Semiparametric Bayesian inference for local extrema of functions in the presence of noise

Meng Li, Zejian Liu, Cheng-Han Yu and Marina Vannucci

Abstract

There is a wide range of applications where the local extrema of a function are the key quantity of interest. However, there is surprisingly little work on methods to infer local extrema with uncertainty quantification in the presence of noise. By viewing the function as an infinite-dimensional nuisance parameter, a semiparametric formulation of this problem poses daunting challenges, both methodologically and theoretically, as (i) the number of local extrema may be unknown, and (ii) the induced shape constraints associated with local extrema are highly irregular. In this article, we build upon a derivative-constrained Gaussian process prior recently proposed by Yu et al. (2023) to derive what we call an encompassing approach that indexes possibly multiple local extrema by a single parameter. We provide closed-form characterization of the posterior distribution and study its large sample behavior under this unconventional encompassing regime. We show that the posterior measure converges to a mixture of Gaussians with the number of components matching the underlying truth, leading to posterior exploration that accounts for multi-modality. Point and interval estimates of local extrema with frequentist properties are also provided. The encompassing approach leads to a remarkably simple, fast semiparametric approach for inference on local extrema. We illustrate the method through simulations and a real data application to event-related potential analysis.

Keywords: Local extrema, Gaussian process, semiparametric, shape-constrained regression, Bernstein-von Mises theorem

1 Introduction

Finding localized features of a smooth function, including local maxima and local minima, plays a pervasive role in statistics, with wide-ranging scientific applications such as in biology (Raghuraman et al., 2001), microscopy (Egner et al., 2007; Geisler et al., 2007), and psychology (Luck, 2005). Moreover, localized features provide additional characterizations of the shape of a function that are useful for visualization and interpretation, and lead to insights into optimization, particularly when operated on approximations of the function.

There is a rich literature on shape-constrained regression, where the overwhelming emphasis has been on incorporating restrictions, including monotonicity, convexity, modality, log-concavity, and piecewise constants, within nonparametric modeling of the underlying surface; see, for example, Ramsay (1998); Holmes and Heard (2003); Neelon and Dunson (2004); Meyer (2008); Shively et al. (2009, 2011); Abraham and Khadraoui (2015); Wheeler et al. (2017); Dasgupta et al. (2021). In this article, we contribute to this growing literature by focusing on a distinct perspective, namely, the inference on local extrema that form the key characterization of the shape constraint, while the underlying regression function is less of interest and can be viewed as a nuisance parameter.

There is surprisingly little work on the inference of local extrema with uncertainty quantification in the presence of noise. Notable exceptions include two-step approaches in the spirit of "smooth first, then estimation", where one first employs nonparametric smoothing techniques, then estimates the local extrema of the smoothed estimate. Along this line, Song et al. (2006) used kernel smoothing followed by hypothesis testing to find locations at which the regression function has zero derivatives at a given statistical significance level. Since the test is performed on all locations, a multiple testing issue emerges, even when there are a limited number of local extrema. Schwartzman et al. (2011) and Cheng and Schwartzman (2017) studied false discovery rate control and power consistency for local maxima under a unimodal true peak assumption. However, uncertainty quantification of the detected local maxima is not reported. Alternatively to the two-step approach, Davies et al. (2001) proposed to use the taut string method for piecewise monotone functions, and Kovac (2007)

extended the approach to smooth functions for finding point estimates of local extrema.

In this article, we consider a semiparametric Bayesian method for local extrema in situations where the number of local extrema may be unknown and the associated shape constraints are highly local. These pose daunting challenges to uncertainty quantification in the presence of noise, particularly when there is more than one local extremum point. Here, we build upon a derivative-constrained Gaussian process prior, recently proposed by Yu et al. (2023) for the location of *stationary* points in event-related potentials (ERP), to derive what we call an *encompassing* approach that indexes possibly multiple local extrema by a single parameter. We provide a rigorous theoretical investigation of this unconventional approach, which ensures a proper interpretation of the derived uncertainty quantification in the context of local extrema detection. The encompassing approach is remarkably simple, as it transforms a varying-dimensional model into one dimension, and thus is particularly well suited to address the multiplicity challenge posed by local extrema. We note that we use the Bayes machinery to derive a posterior distribution but employs frequentist properties to characterize its large sample behavior and justify the obtained point and interval estimates.

In our theoretical investigation, we characterize the posterior distribution of the encompassing approach and show an intrinsic connection to unconstrained nonparametric regression, enabling fast implementation without complicated sampling. We show that the posterior measure converges to a mixture of Gaussians with the number of components matching the underlying truth. This interesting phenomenon not only provides theoretical guarantees for the inference on local extrema that accounts for multi-modality of the posterior distribution, but also extends the Bernstein-von Mises (BvM) theorem beyond the traditional semiparametric Bayesian literature to the encompassing paradigm. Classic semiparametric Bayesian BvMs typically assume separable priors on the function and finite-dimensional parameter with fixed dimension, or rely on the parameter of interest being a bounded functional of the regression function (Castillo, 2012; Castillo and Rousseau, 2015); we instead study the limiting posterior distribution under irregular scenarios when the local extrema have unknown dimension and are embedded in the regression function, hence not separable, and the derivative at any fixed point, when viewed as a functional of the regression function, is not

bounded. This large sample characterization of the posterior distribution leads to consistent estimators of the number and location of local extrema. We additionally provide interval estimation for local extrema with frequentist coverage.

Organization. Section 2 introduces the model, shape-constrained priors, and a closed-form characterization of the posterior distribution of local extrema. In Section 3 we provide non-asymptotic bounds for a range of nonparametric quantities related to the posterior distribution, and establish a local asymptotic normality property and multi-modal limiting distribution under the encompassing regime. Consistent point estimators and interval estimators with frequentist coverage are also provided. In Section 4 we carry out simulation studies, and in Section 5 we illustrate the proposed method in event-related potential analysis. Section 6 concludes the paper. All proofs, additional technical results, and additional numerical experiments can be found in the Supplementary Materials.

2 Methods

2.1 Shape-constrained regression

Suppose we observe independent and identically distributed samples $X = \{X_1, \ldots, X_n\} \in \mathcal{X}^n$ and $\boldsymbol{y} = \{y_1, \ldots, y_n\} \in \mathbb{R}^n$ from a distribution $\mathbb{P}_0$ on $\mathcal{X} \times \mathbb{R}$ with n being the sample size and $\mathcal{X} \subset \mathbb{R}$ the sample space for the covariate that is compact. Throughout the paper we focus on one-dimensional sample space for concreteness and ease of notation, and consider $\mathcal{X} = [0, 1]$ without loss of generality. We briefly comment on extensions to the multi-dimensional space in the Discussion section.

We assume a regression model for the input data of the type

$$y_i = f(X_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

with $f : \mathcal{X} \to \mathbb{R}$ and $X_i \overset{\text{iid}}{\sim} \mathbb{P}_X$, where the measures $\mathbb{P}_X$ admit a density p_X with respect to the Lebesgue measure μ on $\mathcal{X}$, and with random noise $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

We make the following assumptions on the true regression function f_0.

Assumption A1. $f_0 \in C^2(\mathcal{X})$.

Assumption A2. f_0 has exactly M local extrema for some finite $M \geq 1$ at $0 < t_1 < \cdots < t_M < 1$.

Assumption A3. $f_0''(t_m) \neq 0$ for $m = 1, \ldots, M$.

Assumption A1 trivially implies that f_0 is bounded since it is continuous on a closed interval. Assumption A2 means that f_0 possesses exactly M local extrema, with $M \geq 1$ finite but unknown, and local extrema do not occur at the boundary of $\mathcal{X}$. Assumptions A1 and A2 lead to a necessary condition for t_m to be a local extremum: $f_0'(t_m) = 0$ for $m = 1, \ldots, M$. Assumption A3 regularizes the curvature of f_0 at each local extremum. Assumptions A2 and A3 indicate that we focus on local extrema that can be identified based on the second derivative test. Unlike some existing work such as Davies et al. (2001), we do not assume that all stationary points (zeros of f_0') are local extrema, and our assumptions do not regularize stationary points that are not local extrema.

Our goal is to make inference on $\{t_1, \ldots, t_M\}$ when M is unknown, with uncertainty quantification. As such, we next proceed to constrained priors on f accounting for local extrema.

2.2 Shape-constrained prior of f on local extrema

The underlying function f is unknown, and its local extrema are encoded in the function derivatives. We follow Yu et al. (2023) and adopt a constrained Gaussian process prior under derivative constraints.

A widely used prior for f is a Gaussian process (GP) with mean 0 and a covariance kernel that determines its key properties. Starting with a covariance kernel $k(\cdot, \cdot) = \sigma^2(n\lambda)^{-1}K(\cdot, \cdot)$, where $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a continuous, symmetric and positive definite bivariate function, we encode the derivative constraint by conditioning this GP prior on $f'(t) = 0$ for an unknown scalar parameter t. Assuming differentiability of $K(\cdot, \cdot)$, let $K_{jl}(x, x') = \partial^{j+l}K(x, x')/\partial x^j \partial x'^l$

for any $j, l \geq 0$. Then by direct calculation, the conditional GP is also a GP with mean 0 and covariance kernel $k_t(x, x') = \sigma^2(n\lambda)^{-1}\{K(x, x') - K_{01}(x, t)K_{11}^{-1}(t, t)K_{10}(t, x')\}$, provided that $K_{11}(t, t) > 0$. The tuning parameter λ possibly depends on the sample size n. Later we will make all assumptions on $K(\cdot, \cdot)$ clear.

Under the constrained prior $\mathrm{GP}(0, k_t)$, the sample path $f(\cdot)$ satisfies $\mathrm{E}(f'(t)) = 0$ and

$$\mathrm{Var}(f'(t)) = \left.\frac{\partial k_t^2(x, x')}{\partial x \partial x'}\right|_{(x,x')=(t,t)} = \sigma^2(n\lambda)^{-1}\{K_{11}(t, t) - K_{11}(t, t)K_{11}^{-1}(t, t)K_{11}(t, t)\} = 0.$$

Hence, it holds that $f'(t) = 0$ almost surely. Note that here we employ the differentiability of sample paths of Gaussian processes with continuously differentiable covariance kernels and the covariance function of f' that is induced by differentiating k_t (e.g., see Ghosal and van der Vaart, 2017, Proposition I.3).

We conclude the specification of all priors by placing a prior $\pi(t)$ on t, which is supported on $\mathcal{X}$. Thus, the marginal prior distribution of f is a mixture of constrained GPs if we integrate out t with respect to its prior distribution.

Like Yu et al. (2023), we use a univariate t to index all possible local extrema. This *encompassing* strategy eliminates the need to specify the number of local extrema, which is particularly useful when M is unknown and possibly greater than one. In addition, it enables unified inference on *all* local extrema through the Bayes machinery. Although practically appealing, this unconventional encompassing regime in a semiparametric setting is not well understood in the literature. Yu et al. (2023), in particular, employ Monte Carlo EM to conduct an empirical exploration of the posterior of t. A specific focus of this article is a rigorous characterization of the induced posterior distribution, both at finite sample size and asymptotically, which is critical to interpret and substantiate such a strategy, while providing insights into how to carry out posterior summary.

2.3 Closed-form posterior distribution of t

Integrating out f in Model (1) with respect to its prior $\mathrm{GP}(0, k_t)$ gives the marginal distribution $\boldsymbol{y}|X, t \sim N(0, \Sigma_t)$ with

$$\begin{aligned}\Sigma_t &= \sigma^2(n\lambda)^{-1}\left\{K(X,X) - K_{01}(X,t)K_{11}^{-1}(t,t)K_{10}(t,X)\right\} + \sigma^2\boldsymbol{I}_n \\ &= \{\sigma^2(n\lambda)^{-1}K(X,X) + \sigma^2\boldsymbol{I}_n\} - \sigma^2(n\lambda)^{-1}K_{01}(X,t)K_{11}^{-1}(t,t)K_{10}(t,X); \end{aligned} \tag{2}$$

here $K_{01}(X,t) = (K_{01}(X_1,t),\ldots,K_{01}(X_n,t))^T$ is a length n column vector, $K_{10}(t,X) = (K_{10}(t,X_1),\ldots,K_{10}(t,X_n)) = [K_{01}(X,t)]^T$ a length n row vector, $K(X,X) = (K(X_i,X_j))_{i,j=1}^n$ an n by n matrix, and $\boldsymbol{I}_n$ the n by n identity matrix. Thus, the marginal likelihood of t, denoted by $\ell(t) = p(\boldsymbol{y} \mid X, t)$, is the density function of $N(0, \Sigma_t)$ evaluated at $\boldsymbol{y}$.

An intriguing observation is that the first term in Equation (2) $\sigma^2(n\lambda)^{-1}K(X,X) + \sigma^2\boldsymbol{I}_n$ does not depend on t and coincides with the covariance matrix under the $\mathrm{GP}(0,k)$ prior. This enables a reformulation of $\ell(t)$ to relate the posterior distribution of t to unconstrained nonparametric regression. Before formally presenting this connection in Proposition 1, we first review standard GP priors without constraints to introduce notation.

Suppose that one uses the unconstrained GP prior $f \sim \mathrm{GP}(0, \sigma^2(n\lambda)^{-1}K)$ as the prior on f without shape constraints. In this article, we may omit explicit mention of the dependence on λ in most cases, like $\phi_{11,m}$ and $\phi_{11}(\cdot)$, except for a few instances such as f_λ and $t_{\lambda,m}$, which we will introduce later. By conjugacy, the posterior distribution of f is also a GP: $f|X, \boldsymbol{y} \sim \mathrm{GP}(\widehat{\mu}_f(\cdot), \widehat{\Sigma}_f(\cdot,\cdot))$, where $\widehat{\mu}_f(x) = K(x,X)[K(X,X) + n\lambda\boldsymbol{I}_n]^{-1}\boldsymbol{y}$ and $\widehat{\Sigma}_f(x,x') = \sigma^2(n\lambda)^{-1}\left\{K(x,x') - K(x,X)[K(X,X) + n\lambda\boldsymbol{I}_n]^{-1}K(X,x')\right\}$. Moreover, the derivative $f'|X, \boldsymbol{y}$ is also a GP with mean $\widehat{\mu}_{f'}(\cdot)$ and covariance $\widehat{\Sigma}_{f'}(\cdot,\cdot)$:

$$\widehat{\mu}_{f'}(x) = \frac{d\widehat{\mu}_f(x)}{dx} = K_{10}(x,X)[K(X,X) + n\lambda\boldsymbol{I}_n]^{-1}\boldsymbol{y}, \tag{3}$$

and $\widehat{\Sigma}_{f'}(x,x') = \frac{\partial^2\widehat{\Sigma}_f(x,x')}{\partial x\partial x'} = \sigma^2(n\lambda)^{-1}\left\{K_{11}(x,x') - K_{10}(x,X)[K(X,X) + n\lambda\boldsymbol{I}_n]^{-1}K_{01}(X,x')\right\}$.

