Gaussian Processes with Errors in Variables: Theory and Computation

Shuang Zhou[*2], Debdeep Pati[†2], Tianying Wang[‡3], Yun Yang [§1], and Raymond J. Carroll [¶4]

[2]Department of Statistics, Texas A&M University
[1]Department of Statistics, University of Illinois, Urbana Champaign
[3]Department of Biostatistics, Columbia University
[4]Department of Statistics, Texas A&M University and School of Mathematical and Physical Sciences, University of Technology Sydney

October 15, 2019

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 the 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 covariates density attain optimal rates of contraction adaptively over a range of Hölder classes, up to logarithmic terms. This improves upon the existing classical frequentist results which require knowledge of the smoothness of the underlying function to deliver optimal risk bounds. 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 demonstrate the empirical performance of our approach and compare it with competitors in a wide range of simulation experiments and a real data example.

---

[*]shuang@stat.tamu.edu
[†]debdeep@stat.tamu.edu (NSF DMS 1613156, 1854731, 1916371)
[‡]tw2696@cumc.columbia.edu (U01-CA057030)
[§]yy84@illinois.edu (NSF DMS 1810831)
[¶]carroll@stat.tamu.edu (U01-CA057030)

1 Introduction

1.1 Overview

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 a unknown regression function f on noise-contaminated covariates W_i's as

$$
\begin{aligned}
Y_i &= f(X_i) + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2), \\
W_i &= X_i + u_i, \quad u_i \sim g_u, \quad X_i \sim p_0 \quad (i = 1, \ldots, n),
\end{aligned}
\tag{1}
$$

where X_i's are the underlying true covariate variables. We denote p_0 as the marginal distribution of X_i, ϵ_i as the centered Gaussian error with unknown standard deviation σ, and g_u as the measurement error distribution. The goal here is to recover the unknown regression function f and the true density function p_0 of the covariate distribution.

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 deconvoluting kernel density estimators was proposed. Fan and Truong (1993) developed a globally consistent deconvolution kernel type estimator in a nonparametric regression problem and showed the optimal rate of convergence is of logarithmic order if the measurement error is normally distributed. 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. Deconvolution based on Fourier-techniques and local linear and polynomial estimators are also popular, such as SIMEX (simulation-extrapolation) Cook and Stefanski (1994) and Carroll et al. (1996, 1999); Delaigle et al. (2009); Delaigle and Hall (2008); Delaigle et al. (2006); Du et al. (2011); Stefanski and Cook (1995).

On the other hand, Bayesian procedures are naturally suited for general nonparametric regression tasks due to their ability to adapt to unknown smoothness and to allow quantifications of uncertainty. That being said, there is a relatively sparse literature on errors-in-variables problem in a Bayesian framework, let alone any theoretical development. Berry et al. (2002) were the first to develop a fully Bayesian procedure for the nonparametric regression problem using smoothing splines and P-splines. Staudenmayer et al. (2008) used the penalized mixture of B-splines to approximate the density and variance function in the heteroscedastic case. Sarkar et al. (2014) proposed a semiparametric Bayesian method based on B-splines for the regression function and in the presence of conditionally heteroscedastic measurement and regression errors. Cervone and Pillai (2015) developed a Bayesian analysis for Gaussian processes (GP) with location errors using hybrid Monte-Carlo techniques. Although the methods have been demonstrated to be very successful numerically, there is a clear dearth of theoretical results justifying these approaches.

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. Flexibility and richness aside, the immense popularity of these methods can be attributed largely to the development of sophisticated computational machinery that has made implementation of these techniques routine in various applied problems. To illustrate further credibility of such methods, frequentist consistency properties have also been given substantial attention in the literature and results of the type

$$E_{p_0}[\,\Pi_n\{d(p_0, p) > \xi_n \mid X^{(n)}\}\,] \to 0 \tag{2}$$

for a sequence of $\xi_n \to 0$, called posterior contraction rates, have been established, where p denotes the unknown parameter, $X^{(n)}$ denotes a set of precise measurements on X, d a distance metric, $\Pi_n\{\,\cdot\,\mid X^{(n)}\}$ the posterior distribution given $X^{(n)}$, and E_{p_0} the expectation with respect to the true probability density p_0 of X. Such posterior convergence results are useful as they imply the frequentist convergence rate ξ_n for the associated Bayes estimators. Optimal rates of posterior convergence in density estimation using mixture models have been illustrated by Ghosal and van ver Vaart (2007); Kruijer et al. (2010); Shen et al. (2013). Bayesian nonparametric density estimation approaches, such as the Dirichlet process Gaussians mixture model, can be readily adapted to the problem of density deconvolution from a practical point of view. In contrast, in our deconvolution context the covariate density of interest is different from the data generating density of the noise-contaminated covariate, making our theoretical investigation of consistency properties of the posterior substantially different and more difficult.

To the best of our knowledge, the only existing results available in the Bayesian literature are Donnet et al. (2018); Sarkar et al. (2013). In the former papers, an adaptive optimal contraction rate is proved in a density deconvolution problem. A formal theoretical justification for the use of Bayesian procedures in the errors-in-variables regression problem is missing. In this paper, we propose a fully Bayesian framework for errors-in-variables regression using Gaussian process prior, and develop a new theoretical framework for studying its frequentist properties including consistency and the quantification of posterior convergence rates. The optimal rate in the errors-in-variables problem with Gaussian error has been proved to be extremely slow, rendering inefficient inference in applications. In fact, the decaying error variance plays a very important role in improving the rate of the convergence. Such observation can be found in Fan (1992), where the measurement error standard deviation is allowed to decrease at the certain rate (the same as the optimal rate of the bandwidth) to enable the contraction rate of the deconvolution estimator to be as fast as that of the ordinary density estimator.

In this paper, we show that in an 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 frequentist estimators to construct test functions with type-I and type-II error bounds of the type $\exp(-Cn\epsilon_n^2)$ for the testing problem

$$H_0 : p = p_0, \quad \text{vs} \quad H_A : p \in \{p : d(p, p_0) > \xi_n\}. \tag{3}$$

Ray (2013) used concentration properties of thresholded wavelet based estimators based on standard results on concentration of Gaussian priors. However, analogous results fo kernel density estimators suited to density deconvolution problems are lacking. One of our key technical contributions is to develop sharp concentration inequalities of the kernel density estimators to construct tests in (3).

On the computational side, although the 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 regression. 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 the need to do matrix inversion. It also preserves the covariance matrix of a Gaussian process, thereby facilitating easy prior elicitation. In addition to the 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.

1.2 Review of nonparametric regression with errors-in-variables

Consider the regression model (1) with errors in variables, where $\{(Y_i, W_i) (i = 1, \ldots, n)\}$ are independent and identical random variables. Recall that Y_i's denote the observed responses and W_i's are contaminated covariates. We assume that Y_i is conditionally independent of W_i given X_i, for $i = 1, \ldots, n$, where the X_i's denote the unknown covariates having density p_0. The error density g_u considered in the existing literature (Fan and Truong, 1993) can be classified into two major types: the ordinary-smooth distributions whose characteristic function has polynomial decay, such as the

gamma and double-exponential distribution; and the super-smooth distributions whose characteristic function has exponential decay such as the Gaussian and Cauchy. The measurement error distribution g_u controls the rate of convergence of the estimators. It is well known that in absence of any replicated proxy per data-point, the optimal rate for a super-smooth 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).

In a classical context, starting with the construction of the deconvoluting kernel based on some suitable kernel function $K(\cdot)$ and the empirical estimator of the Fourier transform of the marginal density p, one can derive the deconvolution kernel density estimator (Fan and Truong, 1993) of regression function f and marginal density p by

$$
\begin{aligned}
\widehat{p}_n(x) &= \frac{1}{nh} \sum_{i=1}^{n} K_n\{(x - W_i)/h\}, & (4) \\
\widehat{f}_n(x) &= \frac{1}{nh} \sum_{i=1}^{n} K_n\{(x - W_i)/h\} Y_i/\widehat{p}_n(x), & (5) \\
K_n(x) &= \frac{1}{2\pi} \int e^{-itx} \frac{\phi_K(t)}{\phi_u(t/h)} dt. & (6)
\end{aligned}
$$

Here K_n is the deconvoluting kernel function, ϕ_K and ϕ_u are the Fourier transforms of the kernel function K and the density of measurement error g_u, respectively. Usually ϕ_K is assumed to be compactly supported to ensure K_n is well defined. Also, to achieve the rate optimality one requires that K is a kth-order kernel function with k being the order of smooth of the unknown regression function. In practice, such deconvolution kernels typically do not admit closed form expressions, and the estimation could suffer from extra errors due to numerical integrations.

1.3 Bayesian nonparametric regression with errors-in-variables

In this article, we consider the normal measurement error distribution $N(0, \delta^2)$ with unknown variance δ^2. In the Bayesian framework, we obtain the posterior distribution of unknown parameters $\theta = (f, p, \delta)$ given the observed values $D_n = \{(Y_i, W_i), i = 1, \ldots, n\}$ via Bayes' rule:

$$
\mathrm{pr}(\theta \mid D_n) = \frac{\mathrm{pr}(D_n \mid \theta)\,\mathrm{pr}(\theta)}{\mathrm{pr}(D_n)}.
$$

This posterior distribution $\mathrm{pr}(\theta \mid D_n)$ can then be used to conduct statistical inference on p and f, such as constructing point estimators and their associated credible intervals or bands. We consider the following generic Bayesian hierarchical model for nonparametric regression with errors

in variables:

$$Y_i = f(X_i) + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2),$$
$$W_i = X_i + u_i, \quad u_i \sim \mathrm{N}(0, \delta^2), \quad X_i \sim p \quad (i = 1, \ldots, n), \tag{7}$$
$$f \sim \Pi_f, \quad p \sim \Pi_p, \quad \sigma^2 \sim \Pi_{\sigma^2}, \quad \delta^2 \sim \Pi_{\delta^2}.$$

Variants of model defined in (7) 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, for our theoretical analysis, we assume δ to be known and let δ^2 decrease to 0 at a certain rate with respect to n. This is equivalent to having replicated proxies per observation which helps to recover the unknown regression function with more accuracy even in the presence of a Gaussian error distribution (Fan and Truong, 1993). For practical purpose we assign an objective prior on σ^2, the details of which can be found in Sections 3.2 and in Appendix.

By assigning proper priors on f and p, we show that the estimation of f and p can be made adaptive, meaning that 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. The details of choosing the specific priors Π_f on the function space and Π_p on the probability space are discussed in the following subsection. Observe that unlike the deconvolution kernel estimator, a Bayesian method does not require explicitly constructing a deconvoluting kernel function K_n, although the existence of such kernel is used in the proof for constructing the test function aforementioned in the introduction. In the following, we describe choices of Π_f which requires specifying a covariance kernel analogous to the kernel K.

1.4 Prior specifications

In this paper, we choose the prior Π_f for f as a Gaussian process prior (Rasmussen and Williams, 2006), 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 function $c(x, x') = \mathrm{cov}\{f(x), f(x')\}$. Therefore, any finite collection of random observation points $\{y_1(x_1), \ldots, y_N(x_N)\}$ at locations $x_1, \ldots, x_N$ has a joint Gaussian distribution given by

$$\{y_1(x_1), \ldots, y_N(x_N)\} \sim \mathrm{N}(m, \Sigma),$$

where $m = \{m(x_1), \ldots, m(x_N)\}$ and Σ is the covariance matrix with $\Sigma_{ij} = \tau^2 c(x_i, x_j)$. The mean function reflects the expected center of the realization, and the covariance function reflects its fluctuation and local dependence. The hyperparameter τ in the covariance function further controls the fluctuation magnitude. We use the notation $f(\cdot) \sim \mathrm{GP}(m(\cdot), \tau^2 c(\cdot, \cdot))$ to denote our function f follows a Gaussian process with mean function m and covariance function $\tau^2 c$. For the regular Gaussian process regression with noise level σ, the predictive formula (Rasmussen and

Williams, 2006) is

$$f(X^*) \mid X, Y, X^* \sim N(\bar{f}^*, \text{cov}\{f(X^*)\}),$$
$$\bar{f}^* = c(X^*, X)\{c(X, X) + \sigma^2 I\}^{-1} Y,$$
$$\text{cov}\{f(X^*)\} = c(X^*, X^*) - c(X^*, X)\{c(X, X) + \sigma^2 I\}^{-1} c(X, X^*),$$

where X, Y are the given data, X^* is the new data point, $f(X^*)$ is the prediction at X^* and $c(X^*, X)$ denotes the covariance matrix of X^* and X. Refer to Rasmussen and Williams (2006) for a detailed explanation of a Gaussian process. The posterior is a multivariate normal involved with the original data and the new data point. Choice of c is crucial to obtain a desirable functional estimation. A squared exponential covariance or more generally, a Matérn covariance kernel is commonly used in practice. The 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 (p, f). As a flexible prior distribution on the density p, we propose to use a Dirichlet process Gaussian mixture prior defined by

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

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 (X, Ω, P) with Ω the Borel sets of X, A Dirichlet process satisfies that for any finite and measurable partition $B_1, \ldots, B_k$ on X, $\{P(B_1), \ldots, P(B_k)\} \sim \text{Dir}\{\alpha G_0(B_1), \ldots, \alpha G_0(B_k)\}$, where $\text{Dir}\{a_1, \ldots, a_k\}$ denotes the Dirichlet distribution with parameters $a_1, \ldots, a_k$. 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 ver Vaart, 2007; Kruijer et al., 2010; Shen et al., 2013). Applying the Gaussian process prior to recover the true regression combined with modeling the marginal density with finite approximation of the Dirichlet process Gaussian mixture prior, we can correct for the bias due to the measurement error.

2 Theoretical Contraction Properties

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 $\|f\|_1 = \int |f(x)|dx$. We also define the supremum norm $\|f\|_\infty = \sup_{x \in S} |f(x)|$, where S is the domain of function f. Assume $\mathcal{C}^\beta[0,1]$ to be the Hölder space of β-smooth functions $f : [0,1] \to \mathbb{R}$ satisfying

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

for some constant $L > 0$. For any probability measure F on $\mathbb{R}$ let $p_{F,\sigma}(x) = \int \phi_\sigma(x-z)dF(z)$ stand for 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 $\mathrm{DP}(\alpha)$ denote the Dirichlet process with base measure α. We denote the posterior distribution by $\Pi_n(\cdot \mid D_n)$ and the prior distribution by $\Pi(\cdot)$. Here σ is the regression noise level and we assume $\sigma = 1$ in the following to simplify notations. Extension to general σ is straightforward.

2.2 Assumptions

Assumption 2.1. *The regression function $f_0 \in \mathcal{C}^\beta[0,1]$ with $\beta > 1/2$.*

We do not assume that β is known 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 hyperprior on the smoothness parameter of the Gaussian process. The lower bound on the smoothness is a common assumption in a random design regression, refer to Baraud (2002); Birgé (1979); Brown et al. (2002) for further discussion on this topic.

