp[-\tilde{\epsilon}_n^{-1/\beta} \{\log(1/\tilde{\epsilon}_n)\}^{2+1/\beta}]$. For each $F \in S_F$,

$$\begin{aligned} H(p_0, p_{F,\sigma}) & \leq H(p_0, \phi_\sigma * h_\sigma) + H(\phi_\sigma * h_\sigma, \phi_\sigma * F') + H(\phi_\sigma * F', p_{F,\sigma}) \\ & \lesssim \sigma^\beta + \tilde{\epsilon}_n^{b'} \{\log(1/\tilde{\epsilon}_n)\}^{1/4} + \tilde{\epsilon}_n^{b'} \lesssim \sigma^\beta. \end{aligned}$$

Also we can show that for every $x \in [-a_0, a_0]$, $p_{F,\sigma}/p_0 \geq A_4 \tilde{\epsilon}_n^{b'}/\sigma$ for some constant $A_4 > 0$, which leads to $\log \|p_0/p_{F,\sigma}\|_\infty \lesssim \log(1/\tilde{\epsilon}_n)$.

E.4 Proof of Lemma 8

To prove Lemma 8, by the definition of the Kullback–Leibler neighborhood defined in Equation (16), it suffices to bound the Kullback–Leibler divergence and the second moment of Kullback–Leibler divergence between g_{f_0, p_0} and $g_{f, p}$ from above, respectively. Based on Lemma 5.3 in van der Vaart and van Zanten (2009) and Lemma 5 in Appendix A.2, we have $\Pi\{KL(p_0, p) \leq \epsilon_n^2\} \geq e^{-n\epsilon_n^2}$ and $\Pi(\|f - f_0\|_\infty < \epsilon_n) \geq e^{-n\epsilon_n^2}$. Then using the convexity of

the Kullback–Leibler divergence with respect to both arguments, we have

$$\begin{aligned}
 & KL(g_{f_0, p_0}, g_{f, p}) \\
 &= KL\left(\frac{1}{2\pi\delta_n} \int e^{-\frac{1}{2}\{y-f_0(x)\}^2 - \frac{1}{2\delta_n^2}(w-x)^2} dP_0, \frac{1}{2\pi\delta_n} \int e^{-\frac{1}{2}\{y-f(x)\}^2 - \frac{1}{2\delta_n^2}(w-x)^2} \frac{p}{p_0} dP_0\right) \\
 &\leq \int KL\left(\frac{1}{2\pi\delta_n} e^{-\frac{1}{2}(y-f_0(x))^2 - \frac{1}{2\delta_n^2}(w-x)^2}, \frac{1}{2\pi\delta_n} e^{-\frac{1}{2}\{y-f(x)\}^2 - \frac{1}{2\delta_n^2}(w-x)^2} \frac{p}{p_0}\right) dP_0 \\
 &= \int \int \frac{1}{2\pi\delta_n} e^{-\frac{1}{2}\{y-f_0(x)\}^2 - \frac{1}{2\delta_n^2}(w-x)^2} \log\left(\frac{e^{-\frac{1}{2}\{y-f_0(x)\}^2} p_0}{e^{-\frac{1}{2}\{y-f(x)\}^2} p}\right) dy dw dP_0 \\
 &= \int [KL\{N(y; f_0, 1), N(y; f, 1)\} + \log(p_0/p)] dP_0 \\
 &\lesssim \|f_0 - f\|_\infty^2 + KL(p_0, p) \lesssim \epsilon_n^2,
 \end{aligned}$$

where P_0 denotes the distribution measure associated with p_0 . Next, we decompose the second moment of the Kullback–Leibler divergence into,

$$\begin{aligned}
 \int g_{f_0, p_0} \left(\log \frac{g_{f_0, p_0}}{g_{f, p}}\right)^2 &= \int_{A_n} g_{f_0, p_0} \left(\log \frac{g_{f_0, p_0}}{g_{f, p}}\right)^2 + \int_{A_n^c} g_{f_0, p_0} \left(\log \frac{g_{f_0, p_0}}{g_{f, p}}\right)^2 \\
 &=: I_1 + I_2,
 \end{aligned} \tag{40}$$

where $A_n = \{y \in \mathbb{R} : |y| \leq \gamma'/\epsilon_n\}$ for some constant $\gamma' > 0$.

We first bound term I_1 in Equation (40), apply the inequality

$$\int_{A_n} g_{f_0, p_0} \left(\log \frac{g_{f_0, p_0}}{g_{f, p}}\right)^2 \leq 2H^2(g_{f_0, p_0}, g_{f, p})(1 + \log \|(g_{f_0, p_0}/g_{f, p}) \mathbb{1}_{A_n}\|_\infty)^2.$$

It is well known that $H^2(g_{f_0, p_0}, g_{f, p}) \leq KL(g_{f_0, p_0}, g_{f, p})$, then to estimate I_1 it remains to estimate the term $\|(g_{f_0, p_0}/g_{f, p}) \mathbb{1}_{A_n}\|_\infty$. By definition,

$$\begin{aligned}
 & \left| \frac{g_{f_0, p_0}(y, w)}{g_{f, p}(y, w)} \right| \mathbb{1}_{A_n} \\
 &\leq \left| \frac{\int_{A_n} e^{-\frac{1}{2}(y-f_0(x))^2} e^{-\frac{1}{2\delta_n^2}(w-x)^2} p(x) dx}{\int_{A_n} e^{-\frac{1}{2}(y-f_0(x))^2} [e^{-\frac{1}{2}(y-f(x))^2} / e^{-\frac{1}{2}(y-f_0(x))^2}] e^{-\frac{1}{2\delta_n^2}(w-x)^2} p(x) dx} \right| \cdot \left\| \frac{p_0}{p} \right\|_\infty \\
 &\leq \left\| \frac{e^{-\frac{1}{2}(y-f_0)^2}}{e^{-\frac{1}{2}(y-f)^2}} \mathbb{1}_{A_n} \right\|_\infty \left\| \frac{p_0}{p} \right\|_\infty.
 \end{aligned}$$