In particular, the marginal posterior variance of the derivative process is

$$\widehat{\sigma}^2_{f'}(x) = \widehat{\Sigma}_{f'}(x,x) = \sigma^2(n\lambda)^{-1}\left\{K_{11}(x,x) - K_{10}(x,X)[K(X,X)+n\lambda \boldsymbol{I}_n]^{-1}K_{01}(X,x)\right\}. \quad (4)$$

We are now in a position to reformulate the posterior $\pi_n(t \mid X, \boldsymbol{y})$ as follows.

Proposition 1. *Suppose $K \in C^2(\mathcal{X},\mathcal{X})$ and $\widehat{\sigma}^2_{f'}(x) > 0$ for any $x \in \mathcal{X}$. Then it holds that*

$$\ell(t) = C\frac{1}{\sqrt{\widehat{\sigma}^2_{f'}(t)/K_{11}(t,t)}}\exp\left\{-\frac{\widehat{\mu}^2_{f'}(t)}{2\widehat{\sigma}^2_{f'}(t)}\right\}, \quad (5)$$

for some constant C that does not depend on t, where $\widehat{\mu}^2_{f'}(\cdot)$ and $\widehat{\sigma}^2_{f'}(\cdot)$ are defined in Equations (3) and (4), respectively. Consequently, the posterior distribution of t under the prior $\pi(t)$ satisfies

$$\pi_n(t \mid X, \boldsymbol{y}) \propto \frac{1}{\sqrt{\widehat{\sigma}^2_{f'}(t)/K_{11}(t,t)}}\exp\left(-\frac{\widehat{\mu}^2_{f'}(t)}{2\widehat{\sigma}^2_{f'}(t)}\right)\cdot\pi(t). \quad (6)$$

The normalizing constant in $\pi_n(t \mid X, \boldsymbol{y})$ can be calculated using routine one-dimensional numerical integration methods, such as the midpoint or trapezoidal rule. The closed-form type of formulation for the posterior $\pi_n(t \mid X, \boldsymbol{y})$ in Proposition 1 is useful on several fronts. Computationally, a close inspection of (6) suggests that evaluating $\pi_n(t \mid X, \boldsymbol{y})$ in various t only requires inverting an n by n matrix $K(X,X)+n\lambda\boldsymbol{I}_n$ once, dramatically reducing the computation in a naive implementation that directly inverts a varying covariance matrix induced by k_t at each t. Theoretically, Proposition 1 turns inference on t into key quantities related to posterior inference of f' with the unconstrained $\mathrm{GP}(0,k)$ prior, namely $\widehat{\mu}^2_{f'}(\cdot)$ and $\widehat{\sigma}^2_{f'}(\cdot)$. We next build on this connection to analyze large sample behavior of $\pi_n(t \mid X, \boldsymbol{y})$.

3 Theoretical results

In this section, we provide theoretical evidence of a multi-modal posterior distribution of t. Standard Bernstein-von Mises (BvM) theorems state that, under certain conditions, the

posterior distribution is close to a normal distribution. In our encompassing approach which indexes local extrema by a univariate t, a single normal approximation is unlikely to hold. Instead, we show that the posterior distribution of t converges to a mixture of Gaussians. This multi-modal limiting distribution along with a derived local asymptotic property delineate key differences between the adopted encompassing approach and existing semiparametric work, and provide support for posterior summary that accounts for such multi-modality.

3.1 Non-asymptotic analysis of key nonparametric quantities

In this section, we derive non-asymptotic error bounds for nonparametric recipes in Proposition 1: $\widehat{\mu}_{f'}(\cdot)$, $\widehat{\sigma}^2_{f'}(\cdot)$, and their high-order derivatives under the supremum norm. These error bounds are needed to study large sample behavior of $\pi_n(t \mid X, \boldsymbol{y})$, and might be of interest in their own right.

To this end, we take an operator-theoretic approach and make extensive use of differentiable kernels and associated properties. We begin with introducing notation and reviewing a few well-known properties; see Wahba (1990); Cucker and Zhou (2007) for details. For any $f \in L^2_{p_X}(\mathcal{X})$, define the following integral operator $L_K(f)(x) = \int_{\mathcal{X}} K(x, x')f(x')d\mathbb{P}_X(x')$, where $x \in \mathcal{X}$. The integral operator L_K is compact, positive definite, and self-adjoint. The spectral theorem ensures the existence of countable pairs of eigenvalues and eigenfunctions $(\mu_i, \psi_i)_{i\in\mathbb{N}} \subset (0, \infty) \times L^2_{p_X}(\mathcal{X})$ of L_K such that $L_K\psi_i = \mu_i\psi_i$, for $i \geq 1$, where $\{\psi_i\}_{i=1}^{\infty}$ form an orthonormal basis of $L^2_{p_X}(\mathcal{X})$ and $\mu_1 \geq \mu_2 \geq \cdots > 0$ with $\lim_{i\to\infty} \mu_i = 0$.

By Moore-Aronszajn Theorem, there is a unique reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ on $\mathcal{X}$ for which the Mercer kernel K is the reproducing kernel. This RKHS can be characterized by a series representation $\mathbb{H} = \left\{f \in L^2_{p_X}(\mathcal{X}) : \|f\|^2_{\mathbb{H}} = \sum_{i=1}^{\infty} f_i^2/\mu_i < \infty, f_i = \langle f, \psi_i\rangle_2\right\}$, equipped with the inner product $\langle f, g\rangle_{\mathbb{H}} = \sum_{i=1}^{\infty} f_i g_i/\mu_i$ for any $f = \sum_{i=1}^{\infty} f_i\psi_i$ and $g = \sum_{i=1}^{\infty} g_i\psi_i$ in $\mathbb{H}$.

We consider a proximate function of f_0 in $\mathbb{H}$, defined as

$$f_\lambda = (L_K + \lambda I)^{-1} L_K f_0 = \sum_{i=1}^{\infty} \frac{\mu_i}{\mu_i + \lambda} f_i\psi_i,$$

where I is the identity operator.

We make some differentiability assumptions on $K(\cdot,\cdot)$:

Assumption B1. $K(\cdot,\cdot) \in C^8(\mathcal{X},\mathcal{X})$, i.e., $K_{jl}(x,x') = \frac{\partial^{j+l}K(x,x')}{\partial x^j \partial x'^l} \in C(\mathcal{X},\mathcal{X})$ for any $j,l \in \mathbb{N}_0$ and $j+l \leq 8$.

Define $\kappa_{jj} = \sup_{x\in\mathcal{X}} K_{jj}(x,x) > 0$ for $j = 0,\ldots,4$ and write $\kappa = \kappa_{00}$. We also define $\kappa_{0j} = \sup_{x,x'\in\mathcal{X}} |K_{0j}(x,x')|$ for $j = 1,\ldots,4$. Under Assumption B1, a direct application of Theorem 4.7 in Ferreira and Menegatto (2012) gives that $f \in C^3(\mathcal{X})$ for any $f \in \mathbb{H}$, and $\|f^{(3)}\|_\infty \leq \sqrt{\kappa_{33}}\|f\|_{\mathbb{H}}$. In particular, we have $f_\lambda \in C^4(\mathcal{X})$ under Assumption B1.

We further define kth order derivatives for $\widehat{\mu}_{f'}(x)$ and $\widehat{\sigma}^2_{f'}(x)$ as follows for $k = 0,1,2,3$, with $k = 0$ corresponding to the original functions:

$$\begin{aligned}
\widehat{\mu}^{(k)}_{f'}(x) &:= \frac{d^k}{dx^k} K_{10}(x,X)[K(X,X)+n\lambda \boldsymbol{I}_n]^{-1}\boldsymbol{y} = K_{k+1,0}(x,X)[K(X,X)+n\lambda \boldsymbol{I}_n]^{-1}\boldsymbol{y}, \\
\widehat{\sigma}^{2(k)}_{f'}(x) &:= \frac{d^k}{dx^k}\sigma^2(n\lambda)^{-1}\left\{K_{11}(x,x) - K_{10}(x,X)[K(X,X)+n\lambda \boldsymbol{I}_n]^{-1}K_{01}(X,x)\right\} \\
&= \sigma^2(n\lambda)^{-1}\sum_{i=0}^{k}\binom{k}{i}\{K_{i+1,k+1-i}(x,x) \qquad (7) \\
&\qquad -K_{i+1,0}(x,X)[K(X,X)+n\lambda \boldsymbol{I}_n]^{-1}K_{0,k+1-i}(x,X)\big\},
\end{aligned}$$

where (7) uses the general Leibniz rule for matrix operation.

The following Lemma 1 establish a range of non-asymptotic error bounds under a high probability event. Let $K_{jl,x}(\cdot) = K_{jl}(x,\cdot)$ and $\varphi_{jl}(x) = (L_K+\lambda I)^{-1}K_{jl,x}(x)$.

Lemma 1. *Under Assumption B1, the following bounds for $k = 0,1,2,3$ holds simultaneously under a high probability event A_n with $\mathbb{P}_0(A_n) \geq 1 - n^{-10}$:*

$$\begin{aligned}
\|\widehat{\mu}^{(k)}_{f'} - f^{(k+1)}_\lambda\|_\infty \leq{}& \frac{\sqrt{\kappa\kappa_{k+1,k+1}}\|f_0\|_\infty\sqrt{10\log n+5}}{\sqrt{n\lambda}}\left(10 + \frac{4\kappa\sqrt{10\log n+5}}{3\sqrt{n\lambda}}\right) \\
&+ \frac{C_2\sqrt{\kappa\kappa_{k+1,k+1}}\sigma\sqrt{10\log n+4}}{\sqrt{n\lambda}},
\end{aligned}$$

$$|\widehat{\sigma}^{2(k)}_{f'}(x) - \sigma^2 n^{-1} \sum_{i=0}^{k} \binom{k}{i} \varphi_{i+1,k+1-i}(x)| \tag{8}$$
$$\leq \sum_{i=0}^{k} \binom{k}{i} \left[\frac{\sqrt{\kappa \kappa_{i+1,i+1}} \kappa_{0,k+1-i} \sigma^2 \sqrt{10 \log n + 4}}{n\sqrt{n}\lambda^2} \left(10 + \frac{4\sqrt{\kappa}\sqrt{10 \log n + 4}}{3\sqrt{n\lambda}} \right) \right].$$

The following assumption allows us to simplify the bounds for $\widehat{\sigma}^{2(k)}_{f'}(x)$ for $k = 0, 1, 2, 3$.

Assumption B2. $\lambda \sup_{x \in \mathcal{X}} |\varphi_{jl}(x)|$ is bounded for $j, l \geq 1$ and $j + l \leq 5$.

Remark 1. Under Assumption B2, Equation (8) yields $\|\widehat{\sigma}^2_{f'}(x) - \sigma^2 n^{-1} \varphi_{11}(x)\|_\infty \lesssim \frac{\sqrt{\log n}}{n\sqrt{n}\lambda^2}$ and $\|\widehat{\sigma}^{2(k)}_{f'}(x)\|_\infty \lesssim \frac{1}{n\lambda}$ for $k = 1, 2, 3$. Hence, $\widehat{\sigma}^2_{f'}(x)$ approximates $\varphi_{11}(x)\sigma^2/n$ with high probability.

We make the following assumption to ease the presentation. The subsequent theory in Theorems 1 and 2 can be generalized for cases where Assumption B3 does not hold, with more complicated expressions that involve $K_{11}(x, x)$ and its derivatives.

Assumption B3. $K_{11}(x, x)$ does not depend on x.

Assumption B3 simplifies the posterior $\pi_n(t \mid X, \boldsymbol{y})$ in Proposition 1. It holds for any stationary kernels; this is because if $K(x, x') = g(x - x')$ for some function $g(\cdot)$, then $K_{11}(x, x) = -g''(0)$, which does not depend on x.

The following assumption is concerned with the error term $f'_\lambda - f'_0$. This is a deterministic function as f_λ does not depend on random draws of covariates and noise.

Assumption C. $\|f'_\lambda - f'_0\|_\infty \lesssim \lambda^{r_1}$ and $\|f''_\lambda - f''_0\|_\infty \lesssim \lambda^{r_2}$ for some $0 < r_1, r_2 \leq 1$.

Assumption C ensures that f'_λ and f''_λ converge to f'_0 and f''_0 under the supremum norm, respectively. The two parameters r_1 and r_2 correspond to approximation properties of f_λ to the function class that f_0 belongs to, and such properties in turn depend on the covariance kernel and smoothness of the function class. Assumption C is typically verifiable via direct calculation for a given problem. For example, if $L_K^{-r'-\frac{1}{2}} f_0 \in L^2_{p_X}(\mathcal{X})$ for some $0 < r' \leq \frac{1}{2}$,

then the one-dimensional case of Theorem 6 in Liu and Li (2023) gives $r_1 = r_2 = r'$; here the function class for f_0 is defined by integral operators.

Assumptions B1-B3 spell out conditions for the kernel and regularization parameter λ, while Assumption C is a generic condition that includes a range of function classes of f_0. In Section 3.5 we provide examples where Assumptions B1-B3 and C hold.

3.2 LAN property

The following Lemma 2 shows existence of local extrema of f_λ and $\widehat{\mu}_f$ within a neighborhood of t_m, which are respectively denoted by $t_{\lambda,m}$ and $\hat{t}_m$ for $m = 1, \ldots, M$.