Assumption 2.2. *The marginal density p_0 of the unobserved covariates X is in $\mathcal{C}^{\beta'}[0,1]$ for $\beta' \geq \beta$, where β is defined in Assumption 2.1. Also, 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 (5) 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. Refer also to Delaigle and Meister (2007) where $f_0 p_0$ and p_0 are assumed to have the same smoothness.

The assumption $\beta' > \beta$ in Assumptions 2.1 and 2.2 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 the minimum of the contraction rates for estimating p_0 and f_0 separately. Although our theory is derived for compactly supported p_0, it can be extended to the unbounded support case with desirable tail conditions (Kruijer et al., 2010) on p_0.

In the Bayesian errors-in-variables model defined in (7), we assign a centered and rescaled Gaussian process prior on f, denoted as $\mathrm{GP}(0, c; A)$, associated with the squared exponential covariance kernel $c(x, x'; A) = \exp\{-A^2\|x - x'\|^2\}$ with the rescaled random variable A satisfying Assumption 2.3 below. This choice is motivated by the fact that a properly scaled squared exponential covariance kernel is known to lead to optimal rate of posterior convergence (van der Vaart et al., 2007; van der Vaart and van Zanten, 2009). We consider a Dirichlet process Gaussian mixture prior on the marginal density p defined as $p_{F,\widetilde{\sigma}}$, with $F \sim \mathrm{DP}(\alpha)$ and $\widetilde{\sigma} \sim G$, where G satisfies the Assumption 2.4 below. For convenience, to derive the frequentist theoretical properties of the Bayesian errors-in-variables model, we assume the response noise level $\sigma = 1$.

Assumption 2.3. *We assume the rescaled parameter A possesses a density m satisfying for sufficiently large $a > 0$,*

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

for constants $C_1, C_2, D_1, D_2 > 0$ and $p, q \ge 0$. For technical reasons 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 some positive constant A_0.

Assumption 2.3 includes the gamma density as a special case when $q = 0$. A similar assumption appears in van der Vaart and van Zanten (2009).

Assumption 2.4. *The Dirichlet process Gaussian mixture prior on the marginal density $p(x)$ defined by $p_{F,\widetilde{\sigma}}$ with $F \sim \mathrm{DP}(\alpha)$ and $\widetilde{\sigma} \sim G$, satisfy the following conditions:*

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

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

The inverse-gamma density on $\widetilde{\sigma}$ satisfies the above assumptions, whereas the inverse-gamma density on $\widetilde{\sigma}^2$ does not. This is a fairly standard assumption in the Bayesian asymptotics literature on the Dirichlet process mixture of Gaussians, 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 (1) we define the marginal likelihood of a random pair (Y, W) by $g_{f,p}(y, w) = (2\pi\delta)^{-1} \int \phi_1\{y - f(x)\}\phi_\delta(w - x)p(x)dx$ and the corresponding distribution is denoted

by $G_{f,p}$, so the posterior distribution can be written as

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

where B is any measurable subset of $\mathcal{P} = \{(f,p) : f : [0,1] \to \mathbb{R}, \text{ a continuous function, } p : [0,1] \to \mathbb{R}, \text{ a density function}\}$.

Theorem 2.5. *Suppose f_0 and p_0 satisfy Assumptions 2.1 and 2.2 respectively, and the prior Π on (f,p) satisfies the Assumptions 2.3 and 2.4. Then for some fixed large constant $M > 0$, sufficiently large n, and the standard deviation of the measurement error δ_n,*

$$\Pi_n\{(f,p) : \|f - f_0\|_1 < M \max(\epsilon_n, \delta_n^\beta), \|p - p_0\|_1 < M \max(\epsilon_n, \delta_n^\beta) \mid Y_{1:n}, W_{1:n}\}$$
$$\to 1 \text{ almost surely in } G_{f_0,p_0},$$

where $\epsilon_n = n^{-\beta/(2\beta+1)}(\log n)^t$ with $t = \max\{(2 \vee q)\beta/(2\beta+1), 1\}$. When $\delta_n \lesssim \epsilon_n^{1/\beta}$, the convergence rate is a multiple of ϵ_n.

The proof of Theorem 2.5 can be found in Appendix. Existing convergence rate results in the frequentist deconvolution literature (Fan and Truong, 1993) require the knowledge of the smoothness of the true covariate density and the regression function to achieve the optimal convergence rate for the regression function. Our Theorem 2.5, on the other hand, achieve minimax optimal rates of posterior convergence adaptively over all smoothness levels (β', β) with $\beta' > \beta$ defined in Assumptions 2.1 and 2.2 as long as we ensure the noise variance is sufficiently small (or the number of replications is sufficiently large). Since the proposed prior distribution does not require any knowledge of the smoothness of either f_0 or p_0. To understand the implication of the convergence rate in Theorem 2.5 let us focus on the case $\beta = 1$. 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 density of X can be recovered from the observations W. This intuitively justifies the rate $\max(\epsilon_n, \delta_n^\beta)$ in this case.

Analyzing the posterior requires upper bounding the numerator of the posterior and lower bounding the denominator (Ghosal et al., 2000). The upper bound of the numerator is obtained by constructing a sequence of test functions using the deconvolution kernel estimator. We also obtain sharp bounds for the Type I and Type II errors by developing large deviation bounds for the estimators. To lower-bound the denominator of the posterior we need both priors on the regression and marginal density to assign enough mass around the true. A single Gaussian prior on the covariates cannot concentrate enough in the 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 Gaussians prior allows borrowing of information, naturally exploits the smoothness and provides adequate concentration.

9

3 Posterior computation

3.1 Sampling from the posterior distribution

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 which depend on the latent covariates and b) sampling from the conditional distribution of the true covariates, which is intractable. Step 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 the covariates. To bypass the $O(n^3)$ 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 is different 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 considering the prediction as missing data imputation (Guinness and Fuentes, 2017; Stroud et al., 2017). However, it is not straightforward to translate these ideas to an errors-in-variables regression problem as the true covariates are contaminated and the true marginal distribution remains unknown. Instead, we consider using a lower dimensional mapping to approximate the Gaussian process based on the random Fourier basis proposed by Rahimi and Recht (2008) which has the same covariance kernel as the original Gaussian process. This avoids computing the inverse of covariance matrix by introducing more parameters in 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 in to approximate the original Gaussian process arbitrarily well; refer to Theorem 3.1. We describe the approximate Gaussian process in the following Section 3.2.

3.2 An approximation of the Gaussian process

We develop a low-rank random Fourier basis projection of a stationary mean Gaussian process $\text{GP}(0, c)$ with the corresponding spectral density ϕ_c defined through $c(h) = \int e^{ihx} \phi_c(x) dx$. For a suitably chosen large integer N, we define

$$\widetilde{f}_N(x) = (2/N)^{1/2} \sum_{j=1}^{N} a_j \cos(w_j x + s_j), \tag{9}$$

where $a_j \sim \text{N}(0,1)$, $w_j \sim \phi_c$ and $s_j \sim \text{Unif}(0, 2\pi)$, for $j = 1, \ldots, N$. The proposed approximation preserves the original covariance kernel function, furthermore it weakly converges to the original Gaussian process. We formalize our results in Theorem 3.1.

Theorem 3.1. *Suppose f is the original Gaussian process $\text{GP}(0, c)$ and $\widetilde{f}_N$ is defined in (9), we have $\widetilde{f}_N \to f$ in distribution as $N \to \infty$. And for any $x, y \in \mathbb{R}$,*

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

The proof of Theorem 3.1 can be found in Appendix. The construction $\widetilde{f}_N$ is related to the random feature map in the Fourier domain (Rahimi and Recht, 2008), used to project the kernel onto a lower-dimension space $\mathbb{R}^N$. For fitting $\widetilde{f}_N$ to the data, it suffices to treat $\{a_j, w_j, s_j \, (j = 1, \ldots, N)\}$ as unknown parameters endowed with independent priors. In practice, larger N leads to a better approximation, but is associated with a heavier computational burden. In the simulations and real data analysis, we find the approximated estimator performs almost as well as the original Gaussian process when N is chosen in the interval $(n/5, n/4)$ according to our numerical experiments.

4 Numerical results

We present numerical results of the proposed method and its variants in the following synthetic examples. Detailed posterior computation steps are in the Section F in the Appendix. We consider the uniform marginal distribution $X \sim \text{Unif}[-3, 3]$ and two regression functions: $f_1(x) = \sin(\pi x/2)/[1 + 2x^2\{\text{sign}(x) + 1\}]$ and $f_2(x) = (x + x^2)/4$. We consider sample size $n = 100, 250, 500$, with independently and identically distributed errors $\epsilon \sim \text{N}(0, \sigma^2)$ with fixed $\sigma = 0.2$. We only present the numerical results of $n = 500$, the results for $n = 100, 250$ were similar. For $n = 500$, we set the measurement error distribution to be $u \sim \text{N}(0, \delta^2)$ with $\delta^2 = 0.001, 0.005, 0.01, 0.1, 0.5, 1$. For each setting, we compare the following methods:

1. GPEV_a: Approximate Gaussian process method described in Section 3.2 with a Dirichlet process Gaussian mixture prior on marginal density.

2. GPEV_n: Approximate Gaussian process method described in Section 3.2 with a single normal prior on the covariates.

3. GPEV$_f$: Full scale Gaussian process model using regular predictive formula, see Rasmussen and Williams (2006) for more details; with a Dirichlet process Gaussian mixture prior on the marginal density.

4. GP: Gaussian process model that ignores the measurement error.

5. decon: Deconvoluting kernel method as in https://github.com/TimothyHyndman/deconvolve.

For $n = 500$, we used $N = 80$ respectively to implement GPEV$_a$ and GPEV$_n$. The choice of the hyperparameters and additional details for the Gibbs sampler can be found in Section F in the Appendix. For the Bayesian methods, the posterior mean denoted as $\widehat{f}$, is the estimator of the unknown regression function f with pointwise 95% credible intervals 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 $\widehat{f}$ with level $\beta \in (0, 1)$,

$$\mathrm{CB}_n(\beta) = \left\{f : \left\|f - \widehat{f}\right\|_\infty \leq r\right\},$$

where the half length r is chosen so that posterior probability of f falling into the credible band is 95%,

$$\Pi_n\big\{f \in \mathrm{CB}_n(\beta) \,\big|\, D_n\big\} = 0.95.$$

From Table 1, as δ^2 increases, we see GPEV$_a$ has the lowest mean squared error compared to the other methods for both f_1 and f_2, for instance, for f_1 with $\delta^2 = 1$, AMSE obtained by GPEV$_a$ is around 7, while the averaged mean squared errors of other methods are approximately twice. GP in particular has much larger mean squared errors, which suggests ignoring measurement error even in the case when the error is small significantly affects the function estimation. For f_2 we find that GP performs well when δ^2 is small, while decon is relatively worse. This can be explained from the fact that the function f_2 is smooth and less complicated, the contamination in covariate does not affect the function values too much in a small window. A careful inspection of our theory shows that the posterior distribution contracts towards the true function at a much slower rate if a single normal prior is placed on the covariate. This is empirically verified from the fact that GPEV$_n$ obtains much larger mean squared errors for larger δ^2, since it fails to estimate the covariates well. Interestingly, we find that GPEV$_a$ obtains smaller AMSE than GPEV$_f$ in many cases. One possible reason lies in the relative poor mixing of hyperparameters of GPEV$_f$. Figure 9 in Appendix shows the last 200 posterior samples of bandwidth parameter (λ) of squared-exponential kernel, where λ is defined in Section F in Appendix. It is evident that the mixing under GPEV$_a$ is much better than mixing with GP. The random Fourier basis representation provides more efficient way to update the smoothness parameter. The computation time of GPEV$_a$, GPEV$_n$, GPEV$_f$, GP, decon for one Markov chain iteration with sample size $n = 500$ for function f_1 are $0.055, 0.051, 9.50, 0.033, 2.00$

Table 1: *Averaged Mean Squared Errors (*AMSE*)* $\mathbb{E}\left[K^{-1}\sum_{k=1}^{K}\{\widehat{f}_j(t_k) - f_j(t_k)\}^2\right]$ *($\widehat{f}_j(\cdot)$ denotes the proposed estimator of $f_j, j = 1, 2$) over a regular grid $(t_1, \ldots, t_K)$ of size $K = 100$ in the interval $[-3, 3]$ and standard errors ($\times 10^2$) over 50 replicated data sets of size $n = 500$*

Function	Method	δ^2					
		0.001	0.005	0.01	0.1	0.5	1
f_1	GPEV$_a$	0·12 (0·04)	0·14 (0·06)	0·19 (0·08)	1·99 (0·66)	6·37 (2·33)	7·16 (3·82)
	GPEV$_f$	0·12 (0·04)	0·14 (0·06)	0·20 (0·08)	2·15 (0·73)	10·57 (2·77)	13·68 (4·94)
	GPEV$_n$	0·12 (0·04)	0·13 (0·05)	0·15 (0·06)	1·40 (0·45)	11·53 (2·85)	20·45 (5·80)
	GP	1·81 (0·09)	1·79 (0·08)	1·79 (0·10)	2·47 (0·27)	8·39 (1·13)	14·50 (1·78)
	decon	0·40 (0·35)	0·36 (0·22)	0·37 (0·21)	1·01 (0·37)	9·60 (1·61)	18·30 (1·62)

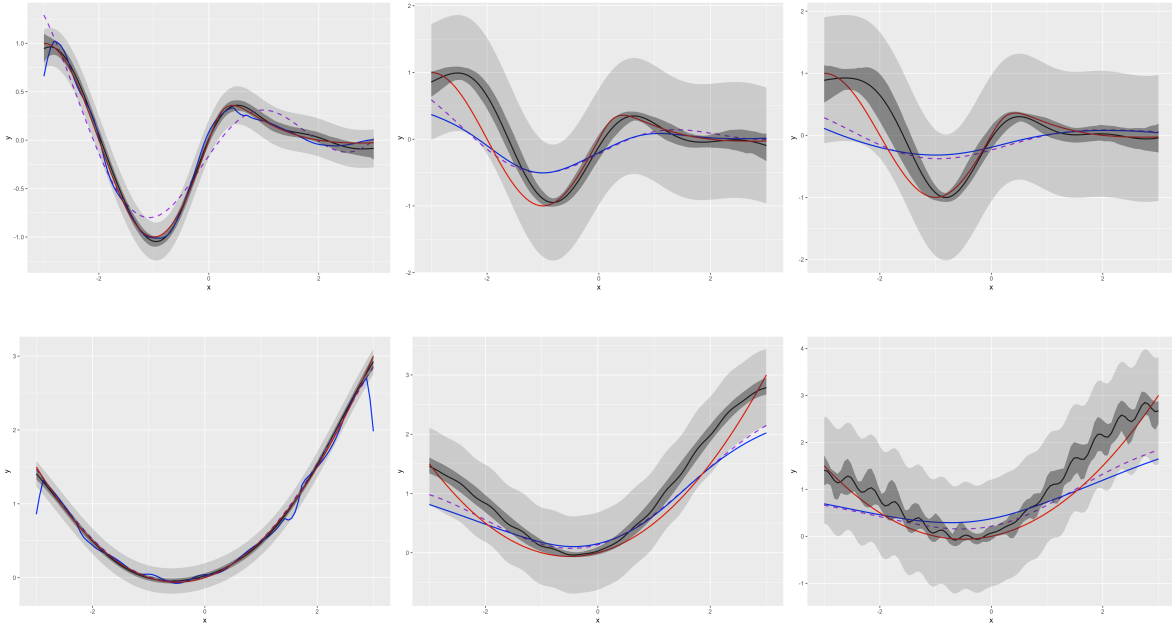| Function	Method	δ^2					
		0.001	0.005	0.01	0.1	0.5	1
f_2	GPEV$_a$	0·11 (0·12)	0·13 (0·11)	0·19 (0·09)	2·62 (0·63)	8·01 (3·05)	11·06 (5·39)
	GPEV$_f$	0·08 (0·04)	0·11 (0·06)	0·21 (0·10)	3·23 (0·72)	9·31 (3·01)	15·52 (12·51)
	GPEV$_n$	0·09 (0·08)	0·10 (0·04)	0·16 (0·18)	1·80 (0·41)	15·72 (2·68)	30·29 (7·15)
	GP	0·07 (0·03)	0·09 (0·04)	0·12 (0·04)	1·07 (0·24)	6·25 (1·02)	13·94 (2·03)
	decon	2·27 (2·09)	2·70 (2·77)	2·55 (2·60)	1·93 (1·54)	8·73 (1·34)	21·16 (2·66)