Based on the Assumption 1, f_0 is a β -smooth function supported on $[0, 1]$ and hence there exists some constant $B'_0 > 0$ such that $\|f_0\|_\infty \leq B'_0$. For $y \in A_n$, we have

$$\begin{aligned}
 \frac{e^{-\frac{1}{2}\{y-f(x)\}^2}}{e^{-\frac{1}{2}\{y-f_0(x)\}^2}} &= e^{\{f(x)-f_0(x)\}\{y-f_0(x)\} - \{f(x)-f_0(x)\}^2/2} \\
 &\geq e^{-\|f-f_0\|_\infty(|y|+\|f_0\|_\infty) - \{f(x)-f_0(x)\}^2/2} \\
 &\geq e^{-\epsilon_n(\gamma'/\epsilon_n+B'_0) - \epsilon_n^2/2} \geq e^{-2\gamma'}.
 \end{aligned}$$

Thus $\|e^{-(y-f_0)^2/2}/e^{-(y-f)^2/2}\mathbf{1}_{A_n}\|_\infty \leq e^{2\gamma'}$. Based on Lemma 5, for any $x \in [0, 1]$ and $p \in \mathcal{P}_n$, we have $\log\|p_0/p\|_\infty \lesssim \log(1/\epsilon_n)$. Therefore, we have shown

$$I_1 = \int_{A_n} g_{f_0, p_0} \{\log(g_{f_0, p_0}/g_{f, p})\}^2 \leq 2\epsilon_n^2 \log^2(1/\epsilon_n). \quad (41)$$

Next we estimate the term I_2 in Equation (40). For all $y \in A_n^c$ and for any fixed $x \in [0, 1]$, we choose $\gamma' > 1$ such that $|y - f_0(x)| \geq |y| - \|f_0\|_\infty > \gamma'/\epsilon_n - B'_0 \geq 1/\epsilon_n$. By Fubini's theorem,

$$\begin{aligned} & \int_{|y|>1/\epsilon_n} g_{f_0, p_0} \left(\log \frac{g_{f_0, p_0}}{g_{f, p}} \right)^2 \\ & \leq \frac{1}{2\pi\delta_n} \int_0^1 \int_{|y-f_0(x)|>1/\epsilon_n} e^{-\frac{1}{2}\{y-f_0(x)\}^2} e^{-\frac{1}{2\delta_n^2}(w-x)^2} \\ & \quad \cdot \left(\log \frac{\int e^{-\frac{1}{2}(y-f_0)^2} e^{-\frac{1}{2\delta_n^2}(w-x)^2} p_0(x) dx}{\int e^{-\frac{1}{2}\{y-f(x)\}^2} e^{-\frac{1}{2\delta_n^2}(w-x)^2} p(x) dx} \right)^2 dy dw p_0(x) dx \\ & \leq \frac{1}{\sqrt{2\pi}} \int_0^1 \int_{|y-f_0(x)|>1/\epsilon_n} e^{-\frac{1}{2}\{y-f_0(x)\}^2} \left(\log \left\| \frac{e^{-(y-f_0)^2/2}}{e^{-(y-f)^2/2}} \right\|_\infty + \log \left\| \frac{p_0}{p} \right\|_\infty \right)^2 dy p_0(x) dx. \end{aligned}$$

Let $z = y - f_0(x)$, we can show that for any $x \in [0, 1]$, $e^{-\{y-f_0(x)\}^2/2+\{y-f(x)\}^2/2} \leq e^{\epsilon_n|z|+\epsilon_n^2/2}$. Then

$$\begin{aligned} & \int_{A_n^c} g_{f_0, p_0} \left(\log \frac{g_{f_0, p_0}}{g_{f, p}} \right)^2 \\ & \leq 4(2\pi)^{-1/2} \int_0^1 \left(\int_{|z|\geq 1/\epsilon_n} e^{-\frac{1}{2}z^2} (\epsilon_n z)^2 dz + \int_{|z|\geq 1/\epsilon_n} e^{-\frac{1}{2}z^2} \log^2(1/\epsilon_n) dz \right) p_0(x) dx \\ & \leq 4(2\pi)^{-1/2} E_0 \left\{ \epsilon_n^2 \int_{t>1/\epsilon_n^2} e^{-t/2} t^{1/2} dt + \log^2(1/\epsilon_n) P(|Z| \geq 1/\epsilon_n) \right\} \\ & \leq 4(2\pi)^{-1/2} E_0 \left\{ \epsilon_n^2 \int_{t>1/\epsilon_n^2} e^{-t/4} dt + \log^2(1/\epsilon_n) e^{-\epsilon_n^{-2}/8} \right\} \\ & \lesssim e^{-\epsilon_n^{-2}/8+\log \log(1/\epsilon_n)} < \epsilon_n^2, \end{aligned}$$

where $Z \sim N(0, 1)$ and $E_0(\cdot)$ denotes taking expectation with respect to the measure associated with the density p_0 . The third line in the preceding uses the change of variables letting $t = z^2$.

Combining the above result for I_2 and the result in Equation (41) for I_1 , we have shown $\int g_{f_0, p_0} (\log g_{f_0, p_0}/g_{f, p})^2 \lesssim \epsilon_n^2$. And further we have

$$\begin{aligned} & \left\{ \int g_{f_0, p_0} \log \frac{g_{f_0, p_0}}{g_{f, p}} \lesssim \epsilon_n^2, \int g_{f_0, p_0} \left(\log \frac{g_{f_0, p_0}}{g_{f, p}} \right)^2 \lesssim \epsilon_n^2 \right\} \\ & \quad \supset \{ \|f - f_0\|_\infty \leq \epsilon_n, KL(p_0, p) \leq \epsilon_n^2 \}, \end{aligned}$$

which yields the conclusion in Lemma 8.