Lemma 2. *Under Assumptions A1-A3, B1 and C, for any sufficiently small λ, there exist $\{t_{\lambda,m} : m = 1, \ldots, M\}$ such that each $t_{\lambda,m}$ is a local extremum of f_λ and $|t_{\lambda,m} - t_m| \lesssim \lambda^{r_1}$. Moreover, under A_n, there exist $\{\hat{t}_m : m = 1, \ldots, M\}$ such that each $\hat{t}_m$ is a local extremum of $\widehat{\mu}_f$ and $|\hat{t}_m - t_{\lambda,m}| \lesssim \sqrt{\log n}/(\sqrt{n}\lambda)$.*

Henceforth we work under the high probability event A_n defined in Lemma 1. Let the regularization parameter $\lambda = n^{-\frac{1}{2}+\beta}(\log n)^{\frac{1}{2}+a}$ for some $0 < \beta < \frac{1}{2}$ and $a > 0$. Since $f'_\lambda(t_{\lambda,m}) = 0$, Lemma 1 implies that

$$|n^\beta \widehat{\mu}_{f'}(t_{\lambda,m})| \lesssim (\log n)^{-a}, \tag{9}$$

$$|\widehat{\mu}^{(k)}_{f'}(t_{\lambda,m}) - f^{(k+1)}_\lambda(t_{\lambda,m})| \lesssim n^{-\beta}(\log n)^{-a}, \quad 1 \le k \le 3, \tag{10}$$

$$\left| n\widehat{\sigma}^2_{f'}(t_{\lambda,m}) - \sigma^2\varphi_{11}(t_{\lambda,m}) \right| \lesssim n^{\frac{1}{2}-2\beta}(\log n)^{-1-a}, \tag{11}$$

$$|\widehat{\sigma}^{2(k)}_{f'}(t_{\lambda,m})| \lesssim n^{-\frac{1}{2}-\beta}(\log n)^{-\frac{1}{2}-a}, \quad 1 \le k \le 3. \tag{12}$$

The following Theorem 1 characterizes the marginal likelihood function of t by presenting a local asymptotic normality (LAN) property at $t_{\lambda,m}$, generalizing traditional LAN properties to the considered encompassing semiparametric regime. For $m = 1, \ldots, M$, we denote by

$$\varphi_{11,m} = \varphi_{11}(t_{\lambda,m}), \quad \sigma^{*2}_m = \sigma^2/f''_0(t_m)^2. \tag{13}$$

Theorem 1. *Suppose Assumptions A1-A3, B1-B3 and C hold. Let* $\lambda = n^{-\frac{1}{2}+\beta}(\log n)^{\frac{1}{2}+a}$ *for some* $\frac{1}{4} < \beta < \frac{1}{2}$ *and* $a > 0$. *Suppose* $n^{\frac{3}{2}-4\beta}\varphi_{11,m}^{-2}$ *is bounded for* $m = 1, \ldots, M$. *Then under event* A_n *as* $n \to \infty$, *for any* $m = 1, \ldots, M$, *the marginal likelihood likelihood* $\ell(t)$ *satisfies the LAN property*

$$\log \frac{\ell(t_{\lambda,m} + \frac{u}{n^\beta})}{\ell(t_{\lambda,m})} = n^{1-2\beta}\varphi_{11,m}^{-1}\left(-\frac{u^2}{2\sigma_{n,m}^2} - \frac{\mu_{n,m}u}{\sigma_{n,m}^2}\right) + o(1), \tag{14}$$

where $\mu_{n,m} = \frac{n^\beta \widehat{\mu}_{f'}(t_{\lambda,m})}{\widehat{\mu}'_{f'}(t_{\lambda,m})}$ *and* $\sigma_{n,m}^2 = \frac{n\varphi_{11,m}^{-1}\widehat{\sigma}_{f'}^2(t_{\lambda,m})}{\widehat{\mu}'_{f'}(t_{\lambda,m})^2}$.

Moreover, letting $r = r_1 \wedge r_2$, *we have*

$$|\mu_{n,m}| \lesssim (\log n)^{-a}, \tag{15}$$

$$|\sigma_{n,m}^2 - \sigma_m^{*2}| \lesssim \lambda^r. \tag{16}$$

Remark 2. The LAN property in Theorem 1 exhibits two differences compared to classical LAN properties (Section 7, van der Vaart (2000), Kleijn and van der Vaart (2012)), in part owing to the adopted unconventional encompassing approach along with the semiparametric problem under consideration. First, the inflation term $n^{1-2\beta}\varphi_{11,m}^{-1}$ is absent in classical semiparametric LAN expansions, pointing to complications in the rate calculation that integrate properties of K and the choice of λ through $\varphi_{11,m}$. Second, our expansion is specific to $t_{\lambda,m}$, a local extremum of f_λ, with varying quantities $\mu_{n,m}$ and $\sigma_{n,m}^2$. As a result, the posterior distribution cannot be approximated by a Gaussian after a homogeneous rescaling of the parameter across $\mathcal{X}$; this reminds us of the *localized* feature of local extrema, and suggests localized rescaling within a small interval centered at t_m. Unlike existing work in semiparametric Bayes where the posterior distribution converges to a single Gaussian, a multi-modal posterior distribution in the form of a mixture of normal distributions is expected.

3.3 Multi-modal limiting distribution

We make the following mild assumption on the prior $\pi(t)$:

Assumption D. The prior density $\pi(t)$ satisfies that $\pi(t) \in C(\mathcal{X})$ and has positive density at local extrema, i.e., there holds $\pi(t_m) > 0$ for $m = 1, \ldots, M$.

This assumption on $\pi(t)$ is rather flexible and can be satisfied by most continuous distributions supported on $\mathcal{X}$, such as the beta distribution.

Let $\Pi_n(\cdot \mid X, \boldsymbol{y})$ be the probability measure of $\pi_n(\cdot \mid X, \boldsymbol{y})$, and $\Phi(\cdot \mid \mu, \sigma^2)$ be the cumulative distribution function of the normal distribution $N(\mu, \sigma^2)$. The following theorem shows that $\pi_n(t \mid X, \boldsymbol{y})$ is close to a mixture of normal distributions when the sample size is large, where the component densities are related to the LAN expansion, and the weights are determined by both prior density and curvature of f_0 at t_m.

Theorem 2. *Suppose Assumptions A1-A3, B1-B3, C and D hold, and let* $\lambda = n^{-\frac{1}{2}+\beta}(\log n)^{\frac{1}{2}+a}$ *for some* $\frac{1}{4} < \beta < \frac{1}{2}$ *and* $a > 0$ *such that* $n^{\frac{3}{2}-4\beta}\varphi_{11,m}^{-2}$ *is bounded. Then the following results hold for any* $z \in \mathbb{R}$*:*

(i) Letting $\pi_m = \frac{|f_0''(t_m)|^{-1}\pi(t_m)}{\sum_{m=1}^{M}|f_0''(t_m)|^{-1}\pi(t_m)}$*, we have*

$$\left|\Pi_n(t \le z \mid X, \boldsymbol{y}) - \sum_{m=1}^{M} \pi_m \Phi(z \mid t_m, n^{-1}\varphi_{11,m}\sigma_m^{*2})\right| \to 0$$

in $\mathbb{P}_0$*-probability, where* $\varphi_{11,m}$ *and* σ_m^{*2} *are given by* (13)*.*

(ii) Letting $\Pi'_{n,m}(\cdot \mid X, \boldsymbol{y})$ *be the posterior of* $\sqrt{\frac{n}{\varphi_{11,m}}}(t\boldsymbol{1}_{I_m}(t) - \hat{t}_m + b_n)$ *where* $b_n = n^{-\beta}\log n$*, and* $I_m = [t_m - \zeta_{m-1}, t_m + \zeta_m]$ *with* $\zeta_0 = t_1$*,* $\zeta_m = (t_{m+1} - t_m)/2$ *for* $m = 1, \ldots, M-1$*, and* $\zeta_M = 1 - t_M$*, we have for any* $z \in \mathbb{R}$*, in* $\mathbb{P}_0$*-probability,*

$$\left|\Pi'_{n,m}(t' \le z \mid X, \boldsymbol{y}) - \Phi(z \mid 0, \sigma_m^{*2})\right| \to 0.$$

A few remarks are in order to elucidate the encompassing strategy using Theorem 2.

Remark 3. Part (i) of Theorem 2 shows that the posterior distribution is close to a mixture of normal distributions. The curvature of f at a local extremum t, defined as

$|f''(t)|/(1+\{f'(t)\}^2)^{3/2}$ that reduces to $|f''(t)|$ when $f'(t)=0$, directly affects the variation of the posterior distribution at t. Indeed, the standard derivation of the Gaussian component at t in the limiting distribution is proportional to $1/|f_0''(t)|$, meaning a large curvature leads to a more concentrated normal component in the posterior distribution. Interestingly, this effect of curvature on the component-wise variance is offset by the mixture weight that is also proportional to $1/|f_0''(t)|$. More specifically, the limiting mixture normal density function evaluated at each local extremum t_m, which is $\pi_m/\sqrt{2\pi n^{-1}\varphi_{11,m}\sigma_m^{*2}} \propto \pi(t_m)/\sqrt{\varphi_{11,m}}$, does not depend on the curvature of f at t_m. Hence, the multi-modal posterior distribution of t does not diminish a local extremum with small curvature, at least asymptotically. The prior weight $\pi(t_m)$ plays a direct role in driving the posterior distribution, which enables incorporating prior knowledge. Section 4.1 provides numerical confirmation for these theoretical implications using finite sample illustration; see, in particular, Figure 2. The marginal likelihood function $\ell(t)$, proportional to the posterior density $\pi_n(t \mid X, \boldsymbol{y})$ with a uniform prior, also tends to be multi-modal, as observed in Figure 2.

Remark 4. Part (ii) generalizes the BvM phenomenon to the encompassing semiparametric regime under consideration. In particular, after rescaling and truncation, the posterior distribution weakly converges to a normal distribution with a bias term b_n that is $o(1)$. BvM theorems in weak convergence (i.e., convergence in distribution) are common in the literature (Kim and Lee, 2004; Castillo and Rousseau, 2015; Castillo and Nickl, 2014; Kim, 2006). Our result is different in that the target distribution is multi-modal with varying mean and variance at each component, necessitating a localized truncation at I_m for weak convergence to a single normal. Although not typical, approximating the posterior distribution via a mixture of Gaussians has appeared in the literature; for example, see Castillo et al. (2015) on Bayesian linear regression models. The established results and proofs show other important differences from the semiparametric literature. Castillo (2012) proposed sufficient conditions for a BvM theorem for separated models, where the model parameter takes the form $\eta = (\theta, f)$. In our case, t is an inherited hyperparameter of f rather than an independent parameter in a separated model (e.g., parameters in a location-scale family). Castillo and Rousseau (2015) provided sufficient conditions for a BvM theorem for smooth functionals

of the parameter in general models. Specifically, they considered a model parameterized by $\theta \in \Theta$, and provided a BvM theorem for $\psi(\theta)$ where $\psi : \Theta \to \mathbb{R}^d$ is a smooth functional of interest (see Equation (2.4) in their paper). However, the derivative at any fixed point, when viewed as a functional of the regression function, is not bounded (Conway, 1994, page 13), and thus local extrema may not be expressed as a functional of the regression function even when their number is known.

Remark 5. Part (ii) indicates a bias-variance trade-off regarding the rescaled posterior for estimating t_m. The bias of the centering quantity $\hat{t}_m - b_n - t_m$ is bounded above by λ^{r_1}, in view of Lemma 2 and the constraint that $\frac{1}{4} < \beta < \frac{1}{2}$. On the other hand, the variance $n^{-1}\varphi_{11,m}$ is bounded above by $(n\lambda)^{-1}$ in view of Assumption B2. Therefore, a larger λ (or equivalently, a larger β) corresponds to a smaller variance but a larger bias. An exact rate calculation for both bias and variance can be obtained by considering the special cases in Section 3.5. Note that such rates depend on the underlying regularity parameter of the true function that is typically unknown, impeding their application to parameter tuning. We propose to use an empirical Bayes approach to select λ by maximizing its marginal likelihood function, which shows competitive performance in our simulations; see Section 4 for details.

Verifying the conditions needed for the preceding theorems often amounts to checking Assumptions B1-B3 and Assumption C, which will provide insight into how to choose the kernel hyperparameters; see Section 3.5 for examples.

3.4 Point and interval estimation

The shape of the posterior distribution characterized in Theorem 2 can be used to construct estimators with frequentist properties. In particular, the multi-modality of the limiting posterior distribution provides a basis to overcome the multiplicity challenge of local extrema, and leads to consistent estimators of t_m through posterior exploration. We additionally provide interval estimation that achieves frequentist coverage.

Theorem 3. *Under the same conditions as in Theorem 2, the following results hold.*

(i) For all $\delta > 0$, with $\mathbb{P}_0$-probability tending to one, $\pi_n(t \mid X, \boldsymbol{y})$ has exactly M local maxima $t_{1,n}, \ldots, t_{M,n}$, and for $m = 1, \ldots, M$ there holds $|t_{m,n} - t_m| \leq \delta$ for all sufficiently large n.

(ii) Let $\Delta_n(\cdot) = K_{10}(\cdot, X)[K(X, X) + n\lambda \boldsymbol{I}_n]^{-1} f_0(X)$. For any $\alpha \in (0, 1)$, the following is an asymptotic $1 - \alpha$ confidence interval for $t_m + \Delta_n(t_m)/f_0''(t_m)$:

$$\hat{t}_m \pm z_{\alpha/2}\sigma \sqrt{K_{10}(\hat{t}_m, X)[K(X, X) + n\lambda \boldsymbol{I}_n]^{-2} K_{10}(\hat{t}_m, X)^T} / |\widehat{\mu}'_{f'}(\hat{t}_m)|,$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

In Theorem 3 (ii) the bias term $\Delta_n(t_m)/f_0''(t_m)$ can be shown to be $o(1)$ with $\mathbb{P}_0$-probability tending to one as it is approximating $f_0'(t_m)/f_0''(t_m) = 0$. For implementation, we propose to "plug in" consistent estimators for unknown quantities in $\Delta_n(t_m)/f_0''(t_m)$, including $\widehat{\mu}'_{f'}(\hat{t}_m)$ for $f_0''(t_m)$, $\hat{t}_m$ for t_m, and $\widehat{\Delta}_n(\cdot) = K_{10}(\cdot, X)[K(X, X) + n\lambda \boldsymbol{I}_n]^{-1} \widehat{\mu}_f(X)$ for $\Delta_n(\cdot)$, leading to the following confidence intervals for t_m:

$$\hat{t}_m + \frac{\widehat{\Delta}_n(\hat{t}_m)}{\widehat{\mu}'_{f'}(\hat{t}_m)} \pm z_{\alpha/2} \frac{\sigma \sqrt{K_{10}(\hat{t}_m, X)[K(X, X) + n\lambda \boldsymbol{I}_n]^{-2} K_{10}(\hat{t}_m, X)^T}}{|\widehat{\mu}'_{f'}(\hat{t}_m)|}. \tag{17}$$

These confidence intervals depend on λ, which will be estimated using empirical Bayes, as discussed in Remark 5. In the context of nonparametric regression, Liu and Li (2022) have demonstrated that this choice of λ tends to adapt to the unknown smoothness level of the underlying function, especially when combined with an oversmooth kernel. However, it is worth noting that constructing adaptive confidence intervals with minimax optimal diameter is a more challenging task, as is typically the case in nonparametric inference; see, for example, Giné and Nickl (2021) for more details. We assess finite sample performance of (17) in Section 4, which shows satisfactory coverage.