Figure 1: Predictions for $f_1(x)$ and $f_2(x)$ with $\delta^2 = 0.005$ (left panel), $\delta^2 = 0.5$ (middle panel) and $\delta^2 = 1$ (right panel). The first row shows predictions for $f_1(x)$ and the second row for $f_2(x)$. Sample size $n = 500$. The red line is the true function, the black line is the estimated function using GPEV$_a$, the blue line is for decon, the purple dashed line is for GP. The darker and the lighter shades are the pointwise and simultaneous 95% credible intervals obtained using GPEV$_a$.
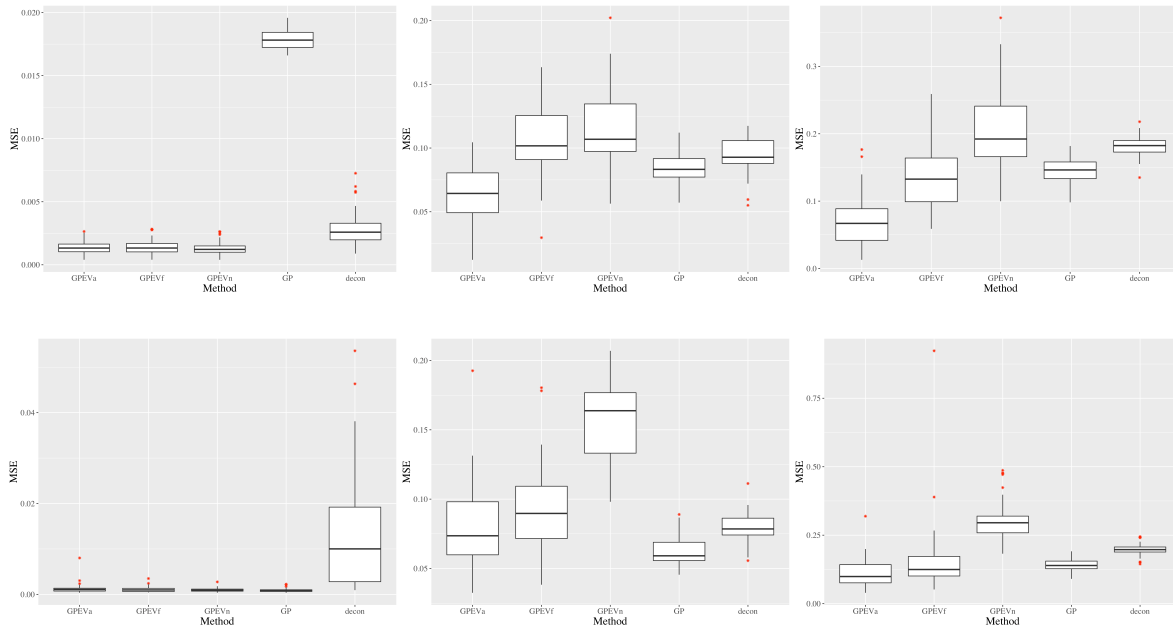
Figure 2: Boxplots of mean squared values for $f_1(x)$ and $f_2(x)$ over methods considered in Section 4 on 50 replicated data sets. First row shows the results for $f_1(x)$ and the second row for $f_2(x)$. Left panel is with $\delta^2 = 0.005$, middle panel with $\delta^2 = 0.5$ and right panel with $\delta^2 = 1$. Sample size $n = 500$. In each panel the methods from left to right are GPEV$_a$, GPEV$_f$, GPEV$_n$, GP and decon.

minutes respectively (with Intel Core i5 / 2.3 GHz processor). It is evident that GPEV$_a$ and GPEV$_n$ achieve substantial speed-up compared to GPEV$_f$, which indicates a strong advantage of employing GPEV$_a$ to the cases with larger sample size. Overall, GPEV$_a$ stands out as a more robust method for different functions as well as sample sizes.

Figure 1 shows the performance in function estimation on $[-3, 3]$. We see that for both f_1 and f_2, when $\delta^2 = 0.01$, all methods perform well, except decon which has an increasingly worse performance as n increases. For f_1 we see that GPEV$_a$ provides good prediction, preserving the function curvature with a slight drift caused by the measurement errors, with the 95% pointwise credible intervals containing the true function. On the other hand, both GP and decon methods are unable to recover the function shape well. A similar pattern is observed for estimating f_2. For small values of δ we see that all GPEV methods work better than GP and decon for both f_1 and f_2 in terms of predictive mean squared errors (MSE) in Figure 2. As δ^2 increases, GPEV$_a$ performs better than all the others, especially for f_1 with larger sample sizes and larger δ^2. For f_2, we observe similar results for decon and GPEV based methods. Figure 10 in Appendix shows the posterior marginal density of the covariates, for sample sizes are 500 and when the true function is f_1. When δ^2 is small, GPEV$_a$ recovers the true marginal distribution Unif$[-3, 3]$ reasonably well. However, as δ^2 increases, the density estimates increasingly deviate from the true marginal distribution confirming our theoretical results.

14

5 A case study

We re-analyzed the real data set studied in Berry et al. (2002) using the proposed methods. As mentioned in Berry et al. (2002), the data set was collected in 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 $\Delta(X) = f(X) - X$. To implement GPEV$_a$, we consider the normal zero-mean measurement error with two choices of the variance, fixed variance $\delta^2 = 0.35$, the estimated value from the study, and unknown variance δ^2 with an objective prior $\Pi(\delta^2) \propto 1/\delta^2$. We choose $N = 60$ and place $\exp(1.5)$ on λ, and treat σ^2 as an unknown parameter and assign the objective prior $\Pi(\sigma^2) \propto 1/\sigma^2$. To update σ^2 in the example, we use step 7 of the Gibbs sampler in Section F in Appendix.

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 group, 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 the groups. We see that for both the treatment and control groups, an objective prior on δ^2 results in similar estimation of $\Delta(X)$ as in the case of fixed δ^2. As for diagnostic checking in algorithm, the mixing of the Markov chains of $\{w_j, s_j, x_j \ (j = 1, \ldots, N)\}$ are good for both cases of δ^2. For more detail, see Figure 11 in Appendix for trace plots and density plots of the posterior samples.
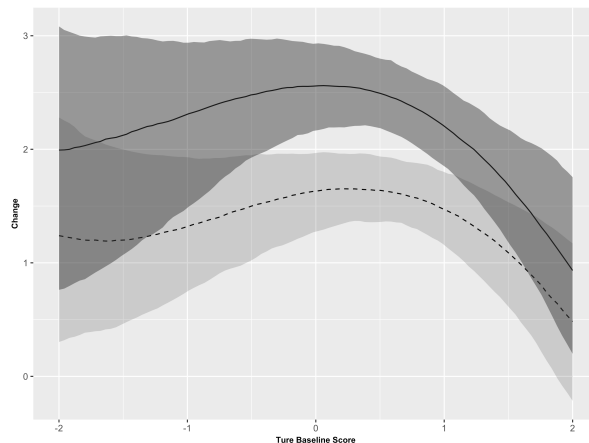

Figure 3: Estimate of $\Delta(X)$ at an equally-spaced grid over $[-2, 2]$ with $\delta^2 = 0.35$. The solid line indicates the treatment group with the darker shade as its 95% pointwise credible intervals and the dashed line indicates the control group with the lighter shade as its 95% pointwise credible intervals.
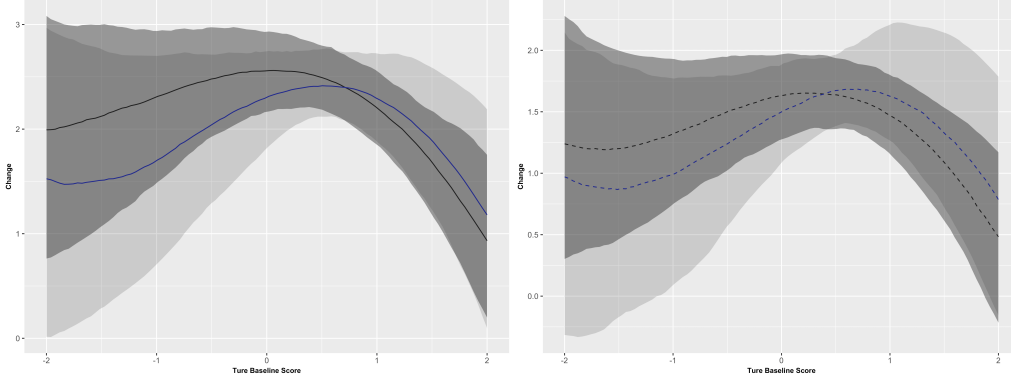
Figure 4: Estimate of $\Delta(X)$ at an equally-spaced grid over $[-2, 2]$ with different choices of δ^2. The left panel (solid lines) indicates the treatment group and the right panel (dashed lines) indicates the control group. The black lines with darker shade are 95% pointwise credible intervals using GPEV$_a$ with $\delta^2 = 0.35$, the blue lines with light shade as 95% pointwise credible intervals using GPEV$_a$ with $\Pi(\delta^2) \propto 1/\delta^2$.

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.

Although our current theoretical results on posterior contraction pertain to the original Gaussian process, our future work will involve extension to the approximate Gaussian process in light of Theorem 3.1.

7 Supplementary Material

R programs for computing our estimator are at Github address https://github.com/szh0u/Gaussian-Process-with-Errors-in-Variables. The data set used in Section 5 is available with those programs.

Acknowledgments

A Summary

Section B introduces the notations used in the proofs and reviews some background knowledge. Section C contains the proof of Theorem 2.5, followed with Section D containing the auxiliary results used to prove Theorem 2.5. The proof of Theorem 3.1 is in Section E. Section F contains the Gibbs sampler for posterior computation, and Section G contains the remaining numerical results for sample sizes $n = 100$ and $n = 250$; trace plots of posterior samples of covariate X as well as hyperparameters of GPEV$_a$ described in Section 4.

B Notations

We first introduce some notations used in the proofs. Denote E_X as the marginal expectation with respect to random variable X; denote $\mathrm{P}_{X,Y}^{f,p}$ as the probability measure of random pair (X,Y) which has joint density (f,p). Let $*$ denote the convolution, say, for two functions f and g, $f * g(x) = \int f(x-t)g(t)\,dt$. Let $\mathbb{1}_C$ denote the indicator function on any set C. Denote the Kullback-Leibler distance between f and g with respect to the Lebesgue measure μ by $KL(f,g) = \int f \log(f/g)\,d\mu$, and define the Kullback-Leibler divergence neighborhood of f_0 as $B_{f_0}(\epsilon) = \{f : \int f_0 \log(f_0/f) \le \epsilon^2, \int f_0 (\log(f_0/f))^2 \le \epsilon^2\}$. Next, we define the kth order kernel function $K(\cdot)$ satisfying,

$$\int K(u)\,du = 1, \quad \int K^2(u)\,du < \infty, \quad \int u^{\lfloor \beta \rfloor} K(u)\,du \ne 0, \tag{10}$$

$$\int u^{i-1} K(u)\,du = 0, \quad \text{for } i = 1, \ldots, \lfloor \beta \rfloor - 1, \ \beta \ge 2. \tag{11}$$

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 $(\mathbb{B}, \|\cdot\|)$, 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

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

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 with 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 $\mathbb{H}_1^a$ 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

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

Set $\mathcal{P}_n = \widetilde{B}_n \otimes \mathcal{F}$, where $\widetilde{B}_n = B_n \cap \mathcal{A}$ with $B_n = M_n \mathbb{H}_1^{a_n} + \epsilon_n \mathbb{B}_1$ and $\mathcal{A}$ as in Assumption 2.3.

C   Proof of Theorem 2.5

Denoting U_n as the set $\{f, p : ||f - f_0||_1 < M\epsilon_n, ||p - p_0||_1 < M\epsilon_n\}$, our target is to show $\Pi_n(U_n^c \mid Y_{1:n}, W_{1:n}) \to 0$ almost surely in G_{f_0,p_0}. We upper bound $\Pi_n(U_n^c \mid Y_{1:n}, W_{1:n})$ by $\Pi_n(f, p : ||f - f_0||_1 > M\epsilon_n \mid Y_{1:n}, W_{1:n}) + \Pi_n(p : ||p - p_0||_1 > M\epsilon_n \mid Y_{1:n}, W_{1:n})$. The second part is well-studied in the literature; refer to Shen et al. (2013) and the references therein which show that Dirichlet process Gaussian mixture prior leads to a posterior convergence rate $n^{-\beta'/(2\beta'+1)}$ where β' is the smoothness parameter of p_0. It remains to analyze first term. To that end, define the joint Kullback-Leibler neighborhood around (f_0, p_0) as

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

The following Contraction Theorem provides the sufficient conditions showing $\Pi_n(f, p : ||f - f_0||_1 > M\epsilon_n \mid Y_{1:n}, W_{1:n})$ converges almost surely to zero. A sketch of the proof is provided below.