Appendix F. Posterior Computation: A Gibbs Sampler

In the following, we develop a Gibbs sampler to generate a Markov chain which will eventually converge to the posterior distribution. We focus on the Gaussian process associated with a squared exponential kernel as an illustration (in practice the algorithm can be applied to other kernels as long as they are symmetric). The squared exponential kernel is denoted by $c(x, x') = \exp\{-(x - x')^2/\lambda\}$ associated with a bandwidth parameter λ . Theorem 2 enforces the prior distributions $w_j \sim N(0, 2/\lambda)$, $s_j \sim \text{Unif}[0, 2\pi]$ and $a_j \sim N(0, 1)$ i.i.d. for $j = 1, \dots, N$. To ensure the conditional conjugacy, we place a gamma distribution $\text{Ga}(a_0, b_0)$ on the bandwidth λ with a shape parameter a_0 and a scale parameter b_0 . We place a Dirichlet process mixture of normals prior defined in Equation (8) over the covariate density, given more precisely by

$$X_i \sim \sum_{h=1}^{\infty} \pi_h N(\mu_h, \tau_h^{-1}), \quad (\mu_h, \tau_h) \sim N(\mu_h; \mu_0, \kappa_0 \tau_h^{-1}) \text{Ga}(\tau_h; a_\tau, b_\tau), \quad (42)$$

for $i = 1, \dots, n$. The prior on π_h is expressed as $\pi_h = \nu_h \prod_{l < h} (1 - \nu_l)$ where $\nu_l \sim \text{Beta}(1, \alpha)$. Here we let $\alpha = 1$. Denote the cluster label of X_i by $S_i \in \{1, \dots, K\}$ indicating that X_i is associated with S_i th component in the Dirichlet process Gaussian mixture prior for $i = 1, \dots, n$. Then Equation (42) can be also written as

$$X_i | S_i, \mu, \tau \sim N(\mu_{S_i}, \tau_{S_i}^{-1}), \quad (\mu_{S_i}, \tau_{S_i}) \sim N(\mu_{S_i}; \mu_0, \kappa_0 \tau_{S_i}^{-1}) \text{Ga}(\tau_{S_i}; a_\tau, b_\tau), \\ i = 1, \dots, n.$$

In both simulation studies and the real application, we set the hyperparameters $\mu_0 = 0, \kappa_0 = 1, a_\tau = 1, b_\tau = 1$, and we choose $a_0 = 5, b_0 = 1$ for the hyperprior $\text{Ga}(a_0, b_0)$. We remark that these hyperparameter choices are based on our preliminary numerical experiments. In addition, recall that we assume $\sigma = 0.2$ in simulation studies and we treat σ^2 as an unknown parameter endowed with an objective prior in real application.

As below we provide a complete updating scheme of the Gibbs sampler. We use bold symbols to distinguish the vectors $\mathbf{a}, \mathbf{w}, \mathbf{s}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}, \mathbf{S}, \mathbf{X}, \mathbf{Y}, \mathbf{W}$ accordingly. Then the joint posterior distribution of $\{\mathbf{a}, \mathbf{w}, \mathbf{s}, \lambda, \mathbf{X} | \mathbf{Y}, \mathbf{W}\}$ given observations $\{\mathbf{Y}, \mathbf{W}\}$ can be factorized as

$$[\mathbf{a}, \mathbf{w}, \mathbf{s}, \lambda, \mathbf{X} | \mathbf{Y}, \mathbf{W}] \propto [\mathbf{Y} | \mathbf{X}, \mathbf{a}, \mathbf{w}, \mathbf{s}, \lambda] \times [\mathbf{W} | \mathbf{X}] \times [\mathbf{w} | \lambda] \times [\lambda] \times [\mathbf{a}] \times [\mathbf{s}] \times [\mathbf{X}].$$

The updating scheme runs as follows:

1. Update $[\mathbf{w} | -]$ in a block by sampling $[w_j | -] \propto [\mathbf{Y} | \mathbf{X}, \mathbf{a}, \mathbf{w}, \mathbf{s}, \lambda] N(w_j; 0, 2/\lambda)$ independently using Metropolis-Hasting algorithm for $j = 1, \dots, N$.
2. Update $[\mathbf{s} | -]$ in a block by sampling $[s_j | -] \propto [\mathbf{Y} | \mathbf{X}, \mathbf{a}, \mathbf{w}, \mathbf{s}, \lambda] \text{Unif}[0, 2\pi]$ independently using Metropolis-Hasting algorithm for $j = 1, \dots, N$.
3. Update $[\mathbf{a} | -]$ from a multivariate normal distribution $N(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, with the mean vector $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\Sigma}} \boldsymbol{\Phi}^T \mathbf{Y} / \sigma^2$, and the covariance matrix $\tilde{\boldsymbol{\Sigma}} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi} / \sigma^2 + \mathbf{I}_N)^{-1}$, where $\boldsymbol{\Phi}$ is a $n \times N$ Fourier basis matrix with (i, j) th element $\boldsymbol{\Phi}_{ij} = (2/N)^{1/2} \cos(w_j x_i + s_j)$ for $i = 1, \dots, n, j = 1, \dots, N$. And \mathbf{I}_N denotes a $N \times N$ identity matrix.

4. Update the parameters $[\mathbf{S}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi} \mid -]$ associated with the Dirichlet process Gaussian mixture prior as in Ishwaran and James (2001) with the number of mixture components truncated at 20.
5. Update $[\mathbf{X} \mid -]$ in a block by sampling

$$[X_i | S_i, X_{-i}, -] \propto N(Y_i; \boldsymbol{\Phi}_i^T \mathbf{a}, \sigma^2) N(W_i; X_i, \delta^2) N(X_i; \mu_{S_i}, \tau_{S_i})$$

using Metropolis-Hasting algorithm for $i = 1, \dots, n$. Here $\boldsymbol{\Phi}_i^T$ denotes the i th row of the matrix $\boldsymbol{\Phi}$ defined in Step 3.