When the error variance σ^2 is unknown, one may substitute σ^2 in the original derivative-constraint GP prior with an estimate $\hat{\sigma}_n^2$. The established results in Theorems 1, 2, and 3 hold with any estimator $\hat{\sigma}_n^2$ that converges to σ^2 in mean square. In particular, we can estimate σ^2 by the maximum marginal likelihood estimator which has been shown to be

mean square consistent under various settings (Yoo and Ghosal, 2016; Liu and Li, 2022). Denote the induced posterior measure of t by $\Pi_{n,\hat{\sigma}_n^2}(\cdot \mid X, \boldsymbol{y})$, and take Part (i) in Theorem 2 as an example. Let $\mathcal{B}_n$ be a shrinking neighborhood of σ^2 such that $\mathbb{P}_0(\hat{\sigma}_n^2 \in \mathcal{B}_n) \to 1$. Conditional on $\mathcal{B}_n$, Equation (8) becomes

$$
\begin{aligned}
&|\widehat{\sigma}_{f'}^{2(k)}(x) - (\sigma^2 + o(1))n^{-1}\sum_{i=0}^{k}\binom{k}{i}\varphi_{i+1,k+1-i}(x)| \\
&\quad \le \sum_{i=0}^{k}\binom{k}{i}\left[\frac{\sqrt{\kappa\kappa_{i+1,i+1}}\kappa_{0,k+1-i}(\sigma^2+o(1))\sqrt{10\log n+4}}{n\sqrt{n}\lambda^2}\left(10+\frac{4\sqrt{\kappa}\sqrt{10\log n+4}}{3\sqrt{n\lambda}}\right)\right],
\end{aligned}
$$

and all the established inequalities in the proof of Theorem 2 hold uniformly over $\sigma^2 \in \mathcal{B}_n$. In particular, there holds $\sup_{\sigma^2\in\mathcal{B}_n}\left|\Pi_n(t \le z \mid X, \boldsymbol{y}) - \sum_{m=1}^{M}\pi_m\Phi(z \mid t_m, n^{-1}\varphi_{11,m}\sigma_m^{*2})\right| \to 0$ in $\mathbb{P}_0$-probability, yielding $\left|\Pi_{n,\hat{\sigma}_n^2}(t \le z \mid X, \boldsymbol{y}) - \sum_{m=1}^{M}\pi_m\Phi(z \mid t_m, n^{-1}\varphi_{11,m}\sigma_m^{*2})\right| \to 0$.

3.5 Applications to special function classes and GP kernels

In this section, we provide examples under which various assumptions and Theorems 1 and 2 hold. We focus on covariance kernels that possess regularized eigenfunctions as follows.

Assumption E. The eigenfunction $\psi_i \in C^s(\mathcal{X})$ for all $i \in \mathbb{N}$ and some $s \in \mathbb{N}$. Moreover, there exists a constant $C > 0$ such that $\|\psi_i^{(s)}\|_\infty \le Ci^s$ for any $i \in \mathbb{N}$ and $\sum_{i=1}^{\infty}\psi_i'(x)^2$ diverges for any $x \in \mathcal{X}$.

In particular, the Fourier basis satisfies Assumption E for any finite s.

Example 1. Stationary kernels with polynomially decaying eigenvalues. Let $K^{\alpha,s}$ be a stationary covariance kernel whose eigenfunctions satisfy Assumption E and eigenvalues decay at a polynomial rate, that is, $\mu_i \asymp i^{-2\alpha}$ for $i \in \mathbb{N}$ and some $\alpha > 0$.

We assume that the true regression function f_0 lies in the Hölder class:

$$
H^\alpha(\mathcal{X}) = \left\{ f \in L^2_{p_X}(\mathcal{X}) : \|f\|^2_{H^\alpha(\mathcal{X})} = \sum_{i=1}^{\infty} i^\alpha |f_i| < \infty, f_i = \langle f, \psi_i\rangle_2 \right\}.
$$

Any function in $H^\alpha(\mathcal{X})$ has continuous derivatives up to order $\lfloor\alpha\rfloor$ and the $\lfloor\alpha\rfloor$th derivative is Lipschitz continuous of order $\alpha - \lfloor\alpha\rfloor$. Using the error bounds for derivatives of $f_\lambda - f_0$ in Lemma 13 of Liu and Li (2023) to verify Assumption C, the following corollary provides an example for Theorems 1 and 2 to hold.

Theorem 4. *Suppose $f_0 \in H^\alpha(\mathcal{X})$ and $K^{\alpha,s}$ is used in the GP prior for $\alpha > 9/2$ and $s \geq 4$. Then Assumptions B1-B3 and C hold, and Theorems 1 and 2 hold for any $\beta \in [\frac{3}{8}, \frac{1}{2})$.*

Example 2. Stationary kernels with exponentially decaying eigenvalues. We consider stationary kernels $K_{\gamma,s}$ with eigenvalues $\mu_i \asymp e^{-2\gamma i}$ for $i \in \mathbb{N}$ and some $\gamma > 0$, and eigenfunctions satisfying Assumption E. The well-known squared exponential kernel can be approximately viewed as an example of $K_{\gamma,s}$, with a closed-form eigendecomposition with respect to Gaussian sampling on the real line (Rasmussen and Williams, 2006; Pati and Bhattacharya, 2015).

We assume f_0 belongs to the analytic-type function class $A_\gamma(\mathcal{X})$:

$$A_\gamma(\mathcal{X}) = \left\{ f \in L^2_{p_X}(\mathcal{X}) : \|f\|^2_{A_\gamma(\mathcal{X})} = \sum_{i=1}^{\infty} e^{\gamma i} |f_i| < \infty, f_i = \langle f, \psi_i \rangle_2 \right\}.$$

We first verify Assumption C in the following Lemma 3.

Lemma 3. *Suppose that $f_0 \in A_\gamma(\mathcal{X})$, and $K_{\gamma,s}$ is used in the GP prior for some $\gamma > \frac{k}{e}$ and $s \geq k$. Then there holds $\|f_\lambda^{(k)} - f_0^{(k)}\|_\infty \lesssim \lambda^{\frac{1}{2} - \frac{k}{2e\gamma}}$.*

This yields another example for our theory to hold using kernels with exponentially decaying eigenvalues, formulated in the following corollary.

Theorem 5. *Suppose $f_0 \in A_\gamma(\mathcal{X})$ and $K_{\gamma,s}$ is used in the GP prior for $\gamma > \frac{5}{2e}$ and $s \geq 4$. Then Assumptions B1-B3 and C hold, and Theorems 1 and 2 hold for any $\beta \in [\frac{3}{8}, \frac{1}{2})$.*

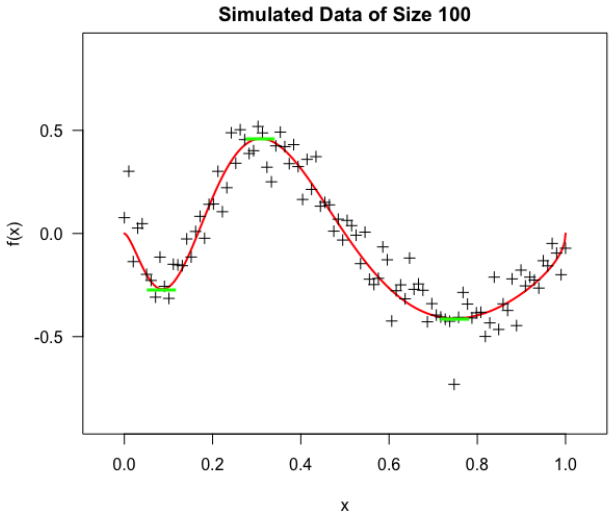


Figure 1: Simulated data with $n = 100$. Observations are marked by "+". The red curve is the true regression function, and green lines indicate the location of three local extrema.

4 Simulation

In this section we carry out simulation studies to illustrate the convergence of the posterior distribution $\pi_n(t \mid X, \boldsymbol{y})$ to Gaussian mixtures, and to assess the performance of the proposed method relative to competing methods.

We use a Doppler-type regression function $f(x) = \sqrt{x(1-x)}\sin(2\pi/(x+0.5))$ for $x \in \mathcal{X} = [0,1]$, which has three local extrema at $t_1 = 0.0863$, $t_2 = 0.3096$ and $t_3 = 0.7491$. We add iid zero-mean Gaussian noise to f with standard deviation $\sigma = 0.1$, observed at equal-spaced $\{x_i\}_{i=1}^n$ in the unit interval $[0,1]$. We vary the sample size $n = 100, 500, 1000$. Figure 1 shows one simulated dataset of sample size 100. Each simulation scenario is replicated 100 times.

We use two prior distributions Beta(1, 1) (the uniform distribution) and Beta(2,3) on t to study the sensitivity of posterior inference to the prior specification. We use the squared exponential kernel function for K, that is, $K(x, x') = \exp\{-(x-x')^2/(2h^2)\}$. For each simulated dataset, we select parameters other than t, which include λ, h, and σ^2, via an empirical Bayes approach by maximizing the unconstrained marginal likelihood function, i.e., the multivariate normal density $N(0, \sigma^2(n\lambda)^{-1}K(X,X) + \sigma^2\boldsymbol{I}_n)$ evaluated at $\boldsymbol{y}$. This is motivated by the excellent performance of empirical Bayes in a variety of settings (Yoo

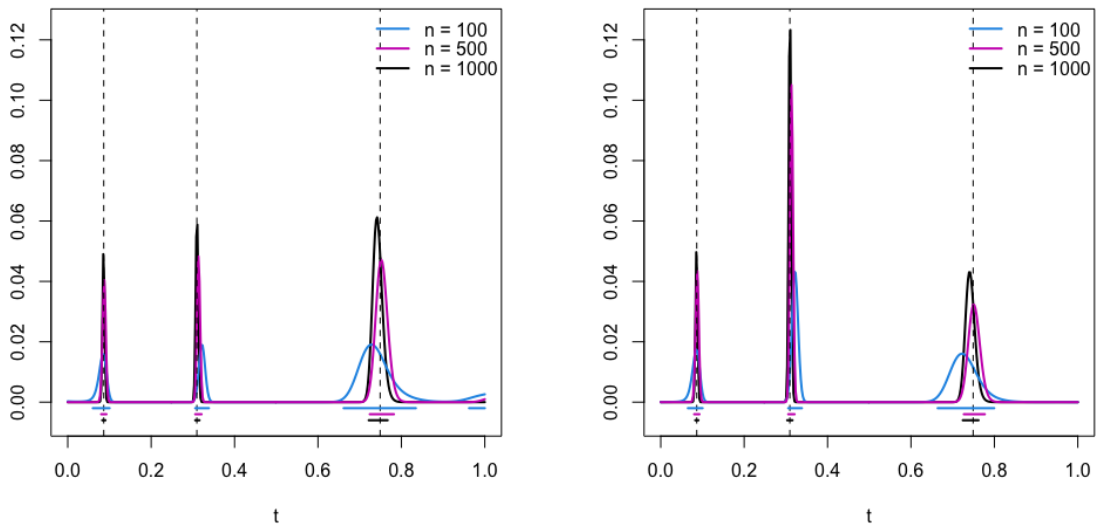
and Ghosal, 2016; Liu and Li, 2022). We use the midpoint rule to calculate the normalizing constant in $\pi_n(t \mid X, \boldsymbol{y})$.

4.1 Finite sample size behavior of the posterior distribution

The proposed method, labeled as DGP, does not require any sampling to obtain the posterior distribution owing to the closed-form expression in Proposition 1. Figure 2 shows the posterior distribution of t with various sample sizes and the two beta priors, each based on one simulated dataset. We can see that the posterior distribution possesses three mixture components at all sample sizes and for both beta priors, matching the true number of local extrema $M = 3$. At each sample size such as $n = 1000$, the posterior distribution tends to have three modes concentrating around (t_1, t_2, t_3), which aligns with the established Theorem 2 that deciphers the limiting behavior of the posterior distribution.

The curvature at a local extremum point t is $|f''(t)|$, which is $(111.04, 44.55, 11.91)$ for (t_1, t_2, t_3), respectively. Figure 2 indicates that the variability of each mixture component decreases substantially as the curvature increases, confirming Theorem 2 in which we show that the standard deviation of each Gaussian component in the limiting distribution is inversely proportional to the curvature at the corresponding local extremum. For example, t_1 with the highest curvature exhibits the least variation, while t_3 with the lowest curvature has the most variation in the posterior distribution, as in Figure 2.