Theorem C.1. *(Contraction Theorem) Consider model* (1) *and under the conditions in Theorem 2.5, let* $\mathcal{U}_n = \{||f - f_0||_1 > M\epsilon_n\}$. *If there exists a sequence of* $\epsilon_n \to 0$ *and* $n\epsilon_n^2 \to \infty$ *and a sequence of sieve* $\mathcal{P}_n \subset \mathcal{P}$ *and a sequence of test functions* $\phi_n = \mathbb{1}_{\{||\widehat{f}_n - f_0||_1 > (M-1)\epsilon_n\}}$ *satisfying the following conditions,*

$$G_{f_0,p_0} \phi_n \leq e^{-(C+4)n\epsilon_n^2}, \qquad \sup_{(f,p) \in \mathcal{P}_n \cap \mathcal{U}_n} G_{f,p} (1 - \phi_n) \leq e^{-(C+4)n\epsilon_n^2}, \tag{12}$$

$$\Pi\{B_{f_0,p_0}(\epsilon_n)\} \geq e^{-n\epsilon_n^2}, \tag{13}$$

$$\Pi(\mathcal{P}_n^c) \leq e^{-(C+4)n\epsilon_n^2}. \tag{14}$$

18

then $\Pi_n(\mathcal{U}_n^c \mid Y_{1:n}, W_{1:n}) \to 0$ *almost surely in* G_{f_0, p_0}, *for M as in Theorem 2.5.*

Proof. (Sketch) Define the set

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

Under the conditions in Theorem 2.5, 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 ϕ_n,

$$\Pi_n(\mathcal{U}_n^c \mid Y_{1:n}, W_{1:n}) \leq G_{f_0, p_0} \phi_n + G_{f_0, p_0}(C_n^c) + G_{f_0, p_0} \Pi(\mathcal{P}_n^c \mid Y_{1:n}, W_{1:n}) \mathbb{1}_{C_n}$$
$$+ G_{f_0, p_0} \Pi(\mathcal{U}_n \cap \mathcal{P}_n \mid Y_{1:n}, W_{1:n}) (1 - \phi_n) \mathbb{1}_{C_n}.$$

From (13) and (14), the third term goes to 0. From (12) and (13), the first and the fourth terms go to 0. □

We now discuss below how Contraction Theorem is employed to prove Theorem 2.5. We prove several auxiliary results in Appendix D which are useful to verify (12)-(14). The steps are

- (13) **of Theorem C.1:** Follows from Lemma D.4 under the conditions of Theorem 2.5.

- (14) **of Theorem C.1:** Follows from Lemma D.1 under the conditions of Theorem 2.5.

- (12) **of Theorem C.1:** For model (1), $\widehat{p}_n$ and $\widehat{f}_n$ are defined in (4) and (5), and $f, p \in \mathcal{P}_n$. It suffices to estimate $P_{Y,W,X}^{f_0,p_0}(||\widehat{f}_n - f_0||_1 > \epsilon_n)$ and $P_{Y,W,X}^{f,p}(||\widehat{f}_n - f||_1 > \epsilon_n)$. Following a similar line of argument in Meister (2009), for any marginal density p_0 satisfying Assumption 2.2 and $p \in \mathcal{P}_n \cup p_0$, define $\Delta p = (\widehat{p}_n - p)/p$ and for $f \in \mathcal{P}_n \cup f_0$ we have

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

By Assumption 2.2, p_0 is lower-bounded by some constant $B^{-1} > 0$. Then applying the inequality (16) in Lemma D.2, for any constant $\epsilon_0 > 0$ we have $||\widehat{p}_n - p||_\infty < \epsilon_0$ with probability at least $1 - e^{-n\epsilon_0 h_n^2}$. Thus for $p \in \mathcal{P}_n$, $||p - p_0||_\infty \leq ||\widehat{p}_n - p_0||_\infty + ||\widehat{p}_n - p||_\infty \leq 2\epsilon_0$ with probability at least $1 - e^{-n\epsilon_0 h_n^2}$. 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 $p \in \mathcal{P}_n$ we have

$$||\widehat{f}_n - f||_1 \leq \frac{1}{B_1} ||\widehat{f}_n \widehat{p}_n - fp||_1 \left(\left\| \frac{\Delta p}{\Delta p + 1} \right\|_\infty + 1 \right) + ||f||_\infty \left\| \frac{\Delta p}{\Delta p + 1} \right\|_1. \tag{15}$$

Since $|||\Delta p||_\infty \leq \epsilon_0/B_1$ with high probability, 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 $p = p_0$, we bound $||\Delta p/(\Delta p+1)||_1 \leq 2B||\widehat{p}_n - p_0||_1$. Combining the above results and (15), we obtain,

$$\mathrm{pr}(||\widehat{f}_n - f||_1 > \epsilon_n) \leq \mathrm{pr}(\widehat{f}_n \cdot \widehat{p}_n - f \cdot p||_1 > B_1\epsilon_n/4) + \mathrm{pr}\{||\widehat{p}_n - p||_1 > B_1\epsilon_n/(4||f||_\infty)\}$$
$$+ \mathrm{pr}(||\widehat{p}_n - p||_\infty > \epsilon_0).$$

Since we assume f_0 and $f \in \mathcal{P}_n$ are bounded, applying Lemma D.2 yields (12).

Thus $\Pi_n(\mathcal{U}_n^c \mid Y_{1:n}, W_{1:n}) \to 0$ as $n \to \infty$ almost surely in G_{f_0,p_0}.

D   Auxiliary results

Lemma D.1. *Suppose Assumptions 2.1, 2.2, 2.3 and 2.4 hold, by taking $M_n = a_n = \epsilon_n^{-1/\beta}$, $H \precsim n\epsilon_n^2, m^{\tau_1} \precsim n, \underline{\sigma} \precsim n^{-1/2\tau_2}$ and $\bar{\sigma}^{2\tau_3} \precsim e^n$, we have $\Pi(\mathcal{P}_n^c) \leq e^{-n\epsilon_n^2}$ with $\epsilon_n = n^{-\beta/(2\beta+1)}(\log n)^t, t = \max\{(2 \vee q)\beta/(2\beta+1), 1\}$.*

Lemma D.2. *For model (1), $\widehat{p}_n$ and $\widehat{f}_n$ defined in (4) and (5) and $f, p \in \mathcal{P}_n$ for any small constant $\epsilon_0 > 0$,*

$$P_{W,X}^p(||\widehat{p}_n - p||_\infty > \epsilon_0) \leq e^{-C_1 n\epsilon_0 h_n^2}, \tag{16}$$

$$P_{W,X}^p(||\widehat{p}_n - p||_1 > \epsilon_n) \leq e^{-n\epsilon_n^2}, \tag{17}$$

$$P_{Y,W,X}^{f,p}(||\widehat{f}_n \widehat{p}_n - f p||_1 > \epsilon_n) \leq e^{-n\epsilon_n^2}, \tag{18}$$

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 D.3. *Suppose Assumptions 2.2, 2.3 and 2.4 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 D.4. *Under the conditions in Theorem 2.5 and suppose Lemma D.3 hold, for sufficiently large n,*

$$\Pi\{B_{(f_0,g_0)}(\epsilon_n)\} \geq e^{-n\epsilon_n^2}, \tag{19}$$

where $\epsilon_n = n^{-\beta/(2\beta+1)}(\log n)^t$ with $t = \max\{(2 \vee q)\beta/(2\beta+1), 1\}$.

Lemma D.5. *(Theorem 7.3 in Bousquet (2003)) Suppose $\mathcal{G}$ is a countable set of functions $g : \mathcal{X} \to \mathbb{R}$ and assume all functions $g \in \mathcal{G}$ are measurable, squared-integrable and satisfy $E\{g(X_k)\} = 0$. Assume $\sup_{g\in\mathcal{G}} \mathrm{ess}\sup g$ 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*

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

Lemma D.6. *(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 $t > 0$,*

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

D.1 Proof of Lemma D.1

Based on the definition of sieves $\mathcal{P}_n$, $\mathcal{P}_n^c = (B_n^c \otimes \mathcal{F}) \cup (B_n \otimes \mathcal{F}^c) \cup (B_n^c \otimes \mathcal{F}^c)$, thus $\Pi(\mathcal{P}_n^c) \leq 2\{\Pi(B_n^c) + \Pi(\mathcal{F}^c)\}$. First we bound the second term in the right hand, by the Assumptions 2.3 and 2.4,

$$\Pi(\mathcal{F}^c) \leq H\bar{\alpha}([-m,m]^c) + pr(\widetilde{\sigma} \notin [\underline{\sigma}, \bar{\sigma}]) + pr\left(\sum_{h>H} \pi_h > \epsilon\right)$$

$$\leq He^{-b_1 m^{\tau_1}} + c_2 \bar{\sigma}^{-2\tau_3} + c_1 e^{-b_2 \underline{\sigma}^{-2\tau_2}} + \left(\frac{e|\alpha|}{H}\log\frac{1}{\epsilon}\right)^H$$

Choosing $m^{\tau_1} \precsim n, \underline{\sigma} \precsim n^{-1/2\tau_2}$ and $\bar{\sigma}^{2\tau_3} \precsim e^n$ with $\epsilon = \epsilon_n$ for same ϵ_n in Theorem 2.5, the first three terms in the second inequality can be bounded by a multiple of e^{-n} and by taking $H \precsim n\epsilon_n^2$ the last term can be bounded by,

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

Thus $\Pi(\mathcal{P}_n^c) \precsim e^{-c_4 n\epsilon_n^2}$ for every $c_4 > 0$. Now we bound $\Pi(\widetilde{B}_n^c)$. By definition, $\Pi(\widetilde{B}_n^c) = \Pi(B_n^c \mid \mathcal{A}) \leq \Pi(B_n^c)/pr(\mathcal{A})$, with $\mathcal{A}$ defined in the Assumption 2.3. By the fact $E(\|f\|_\infty) < \infty$, and $\sigma_f^2 = \sup_{x \in [0,1]} E\{f(x)\}^2 < \infty$, applying Borell's inequality in Lemma D.6, we have $pr(\mathcal{A}) = pr(\|f\|_\infty < A_0) \geq 1 - e^{-A_0^2/2\sigma_f^2} \geq a_0$, for some constants $A_0 > 0$ and $a_0 \in (0,1)$. Thus $\Pi(\widetilde{B}_n^c) \precsim \Pi(B_n^c) \precsim e^{-n\epsilon_n^2}$ if $M_n^2 \precsim n\epsilon_n^2$ and $a_n^2 \precsim n\epsilon_n^2$, more details can be found in the proof of Theorem 3.1 in van der Vaart and van Zanten (2009).

D.2 Proof of Lemma D.2

To prove Lemma D.2 we will prove the third assertion (18) in detail and discuss the key elements in the proof of the first two assertions (16)-(17) since they follow the similar line of argument. The key steps of the proof are application of Talagrand's inequality stated in Lemma D.5 and the fact that $\|K_n\|_1$ is bounded which is discussed in the following proposition.

Proposition D.7. *For any kernel function K satisfying (10)-(11) and K_n defined in (6), 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 (10)-(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 compact support. 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

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

thus $\int_{|s|\leq a}|K_n(s)|\,ds \precsim \exp(\delta_n^2/2h_n^2)$. 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

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

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. $\qquad\square$

Proposition D.8. *For $\widehat{p}_n$ and $\widehat{f}_n$ defined in (4) and (5) and for any $f, p \in \mathcal{P}_n$ we have*

$$||E_{W,X}(\widehat{p}_n) - p||_1 \precsim \epsilon_n, \quad ||E_{Y,W,X}(\widehat{f}_n\widehat{p}_n) - fp||_1 \precsim \epsilon_n, \tag{20}$$

with ϵ_n in Theorem 2.5.

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 $p = \phi_{\widetilde{\sigma}}*F$, by Cauchy-Schwarz inequality we have $||K_{h_n}*p - p||_1 \leq ||K_{h_n}*p - p||_2$. Again, we consider the kernel function K with the Fourier transform $\phi_K(t)$, by Parseval's theorem,

$$||K_{h_n}*p - p||_2^2 = \int |2\pi\phi_K(h_n t) - 1|^2|\widehat{p}(t)|^2 dt = \int_{|t|>1/h_n}|\widehat{F}(t)|^2|\widehat{\phi}_{\widetilde{\sigma}}(t)|^2 dt$$

$$\leq \int_{|t|>1/h_n}|\widehat{\phi}_\sigma(t)|^2 dt \leq (h_n/\underline{\sigma}^2)e^{-(\underline{\sigma}/h_n)^2/2} \precsim h_n^{-1}(\log n)^{-t_3}e^{-K^2(\log n)^{2t_3}/2} \precsim \epsilon_n^2,$$

for all $\widetilde{\sigma} \geq \underline{\sigma}$. Let $h_n \asymp \epsilon_n^{1/\beta}$ with ϵ_n in Theorem 2.5 and from Lemma D.1 we have $\underline{\sigma} \precsim n^{-1/2\tau_2}$, we choose τ_2 such that $\underline{\sigma} = Kh_n(\log n)^{t_3}$ for some constants $K^2/2 > 1$ and $t_3 > 1/2$. Now we consider the bias term of $\widehat{f}_n\widehat{p}_n$. By triangle inequality

$$||K_{h_n}*(fp) - fp||_1 \leq ||K_{h_n}*(fp) - pK_{h_n}*f||_1 + ||pK_{h_n}*f - fp||_1. \tag{21}$$

By Cauchy-Schwarz inequality, the first term of the right hand side of (21) can be bounded as

$$||K_{h_n} * (fp) - p K_{h_n} * f||_1 = \int |K_{h_n}(x-y)\{p(y) - p(x)\}f(y)\,dy|\,dx$$

$$\leq ||K_{h_n} * p - p||_2\, ||f||_2 \precsim ||K_{h_n} * p - p||_2,$$

since $||f||_2 \leq ||f||_\infty \leq A_0$. The second term of the right hand side of (21) can be bounded

$$||p K_{h_n} * f - fp||_1 \leq ||p||_1 ||K_{h_n} * f - f||_\infty = ||K_{h_n} * f - f||_\infty \precsim \epsilon_n,$$

the last equation holds from the properties of higher order kernel as in Lemma 4.3 of van der Vaart and van Zanten (2009). □

Now we prove the inequality (18). By triangle inequality,

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

First we estimate $\mathrm{pr}(I_{1,n} > \epsilon_n)$. By definition

$$\widehat{f}_n \widehat{p}_n - E_{Y,W|X}(\widehat{f}_n \widehat{p}_n) = \frac{1}{2\pi n h_n} \sum_{j=1}^n \int e^{-\frac{itx}{h_n}} \left\{ e^{itW_j/h_n} Y_j - E_{W|X}\left(e^{itW_j/h_n}\right) E_{Y|X}(Y_j) \right\} \frac{\phi_K(t)}{\phi_u(t/h_n)} dt$$

$$= \frac{1}{2\pi n h_n} \sum_{j=1}^n \int e^{-\frac{it(x-W_j)}{h_n}} \frac{\phi_K(t)}{\phi_u(t/h_n)} dt \{Y_j - E_{Y|X}(Y_j)\}$$

$$+ \frac{1}{2\pi n h_n} \sum_{j=1}^n \int e^{-\frac{itx}{h_n}} \left\{ e^{itW_j/h_n} - E_{W|X}\left(e^{itW_j/h_n}\right) \right\} \frac{\phi_K(t)}{\phi_u(t/h_n)} dt\, E_{Y|X}(Y_j)$$

$$:= T_{1,n} + T_{2,n}.$$

First, we estimate $\mathrm{pr}(||T_{2,n}||_1 > \epsilon_n/2)$. 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 $h \in L_\infty[0,1]$ and $||T_{2,n}||_1 = ||T||_{\mathcal{F}_1}$, where $\mathcal{F}_1 \subset L_\infty[0,1]$ is countable and dense. Thus we have