6. Update $[\lambda \mid -]$ from a gamma distribution $\text{Ga}(\hat{a}, \hat{b})$ with $\hat{a} = a_0$ and $\hat{b} = b_0 / (1 + b_0 \sum_{j=1}^n w_j^2 / 4)$.
7. Update $[\sigma^2 \mid -]$ from a inverse-gamma distribution $\text{IG}(a_\sigma, b_\sigma)$ with $a_\sigma = n/2$ and $b_\sigma = (\mathbf{Y} - \boldsymbol{\Phi} \mathbb{1}_N)^T (\mathbf{Y} - \boldsymbol{\Phi} \mathbb{1}_N) / 2$, where $\mathbb{1}_N$ denotes a $n \times 1$ vector of ones. (This step will be implemented only in the real example of Section 5.)

In particular, in Metropolis-Hasting algorithm used for updating $\{w_j\}$ in Step 1, we consider a random walk proposal $w_j^{\text{prop}} \sim N(w_j^{\text{cur}}, 1/4)$ for $j = 1, \dots, N$, where w_j^{cur} denotes the current state and the proposal variance is tuned to obtain average pointwise acceptance rate around 0.7. In Metropolis-Hasting algorithm used for updating $\{s_i\}$ in Step 2, we consider the independence proposal $s_i^{\text{prop}} \sim \text{Unif}[0, 2\pi]$ for $i = 1, \dots, n$. We note that the averaged pointwise acceptance rate for s_i is around 0.6. Finally, to update $\{x_i\}$ in Step 5, we use an adaptive proposal $x_i^{\text{prop}} \sim N(W_i / \delta^2 + \mu_{S_i} \tau_{S_i}, 1 / (1 / \delta^2 + \tau_{S_i}))$ for $i = 1, \dots, n$ with the averaged acceptance rate around 0.8.

Construction of spontaneous credible bands. We provide one example of constructing the spontaneous credible bands (CB) with $\gamma = 0.95$ for out-of-sample prediction of some model $f(x, \theta)$ evaluated at a test data set x_{test} of size n_t , based on L number of posterior samples $\{\theta^{(l)}, l = 1, \dots, L\}$ of parameter θ associated with the model f . Denote by $f^{(l)}(x) = f(x; \theta^{(l)})$ for $l = 1, \dots, L$ and let $\hat{f}(x) = (1/L) \sum_{l=1}^L f(x; \theta^{(l)})$ denote the posterior estimate of the function. Then, for each $l = 1, \dots, L$, we first calculate the maximum distance between the functions $\hat{f}(x)$ and $f^{(l)}(x)$ over the test data points, defined as $d_l = \max_{i=1, \dots, n_t} |f^{(l)}(x_{\text{test}, i}) - \hat{f}(x_{\text{test}, i})|$. To find the simultaneous CB, we find the 95% quantile of $\{d_l\}$ denoted by $d_{95\%}$ and take $d_{95\%}$ as the half range of the simultaneous CB. Then we define the spontaneous 95% credible band as $[\hat{f}(x_{\text{test}}) - d_{95\%}, \hat{f}(x_{\text{test}}) + d_{95\%}]$.

Appendix G. Additional Numerical Results

In this section, we provide additional numerical results for $n = 250$ under the same setting in Section 4, refer to Table 2, Figure 5 and Figure 6. We include the AMSE values for estimating the true locations for GPEV_a and GPEV_n in Table 3 under all three settings of sample sizes. We also collect diagnostic summaries under the settings in Section 4 including the mixing of the Markov chain of hyperparameter associated with covariance kernel in Figures 7, marginal posterior density plot of covariate based on GPEV_a (Figure 8) and effective sample sizes for estimated function values over training data points for GPEV_a

and $GPEV_f$ in Figure 9. At last, we provide trace plots and density plots of parameters associated with $GPEV_a$ (Figure 10) for the real application in Section 5.

n	Method	δ^2					
		0.01	0.2	0.4	0.6	0.8	1
250	$GPEV_a$	0.23 (0.07)	0.86 (0.51)	2.10 (2.33)	3.44 (4.83)	3.60 (4.86)	4.86 (6.21)
	$GPEV_f$	0.21 (0.07)	0.78 (0.46)	1.62 (0.98)	2.80 (2.84)	2.94 (3.44)	4.26 (4.91)
	$GPEV_n$	0.24 (0.09)	4.24 (1.25)	10.41 (2.78)	14.43 (3.61)	18.28 (4.72)	20.23 (4.77)
	GP	2.31 (0.15)	4.44 (0.62)	7.38 (1.13)	10.06 (1.50)	12.28 (1.72)	14.29 (1.85)
	decon	0.48 (0.27)	2.99 (0.94)	7.45 (1.62)	11.91 (2.01)	15.57 (1.99)	18.17 (1.77)

Table 2: Averaged mean squared errors (AMSE) defined as $\mathbb{E}[K^{-1} \sum_{k=1}^K \{ \hat{f}(t_k) - f(t_k) \}^2]$ ($\hat{f}(\cdot)$ denotes the proposed estimator of f , $\mathbb{E}(\cdot)$ denotes taking average over replicates) on an evenly spaced grid (t_1, \dots, t_K) of size $K = 100$ over the interval $[-3, 3]$ and standard errors ($\times 10^2$) over 50 replicated data sets of size $n = 250$.