As n increases, the mixture components in the posterior distribution are more bell-shaped. When the sample size is small, such as $n = 100$, the mixture component may be skewed. This is particularly the case for the first (left skewed) and the third mixture component (right skewed) when the Beta(1,1) prior is used. A closer inspection of Figure 2 (a) indicates a *boundary effect* when $n = 100$, that is, there appears to be a small bump near the boundary. Such boundary effects are reasonable as there are sparser data near the boundary, lacking information outside the range $\mathcal{X}$, and that the right boundary point $t = 1$ indeed gives the largest function value on $(t_3, 1]$. Both skewed mixture components and boundary effects are much mitigated when the sample size increases to 500 and 1000. In addition, the Beta(2, 3)



(a) Beta(1, 1) prior (b) Beta(2, 3) prior

Figure 2: Effect of n on the posterior distribution of t with beta prior distributions. Vertical dashed lines indicate the true locations of local extrema. The intervals in each plot are the 95% highest posterior density regions. Each posterior density function is based on one simulated dataset.

prior tends to zero out the boundary bumps even when the sample size is as small as 100. The posterior distributions corresponding to the two beta priors have various density values at their local peaks, which are also suggested by Theorem 2. We remark that, however, an appropriate posterior summary method such as highest posterior density regions may lead to interval estimates that are less sensitive to the priors; see the interval estimates in Figure 2, and the next section, particularly Table 1, for more details about point estimates.

The proposed Bayesian approach has the advantage of allowing users to incorporate any prior knowledge, if available, about the location of local extrema. For example, one may use a suitable prior distribution to rule out the possibilities of local extrema near the boundary. This does not necessarily mean that local extrema are not located near the boundary, but rather that such local extrema are not desired.

4.2 Comparison with other methods

We use the 95% highest posterior density region (HPDR) of the posterior distribution of t for posterior summary, which consists of a number of disjoint intervals enclosing local

modes. We use the number of segments in the HPDR to estimate M, and the corresponding posterior mode within each segment to estimate local extrema. This posterior summary is a reasonable strategy according to our asymptotic characterization of the posterior distribution of t, which approximates a mixture of Gaussians with the number and location of the mixture components matching the local extrema of the underlying regression function. We expect this method to estimate M correctly with high probability according to Theorem 3 (i).

For comparison, we implement another three methods: the smoothed taut string (STS) method proposed by Kovac (2007), the original taut string (TS) method in Davies et al. (2001), and the nonparametric kernel smoothing (NKS) method proposed by Song et al. (2006). STS and TS estimate the number of local extrema by minimizing a loss function of the corresponding taut string, and do not provide uncertainty quantification about local extrema. Since STS is an improved version of TS, and we find that these two methods lead to similar numerical performance in our experiments, in this section we omit the results of TS and use STS to represent taut string-based methods. NKS first estimates the regression function, denoted by $\hat{f}(x)$, then chooses a set of x's such that the confidence interval of $f'(x)$ contains zero. Within this set, point estimates of stationary points are obtained by locating those at which $\hat{f}'(x)$ are closest to zero, denoted by $\{x_m^*\}_{m=1}^{\hat{M}}$, which are a subset of $\{x_i\}_{i=1}^n$ by design. For interval estimation, NKS inverts the lower and upper limits of the 95% confidence band for $\hat{f}'(x_m^*)$. Such intervals may not exist, and if one of the upper or lower bounds can be found, they further assume asymptotic normality and construct a symmetric interval based on x_m^* and the available bound. In contrast, interval estimation in the semiparametric Bayesian approach through HPDRs is computationally more straightforward and conceptually more coherent. STS and TS are implemented in the R package `ftnonpar`, and we implement NKS using the R code provided by the authors of Song et al. (2006). The bandwidth parameter in NKS is chosen by minimizing asymptotic mean integrated squared error.

4.2.1 Estimation of M

Figure 3 plots the estimated number of local extrema by each method at various sample sizes. The plot with $n = 1000$ is similar to the one with $n = 500$, and is thus omitted here.

We can see that the semiparametric Bayesian method DGP with both priors and the STS method tend to capture the true number of local extrema as the sample size increases to 500. When the sample size is 100, DGP with the Beta(2,3) prior gives the largest frequency at the true number $M = 3$ among all methods, while STS and DGP with the Beta(1,1) prior either underestimate or overestimate M in about half of the 100 simulations, although mostly by one. Figure 3 reinforces that the estimation accuracy of M is greatly improved by using the Beta(2, 3) prior to remove trivial points around the boundary, which gives the most accurate estimate of M for both $n = 100$ and $n = 500$. It is reassuring that the effect of prior distributions for DGP is diminished as n increases from 100 to 500. For both DGP and STS, a sample size over 500 appears large enough to ensure an accurate estimate of the number of local extrema, at least under the simulation setting. NKS does not exhibit a clear convergence behavior as DGP and STS do. It overestimates M considerably more than DGP and STS, and such overestimation persists when the sample size increases to 500 and 1000. This confirms that the presence of multiple local extrema poses challenges to NKS, as commented by Song et al. (2006), and suggests that multiple testing correction is particularly needed for the two-step approach, while appealing performance of the unified approach DGP does not hinge on such a correction.

Additional experiments are included in the supplementary material to investigate the effects of noise standard deviation and credible levels, which show quite robust performance of the Beta(2,3) prior in estimating M for a wide range of credible levels. A highly fluctuated regression function with large M is also considered.

4.2.2 Point estimation of local extrema

We now turn to comparing the estimates $\hat{t}_i$ for $i = 1, \ldots, \hat{M}$ for each method. Since $\hat{M}$ might deviate from $M = 3$, as suggested in Figure 3, we adopt the following convention to align the estimated local extrema with the true t_i for $i = 1, 2, 3$ for all methods. We consider three intervals (b_0, b_1), (b_1, b_2), and (b_2, b_3), where $b_0 = 0$, $b_3 = 1$, and $b_i = (t_i + t_{i+1})/2$ for $i = 1, 2$. Then for each method, we collect all estimated local extrema that fall into each interval. If the ith interval ($i = 1, 2, 3$) contains more than one estimate, we use the average of all

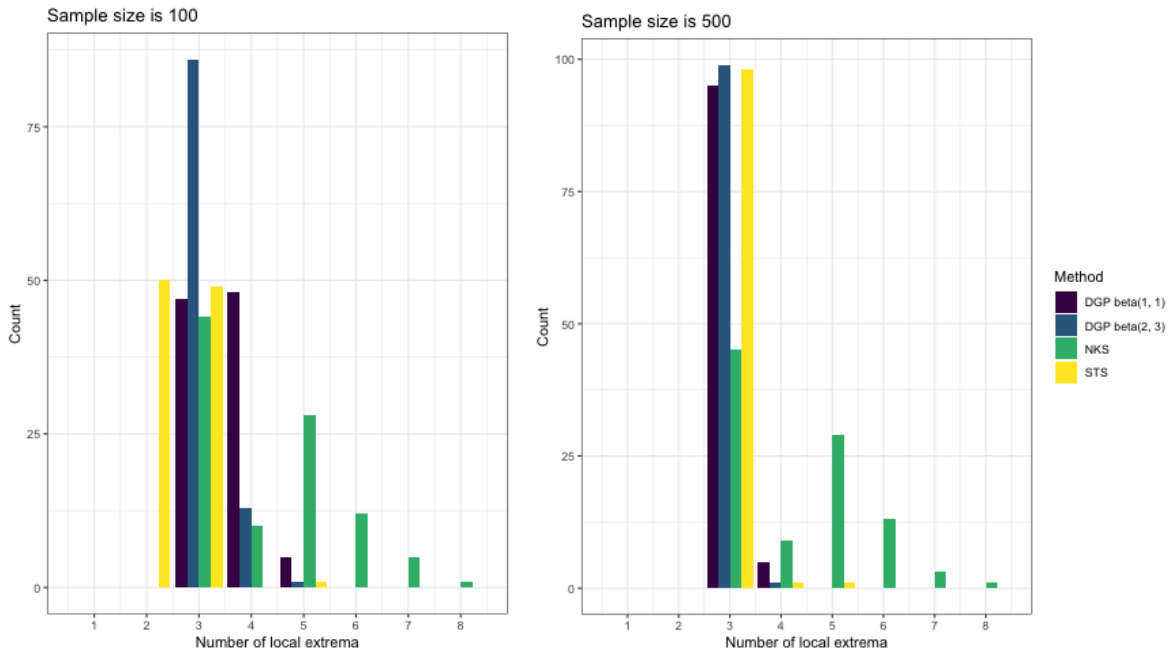


Figure 3: Frequency of the estimated number of local extrema by each method across 100 replicated simulations. Color code: yellow for STS, green for NKS, blue for DGP with the Beta(2,3) prior, and purple for DGP with the Beta(1, 1) prior. The plot with $n = 1000$ is similar to the one with $n = 500$, and is thus omitted here.

local extrema within the interval as the estimate of t_i; if an interval contains no estimates, we put an NA to indicate missingness. Performance of each method in estimating t_i for $i = 1, 2, 3$ is compared by calculating the root mean squared error (RMSE), averaged across 100 simulations excluding NAs. The number of simulations in which a method gives zero or multiple local extrema within each interval is reported in Table 2.

Table 1 reports the RMSE for estimated local extrema by all methods. The semiparametric Bayesian method DGP, with the Beta(1,1) or the Beta(2,3) prior, gives the smallest RMSEs in nearly all cases, with only one exception for t_3 at $n = 100$ when STS is slightly better. For t_1 and t_2, DGP often reduces the RMSEs of STS and NKS by over half or more, consistently across all sample sizes. For DGP, the two priors yield similar RMSEs in most cases, indicating that point estimates of local extrema tend to be minimally affected by the prior specification. Table 2 indicates that NKS produces multiple local extremum estimates in each interval much more often than DGP and STS, especially for t_3 with small curvature. When the sample size is 100, DGP leads to multiple local extrema in 14 (Beta(2,3) prior) and 15 (Beta(1, 1) prior)

out of 100 simulations, while STS misses estimates within $(0, b_1)$ for 50 simulations. Both DGP and STS estimate local extrema that align well with the true local extrema when n increases to 500 and 1000. It is worth mentioning that all methods give one local extremum in the interval (b_1, b_2) for almost all simulations (Table 2), which provides a scenario that eliminates the need to account for zero or multiple estimates; Table 1 shows that in this scenario that corresponds to estimating t_2 the proposed DGP achieves the smallest RMSEs, suggesting superior performance of DGP.

Table 1: Comparison of various methods using root mean square error (RMSE). The reported RMSEs are multiplied by 100 for easy comparison. The smallest and second smallest RMSEs in each column are marked in bold. All RMSEs are averaged across 100 repeated simulations.

Method	$n = 100$			$n = 500$			$n = 1000$		
	t_1	t_2	t_3	t_1	t_2	t_3	t_1	t_2	t_3
DGP Beta(1,1)	**0.67**	**0.88**	4.13	**0.29**	**0.54**	**2.11**	**0.25**	**0.46**	**1.31**
DGP Beta(2,3)	**0.65**	**0.88**	**3.85**	**0.30**	**0.54**	**2.00**	**0.24**	**0.45**	**1.30**
STS	1.65	1.53	**3.14**	0.78	1.22	2.56	0.75	1.11	1.90
NKS	1.68	1.46	4.98	1.54	1.08	3.30	1.90	1.07	2.25

Table 2: Number of simulations with missing or multiple estimated local extrema in each interval. The three intervals, indexed by t_1, t_2, and t_3 in the table, are $(0, (t_1 + t_2)/2), ((t_1 + t_2)/2, (t_2 + t_3)/2), ((t_2 + t_3)/2, 1)$, respectively. The number of simulations with missingness, if non-zero, is reported as the second number in a pair; otherwise if there is no missingness, we only report the number of simulations with multiple estimates.

Method	$n = 100$			$n = 500$			$n = 1000$		
	t_1	t_2	t_3	t_1	t_2	t_3	t_1	t_2	t_3
DGP Beta(1,1)	0	0	15	0	0	5	0	0	1
DGP Beta(2,3)	0	0	14	0	0	1	0	0	0
STS	(0, 50)	0	1	1	0	1	1	0	0
NKS	19	1	47	14	3	45	20	0	40

We note that there are a few noticeable differences in our implementation of the encompassing strategy with respect to Yu et al. (2023). In our estimation approach, we fix the hyperparameters σ, τ and h at the values that maximize the marginal unconstrained likelihood, while in the Monte Carlo Expectation Maximization (MCEM) approach of Yu et al. (2023) σ is sampled in each MC E-step while τ and h are iteratively updated in the M-step given their previous values and the samples of t and σ drawn in the E-step. Furthermore,

while we make use of the analytic form of the posterior distribution of t, the MCEM approach of Yu et al. (2023) draws posterior samples of t. One clear advantage of our implementation is that it is computationally much faster than the MCEM method. (Using $n = 100$ as an example, our implementation was 100 times faster than MCEM on a regular PC, at a magnitude of 32 seconds versus 3200 seconds for completing 100 simulations.) When applying the MCEM method to the simulation study, we found overall similar values in the selected parameters. For example, estimates for σ, τ and h, averaged over 100 simulated datasets, were 0.10, 0.30 and 0.13, respectively, for our approach and 0.14, 0.37 and 0.13, respectively, for the MCEM.

4.3 Interval estimation of local extrema

We now assess the proposed interval estimates in (17). Within each HPDR segment under the Beta(1,1) prior, we estimate $\hat{t}_m$ by finding the local extremum of $\hat{\mu}_f$; if multiple local extrema are found, then the average is used. We assess the coverage of interval estimators conditional on $\hat{M} = 3$. This conditional event tends to occur with probability one given the consistency of $\hat{M}$ and indeed has a high probability in finite sample settings as observed in Section 4.2.1. We consider three confidence levels $1 - \alpha$ for $\alpha \in \{0.1, 0.05, 0.1\}$. In addition to (marginal) confidence intervals for each t_i, we also obtain a joint confidence set for $\{t_1, t_2, t_3\}$ using the Bonferroni correction. We compare the empirical coverage of three marginal confidence intervals and one joint confidence set with the confidence level.

Table 3 shows that the empirical coverage is close to the nominal level when n increases to 500, for both marginal confidence intervals and joint confidence sets. We observe no significant derivation of the observed coverage from the confidence level relative to the standard errors when $n \in \{500, 1000\}$, indicating satisfactory coverage of the proposed interval estimates in this finite sample setting. In results not reported here, changing the prior to Beta(2,3) when deriving HPDR leads to a similar coverage for both marginal confidence intervals and joint confidence sets.