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

and $||n T_{2,n}||_1 = \sup_{k \in \mathcal{K}} |\sum_{j=1}^n k(W_j, X_j)|$. To apply Lemma D.5, 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 2.3 and 2.4 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 variable $s = (x-u)/h_n$, and for any fixed

positive constant a, $\int_0^1 |K_n\{(x-u)/h_n\}/h_n| \, dx \le \int |K_n(s)| ds \le C'$ for some constant C'. The second inequality holds by Proposition D.7. Since $W \mid X \sim N(X, \delta_n^2)$,

$$E_{W|X}\left\{K_n\left(\frac{x-W}{h_n}\right)\right\} = \frac{1}{2\pi} \int E_{W|X}\left(e^{-it[\{x-X-(W-X)\}/h_n]}\right) \frac{\phi_K(t)}{\phi_u(t/h_n)} dt$$

$$= \frac{1}{2\pi} \int e^{-it(x-X/h_n)} \phi_K(t) dt = K\{(x-X)/h_n\}.$$

Again by change of variable $r = (x-X)/h_n$, 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 $\|k\|_\infty \le K_1$ for any $k \in \mathcal{K}$, then $\sup_{k\in\mathcal{K}} \|k\|_\infty \precsim \max\{1, \exp(\delta_n^2/2h_n^2)\}$. Next we estimate the term $\sigma_\mathcal{K}^2$. For any $k \in \mathcal{K}$ and $W \mid X \sim N(X, \delta_n^2)$,

$$k(W,X)^2 = \frac{1}{h_n^2}\left(\int_0^1 \left[K_n\left(\frac{x-u}{h_n}\right) - E_{W|X}\left\{K_n\left(\frac{x-W}{h_n}\right)\right\}\right] f(X)h(x)dx\right)^2$$

$$\precsim \frac{1}{h_n^2}\left\{\int_0^1 K_n\left(\frac{x-u}{h_n}\right)dx\right\}^2 + \frac{1}{h_n^2}\left\{\int_0^1 E_{W|X}K_n\left(\frac{x-W}{h_n}\right)dx\right\}^2$$

$$\precsim \max\{1, \exp(\delta_n^2/h_n^2)\}.$$

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

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

$$E_{W|X}\left(\sup_{k\in\mathcal{K}}\left|\sum_{j=1}^n k(W_j, X_j)\right|\right) \le \left[E_{W|X}\left\{\sup_{k\in\mathcal{K}}\left|\sum_{j=1}^n k(W_j, X_j)\right|\right\}^2\right]^{1/2}$$

$$\precsim \left(\frac{1}{h_n^2}\sum_{j=1}^n E_{W|X}\left[\int\left|K_n\left(\frac{x-W_j}{h_n}\right) - E_{W|X}\left\{K_n\left(\frac{x-W_j}{h_n}\right)\right\}\right|dx\right]^2\right)^{1/2}$$

$$\precsim n^{1/2}\max\{1, \exp(\delta_n^2/2h_n^2)\}.$$

To apply the Lemma D.5, we choose $\delta_n \asymp h_n$ and same ϵ_n in Theorem 2.5, we have $\exp(\delta_n^2/2h_n^2) = O(1)$. By choosing $t = n\epsilon_n^2$, we have $n^{1/2} + (2(n+n^{1/2})n\epsilon_n^2)^{1/2} + n\epsilon_n^2/3 \precsim n\epsilon_n$.

We now discuss bounding the probability $\text{pr}(\|T_{1,n}\|_1 > \epsilon_n/2)$. Recall that

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

with $\widetilde{Y}_j \sim N(0,1)$ independently and identically for $j = 1, \ldots, n$, since $Y_j \mid X_j \sim N(f(X_j), 1)$ for $j = 1, \ldots, n$. Again by Hahn-Banach theorem there exists the countable and dense set $\mathcal{T} \in L_\infty[0,1]$ and the class of bounded linear functionals on $L_\infty[0,1]$,

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

24

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

To apply Lemma D.6, it suffices to estimate $\sigma_{\mathcal{Q}}^2 = \sup_{q\in\mathcal{Q}} E_{Y|X}\{\sum_{j=1}^n \widetilde{q}(W_j)\}^2$ and $E_{Y|X}\sup_{q\in\mathcal{Q}}\|q\|_\infty$. Again by change of variable and the fact $\|t\|_\infty \leq 1$ we have

$$E_{Y|X}\left\{\sum_{j=1}^n \widetilde{q}(W_j)\right\}^2 = \frac{1}{h_n^2}\sum_{j=1}^n\left\{\int_0^1 K_n\left(\frac{x-W_j}{h_n}\right)t(x)\,dx\right\}^2 \leq \frac{1}{h_n^2}\sum_{j=1}^n\left\{\int_0^1 K_n\left(\frac{x-W_j}{h_n}\right)dx\right\}^2$$

$$\leq \sum_{j=1}^n\left(\int |K_n(u)|du\right)^2 \precsim n\max\{1,\exp(\delta_n^2/h_n^2)\}.$$

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

$$E_{Y|X}\left(\sup_{q\in\mathcal{Q}}\|q\|_\infty\right) = E_{Y|X}(\|nT_{1,n}\|_1) \leq \{E_{Y|X}(\|nT_{1,n}\|_1^2)\}^{1/2}$$

$$\leq \|[E_{Y|X}\{(nT_{1,n})^2\}]^{1/2}\|_1 = \int\left\{\frac{1}{h_n^2}\sum_{j=1}^n K_n\left(\frac{x-W_j}{h_n}\right)^2\right\}^{1/2}dx.$$

the last equation in the second line holds since Y_j's are independent, by Jensen's inequality and change of variable 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 $a' > 0$, $\int K_n(u)^2 du \leq \int_{|u|>a'}(u^4/a'^4)K_n(u)^2\,du + \int_{|u|\leq a'}K_n(u)^2 du$. It has been shown in the proof of Proposition D.7 that $\int u^4 K_n(u)^2 du \precsim \exp(\delta_n^2/h_n^2)$, it is easy to see that $\int K_n(u)^2 du \precsim \max\{1,\exp(\delta_n^2/h_n^2)\}$. Thus we have $E_{Y|X}\sup_{q\in\mathcal{Q}}\|q\|_\infty \precsim n^{1/2}\max\{1,\exp(\delta_n^2/h_n^2)\}/\sqrt{h_n}$. Applying Borell's inequality in Lemma D.6 by choosing $x = n\epsilon_n$, $\delta_n \asymp h_n \asymp \epsilon_n^{1/\beta}$, $\epsilon_n = n^{-\beta/(2\beta+1)}(\log n)^t$ and $t = \max\{(2\vee q)\beta/(2\beta+1),1\}$, we have shown that $\mathrm{pr}(\|T_{1,n}\|_1 > \epsilon_n/2) < e^{-n\epsilon_n^2/8}$.

Now we estimate the probability $\mathrm{pr}(I_{2,n} > \epsilon_n)$. Recall that $I_{2,n} = E_{Y,W|X}(\widehat{f}_n\widehat{p}_n) - E_{Y,W,X}(\widehat{f}_n\widehat{p}_n)$, by simple calculation we can show that $E_{Y,W|X}(\widehat{f}_n\widehat{p}_n) = \sum_{j=1}^n K\{(x-X_j)/h_n\}f(X_j)/(nh_n)$. Thus 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 the class of bounded linear functionals

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

and we have $nI_{2,n} = \sup_{l\in\mathcal{L}}\|\sum_{j=1}^n l(X_j)\|_\infty$. To apply the Talagrand's inequality we need to bound the following quantities. First we bound $\sup_{l\in\mathcal{L}}\|l(u)\|_\infty \leq \int |K\{(x-X_j)/h_n\}/h_n|\,dx\,\|f\|_\infty$. Since we can bound $\int|K(u)|du$ above by some constant $K_3 > 0$, by change of variable and Assumption 2.3 we have $\sup_{l\in\mathcal{L}}\|l(u)\|_\infty \leq K_4$, for some constant $K_4 > 0$.

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

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

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

$$E_X \sup_{l \in \mathcal{L}} \left| \sum_{j=1}^{n} l(X_j) \right| \le \left\{ E_X \left(\sup_{l \in \mathcal{L}} \left| \sum_{j=1}^{n} l(X_j) \right| \right)^2 \right\}^{1/2}$$

$$\le \frac{1}{h_n} \left(2n \left[E_X \left\{ \int K\left(\frac{x-X}{h_n} \right) dx \right\}^2 + \left\{ \int E_X K\left(\frac{x-X}{h_n} \right) dx \right\}^2 \right] \right)^{1/2} \|f\|_\infty$$

$$\precsim (n/h_n)^{1/2}.$$

Choosing $h_n \asymp \epsilon_n^{1/\beta}$ and same ϵ_n in the Theorem 2.5, and applying Talagrand's inequality yields the result $\mathrm{pr}(I_{2,n} > \epsilon_n/2) \le e^{-n\epsilon_n^2/8}$.

Finally, it is easy to see $I_{3,n} \le \epsilon$ by Proposition D.8. Combining the results of $I_{1,n}, I_{2,n}$ and $I_{3,n}$, we prove the inequality (18). Inequality (17) also holds since it can be seen as a special case of inequality (18) when taking the regression function $f(x) \equiv c$ for some constant $c > 0$.

The proof of inequality (16) follows the same line of arguments. Let $P_{1,n} = \widehat{p}_n - E_{W|X}(\widehat{p}_n)$, $P_{2,n} = E_{W|X}(\widehat{p}_n) - E_{W,X}(\widehat{p}_n)$ and $P_{3,n} = E_{W,X}(\widehat{p}_n) - p$ respectively. First, we estimate $\mathrm{pr}(\|P_{1,n}\|_\infty > \epsilon_0/2)$. The difference is that we consider the empirical process directly in $\| \cdot \|_\infty$. Since 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,

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

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

$$\sup_{x \in T} \|m_x\|_\infty \precsim h_n^{-1} \exp(\delta_n^2/2h_n^2),$$

$$\sup_{x \in T} E_{W|X}[m_x(W)]^2 \precsim h_n^{-2} \exp(\delta_n^2/h_n^2),$$

$$E_{W|X} \sup_{x \in T} | \sum_{j=1}^{n} m_x(W_j)| \precsim n^{1/2} h_n^{-1} \exp(\delta_n^2/h_n^2).$$

Therefore choosing $\delta_n = o(h_n)$ and $h_n = o(\epsilon_n)$ with same ϵ_n in Theorem 2.5 for any $\epsilon_0 > 0$ taking $t = \epsilon_0 n h_n^2$ we have

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

By applying Lemma D.5 we have $\mathrm{pr}(\|\widehat{p}_n - E_{W|X}(\widehat{p}_n)\|_\infty > \epsilon_0) \le e^{-\epsilon_0 n h_n^2}$. Similarly, for $P_{2,n} = E_{W|X}(\widehat{p}_n) - E_{W,X}(\widehat{p}_n) = \sum_{j=1}^{n} \widetilde{g}_x(X_j)/(nh_n)$, where $\widetilde{g}_x(u) = K\{(x-u)/h_n\} - E_X[K\{(x-X)/h_n\}]$ for any $x \in T$. Construct the class $\mathcal{G} = \{g_x, x \in T\}$ 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 completes the proof.

D.3 Proof of Lemma D.3

The Kullback-Leibler neighborhood around f_0 has been studied extensively in literature. We give a brief argument mentioning the difference in our case, refer to Shen et al. (2013) for extended proof. Under the Assumption 2.2, 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 $H(p_0, \phi_\sigma * h_\sigma) \precsim \sigma^\beta$, for some constant $a_0 > 0$. Fix $\sigma^\beta = \widetilde{\epsilon}_n \{\log(1/\widetilde{\epsilon}_n)\}^{-1}$ and find $b' > \max(1, 1/(2\beta))$ such that $\widetilde{\epsilon}_n^{b'} \{\log(1/\widetilde{\epsilon}_n)\}^{5/4} \leq \widetilde{\epsilon}_n$. By Lemma 2 of Ghosal and van ver 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 $H(\phi_\sigma * h_\sigma, \phi_\sigma * F') \leq \widetilde{\epsilon}_n^{b'}\{\log(1/\widetilde{\epsilon}_n)\}^{1/4}$. We construct the partition $\{U_1, \ldots, U_M\}$ in the flavor of $c\sigma\widetilde{\epsilon}_n^{b'} \leq \alpha(U_j) \leq 1$ for $j = 1, \ldots, M$, and $M \precsim \widetilde{\epsilon}_n^{1/\beta}\{\log(1/\widetilde{\epsilon}_n)\}^{1+1/\beta}$. Further denote the set S_F of probability measure F with $\sum_{j=1}^M |F(U_j) - p_j| \leq 2\widetilde{\epsilon}_n^{2b'}$ and $\min_{1 \leq j \leq M} F(U_j) \geq \widetilde{\epsilon}_n^{4b'}/2$ for sufficiently large n. Then $\Pi(S_F) \succsim \exp[-\widetilde{\epsilon}_n^{-1/\beta}\{\log(1/\widetilde{\epsilon}_n)\}^{2+1/\beta}]$. For each $F \in S_F$,

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

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

D.4 Proof of Lemma D.4

To prove Lemma D.4, it suffice to upper bound the Kullback-Leibler divergence between g_{f_0,p_0} and $g_{f,p}$ as well as the second moment of Kullback-Leibler divergence. From Lemma D.1 and Lemma 5.3 in van der Vaart and van Zanten (2009), 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

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

27

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

$$\int g_{f_0,p_0} \left(\log \frac{g_{f_0,p_0}}{g_{f,p}} \right)^2 = \int_{A_n} g_{f_0,p_0} \left(\log \frac{g_{f_0,p_0}}{g_{f,p}} \right)^2 + \int_{A_n^c} g_{f_0,p_0} \left(\log \frac{g_{f_0,p_0}}{g_{f,p}} \right)^2 \tag{22}$$

$$=: I_1 + I_2.$$

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

For I_1 in (22), we apply the inequality

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

Since $H^2(g_{f_0,p_0}, g_{f,p}) \le KL(g_{f_0,p_0}, g_{f,p})$, we need to estimate the term $\|(g_{f_0,p_0}/g_{f,p})\mathbb{1}_{A_n}\|_\infty$. By definition,

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

$$\le \left\| \frac{e^{-\frac{1}{2}(y-f_0)^2}}{e^{-\frac{1}{2}(y-f)^2}} \mathbb{1}_{A_n} \right\|_\infty \left\| \frac{p_0}{p} \right\|_\infty.$$

Based on the Assumption 2.1 f_0 is β-smooth function supported on $[0,1]$ and hence there exists some constant $B_0' > 0$ such that $\|f_0\|_\infty \le B_0'$. For $y \in A_n$,

$$\frac{e^{-\frac{1}{2}\{y-f(x)\}^2}}{e^{-\frac{1}{2}\{y-f_0(x)\}^2}} \mathbb{1}_{A_n} = e^{\{f(x)-f_0(x)\}\{y-f_0(x)\}-\{f(x)-f_0(x)\}^2/2} \mathbb{1}_{A_n}$$

$$\ge e^{-\|f-f_0\|_\infty(|y|+\|f_0\|_\infty)-\{f(x)-f_0(x)\}^2/2} \mathbb{1}_{A_n}$$

$$\ge e^{-\epsilon_n(\gamma'/\epsilon_n+B_0')-\epsilon_n^2/2} \ge e^{-2\gamma'}.$$

Thus $\|e^{-(y-f_0)^2/2}/e^{-(y-f)^2/2}\mathbb{1}_{A_n}\|_\infty \le e^{2\gamma'}$ and based on the results from Lemma D.1 for any $x \in [0,1]$ and $p \in \mathcal{P}_n$, we have $\log \|p_0/p\|_\infty \precsim \log(1/\epsilon_n)$. Therefore $\int_{A_n} g_{f_0,p_0}\{\log(g_{f_0,p_0}/g_{f,p})\}^2 \le 2\epsilon_n^2 \log^2(1/\epsilon_n)$.