n	Method	δ^2					
		0.01	0.2	0.4	0.6	0.8	1
100	$GPEV_a$	0.92 (0.13)	13.27 (2.04)	26.41 (4.85)	37.88 (6.11)	47.68 (8.93)	57.79 (10.41)
	$GPEV_n$	0.94 (0.13)	12.86 (2.12)	34.17 (4.97)	50.73 (7.76)	64.25 (11.2)	76.86 (10.57)
250	$GPEV_a$	0.89 (0.09)	12.36 (1.52)	23.70 (2.95)	33.56 (4.11)	42.98 (5.16)	52.07 (6.38)
	$GPEV_n$	0.91 (0.09)	15.95 (1.59)	33.10 (3.41)	48.49 (4.47)	62.64 (5.80)	74.40 (6.54)
n	Method	δ^2					
		0.001	0.005	0.01	0.1	0.5	1
500	$GPEV_a$	0.098 (0.006)	0.46 (0.03)	0.88 (0.05)	6.54 (0.53)	28.04 (2.28)	50.63 (4.36)
	$GPEV_n$	0.098 (0.006)	0.47 (0.03)	0.89 (0.05)	7.61 (0.68)	40.26 (2.90)	74.42 (5.69)

Table 3: Averaged mean squared errors ($\times 10^2$) with standard errors ($\times 10^2$) in estimating the true locations defined as $\mathbb{E}[n^{-1} \sum_{k=1}^n (\hat{x}_k - x_k^*)^2]$, where $\{\hat{x}_k\}$ denote the posterior estimate (mean) of covariates and $\{x_k^*\}$ denote the true locations, and $\mathbb{E}(\cdot)$ denotes taking average over replicates. The AMSE values and standard deviations are averaged over 50 replicated data sets of size $n = 100, 250, 500$ separately.

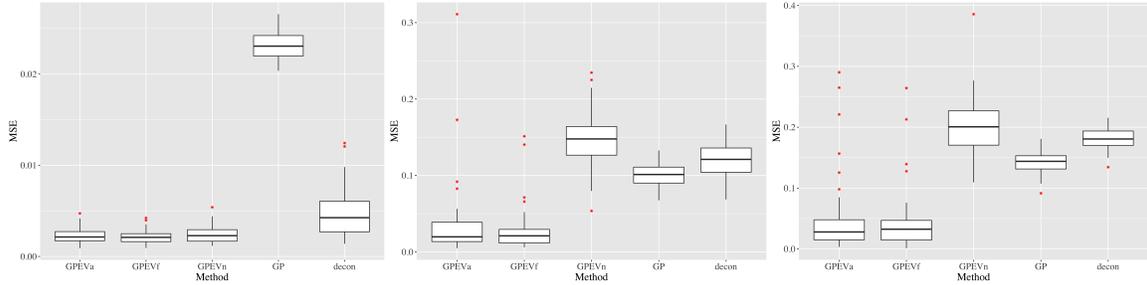


Figure 5: Boxplots of mean squared errors for compared methods in Section 4 over 50 replicated data sets of size $n = 250$ with $\delta^2 = 0.01$ (left panel), $\delta^2 = 0.6$ (middle panel) and $\delta^2 = 1$ (right panel). In each panel the compared methods from left to right are $GPEV_a$, $GPEV_f$, $GPEV_n$, GP and decon.

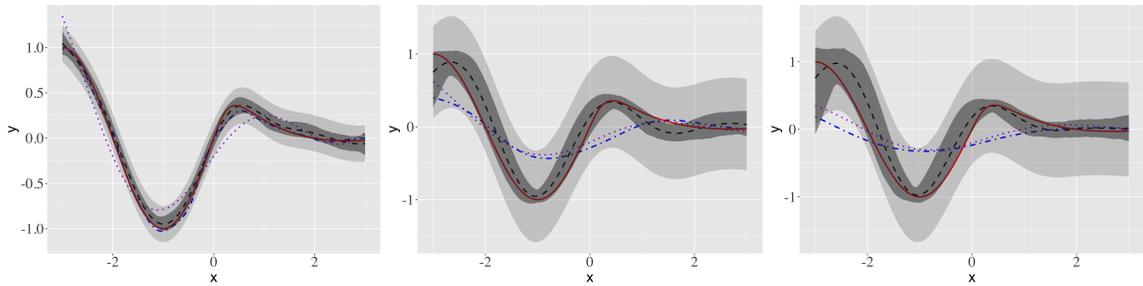


Figure 6: Out-of-sample predictions of $f(x)$ for $\delta^2 = 0.01$ (left panel), $\delta^2 = 0.6$ (middle panel) and $\delta^2 = 1$ (right panel) with sample size $n = 250$. The red solid line stands for the true function, the black dashed line stands for the predictive curve based on $GPEV_a$, the blue dot-dashed line is based on decon and the purple dotted dashed line is based on GP. The darker and the lighter shades are the pointwise and simultaneous 95% credible intervals of $GPEV_a$, respectively.