Table 3: Coverage of confidence intervals at various confidence levels $1-\alpha$ for $\alpha \in \{0.1, 0.05, 0.01\}$. The first three blocks report the coverage of marginal confidence intervals for each local extremum, while the last block is the coverage of joint confidence sets using the Bonferroni correction. For each of the three rows, the maximum standard errors are 0.07, 0.04, and 0.04, respectively.

	t_1			t_2			t_3			Joint		
	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
$n = 100$	0.81	0.91	0.98	0.75	0.81	0.87	0.74	0.81	0.83	0.66	0.72	0.77
$n = 500$	0.91	0.94	1	0.84	0.90	0.99	0.90	0.94	1	0.85	0.94	1
$n = 1000$	0.86	0.95	0.99	0.86	0.92	0.99	0.88	0.94	0.99	0.85	0.93	0.99

5 Real data application

In this section, we show an application of our method to the analysis of event-related potentials (ERP), which represent electroencephalogram (EEG) recorded in response to stimuli. Primary statistical analyses of an ERP waveform focus on estimating the amplitude (microvolts) and latency (milliseconds) of specific peaks and dips, also called ERP components, as these have been shown to be associated with human sensory and cognitive functions (Luck, 2005). Although ERPs have been extensively used in psychology and the cognitive science community, research in statistical modeling for latency estimation with uncertainty quantification is not mature yet and still under development. Here we show how our methodology can be applied to derive posterior distributions of ERP component latencies, an important information when making scientific discoveries based on ERP data (Yu et al., 2023).

We use ERP data publicly available at `http://dsenturk.bol.ucla.edu/supplements.html`. The dataset consists of ERP signals of a single subject with autism spectrum disorder (ASD), evaluated at one electrode, one condition, and 72 trials, each having 250 time points. Figure 4 shows the time series of all 72 trials and the grand average time course averaged over all 72 trials. Two ERP components, N1, typically within the window [100, 250] msec, and P3, typically within the window [190, 350] msec, are the main interest of the study. We therefore restrict our analysis to the time window [100, 350] msec.

Since EEG signals are typically noisy, traditionally neuroscientists average signals across

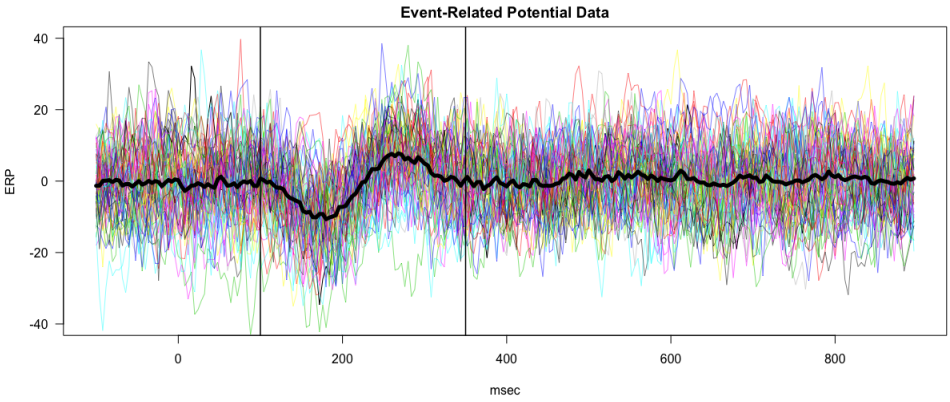


Figure 4: ERP data: 72 time series each corresponding to one single trial and the grand average time course averaged over all 72 trials. The time epoch between the two vertical lines defines the search window for components N1 and P3.

trials to obtain an overall or grand average ERP waveform, which they then visually inspect to determine the amplitude size and latency location of the ERP components. Following such practice, we first applied our method to the grand average time course. As in the simulations, empirical Bayes estimates of the parameters other than t were obtained by maximizing the marginal likelihood, as $(\sigma, \tau, h) = (0.642, 6.385, 0.053)$, with a uniform prior on t. Curve fitting and posterior distribution of the latency are shown in Figure 5. We note that the observations, i.e. the grand average over all trials, and the fitted curve appear to have a similar smoothness level as in the preceding simulation section; for example, compare Figure 5 (top row) with Figure 1. In addition to generating a smooth fitted ERP curve along with 95% credible intervals for amplitude estimation, our model-based approach provides a full posterior distribution of latency locations for ERP components. The 95% credible intervals for the N1 and P3 latencies are [174.58, 178.32] msec and [266.58, 270.93] msec, respectively.

We also investigated the robustness of the results to the smoothness of the ERP waveform. Indeed, since our model explicitly accounts for errors in the data, some of the excessive averaging, which is routinely done to obtain smooth curves, can be avoided. When fewer trials are averaged, we expect posterior distributions with larger variation. As an example, the left panel of Figure 6 shows the estimated waveform and the posterior density of latency when only the first 2 trials are averaged. The N1 and P3 latencies are still identified, though, as expected, with larger uncertainty. In particular, the 95% credible intervals for N1 and P3

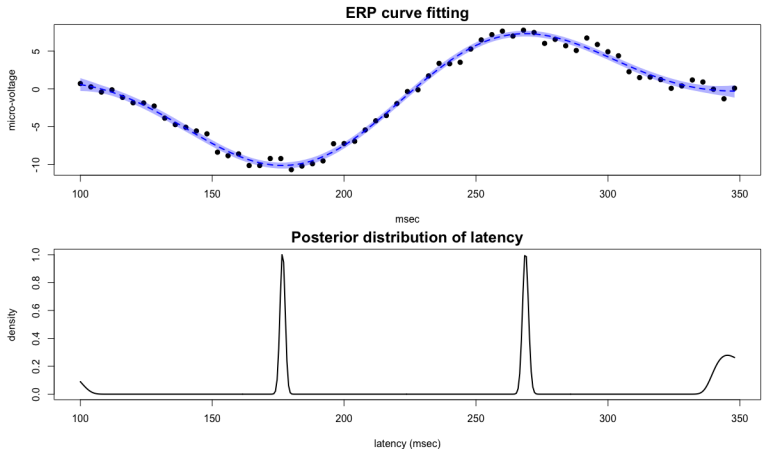


Figure 5: ERP data: Curve fitting and posterior density of the latency of the grand ERP waveform averaged over 72 trials.

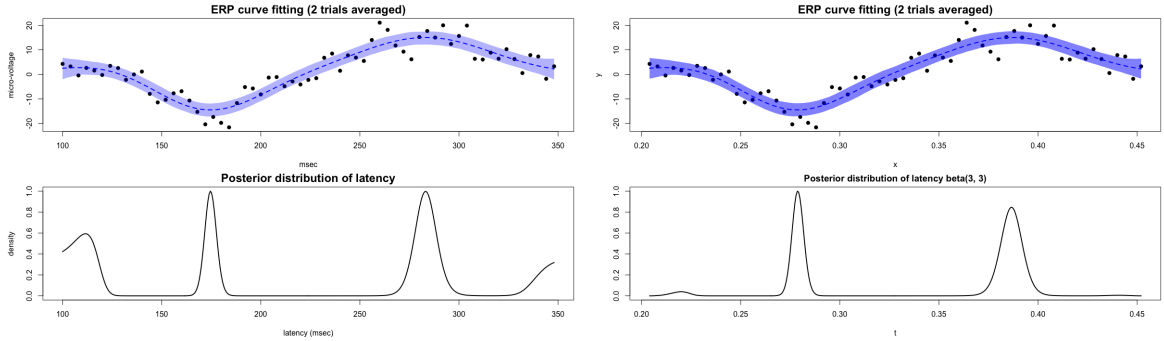


Figure 6: ERP data: Curve fitting and posterior density of the latency of the ERP waveform averaged over the first two trials, with a uniform prior (left panel) and a Beta(3,3) prior (right panel) on the latency.

are [168.99, 180.80] msec and [271.55, 295.17] msec, respectively. Furthermore, as an effect of the smaller level of averaging, some smaller modes at the extremes of the interval are now more pronounced. These are spurious effects and can be avoided by utilizing a prior distribution on the latency that discourages local extrema at the endpoints of the interval. For example, the right panel of Figure 6 shows the inference using a Beta(3, 3) prior.

6 Discussion

In this article, we have studied an encompasssing semiparametric Bayesian approach for identifying multiple local extrema of an unknown function. We have shown that the pos-

terior distribution is connected to unconstrained nonparametric regression in a closed-form characterization, enabling fast computation. We have established local asymptotic normality properties and convergence to Gaussian mixtures for this unconventional strategy, indicating multi-modality of the posterior distribution and substantiating the use of the highest posterior density region for posterior exploration. Our simulations have suggested superior performance of this encompassing semiparametric method relative to existing methods.

Although we have focused on Gaussian processes with stationary covariance functions whose eigenvalues decay at certain rates as special examples, the developed framework of this article, which bases inference on the multi-modal posterior distribution of t with justified asymptotic properties, can be extended to other nonparametric priors, including Gaussian processes with other covariance functions and random series priors. Similar to the flexibility encoded in covariance kernels of Gaussian processes, the rich menu for basis functions in random series priors allows flexible shapes of the underlying functions; for example, wavelets might be better suited for spiky functions in certain applications such as mass spectrometry (Liu et al., 2020), and B-splines for locally supported functions (Wang et al., 2023). In these generalizations, one needs to verify the conditions in Theorem 2 for the adopted nonparametric prior, with technical challenges including deriving the approximate properties of relevant estimates as in Lemma 1 and Assumption C, and selecting hyperparameters such as the number of basis functions in the context of local extrema detection.

Throughout the paper we have focused on a one-dimensional sample space. It may be argued that the encompassing strategy studied in this article generalizes to d-dimensional compact sample spaces for any $d \geq 1$ by using GP prior counterparts supported on d-dimensional $\mathcal{X}$. However, the main challenges in multivariate settings include the need to theoretically study multivariate posterior distributions with multi-modality, and develop computationally efficient algorithms for posterior exploration.

There are several other interesting future directions to pursue. Firstly, Assumption A3 can be relaxed to allow local extrema based on high-order derivative tests, and we envision the developed arguments in this article are largely applicable with the LAN expansion extended

to its higher-order counterpart. Secondly, one may study the encompassing strategy with potentially different posterior exploration methods when there are many or even a diverging number of local extrema, and compare its performance with alternative approaches. Finally, one substantial challenge the proposed method overcomes is the multiplicity of local extrema with unknown dimensions and locations. With given M, which is a different setting, further efficiency gain might be possible by incorporating this knowledge into the method. In this case, it is also interesting to study the optimal rate for estimating the M-dimensional local extrema, and compare the proposed estimator with the optimal rate.

References

Abraham, C. and Khadraoui, K. (2015). Bayesian regression with B-splines under combinations of shape constraints and smoothness properties. *Statistica Neerlandica*, 69:150–170.

Carreira-Perpinán, M. A. and Williams, C. K. (2003). On the number of modes of a Gaussian mixture. In *International Conference on Scale-Space Theories in Computer Vision*, pages 625–640. Springer.

Castillo, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields*, 152(1-2):53–99.

Castillo, I. and Nickl, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *The Annals of Statistics*, 42(5):1941–1969.

Castillo, I. and Rousseau, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics*, 43(6):2353–2383.

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. W. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43:1986–2018.

Cheng, D. and Schwartzman, A. (2017). Multiple testing of local maxima for detection of peaks in random fields. *The Annals of Statistics*, 45(2):529–556.

Conway, J. (1994). *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer New York.

Cucker, F. and Zhou, D.-X. (2007). *Learning Theory: An Approximation Theory Viewpoint*, volume 24. Cambridge University Press.

Dasgupta, S., Pati, D., Jermyn, I. H., and Srivastava, A. (2021). Modality-constrained density estimation via deformable templates. *Technometrics*, 63(4):536–547.

Davies, P., Kovac, A., et al. (2001). Local extremes, runs, strings and multiresolution. *The Annals of Statistics*, 29(1):1–65.

Devroye, L., Mehrabian, A., and Reddad, T. (2018). The total variation distance between high-dimensional Gaussians with the same mean. *arXiv preprint arXiv:1810.08693*.

Egner, A., Geisler, C., Von Middendorff, C., Bock, H., Wenzel, D., Medda, R., Andresen, M., Stiel, A. C., Jakobs, S., Eggeling, C., Schönle, A., and Hell, S. W. (2007). Fluorescence nanoscopy in whole cells by asynchronous localization of photoswitching emitters. *Biophysical journal*, 93(9):3285–3290.

Ferreira, J. C. and Menegatto, V. A. (2012). Reproducing properties of differentiable Mercer-like kernels. *Mathematische Nachrichten*, 285(8-9):959–973.

Geisler, C., Schönle, A., Von Middendorff, C., Bock, H., Eggeling, C., Egner, A., and Hell, S. (2007). Resolution of $\lambda/10$ in fluorescence microscopy using fast single molecule photoswitching. *Applied Physics A*, 88(2):223–226.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, Cambridge.

Giné, E. and Nickl, R. (2021). *Mathematical Foundations of Infinite-dimensional Statistical Models*. Cambridge university press.

Holmes, C. C. and Heard, N. A. (2003). Generalized monotonic regression using random change points. *Statistics in Medicine*, 22:623–638.

Kim, Y. (2006). The Bernstein–von Mises theorem for the proportional hazard model. *The Annals of Statistics*, 34(4):1678–1700.

Kim, Y. and Lee, J. (2004). A Bernstein-von Mises theorem in the nonparametric right-censoring model. *The Annals of Statistics*, 32(4):1492–1512.

Kleijn, B. J. K. and van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Kovac, A. (2007). Smooth functions and local extreme values. *Computational Statistics and Data Analysis*, 51(10):5155–5171.

Liu, Y., Li, M., and Morris, J. S. (2020). Function-on-scalar quantile regression with application to mass spectrometry proteomics data. *Annals of Applied Statistics*, 14(2):521–541.

Liu, Z. and Li, M. (2022). Optimal plug-in Gaussian processes for modelling derivatives. *arXiv preprint arXiv:2210.11626*.

Liu, Z. and Li, M. (2023). On the estimation of derivatives using plug-in kernel ridge regression estimators. *Journal of Machine Learning Research*. In press.

Luck, S. J. (2005). *An Introduction to the Event-Related Potential Technique*. The MIT Press.

Meyer, M. C. (2008). Inference using shape-restricted regression splines. *Annals of Applied Statistics*, 2:1013–1033.

Neelon, B. and Dunson, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics*, 60:398–406.