Next we estimate I_2 in (22). For all $y \in A_n^c$ and for any fixed $x \in [0,1]$ we choose $\gamma' > 1$ such that $|y - f_0(x)| \ge |y| - \|f_0\|_\infty > \gamma'/\epsilon_n - B_0' \ge 1/\epsilon_n$. From Fubini's theorem,

$$\int_{|y|>1/\epsilon_n} g_{f_0,p_0} \left(\log \frac{g_{f_0,p_0}}{g_{f,p}} \right)^2$$

$$\le \frac{1}{2\pi\delta_n} \int_0^1 \int_{|y-f_0(x)|>1/\epsilon_n} e^{-\frac{1}{2}\{y-f_0(x)\}^2} e^{-\frac{1}{2\delta_n^2}(w-x)^2} \left(\log \frac{\int e^{-\frac{1}{2}(y-f_0)^2} e^{-\frac{1}{2\delta_n^2}(w-x)^2} p_0(x)\, dx}{\int e^{-\frac{1}{2}\{y-f(x)\}^2} e^{-\frac{1}{2\delta_n^2}(w-x)^2} p(x)\, dx} \right)^2 dy\, dw\, p_0(x)\, dx$$

$$\le (2\pi)^{-1/2} \int_0^1 \int_{|y-f_0(x)|>1/\epsilon_n} e^{-\frac{1}{2}\{y-f_0(x)\}^2} \left(\log \left\| \frac{e^{-(y-f_0)^2/2}}{e^{-(y-f)^2/2}} \right\|_\infty + \log \left\| \frac{p_0}{p} \right\|_\infty \right)^2 dy\, p_0(x)\, dx.$$

28

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

$$\int_{A_n^c} g_{f_0,p_0} \left(\log \frac{g_{f_0,p_0}}{g_{f,p}} \right)^2 \le 4(2\pi)^{-1/2} \int_0^1 \left(\int_{|z| \ge 1/\epsilon_n} e^{-\frac{1}{2}z^2} (\epsilon_n z)^2 \, dz + \int_{|z| \ge 1/\epsilon_n} e^{-\frac{1}{2}z^2} \log^2(1/\epsilon_n) dz \right) p_0(x) \, dx$$

$$\le 4(2\pi)^{-1/2} P_0 \left\{ \epsilon_n^2 \int_{t > 1/\epsilon_n^2} e^{-t/2} t^{1/2} \, dt + \log^2(1/\epsilon_n) \mathrm{pr}(|Z| \ge 1/\epsilon_n) \right\}$$

$$\le 4(2\pi)^{-1/2} P_0 \left\{ \epsilon_n^2 \int_{t > 1/\epsilon_n^2} e^{-t/4} \, dt + \log^2(1/\epsilon_n) e^{-\epsilon_n^{-2}/8} \right\}$$

$$\precsim e^{-\epsilon_n^{-2}/8 + \log\log(1/\epsilon_n)} < \epsilon_n^2.$$

Combining results of I_1 and I_2, we can show $\int g_{f_0,p_0} (\log g_{f_0,p_0}/g_{f,p})^2 \precsim \epsilon_n^2$. And further we have

$$\left\{ \int g_{f_0,p_0} \log \frac{g_{f_0,p_0}}{g_{f,p}} \precsim \epsilon_n^2, \int g_{f_0,p_0} \left(\log \frac{g_{f_0,p_0}}{g_{f,p}} \right)^2 \precsim \epsilon_n^2 \right\} \supset \{ \|f - f_0\|_\infty \le \epsilon_n, \ KL(p_0, p) \le \epsilon_n^2 \},$$

which yields the conclusion.

E   Proof of Theorem 3.1

We first compute the expectation and covariance of $\widetilde{f}_N$. For any $x \in \mathbb{R}$ the expectation is

$$E\{\widetilde{f}_N(x)\} = (2/N)^{-1/2} \sum_{j=1}^N \int \int \frac{1}{2\pi} \cos(w_j x + s_j) \phi_c(w_j) \, dw_j \, ds_j$$

$$= (2/N)^{-1/2} \sum_{j=1}^N \int \frac{1}{2\pi} \left\{ \cos(w_j x) \int_{-\pi}^{-\pi} \cos s_j ds_j - \sin(w_j x) \int_{-\pi}^{-\pi} \sin s_j ds_j \right\} \phi_c(w_j) \, dw_j = 0.$$

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

$$\mathrm{cov}\{\widetilde{f}_N(x), \widetilde{f}_N(y)\} = (2/N) \sum_{j=1}^N \mathrm{cov}\{\cos(w_j x + s_j), \cos(w_j x + s_j)\} = 2E_{w,s} \cos(xw + s)^2$$

$$= \frac{1}{2\pi} \int_w \int_{-\pi}^\pi [\cos\{(x+y)w + 2s\} + \cos\{(x-y)w\}] \phi_c(w) ds dw$$

$$= \frac{1}{2\pi} \int_w \int_{-\pi}^\pi [\cos\{(x+y)w\} \sin(2s) + \sin\{(x+y)w\} \cos(2s)] \, ds \, \cos\{(x-y)w\} \phi_c(w) \, dw$$

$$= \frac{1}{2\pi} \int_w \cos\{(x-y)w\} \phi_c(w) \, dw = c(x, y).$$

Now we 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 $(x_1, \ldots, x_k)'$ of $[0, 1]$

with integer $k > 0$, applying multivariate central limit theorem with the above moment results, we obtain as $N \to \infty$,

$$\{\widetilde{f}_N(x_1), \ldots, \widetilde{f}_N(x_k)\} \to N(0, c_{k,k}),$$

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

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

$$\mathrm{pr}\{\widetilde{f}(x_0) \in K\} > 1 - \epsilon. \tag{23}$$

For any $x_0 \in [0, 1]$, we upper bound $\widetilde{f}(x_0)$ as

$$|\widetilde{f}_N(x_0)| \le (2/N)^{1/2} \sum_{i=1}^{N} |a_j|,$$

for $a_j \sim \mathrm{N}(0, 1)$, $j = 1, \ldots, N$. With the well-known result that $|a_j|$ is a sub-gaussian random variable. For any $t > 0$, we have

$$\mathrm{pr}\{|\widetilde{f}_N(x_0)| \ge t\} \le \mathrm{pr}\left\{ (2/N)^{1/2} \sum_{i=1}^{N} |a_j| \ge t\right\} \le 2\exp(-ct^2)$$

for some constant $c > 0$. For any $\epsilon > 0$, we choose $t = \{2\log(1/\epsilon)\}^{1/2}$ and $K = \{|\widetilde{f}(x_0)| \le t\}$, then (23) holds, thus we show the tightness of $\widetilde{f}_N(x_0)$ for any $x_0 \in [0, 1]$.

Third, we show $\widetilde{f}_N$ is asymptotically uniformly d–eqicontinuous, where d is the Euclidean norm and $d(x, y) = |x - y|$ for $x, y \in \mathbb{R}$. The definition is, for any $\epsilon, \eta > 0$, there exists $\gamma > 0$ such that,

$$\mathrm{pr}\left\{ \sup_{d(x,y)<\gamma} |\widetilde{f}_N(x) - \widetilde{f}_N(y)| > \epsilon \right\} < \eta. \tag{24}$$

Recall $a_j \sim \mathrm{N}(0, 1)$ and $w_j \sim \mathrm{N}(0, 2/\lambda)$, a_j and w_j are independent. In the following we first show $a_j w_j$ is sub-exponential random variable with parameters $(16/\lambda, (4/\lambda)^{1/2})$. By the definition of the sub-exponential random variables and the properties of sub-Gaussian random variable, we show that $(a_j w_j)$ is a sub-exponential random variable. For any $t > 0$,

$$E\{\exp(ta_j w_j)\} = E[E\{\exp(ta_j w_j)\} \mid w_j] = 1 + \sum_{k=1}^{\infty} \frac{(t/2)^2 E(w_j^{2k})}{k!}$$

$$= 1 + 2\sum_{k=1}^{\infty} (2t^2/\lambda)^k = 1 + 2\frac{2t^2/\lambda}{1 - 2t^2/\lambda}$$

$$\le 1 + 8t^2/\lambda \le \exp\left\{ \frac{t^2}{2}(16t^2/\lambda)\right\}.$$

The last two inequalities hold based on the inequality $x/(1-x) \le 2x$ for $x \le 1/2$ and $1+x \le \exp(x)$

separately. Since both a_j and w_j are symmetric about 0, we have

$$\mathrm{pr}(|a_j w_j| \geq t) = 2\mathrm{pr}(a_j w_j \geq t),$$

Thus $|a_j w_j|$ is also a sub-exponential random variable. Note that,

$$\sup_{|x-y|<\gamma} |\widetilde{f}_N(x) - \widetilde{f}_N(y)| \leq (2/N)^{1/2} \sum_{j=1}^{N} \gamma \, |a_j w_j|.$$

Also, $E(|a_j w_j|) = (2/\lambda)^{1/2}$. Applying Bernstein inequality, we have for any $t > 0$,

$$\mathrm{pr}\left\{ \sup_{|x-y|<\gamma} |\widetilde{f}_N(x) - \widetilde{f}_N(y)| \geq (2N)^{1/2} \gamma t \right\} \leq \mathrm{pr}\left[(2/N)^{1/2} \sum_{i=1}^{N} \gamma \left\{ |a_j w_j| - E(|a_j w_j|) \right\} \geq (2N)^{1/2} \gamma t \right]$$

$$\leq \exp\left(-\frac{Nt^2}{2/\lambda + (4/\lambda)^{1/2} t/3} \right).$$

For any $\epsilon, \eta > 0$. Choose t such that

$$\exp\left(-\frac{Nt^2}{2/\lambda + (4/\lambda)^{1/2} t/3} \right) = \eta,$$

and choose γ such that $(2N)^{1/2} \gamma t = \epsilon$. With such γ, t, $\widetilde{f}_N$ satisfies (24). Hence the proof of weak convergence of $\widetilde{f}_N$ to the original Gaussian procss is completed.

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 $[\theta \mid D_n]$. We start with the Gaussian process associated with an exponential squared kernel as an illustration (in practice the algorithm can be applied to other kernels as long as they are symmetric). The exponential squared kernel is $C(x, x') = \exp\{-(x - x')^2/\lambda\}$ with bandwidth parameter λ. Theorem 3.1 enforces the prior distributions $w_j \sim \mathrm{N}(0, 2/\lambda)$, $s_j \sim \mathrm{Unif}\,(0, 2\pi)$ and $a_j \sim \mathrm{N}(0, 1)$ identically and independently for $j = 1, \ldots, N$. To ensure conditional conjugacy, we place a gamma distribution $\mathrm{Ga}(a_0, b_0)$ on bandwidth λ with shape parameter a_0 and scale parameter b_0. We place a Dirichlet process mixture of normals prior defined in (8), given more precisely by

$$X_i \sim \sum_{h=1}^{\infty} \pi_h \mathrm{N}(\mu_h, \tau_h^{-1}), \quad (\mu_h, \tau_h) \sim \mathrm{N}(\mu_h; \mu_0, \kappa_0 \tau_h^{-1}) \mathrm{Ga}(\tau_h; a_\tau, b_\tau), \tag{25}$$

on the density of X. The prior on π_h is expressed as $\pi_h = \nu_h \prod_{l<h}(1 - \nu_l)$ where $\nu_l \sim \mathrm{Beta}(1, \alpha)$. Here $\alpha = 1$. Denote the cluster label $S_i \in \{1, \ldots, K\}$ for X_i indicating that each X_i is associated with S_ith component in the Dirichlet process Gaussian mixture prior. Then (26) can be also written

as

$$X_i \mid S_i, \mu, \tau \sim \mathrm{N}(\mu_{S_i}, \tau_{S_i}^{-1}), \quad (\mu_{S_i}, \tau_{S_i}) \sim \mathrm{N}(\mu_{S_i}; \mu_0, \kappa_0 \tau_{S_i}^{-1}) \mathrm{Ga}(\tau_{S_i}; a_\tau, b_\tau). \tag{26}$$

In the simulation studies and the real data example, we fix $\mu_0 = 0, \kappa_0 = 1, a_\tau = 1, b_\tau = 1$. We put a gamma prior on λ with hyperparameters $a_0 = 5, b_0 = 1$, chosen based on our numerical experiments. In addition, we assume $\sigma = 0.2$ in the simulation studies. To describe the full conditional distributions, we use symbols a, w, s, X to denote the corresponding vectors. Then the joint full posterior distribution for $\{a, w, s, \lambda, X\}$ can be factored as

$$[a, w, s, \lambda, X \mid Y, W] \propto [Y \mid X, a, w, s, \lambda] \times [W \mid X] \times [w \mid \lambda] \times [\lambda] \times [a] \times [s] \times [X].$$

The full conditional distributions are as follows:

1. Update $[w \mid -]$ in a block by sampling $[w_j \mid -] \propto [Y \mid X, a, w, s, \lambda] \, \mathrm{N}(w_j; 0, 2/\lambda)$ independently using Metropolis-Hastings algorithm.

2. Update $[s \mid -]$ in a block by sampling $[s_j \mid -] \propto [Y \mid X, a, w, s, \lambda] \, \mathrm{Unif}\,[0, 2\pi]$ independently using Metropolis-Hastings algorithm.

3. Update $[a \mid -]$ from a multivariate normal distribution $\mathrm{N}(\widetilde{\mu}, \widetilde{\Sigma})$, with mean vector $\widetilde{\mu} = \widetilde{\Sigma} \, \Phi^{\mathrm{T}} Y / \sigma^2$, and $\widetilde{\Sigma} = (\Phi^{\mathrm{T}} \Phi / \sigma^2 + I_N)^{-1}$, where Φ is a $n \times N$ matrix with (i, j)th element $\Phi_{i,j} = (2/N)^{1/2} \cos(w_j x_i + s_j)$, and I_N is $N \times N$ identity matrix.

4. Update the parameters $[S, \mu, \tau, \pi \mid -]$ of the density in Dirichlet process Gaussian mixture prior as in Ishwaran and James (2001) with the number of mixture components truncated at 20.