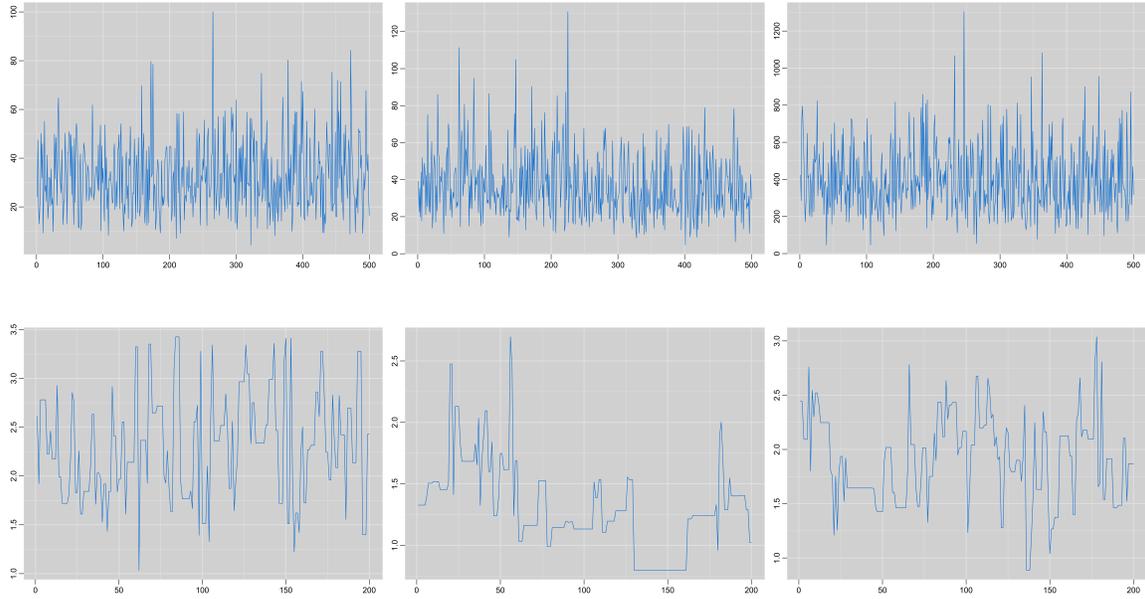


Figure 7: Trace plots of posterior samples of λ from $GPEV_a$ (first row) and $GPEV_f$ (second row) with sample size $n = 100$. In each row, the values of δ^2 are 0.01 (left panel), 0.6 (middle panel) and 1 (right panel).

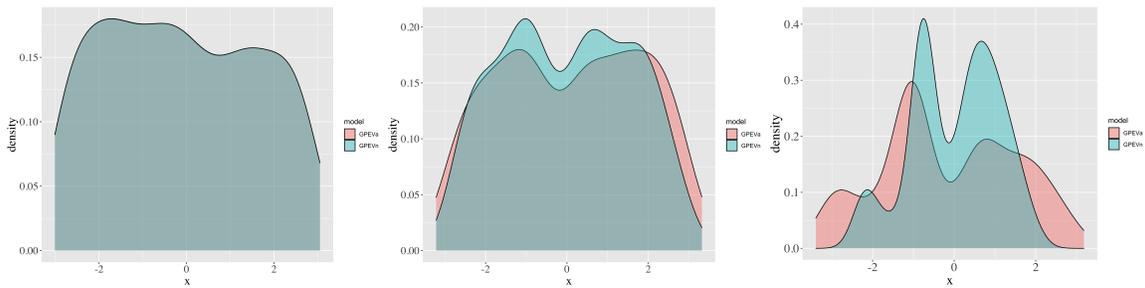


Figure 8: Marginal posterior density plots of the covariate based on $GPEV_a$ (red) and $GPEV_n$ (green) with $n = 500$. The values of δ^2 are 0.001 (left panel), 0.1 (middle panel), 0.5 (right panel).

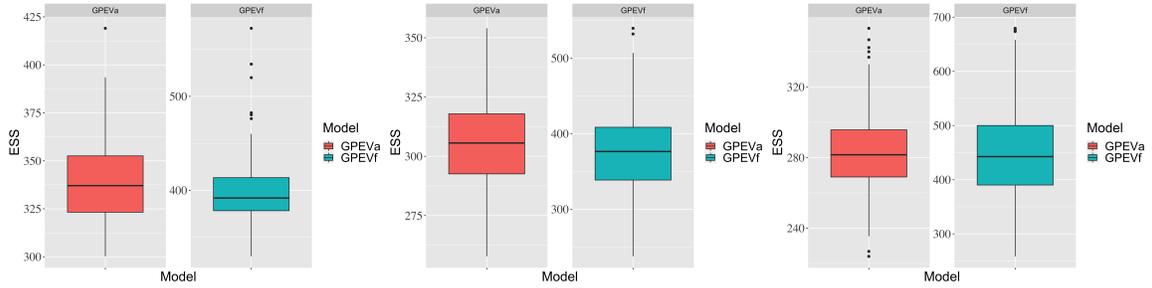


Figure 9: Boxplots of effective sample sizes of function value estimated over training data points based on $GPEV_a$ and $GPEV_f$ over replicated data sets of sizes $n = 100$ (left panel), $n = 250$ (middle panel), and $n = 500$ (right panel). The effective sample sizes are averaged over 50 replicates with $\delta^2 = 0.01$ for $n = 100, 250$ and $\delta^2 = 0.001$ for $n = 500$.

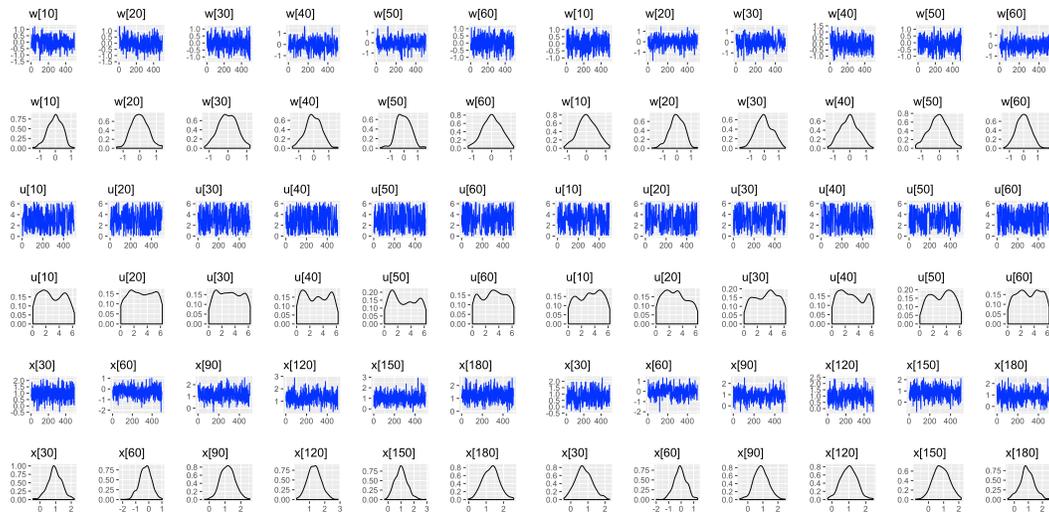


Figure 10: Trace plots and density plots of the 500 posterior samples of a subset of $\{w_j, s_j, x_j\}$ from treatment group with $\delta^2 = 0.35$ (left panel) and with unknown δ^2 (right panel) in the data example in Section 5.