Pati, D. and Bhattacharya, A. (2015). Adaptive Bayesian inference in the Gaussian sequence model using exponential-variance priors. *Statistics & Probability Letters*, 103:100–104.

Raghuraman, M., Winzeler, E. A., Collingwood, D., Hunt, S., Wodicka, L., Conway, A., Lockhart, D. J., Davis, R. W., Brewer, B. J., and Fangman, W. L. (2001). Replication dynamics of the yeast genome. *Science*, 294(5540):115–121.

Ramsay, J. O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):365–375.

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

Schwartzman, A., Gavrilov, Y., and Adler, R. J. (2011). Multiple testing of local maxima for detection of peaks in 1D. *The Annals of Statistics*, 39(6):3290–3319.

Shively, T. S., Sager, T. W., and Walker, S. G. (2009). A Bayesian approach to nonparametric monotone function estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):159–175.

Shively, T. S., Walker, S. G., and Damien, P. (2011). Nonparametric function estimation subject to monotonicity, convexity and other shape constraints. *Journal of Econometrics*, 161:166–181.

Song, P., Gao, X., Liu, R., and Le, W. (2006). Nonparametric inference for local extrema with application to oligonucleotide microarray data in yeast genome. *Biometrics*, 62(2):545–554.

van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM.

Wang, Z., Magnotti, J., Beauchamp, M. S., and Li, M. (2023). Functional group bridge for simultaneous regression and support estimation. *Biometrics*, 79(2):1226–1238.

Wheeler, M. W., Dunson, D. B., and Herring, A. H. (2017). Bayesian local extremum splines. *Biometrika*, 104(4):939–952.

Yoo, W. W. and Ghosal, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *The Annals of Statistics*, 44(3):1069–1102.

Yu, C.-H., Li, M., Noe, C., Fischer-Baum, S., and Vannucci, M. (2023). Bayesian inference for stationary points in Gaussian process regression models for event-related potentials analysis. *Biometrics*, 79(2):629–641.

Supplementary material for "Semiparametric Bayesian inference for local extrema of functions in the presence of noise"

In this supplementary material, we present proofs of all results in the main paper, additional technical lemmas, and additional numerical experiments.

A Proofs

A.1 Proof of Proposition 1

By Bayes' theorem, it suffices to show that the likelihood takes the form of (5). Recall that $\boldsymbol{y}|X,t \sim N(0,\Sigma_t)$, where $\Sigma_t = \sigma^2(n\lambda)^{-1}(A+B)$ with $A = K(X,X) + n\lambda \boldsymbol{I}_n$ and $B = -K_{01}(X,t)K_{11}^{-1}(t,t)K_{10}(t,X) = -\boldsymbol{a}\boldsymbol{a}^T$ by letting $\boldsymbol{a} = K_{01}(X,t)K_{11}^{-1/2}(t,t)$. Note that the condition $\widehat{\sigma}^2_{f'}(t) > 0$ for any t ensures $K_{11}(t,t) > 0$ in view of (4).

In view of the Sherman–Morrison formula, we have $\det(A+B) = (1-\boldsymbol{a}^TA^{-1}\boldsymbol{a})\det(A)$ and $(A+B)^{-1} = A^{-1} + \frac{A^{-1}\boldsymbol{a}\boldsymbol{a}^TA^{-1}}{1-\boldsymbol{a}^TA^{-1}\boldsymbol{a}}$, assuming $1-\boldsymbol{a}^TA^{-1}\boldsymbol{a} \neq 0$. Substituting these two identities into the multivariate normal density $\ell(t)$ yields

$$
\begin{aligned}
\ell(t) &= \{2\pi\sigma^2(n\lambda)^{-1}\}^{-n/2}\det(A+B)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2(n\lambda)^{-1}}\boldsymbol{y}^T(A+B)^{-1}\boldsymbol{y}\right\} \\
&= \{2\pi\sigma^2(n\lambda)^{-1}\}^{-n/2}(\det(A))^{-1/2}\exp\left\{-\frac{\boldsymbol{y}^TA^{-1}\boldsymbol{y}}{2\sigma^2(n\lambda)^{-1}}\right\} \\
&\quad \cdot (1-\boldsymbol{a}^TA^{-1}\boldsymbol{a})^{-1/2}\exp\left\{-\frac{\boldsymbol{y}^TA^{-1}\boldsymbol{a}\boldsymbol{a}^TA^{-1}\boldsymbol{y}}{2\sigma^2(n\lambda)^{-1}(1-\boldsymbol{a}^TA^{-1}\boldsymbol{a})}\right\} \\
&= C\{\sigma^2(n\lambda)^{-1}(1-\boldsymbol{a}^TA^{-1}\boldsymbol{a})\}^{-1/2}\exp\left\{-\frac{\boldsymbol{y}^TA^{-1}\boldsymbol{a}K_{11}(t,t)\boldsymbol{a}^TA^{-1}\boldsymbol{y}}{2\sigma^2(n\lambda)^{-1}K_{11}(t,t)(1-\boldsymbol{a}^TA^{-1}\boldsymbol{a})}\right\},
\end{aligned}
$$

where

$$
C = \{2\pi\sigma^2(n\lambda)^{-1}\}^{-n/2}(\det(A))^{-1/2}\exp\left\{-\frac{\boldsymbol{y}^TA^{-1}\boldsymbol{y}}{2\sigma^2(n\lambda)^{-1}}\right\}\cdot\{\sigma^2(n\lambda)^{-1}\}^{1/2}
$$

does not depend on t. The proof is completed by noticing that $\widehat{\mu}_{f'}(t) = K_{11}^{1/2}(t,t)\boldsymbol{a}^TA^{-1}\boldsymbol{y}$ and $\widehat{\sigma}^2_{f'}(t) = \sigma^2(n\lambda)^{-1}K_{11}(t,t)(1-\boldsymbol{a}^TA^{-1}\boldsymbol{a})$. This completes the proof.

A.2 Proof of Lemma 1

A one-dimensional version of Theorem 4 in Liu and Li (2023) shows that

$$\|\widehat{\mu}_{f'}^{(k)} - f_{\lambda}^{(k+1)}\|_{\infty} \leq \frac{\sqrt{\kappa\kappa_{k+1,k+1}}\|f_0\|_{\infty}\sqrt{\log(9/\delta)}}{\sqrt{n}\lambda}\left(10 + \frac{4\kappa\sqrt{\log(9/\delta)}}{3\sqrt{n\lambda}}\right) + \frac{C_2\sqrt{\kappa\kappa_{k+1,k+1}}\sigma\sqrt{\log(3/\delta)}}{\sqrt{n}\lambda}, \quad 0 \leq k \leq 3, \tag{18}$$

For any bounded $f \in L_{p_X}^2(\mathcal{X})$, we define a bias of estimators of f by matrix and integral operation as

$$\begin{aligned} E(K, X, f) &= (L_{K,X} + \lambda I)^{-1} L_{K,X} f - (L_K + \lambda I)^{-1} L_K f \\ &= K(\cdot, X)[K(X, X) + n\lambda \boldsymbol{I}_n]^{-1} f(X) - (L_K + \lambda I)^{-1} L_K f, \end{aligned}$$

which belongs to $\mathbb{H}$. Consider any $j, l \geq 1$ and $j + l \leq 5$, taking $f = K_{0l,x}$ yields

$$\partial^j E(K, X, K_{0l,x}) = K_{j0}(\cdot, X)[K(X, X) + n\lambda \boldsymbol{I}_n]^{-1} K_{0l}(X, x) - \partial^j (L_K + \lambda I)^{-1} L_K K_{0l,x}.$$

Thus,

$$\begin{aligned} \partial^j E(K, X, K_{0l,x})(x) &= K_{j0}(x, X)[K(X, X) + n\lambda \boldsymbol{I}_n]^{-1} K_{0l}(X, x) - \partial^j (L_K + \lambda I)^{-1} L_K K_{0l,x}(x) \\ &= K_{j0}(x, X)[K(X, X) + n\lambda \boldsymbol{I}_n]^{-1} K_{0l}(X, x) - (L_K + \lambda I)^{-1} L_K K_{jl,x}(x) \end{aligned}$$

We write $(L_K + \lambda I)^{-1} L_K K_{jl,x}(x) = K_{jl,x}(x) - \lambda (L_K + \lambda I)^{-1} K_{jl,x}(x) = K_{jl}(x, x) - \lambda \varphi_{jl}(x)$. Then, by Theorem 16 in Liu and Li (2023) we have that for any $\delta \in (0, 1)$, with $\mathbb{P}_0$-probability at least $1 - \delta$ it holds

$$\begin{aligned} & |K_{j0}(x, X)[K(X, X) + n\lambda \boldsymbol{I}_n]^{-1} K_{0l}(X, x) - K_{jl}(x, x) + \lambda \varphi_{jl}(x)| \\ \leq & \|\partial^j E(K, X, K_{0l,x})\|_{\infty} \\ \leq & \frac{\sqrt{\kappa\kappa_{jj}}\|K_{0l,x}\|_{\infty}\sqrt{\log(3/\delta)}}{\sqrt{n}\lambda}\left(10 + \frac{4\sqrt{\kappa}\sqrt{\log(3/\delta)}}{3\sqrt{n\lambda}}\right) \end{aligned}$$

$$= \frac{\sqrt{\kappa\kappa_{jj}}\kappa_{0l}\sqrt{\log(3/\delta)}}{\sqrt{n}\lambda}\left(10 + \frac{4\sqrt{\kappa}\sqrt{\log(3/\delta)}}{3\sqrt{n\lambda}}\right).$$

In view of (7), $\widehat{\sigma}^{2(k)}_{f'}(x)$ is a linear combination of quadratic forms

$$K_{jl}(x,x) - K_{j0}(\cdot, X)[K(X,X) + n\lambda \boldsymbol{I}_n]^{-1}K_{0l}(X,x).$$

Therefore, for any $0 \le k \le 3$, we have

$$\begin{aligned} &|\widehat{\sigma}^{2(k)}_{f'}(x) - \sigma^2 n^{-1}\sum_{i=0}^{k}\binom{k}{i}\varphi_{i+1,k+1-i}(x)| \qquad (19)\\ &\quad \le \sum_{i=0}^{k}\binom{k}{i}\left[\frac{\sqrt{\kappa\kappa_{i+1,i+1}}\kappa_{0,k+1-i}\sigma^2\sqrt{\log(3/\delta)}}{n\sqrt{n}\lambda^2}\left(10 + \frac{4\sqrt{\kappa}\sqrt{\log(3/\delta)}}{3\sqrt{n\lambda}}\right)\right]. \end{aligned}$$

The above (18) and (19) can hold simultaneously with $\mathbb{P}_0$-probability $1-8\delta$. Let $8\delta = n^{-10}$, and A_n be the corresponding event. We immediately have $\mathbb{P}_0(A_n) \ge 1 - n^{-10}$ with $\log(3/\delta) \le 10\log n + 4$ and $\log(9/\delta) \le 10\log n + 5$ in the upper bound. This completes the proof.

A.3 Proof of Lemma 2

First we prove that for any local extremum t_m of f_0, there exists a local extremum $t_{\lambda,m}$ of f_λ such that $t_{\lambda,m} \to t_m$ as $\lambda \to 0$. There exists $\delta > 0$ such that for any $0 < \epsilon < \delta$, it holds that $f_0'(t_m - \epsilon) < 0$, $f_0'(t_m + \epsilon) > 0$ and $f_0''(t_m \pm \epsilon) \neq 0$ without loss of generality. By Assumption C, we have

$$|f_\lambda'(t_m - \epsilon) - f_0'(t_m - \epsilon)| \lesssim \lambda^{r_1}.$$

Hence, for sufficiently small λ, it holds $f_\lambda'(t_m - \delta/2) < 0$. Similarly, we have $f_\lambda'(t_m + \delta/2) > 0$. According to the continuity of f_λ, there exists a $t_{\lambda,m} \in (t_m - \delta/2, t_m + \delta/2)$ such that $f_\lambda'(t_{\lambda,m}) = 0$. It can also be shown that $f_\lambda''(t) \neq 0$ for any $t \in (t_m - \delta/2, t_m + \delta/2)$ and sufficiently small λ, which implies $f_\lambda''(t_{\lambda,m}) \neq 0$. Finally, we have $t_{\lambda,m} \to t_m$ as $\delta \to 0$ and $\lambda \to 0$.

Again by Assumption C we can see that

$$|f'_\lambda(t_{\lambda,m}) - f'_0(t_{\lambda,m})| \lesssim \lambda^{r_1}.$$

Since $f'_\lambda(t_{\lambda,m}) = 0$, in view of the mean value theorem, we have

$$|f'_0(t_m) + f''_0(\xi_1)(t_{\lambda,m} - t_m)| \lesssim \lambda^{r_1},$$

where ξ_1 lies between and $t_{\lambda,m}$ and t_m. Since $\xi_1 \to t_m$ and $f''_0(t_m) \neq 0$ by Assumption A, we obtain

$$|t_{\lambda,m} - t_m| \lesssim \lambda^{r_1}.$$

Under A_n, the existence and convergence rate of $\hat{t}_m$ can be shown similarly by applying (8). This completes the proof.