5. Update $[X \mid -]$ in a block by sampling

$$[X_i | S_i, X_{-i}, -] \propto \mathrm{N}(Y_i; \Phi_{i,1:N} a, \sigma^2) \mathrm{N}(W_i; X_i, \delta) \mathrm{N}(X_i; \mu_{S_i}, \tau_{S_i})$$

using Metropolis-Hastings algorithm.

6. Update $[\lambda \mid -] \propto \mathrm{Ga}(\widehat{a}, \widehat{b})$ with $\widehat{a} = a_0$ and $\widehat{b} = b_0/(1 + b_0 \sum_{j=1}^n w_j^2/4)$.

7. Update $[\sigma^2 \mid -] \propto \mathrm{IG}(a_{\sigma^2}, b_{\sigma^2})$, where $a_{\sigma^2} = n/2$ and $b_{\sigma^2} = (Y - \Phi \mathbb{1}_N)^{\mathrm{T}}(Y - \Phi \mathbb{1}_N)/2$, where $\mathbb{1}_N$ denotes the $n \times 1$ vector with all elements to be 1.

We use random walk proposal $w_j^{\mathrm{prop}} \sim \mathrm{N}(w_j^{\mathrm{cur}}, 1/4)$. The proposal variance is tuned to obtain average pointwise acceptance rate around 0.7; to sample from the full conditional of s_i we consider the independence proposal $s_i^{\mathrm{prop}} \sim \mathrm{Unif}\,(0, 2\pi)$. We noted that the averaged point-wise acceptance rate for s_i is around 0.6. Finally, to sample from the full conditional distribution of x_i, we use an adaptive proposal $x_i^{\mathrm{prop}} \sim \mathrm{N}(W_i/\delta^2 + \mu_{S_i} \tau_{S_i}, 1/(1/\delta^2 + \tau_{S_i}))$ with average acceptance rate around 0.8.

G Additional numerical results

Table 2: *Averaged Mean Squared Errors (*AMSE*)* $\mathbb{E}\, K^{-1} \sum_{k=1}^{K} \{\, \widehat{f}_j(t_k) - f_j(t_k)\,\}^2$ *($\widehat{f}_j(\cdot)$ denotes the proposed estimator of $f_j, j = 1, 2$) over a regular grid $(t_1, \ldots, t_K)$ of size $K = 100$ in the interval $[-3, 3]$ and standard errors ($\times 10^2$) over 50 replicated data sets of size $n = 100$*

Function	Method	δ^2					
		0·01	0·2	0·4	0·6	0·8	1
	GPEV$_a$	0·55 (0·24)	4·56 (2·74)	7·59 (4·13)	9·53 (6·64)	11·89 (8·06)	13·48 (11·16)
	GPEV$_f$	0·57 (0·26)	4·42 (2·4)	8·37 (3·97)	11·81 (4·81)	13·44 (7·58)	15·89 (8·04)
f_1	GPEV$_n$	0·55 (0·24)	5·58 (2·28)	12·62 (5·47)	18·08 (7·43)	20·73 (8·85)	22·46(7·71)
	GP	3·31(0·36)	5·84 (1·15)	8·73 (1·83)	11·42 (1·83)	13·97 (2·94)	15·7 (3·24)
	decon	1·1(0·61)	4·56 (1·51)	9·13 (2·72)	13·65 (3·49)	17·51 (3·28)	19·95 (3·2)
Function	Method	δ^2					
		0·01	0·2	0·4	0·6	0·8	1
	GPEV$_a$	0·55 (0·31)	3·88 (2·18)	6·43 (4·53)	8·23 (5·05)	9·84 (4·78)	12·99 (9·27)
	GPEV$_f$	0·58 (0·34)	4·4 (2·41)	6·64 (3·49)	9·38 (0·05)	15·19 (14·65)	17·19 (9·73)
f_2	GPEV$_n$	0·54 (0·3)	5·83 (2·62)	13·45 (4·05)	20·04 (5·28)	24·05 (6·51)	29·84 (9·32)
	GP	0·57 (0·29)	3·15 (1·09)	5·82 (1·95)	8·4 (3·00)	10·93 (3·87)	13·67 (4·68)
	decon	4·01 (4·74)	5·73 (3·18)	8·9 (2·86)	13·68 (4·69)	18·19 (4·83)	22·28 (5·06)

Table 3: *Averaged Mean Squared Errors (*AMSE$)$ $\mathbb{E} K^{-1} \sum_{k=1}^{K} \{ \widehat{f}_j(t_k) - f_j(t_k) \}^2$ $(\widehat{f}_j(\cdot)$ denotes the proposed estimator of $f_j, j = 1, 2)$ over a regular grid $(t_1, \ldots, t_K)$ of size $K = 100$ in the interval $[-3, 3]$ and standard errors $(\times 10^2)$ over 50 replicated data sets of size $n = 250$*

Function	Method	δ^2					
		0·01	0·2	0·4	0·6	0·8	1
	GPEV$_a$	0·25 (0·09)	4·08 (1·51)	5·75 (3·15)	6·16 (3·73)	6·03 (4·46)	8·90 (6·84)
	GPEV$_f$	0·26 (0·10)	4·45 (1·31)	7·37 (2·91)	9·37 (3·11)	11·33 (4·23)	13·94 (7·79)
f_1	GPEV$_n$	0·23 (0·09)	4·31 (1·37)	9·41 (3·06)	13·82 (3·79)	17·04 (4·48)	21·93 (6·86)
	GP	2·31 (0·16)	4·67 (0·52)	7·68 (0·98)	10·48 (1·27)	13·08 (1·76)	15·01 (2·13)
	decon	0·58 (0·36)	3·17 (0·81)	7·80 (1·43)	12·87 (1·76)	16·80 (1·92)	18·89 (1·89)

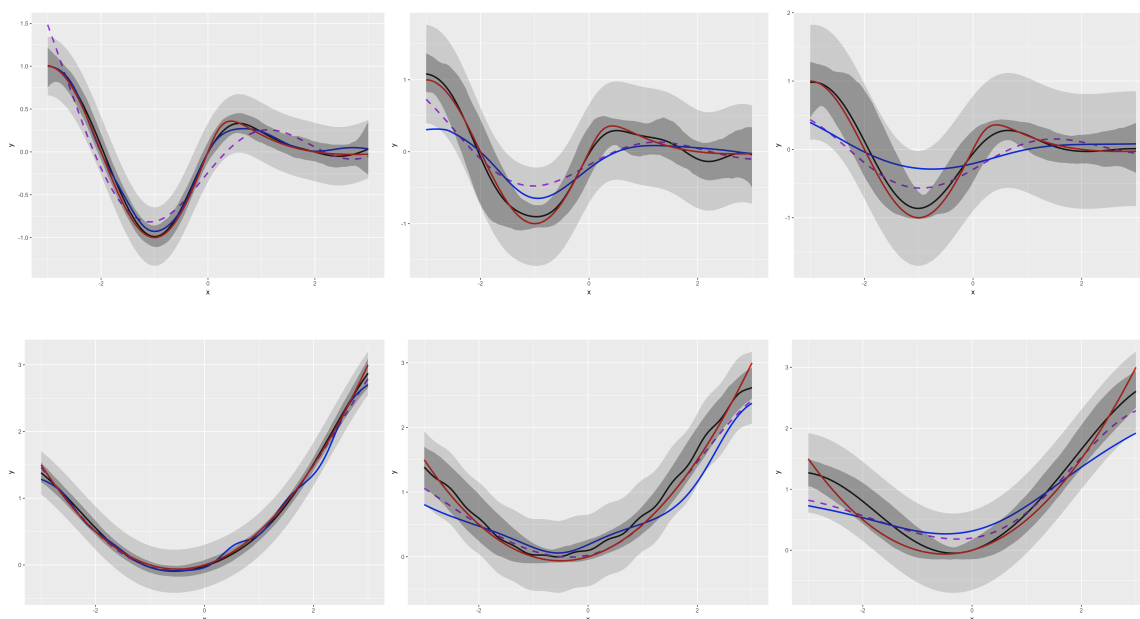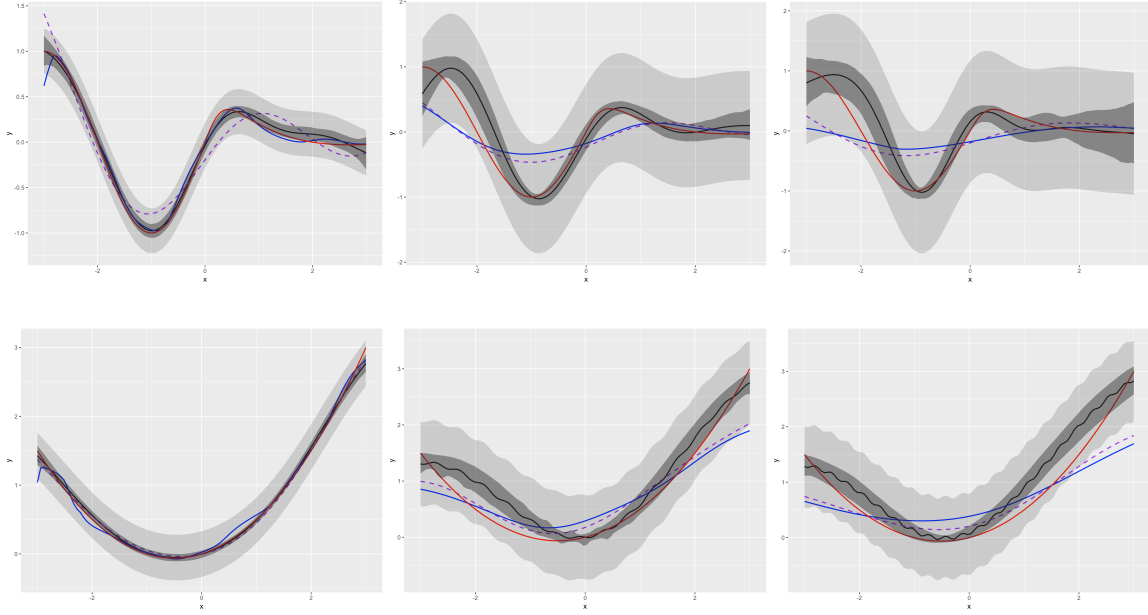| Function	Method	δ^2					
		0·01	0·2	0·4	0·6	0·8	1
	GPEV$_a$	0·25 (0·08)	4·18 (1·56)	6·70 (1·94)	7·87 (3·70)	10·46 (5·11)	11·15 (6·32)
	GPEV$_f$	0·26 (0·09)	4·88 (1·66)	7·39 (2·65)	9·50 (4·18)	13·55 (7·98)	18·67 (11·95)
f_2	GPEV$_n$	0·24 (0·08)	5·53 (1·6)	12·25 (3·44)	21·29 (5·05)	25·58 (4·96)	28·54 (6·49)
	GP	0·23 (0·09)	2·44 (0·57)	4·73 (1·08)	7·62 (1·66)	10·55 (2·04)	13·36 (2·34)
	decon	1·87 (2·19)	3·24 (0·89)	6·70 (1·71)	11·63 (2·21)	16·33 (2·56)	21·25 (2·67)

Figure 5: Estimation of $f_1(x)$ and $f_2(x)$ with $\delta^2 = 0.01$ (left panel), $\delta^2 = 0.6$ (middle panel) and $\delta^2 = 1$ (right panel). The first row shows estimation of $f_1(x)$ and the second row is for $f_2(x)$. Sample size $n = 100$. The red line is the true function, the black line is the estimated function using GPEV$_a$, the blue line is for decon, the purple dashed line is for GP. The darker and the lighter shades are the pointwise and simultaneous 95% credible intervals of GPEV$_a$.
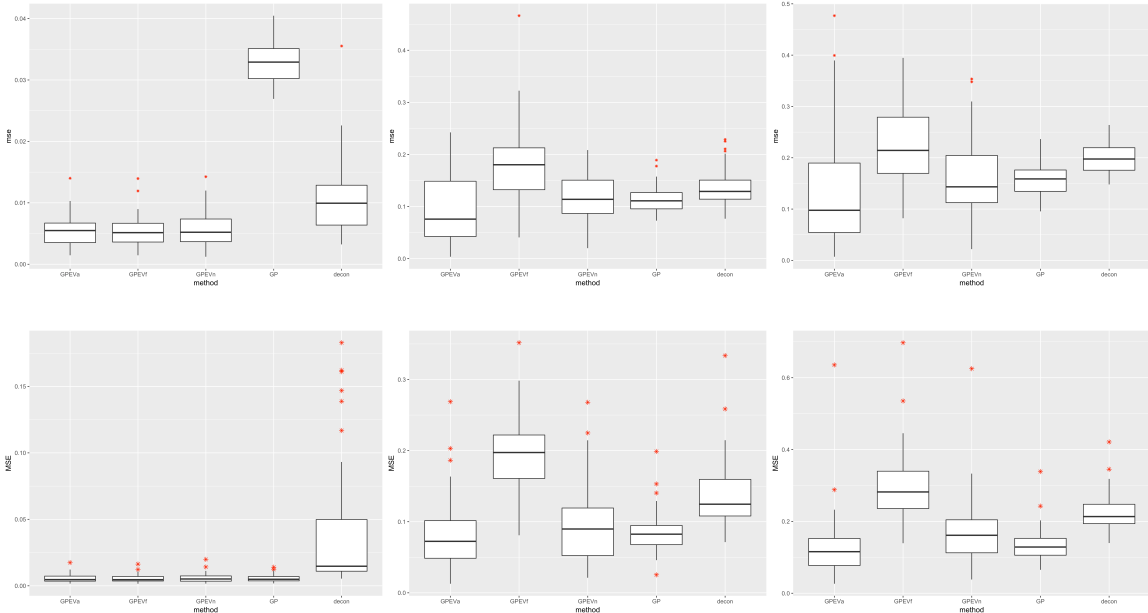
Figure 6: Estimation of $f_1(x)$ and $f_2(x)$ with $\delta^2 = 0.01$ (left panel), $\delta^2 = 0.6$ (middle panel) and $\delta^2 = 1$ (right panel). The first row shows predictions for $f_1(x)$ and the second row for $f_2(x)$. Sample size $n = 250$. The red line is the true function, the black line is GPEV$_a$, the blue line is decon, the purple dashed line is GP. The darker and the lighter shades are the pointwise and simultaneous 95% credible intervals of GPEV$_a$.
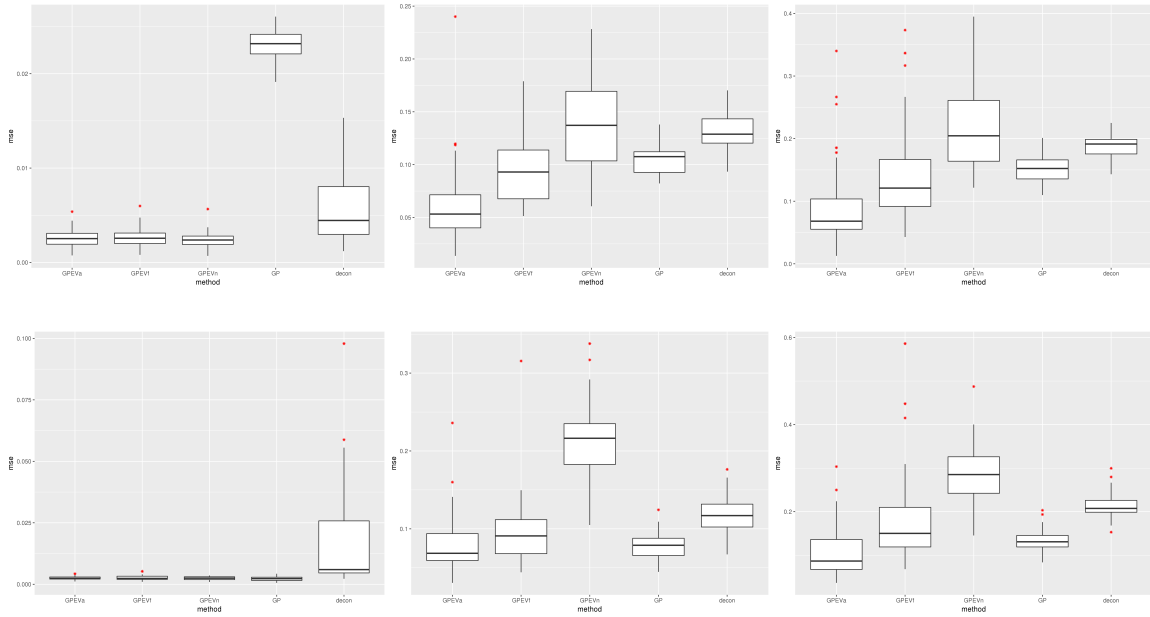