References

- Robert J Adler. An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes. *Lecture Notes-Monograph Series*, 12, 1990.
- Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velinger, and Amir Zandieh. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International conference on machine learning*, pages 253–262. PMLR, 2017.
- Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.
- Sudipto Banerjee, Alan E Gelfand, Andrew O Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, 70(4):825–848, 2008.
- Yannick Baraud. Model selection for regression on a random design. *ESAIM: Probability and Statistics*, 6:127–146, 2002.
- Scott M Berry, Raymond J Carroll, and David Ruppert. Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97(457):160–169, 2002.
- Lucien Birgé. *Sur un théoreme de minimax et son application aux tests*. Univ. de Paris-Sud, Dép. de Mathématique, 1979.
- Olivier Bousquet. Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic Inequalities and Applications*, pages 213–247, 2003.
- Lawrence D Brown, T Tony Cai, Mark G Low, and Cun-Hui Zhang. Asymptotic equivalence theory for nonparametric regression with random design. *Annals of statistics*, 30(3):688–707, 2002.
- Raymond J Carroll and Peter Hall. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186, 1988.
- Raymond J Carroll, Helmut Küchenhoff, F Lombard, and Leonard A Stefanski. Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, 91(433):242–250, 1996.
- Raymond J Carroll, Jeffrey D Maca, and David Ruppert. Nonparametric regression in the presence of measurement error. *Biometrika*, 86(3):541–554, 1999.
- Daniel Cervone and Natesh S Pillai. Gaussian process regression with location errors. *arXiv preprint arXiv:1506.08256*, 2015.
- Christophe Chesneau. On adaptive wavelet estimation of the regression function and its derivatives in an errors-in-variables model. 2010.

- Michaël Chichignoud, Van Ha Hoang, Thanh Mai Pham Ngoc, and Vincent Rivoirard. Adaptive wavelet multivariate regression with errors in variables. *Electronic journal of statistics*, 11(1):682–724, 2017.
- Fabienne Comte and Claire Lacour. Anisotropic adaptive kernel deconvolution. In *Annales de l'IHP Probabilités et statistiques*, volume 49, pages 569–609, 2013.
- John R Cook and Leonard A Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328, 1994.
- Aurore Delaigle. An alternative view of the deconvolution problem. *Statistica Sinica*, pages 1025–1045, 2008.
- Aurore Delaigle. Nonparametric kernel methods with errors-in-variables: Constructing estimators, computing them, and avoiding common mistakes. *Australian & New Zealand Journal of Statistics*, 56(2):105–124, 2014.
- Aurore Delaigle and Irène Gijbels. Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Annals of the Institute of Statistical Mathematics*, 56(1):19–47, 2004a.
- Aurore Delaigle and Irène Gijbels. Practical bandwidth selection in deconvolution kernel density estimation. *Computational Statistics & Data Analysis*, 45(2):249–267, 2004b.
- Aurore Delaigle and Peter Hall. Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *Journal of the American Statistical Association*, 103(481):280–287, 2008.
- Aurore Delaigle and Alexander Meister. Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *Journal of the American Statistical Association*, 102(480):1416–1426, 2007.
- Aurore Delaigle, Peter Hall, and Peihua Qiu. Nonparametric methods for solving the Berkson errors-in-variables problem. *Journal of the Royal Statistical Society, Series B*, 68(2):201–220, 2006.
- Aurore Delaigle, Jianqing Fan, and Raymond J Carroll. A design-adaptive local polynomial estimator for the errors-in-variables problem. *Journal of the American Statistical Association*, 104(485):348–359, 2009.
- Sophie Donnet, Vincent Rivoirard, Judith Rousseau, and Catia Scricciolo. Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures. *Bernoulli*, 24(1):231–256, 2018.
- Lilun Du, Changliang Zou, and Zhaojum Wang. Nonparametric regression function estimation for errors-in-variables models with validation data. *Statistica Sinica*, 21(3):1093–1113, 2011.

- Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.
- Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19(3):1257–1272, 1991.
- Jianqing Fan. Deconvolution with supersmooth distributions. *Canadian Journal of Statistics*, 20(2):155–169, 1992.
- Jianqing Fan and Ja-Yong Koo. Wavelet deconvolution. *IEEE transactions on information theory*, 48(3):734–747, 2002.
- Jianqing Fan and Young K Truong. Nonparametric regression with errors in variables. *Annals of Statistics*, 21(4):1900–1925, 1993.
- Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230, 1973.
- Andrew O Finley, Huiyan Sang, Sudipto Banerjee, and Alan E Gelfand. Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis*, 53(8):2873–2884, 2009.
- Reinhard Furrer, Marc G Genton, and Douglas Nychka. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523, 2006.
- Fengnan Gao and Aad W van der Vaart. Posterior contraction rates for deconvolution of dirichlet-laplace mixtures. *Electronic Journal of Statistics*, 10(1):608–627, 2016.
- Subhashis Ghosal and Aad W van Der Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Annals of Statistics*, pages 1233–1263, 2001.
- Subhashis Ghosal and Aad van ver Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Annals of Statistics*, 35(2):697–723, 2007.
- Subhashis Ghosal, Jayanta K Ghosh, and Aad W van der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.
- Yongtao Guan. A composite likelihood approach in fitting spatial point process models. *Journal of the American Statistical Association*, 101(476):1502–1512, 2006.
- Joseph Guinness and Montserrat Fuentes. Circulant embedding of approximate covariances for inference from Gaussian data on large lattices. *Journal of Computational and Graphical Statistics*, 26(1):88–97, 2017.
- Peter Hall and Yanyuan Ma. Semiparametric estimators of functional measurement error models with unknown error. *Journal of the Royal Statistical Society: Series B*, 69(3):429–446, 2007.

- Peter Hall and Alexander Meister. A ridge-parameter approach to deconvolution. *Annals of Statistics*, 35(4):1535–1558, 2007.
- Patrick J Heagerty and Subhash R Lele. A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, 93(443):1099–1111, 1998.
- Dimitrios A Ioannides and Philippos D Alevizos. Nonparametric regression with errors in variables and applications. *Statistics & Probability Letters*, 32(1):35–43, 1997.
- Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.
- Jan Johannes. Deconvolution with unknown error distribution. *Annals of Statistics*, 37(5A):2301–2323, 2009.
- Maria Kalli, Jim E Griffin, and Stephen G Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011.
- Johanna Kappus and Gwennaëlle Mabon. Adaptive density estimation in deconvolution problems with unknown error distribution. *Electronic journal of statistics*, 8(2):2879–2904, 2014.
- Cari G Kaufman, Mark J Schervish, and Douglas W Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555, 2008.
- Bartek T Knapik, Aad W van der Vaart, and J Harry van Zanten. Bayesian inverse problems with Gaussian priors. *Annals of Statistics*, 39(5):2626–2657, 2011.
- Willem Kruijer, Judith Rousseau, and Aad W van der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.
- Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random fourier features. In *International conference on machine learning*, pages 3905–3914. PMLR, 2019.
- Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan AK Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148, 2021.
- Albert Y Lo. On a class of Bayesian nonparametric estimates: I. density estimates. *Annals of Statistics*, 12(1):351–357, 1984.
- Steven N MacEachern and Peter Müller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.
- Alexander Meister. *Deconvolution Problems in Nonparametric Statistics*. Lecture Notes in Statistics 193. Springer, Berlin, 2009.

- Iain Murray and Ryan Prescott Adams. Slice sampling covariance hyperparameters of latent Gaussian models. *arXiv preprint arXiv:1006.0868*, 2010.
- Radford M Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.
- Michael H Neumann. Deconvolution from panel data with unknown error distribution. *Journal of Multivariate Analysis*, 98(10):1955–1968, 2007.
- Debdeep Pati, Anirban Bhattacharya, and Guang Cheng. Optimal Bayesian estimation in random covariate design with a rescaled Gaussian process prior. *Journal of Machine Learning Research*, 16:2837–2851, 2015.
- Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008a.
- Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 555–561. IEEE, 2008b.
- Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21, 2008c.
- Carl Edward Rasmussen and Christopher KI Williams. *Gaussian Process for Machine Learning*. MIT Press, 2006.
- Kolyan Ray. Bayesian inverse problems with non-conjugate priors. *Electronic Journal of Statistics*, 7:2516–2549, 2013.
- Judith Rousseau and Catia Scricciolo. Wasserstein convergence in Bayesian deconvolution models. *arXiv preprint arXiv:2111.06846*, 2021.
- Abhra Sarkar, Bani K Mallick, John Staudenmayer, Debdeep Pati, and Raymond J Carroll. Bayesian semiparametric density deconvolution in the presence of conditionally heteroscedastic measurement errors. *Journal of Computational and Graphical Statistics*, 23(4):1101–1125, 2014.
- Weining Shen, Surya T Tokdar, and Subhashis Ghosal. Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013.
- Bharath Sriperumbudur and Zoltán Szabó. Optimal rates for random fourier features. *Advances in neural information processing systems*, 28, 2015.
- John Staudenmayer, David Ruppert, and John P Buonaccorsi. Density estimation in the presence of heteroscedastic measurement error. *Journal of the American Statistical Association*, 103(482):726–736, 2008.
- Leonard A Stefanski and Raymond J Carroll. Deconvolving kernel density estimators. *Statistics*, 21(2):169–184, 1990.

- Leonard A Stefanski and James R Cook. Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 90(432):1247–1256, 1995.
- Jonathan R Stroud, Michael L Stein, and Shaun Lysen. Bayesian and maximum likelihood estimation for Gaussian processes on an incomplete lattice. *Journal of Computational and Graphical Statistics*, 26(1):108–120, 2017.
- Ya Su, Anirban Bhattacharya, Yan Zhang, Nilanjan Chatterjee, and Raymond J Carroll. Nonparametric Bayesian deconvolution of a symmetric unimodal density. *arXiv preprint arXiv:2002.07255*, 2020.
- Danica J Sutherland and Jeff Schneider. On the error of random fourier features. *arXiv preprint arXiv:1506.02785*, 2015.
- Aad W van der Vaart and Harry van Zanten. Bayesian inference with rescaled Gaussian process priors. *Electronic Journal of Statistics*, 1:433–448, 2007.
- Aad W van der Vaart and J Harry van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pages 200–222. Institute of Mathematical Statistics, 2008.
- Aad W van der Vaart and J Harrys van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Annals of Statistics*, 37(5B):2655–2675, 2009.
- Aad W van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Science & Business Media, 1996.
- James Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Deisenroth. Efficiently sampling functions from gaussian process posteriors. In *International Conference on Machine Learning*, pages 10292–10302. PMLR, 2020.
- Andrew TA Wood and Grace Chan. Simulation of stationary Gaussian processes in $[0, 1]^d$. *Journal of Computational and Graphical Statistics*, 3(4):409–432, 1994.
- Yun Yang and David B Dunson. Bayesian manifold regression. *Annals of Statistics*, 44(2):876–905, 2016.
- Zitong Yang, Yu Bai, and Song Mei. Exact gap between generalization error and uniform convergence in random feature models. In *International Conference on Machine Learning*, pages 11704–11715. PMLR, 2021.
- Jian Zhang, Avner May, Tri Dao, and Christopher Ré. Low-precision random fourier features for memory-constrained kernel approximation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1264–1274. PMLR, 2019.