Figure 7: Boxplots for $f_1(x)$ and $f_2(x)$ over five methods mentioned in Section 4 on 50 replicated data sets. First row shows the results for $f_1(x)$ and the second row for $f_2(x)$. $\delta^2 = 0.01$ (left panel), $\delta^2 = 0.6$ (middle panel) and $\delta^2 = 1$ (right panel). Sample size $n = 100$. In each panel the displayed methods from left to right are GPEV$_a$, GPEV$_f$, GPEV$_n$, GP and decon.
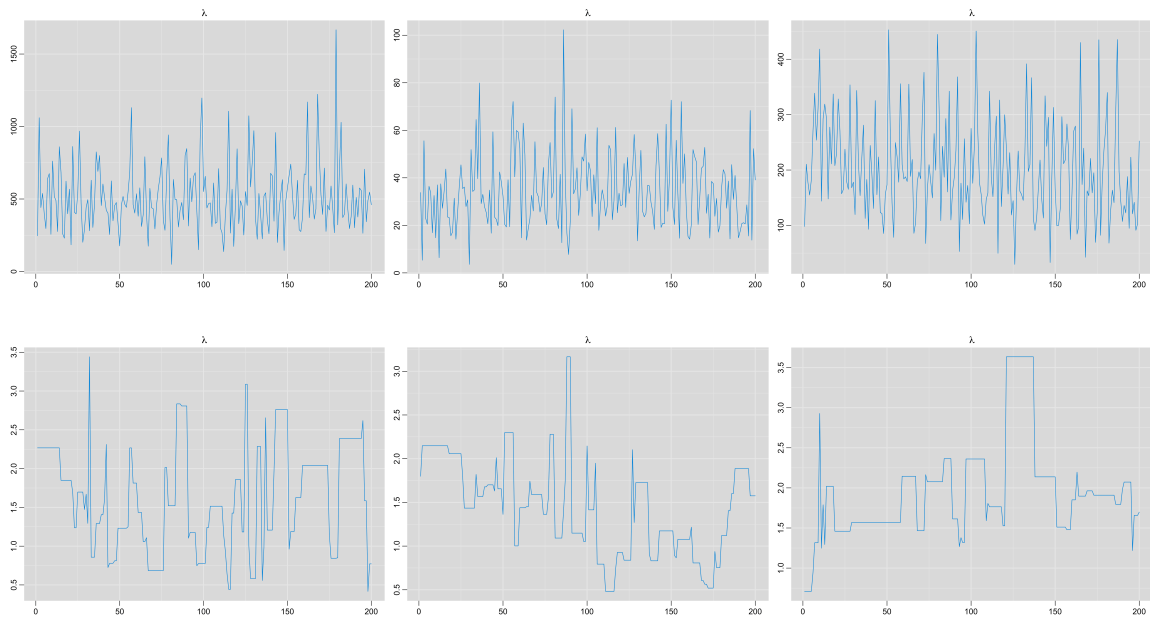
35

Figure 8: Boxplots for $f_1(x)$ and $f_2(x)$ over five methods mentioned in Section 4 on 50 replicated data sets. First row shows the results for $f_1(x)$ and the second row for $f_2(x)$. $\delta^2 = 0.01$ (left panel), $\delta^2 = 0.6$ (middle panel) and $\delta^2 = 1$ (right panel). Sample size $n = 250$. In each panel the displayed methods from left to right are GPEV_a, GPEV_f, GPEV_n, GP and decon.

Figure 9: Trace plots of last 200 posterior samples of λ of GPEV_a (first row) and GPEV_f (second row) modeling f_1 when 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).
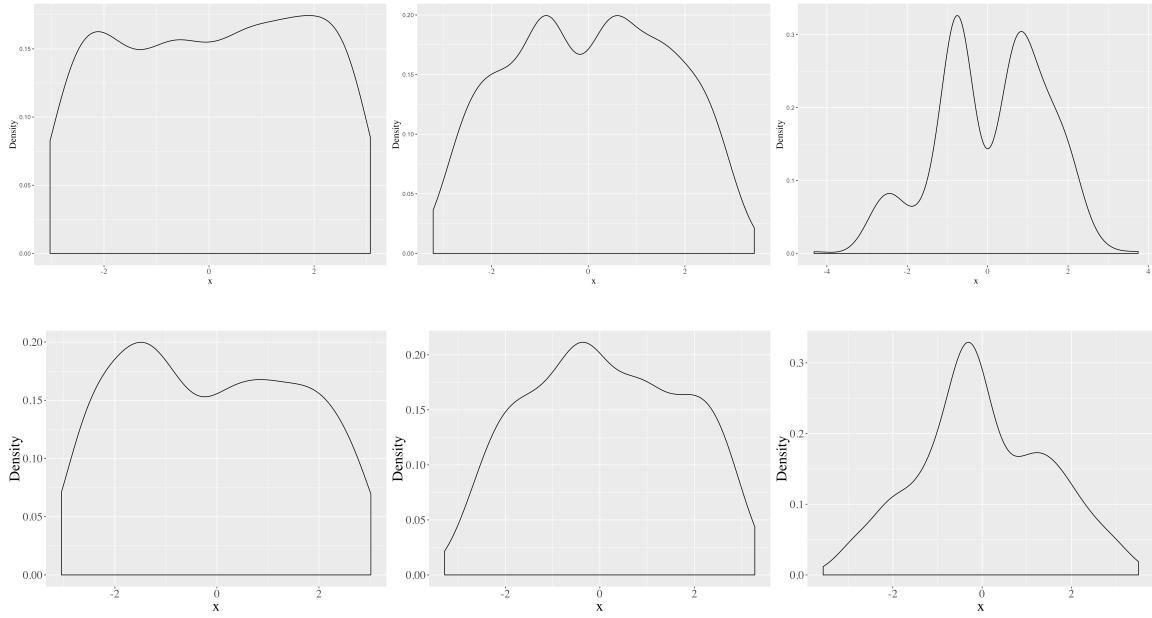
36

Figure 10: Density of posterior samples of covariates of GPEV$_a$ estimating $f_1(x)$ (first row) and $f_2(x)$ (second row) when $n = 500$. The value of δ^2 are 0.005 (left panel), 0.1 (middle panel), 0.5 (right panel).
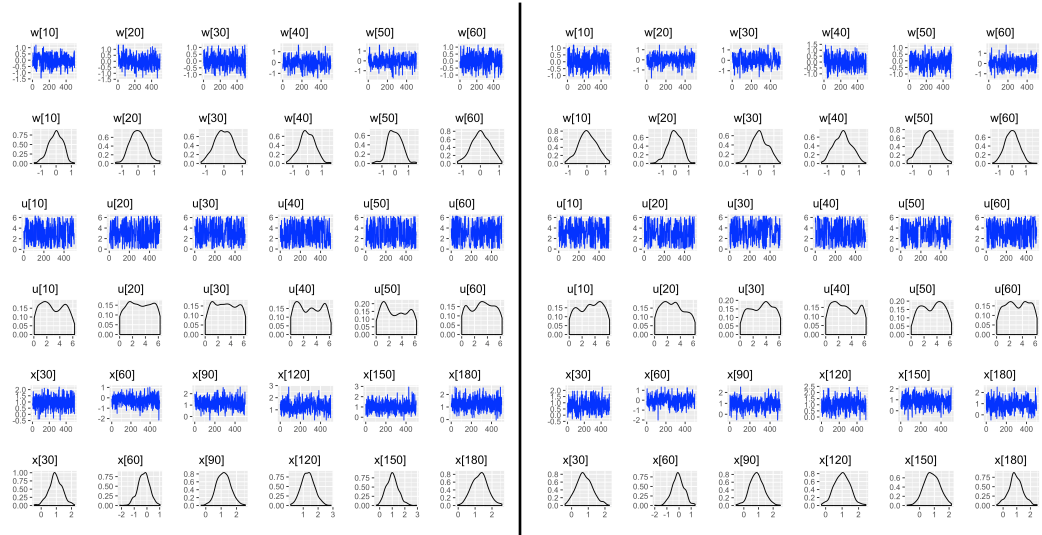

Figure 11: Trace plots and density plots of the 500 posterior samples of a subset of $\{w_j, s_j, x_j\}$ from treatment group with $\delta^2 = 0.35$ (left panel) and with unknown δ^2 (right panel) in the data example.

References

Adler, R. J. (1990). An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes. *Lecture Notes-Monograph Series*, 12.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, 70(4):825–848.

Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM: Probability and Statistics*, 6:127–146.

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97(457):160–169.

Birgé, L. (1979). *Sur un théoreme de minimax et son application aux tests*. Univ. de Paris-Sud, Dép. de Mathématique.

Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. In Giné, E., Houdré, C., and Nualart, D., editors, *Stochastic Inequalities and Applications*, pages 213–247, Basel. Birkhäuser Basel.

Brown, L. D., Cai, T. T., Low, M. G., Zhang, C.-H., et al. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *Annals of statistics*, 30(3):688–707.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186.

Carroll, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, 91(433):242–250.

Carroll, R. J., Maca, J. D., and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika*, 86(3):541–554.

Cervone, D. and Pillai, N. S. (2015). Gaussian process regression with location errors. *arXiv preprint arXiv:1506.08256*.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328.

Delaigle, A. (2014). 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.

Delaigle, A., Fan, J., and Carroll, R. J. (2009). A design-adaptive local polynomial estimator for the errors-in-variables problem. *Journal of the American Statistical Association*, 104(485):348–359.

Delaigle, A. and Hall, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *Journal of the American Statistical Association*, 103(481):280–287.

Delaigle, A., Hall, P., and Qiu, P. (2006). Nonparametric methods for solving the Berkson errors-in-variables problem. *Journal of the Royal Statistical Society, Series B*, 68(2):201–220.

Delaigle, A. and Meister, A. (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *Journal of the American Statistical Association*, 102(480):1416–1426.

Donnet, S., Rivoirard, V., Rousseau, J., and Scricciolo, C. (2018). Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures. *Bernoulli*, 24(1):231–256.

Du, L., Zou, C., and Wang, Z. (2011). Nonparametric regression function estimation for errors-in-variables models with validation data. *Statistica Sinica*, 21(3):1093–1113.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19(3):1257–1272.

Fan, J. (1992). Deconvolution with supersmooth distributions. *Canadian Journal of Statistics*, 20(2):155–169.

Fan, J. and Truong, Y. K. (1993). Nonparametric regression with errors in variables. *Annals of Statistics*, 21(4):1900–1925.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis*, 53(8):2873–2884.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statisics*, 28(2):500–531.

Ghosal, S. and van ver Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Annals of Statistics*, 35(2):697–723.

Guan, Y. (2006). A composite likelihood approach in fitting spatial point process models. *Journal of the American Statistical Association*, 101(476):1502–1512.

Guinness, J. and Fuentes, M. (2017). Circulant embedding of approximate covariances for inference from Gaussian data on large lattices. *Journal of Computational and Graphical Statistics*, 26(1):88–97.

Hall, P. and Ma, Y. (2007). Semiparametric estimators of functional measurement error models with unknown error. *Journal of the Royal Statistical Society: Series B*, 69(3):429–446.

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, 93(443):1099–1111.

Ioannides, D. and Alevizos, P. (1997). Nonparametric regression with errors in variables and applications. *Statistics & Probability Letters*, 32(1):35–43.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Johannes, J. (2009). Deconvolution with unknown error distribution. *Annals of Statistics*, 37(5A):2301–2323.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555.

Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Annals of Statistics*, 39(5):2626–2657.

Kruijer, W., Rousseau, J., van der Vaart, A., et al. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *Annals of Statistics*, 12(1):351–357.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238.

Meister, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. Lecture Notes in Statistics **193**. Springer, Berlin.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Neumann, M. H. (2007). Deconvolution from panel data with unknown error distribution. *Journal of Multivariate Analysis*, 98(10):1955–1968.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian Process for Machine Learning.* MIT Press.

Ray, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electronic Journal of Statistics*, 7:2516–2549.

Sarkar, A., Mallick, B. K., Staudenmayer, J., Pati, D., and Carroll, R. J. (2014). Bayesian semi-parametric density deconvolution in the presence of conditionally heteroscedastic measurement errors. *Journal of Computational and Graphical Statistics*, 23(4):1101–1125.

Sarkar, A., Pati, D., Mallick, B. K., and Carroll, R. J. (2013). Adaptive posterior convergence rates in Bayesian density deconvolution with supersmooth errors. *arXiv preprint arXiv:1308.5427*.

Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640.

Staudenmayer, J., Ruppert, D., and Buonaccorsi, J. P. (2008). Density estimation in the presence of heteroscedastic measurement error. *Journal of the American Statistical Association*, 103(482):726–736.

Stefanski, L. A. and Carroll, R. J. (1990). Deconvolving kernel density estimators. *Statistics*, 21(2):169–184.

Stefanski, L. A. and Cook, J. R. (1995). Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 90(432):1247–1256.

Stroud, J. R., Stein, M. L., and Lysen, S. (2017). Bayesian and maximum likelihood estimation for Gaussian processes on an incomplete lattice. *Journal of Computational and Graphical Statistics*, 26(1):108–120.

van der Vaart, A., van Zanten, H., et al. (2007). Bayesian inference with rescaled Gaussian process priors. *Electronic Journal of Statistics*, 1:433–448.

van der Vaart, A. W. and van Zanten, J. H. (2008). 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.

van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Annals of Statistics*, 37(5B):2655–2675.

van der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and Empirical Processes*, pages 16–28. Springer.

Wood, A. T. and Chan, G. (1994). Simulation of stationary Gaussian processes in $[0, 1]^d$. *Journal of Computational and Graphical Statistics*, 3(4):409–432.