A.4 Proof of Theorem 1

The proof is based on the high probability event A_n defined in Lemma 1. Conditions of Theorem 1 imply $n^{\frac{1}{2}-2\beta}\varphi_{11,m} = o(1)$, yielding $\widehat{\sigma}^2_{f'}(t_{\lambda,m}) > 0$ in view of (11). Invoking the likelihood function $\ell(t)$ in (5), which holds at $t_{\lambda,m}$ and in its small neighborhood, we have

$$\Lambda(t, \Delta t) = \log \frac{\ell(t+\Delta t)}{\ell(t)} = \frac{\ell'(t)}{\ell(t)}\Delta t + \frac{\ell''(t)\ell(t) - \ell'(t)^2}{2\ell(t)^2}(\Delta t)^2 + R_3(\xi)(\Delta t)^3,$$

where $R_3(\xi) = \{2\ell'(t)^3 + \ell'''(\xi)\ell(\xi)^2 - 3\ell'(\xi)\ell''(\xi)\ell(\xi)\}/\{6\ell(\xi)^3\}$ and ξ is bewteen t and $t+\Delta t$. Thus,

$$\begin{aligned}\Lambda(t_{\lambda,m}, \frac{u}{n^\beta}) = &\left[-\frac{\widehat{\sigma}^{2'}_{f'}(t_{\lambda,m})}{2\widehat{\sigma}^2_{f'}(t_{\lambda,m})} - \frac{\widehat{\mu}_{f'}(t_{\lambda,m})\widehat{\mu}'_{f'}(t_{\lambda,m})}{\widehat{\sigma}^2_{f'}(t_{\lambda,m})} + \frac{\widehat{\mu}_{f'}(t_{\lambda,m})^2\widehat{\sigma}^{2'}_{f'}(t_{\lambda,m})}{2\widehat{\sigma}^2_{f'}(t_{\lambda,m})^2}\right]\frac{u}{n^\beta} \\ &+ \frac{1}{2}\left[\frac{\widehat{\sigma}^{2'}_{f'}(t_{\lambda,m})^2}{2\widehat{\sigma}^2_{f'}(t_{\lambda,m})^2} - \frac{\widehat{\sigma}^{2''}_{f'}(t_{\lambda,m})}{2\widehat{\sigma}^2_{f'}(t_{\lambda,m})} + \frac{2\widehat{\mu}_{f'}(t_{\lambda,m})\widehat{\mu}'_{f'}(t_{\lambda,m})\widehat{\sigma}^{2'}_{f'}(t_{\lambda,m})}{\widehat{\sigma}^2_{f'}(t_{\lambda,m})^2}\right. \\ &\left. - \frac{\widehat{\mu}'_{f'}(t_{\lambda,m})^2 + \widehat{\mu}_{f'}(t_{\lambda,m})\widehat{\mu}''_{f'}(t_{\lambda,m})}{\widehat{\sigma}^2_{f'}(t_{\lambda,m})}\right.\end{aligned}$$

$$-\frac{1}{2}\widehat{\mu}_{f'}(t_{\lambda,m})^2\left(\frac{2\widehat{\sigma}^{2'}_{f'}(t_{\lambda,m})^2}{\widehat{\sigma}^2_{f'}(t_{\lambda,m})^3}-\frac{\widehat{\sigma}^{2''}_{f'}(t_{\lambda,m})}{\widehat{\sigma}^2_{f'}(t_{\lambda,m})^2}\right)\Bigg]\frac{u^2}{n^{2\beta}}$$
$$+\frac{1}{6}R_3(\xi)\frac{u^3}{n^{3\beta}}.$$

Based on the rates given by (9), (10), (11) and (12), we obtain

$$|\widehat{\mu}_{f'}(t_{\lambda,m})|\lesssim n^{-\beta}(\log n)^{-a},\quad |\widehat{\mu}^{(k)}_{f'}(t_{\lambda,m})|\lesssim 1,$$
$$|\widehat{\sigma}^2_{f'}(t_{\lambda,m})|\lesssim n^{-1}\varphi_{11,m},\quad |\widehat{\sigma}^{2(k)}_{f'}(t_{\lambda,m})|\lesssim n^{-\frac{1}{2}-\beta}(\log n)^{-\frac{1}{2}-a},$$

for $1\le k\le 3$. Further calculation gives $R_3(\xi)\lesssim\frac{\widehat{\sigma}^{2''}_{f'}(\xi)}{\widehat{\sigma}^2_{f'}(\xi)^2}=O\left(n^{\frac{3}{2}-\beta}(\log n)^{-\frac{1}{2}-a}\varphi^{-2}_{11,m}\right)$. Substituting these into the above $\Lambda(t_{\lambda,m},\frac{u}{n^\beta})$ yields

$$\begin{aligned}\Lambda(t_{\lambda,m},\frac{u}{n^\beta})&=-\frac{\widehat{\mu}_{f'}(t_{\lambda,m})\widehat{\mu}'_{f'}(t_{\lambda,m})}{n^\beta\widehat{\sigma}^2_{f'}(t_{\lambda,m})}u-\frac{\widehat{\mu}'_{f'}(t_{\lambda,m})^2}{2n^{2\beta}\widehat{\sigma}^2_{f'}(t_{\lambda,m})}u^2+o(n^{\frac{3}{2}-4\beta}\varphi^{-2}_{11,m})\\&=-\frac{n^{1-\beta}\widehat{\mu}_{f'}(t_{\lambda,m})\widehat{\mu}'_{f'}(t_{\lambda,m})}{n\widehat{\sigma}^2_{f'}(t_{\lambda,m})}u-\frac{n^{1-2\beta}\widehat{\mu}'_{f'}(t_{\lambda,m})^2}{2n\widehat{\sigma}^2_{f'}(t_{\lambda,m})}u^2+o(n^{\frac{3}{2}-4\beta}\varphi^{-2}_{11,m})\\&=n^{1-2\beta}\varphi^{-1}_{11,m}\left\{-\frac{\widehat{\mu}'_{f'}(t_{\lambda,m})^2u^2}{2n\varphi^{-1}_{11,m}\widehat{\sigma}^2_{f'}(t_{\lambda,m})}-\frac{n^\beta\widehat{\mu}_{f'}(t_{\lambda,m})^2u}{n\varphi^{-1}_{11,m}\widehat{\sigma}^2_{f'}(t_{\lambda,m})}\right\}+o(1),\end{aligned}$$

when $n^{\frac{3}{2}-4\beta}\varphi^{-2}_{11,m}=O(1)$.

Then we study the convergence of $\mu_{n,m}$ and $\sigma^2_{n,m}$. According to (9) and (10), we have

$$|n^\beta\widehat{\mu}_{f'}(t_{\lambda,m})|\lesssim(\log n)^{-a},$$

$$|\widehat{\mu}'_{f'}(t_{\lambda,m})-f''_\lambda(t_{\lambda,m})|\lesssim n^{-\beta}(\log n)^{-a}.$$

In view of Lemma 2, Assumption A2, and Assumption C, we obtain that $\widehat{\mu}'_{f'}(t_{\lambda,m})$ converges to $f''_0(t_m)$, and thus is bounded away from zero and infinity for sufficiently large n. Therefore,

$$|\mu_{n,m}|=\left|\frac{n^\beta\widehat{\mu}_{f'}(t_{\lambda,m})}{\widehat{\mu}'_{f'}(t_{\lambda,m})}\right|\asymp|n^\beta\widehat{\mu}_{f'}(t_{\lambda,m})|\lesssim(\log n)^{-a}.$$

From (11) we have

$$\left| n\varphi_{11,m}^{-1}\widehat{\sigma}_{f'}^2(t_{\lambda,m}) - \sigma^2 \right| \lesssim n^{\frac{1}{2}-2\beta}(\log n)^{-1-a}\varphi_{11,m}^{-1}.$$

Therefore,

$$\begin{aligned}
&\left| \sigma_{n,m}^2 - \frac{\sigma^2}{f_\lambda''(t_{\lambda,m})^2} \right| = \left| \frac{n\varphi_{11,m}^{-1}\widehat{\sigma}_{f'}^2(t_{\lambda,m})}{\widehat{\mu}_{f'}'(t_{\lambda,m})^2} - \frac{\sigma^2}{f_\lambda''(t_{\lambda,m})^2} \right| \\
\asymp &\left| f_\lambda''(t_{\lambda,m})^2 n\varphi_{11,m}^{-1}\widehat{\sigma}_{f'}^2(t_{\lambda,m}) - \widehat{\mu}_{f'}'(t_{\lambda,m})^2\sigma^2 \right| \\
\lesssim &\left| f_\lambda''(t_{\lambda,m})^2 \left[n\varphi_{11,m}^{-1}\widehat{\sigma}_{f'}^2(t_{\lambda,m}) - \sigma^2 \right] \right| + \left| \left[\widehat{\mu}_{f'}'(t_{\lambda,m})^2 - f_\lambda''(t_{\lambda,m})^2 \right] \sigma^2 \right| \\
\lesssim &\, n^{\frac{1}{2}-2\beta}(\log n)^{-1-a}\varphi_{11,m}^{-1} + n^{-\beta}(\log n)^{-a} \\
\lesssim &\, n^{\frac{1}{2}-2\beta}(\log n)^{-\frac{3}{2}}\varphi_{11,m}^{-1}.
\end{aligned} \tag{20}$$

On the other hand,

$$\left| \frac{\sigma^2}{f_\lambda''(t_{\lambda,m})^2} - \frac{\sigma^2}{f_0''(t_m)^2} \right| \asymp |f_0''(t_m)^2 - f_\lambda''(t_{\lambda,m})^2| \lesssim \left[f_0''(t_m) - f_\lambda''(t_{\lambda,m}) \right]. \tag{21}$$

Since $K \in C^8(\mathcal{X}, \mathcal{X})$, we have $f_\lambda \in C^4(\mathcal{X})$. Then the mean value theorem gives

$$f_0''(t_m) - f_\lambda''(t_{\lambda,m}) = f_0''(t_m) - f_\lambda''(t_m) - f_\lambda'''(\xi)(t_{\lambda,m} - t_m).$$

By Assumption C and Lemma 2 we have

$$|f_0''(t_m) - f_\lambda''(t_{\lambda,m})| \lesssim \lambda^{r_2} + \lambda^{r_1} \leq 2\lambda^r. \tag{22}$$

Combining (20), (21) and (22), we obtain

$$\left| \sigma_{n,m}^2 - \frac{\sigma^2}{f_0''(t_m)^2} \right| \lesssim \lambda^r.$$

This completes the proof.

A.5 Proof of Theorem 2

We first present a technical lemma and leave its proof to Section A.10.

Lemma 4. *Suppose Assumption B1 holds and let* $\lambda = n^{-\frac{1}{2}+\beta}(\log n)^{\frac{1}{2}+a}$ *for some* $\frac{1}{4} < \beta < \frac{1}{2}$ *and* $a > 0$. *Under event* A_n, *there exists* $C > 0$ *such that*

$$\left| n^{-\frac{1}{2}} \varphi_{11,m}^{\frac{1}{2}} \ell(t_{\lambda,m}) \Big/ \exp\left(-n^{1-2\beta} \varphi_{11,m}^{-1} \frac{\mu_{n,m}^2}{2\sigma_{n,m}^2} \right) - \frac{C}{|f_0''(t_m)|\sigma_m^*} \right| \lesssim n^{\frac{1}{2}-2\beta} (\log n)^{-1-a}.$$

For any $x \geq 0$, define the error function as

$$\mathrm{Erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

By changing of variable, we have

$$\int_{-A}^{B} a \exp\left(-a^2 \frac{(u+b)^2}{c} \right) du = \sqrt{\frac{\pi c}{2}} \left[\mathrm{Erf}\left(\frac{a(A-b)}{\sqrt{c}} \right) + \mathrm{Erf}\left(\frac{a(B+b)}{\sqrt{c}} \right) \right], \tag{23}$$

where $a, c, A, B > 0$ and $b \in \mathbb{R}$.

A.5.1 Proof of (i)

The proof will follow three steps.

Step 1: According to Theorem 1.3 in Devroye et al. (2018) and Lemma 2, we have that for any $z \in \mathbb{R}$,

$$\begin{aligned}
&\left| \sum_{m=1}^{M} \pi_m \Phi(z \mid t_{\lambda,m}, n^{-1}\varphi_{11,m}\sigma_m^{*2}) - \sum_{m=1}^{M} \pi_m \Phi(z \mid t_m, n^{-1}\varphi_{11,m}\sigma_m^{*2}) \right| \\
&\leq d_{TV}\left(\sum_{m=1}^{M} \pi_m \phi(\cdot \mid t_{\lambda,m}, n^{-1}\varphi_{11,m}\sigma_m^{*2}),\ \sum_{m=1}^{M} \pi_m \phi(\cdot \mid t_m, n^{-1}\varphi_{11,m}\sigma_m^{*2}) \right) \\
&\lesssim \sum_{m=1}^{M} \pi_m |t_{\lambda,m} - t_m| \lesssim \lambda^{r_1} = o(1),
\end{aligned}$$

where d_{TV} is the total variation distance between two distributions. Thus, we only need to

show

$$\left|\Pi_n(t \le z \mid X, \boldsymbol{y}) - \sum_{m=1}^{M} \pi_m \Phi(z \mid t_{\lambda,m}, n^{-1}\varphi_{11,m}\sigma_m^{*2})\right| \to 0$$

for any $z \in \mathbb{R}$ in $\mathbb{P}_0$-probability.

Step 2: We work under the high probability event A_n henceforth in this proof, that is, all convergence rates and bounding integrals only hold under A_n.

Define a sequence of functions

$$\tilde{h}_n(t) = \sum_{m=1}^{M} \ell(t_{\lambda,m})\tilde{\phi}_{n,m}(t)\pi(t_m), \tag{24}$$

where

$$\tilde{\phi}_{n,m}(t) = \exp\left(-\frac{(t - t_{\lambda,m})^2}{2n^{-1}\varphi_{11,m}\sigma_m^{*2}} + n^{1-2\beta}\varphi_{11,m}^{-1}\frac{\mu_{n,m}^2}{2\sigma_{n,m}^2}\right).$$

In this step, we will prove that

$$\left|\int_{-\infty}^{z} \ell(t)\pi(t)dt - \int_{-\infty}^{z} \tilde{h}_n(t)dt\right| \to 0 \tag{25}$$

for any $z \in \mathbb{R}$. That is, $\tilde{h}_n(t)$ approximates the unnormalized limit density where each mixture component is properly rescaled. In line with the LAN condition (14), we expand $\tilde{h}_n(t)$ at $t = t_{\lambda,m} + u/n^{\beta}$ for $m = 1, \ldots, M$, transforming $\tilde{\phi}_{n,m}(t)$ to

$$\nu_{n,m}(u) = \exp\left(n^{1-2\beta}\varphi_{11,m}^{-1}\left(-\frac{u^2}{2\sigma_m^{*2}} + \frac{\mu_{n,m}^2}{2\sigma_{n,m}^2}\right)\right).$$

We consider three cases for z: (1) $z \le 0$, (2) $0 < z \le 1$, and (3) $z > 1$.

Case (1) ($z \le 0$). Since $\ell(t) = 0$ in $\mathbb{R}\backslash[0,1]$, the left hand side of (25) becomes

$$\begin{aligned}\int_{-\infty}^{z} \tilde{h}_n(t)dt &\le \sum_{m=1}^{M} \int_{-\infty}^{z} \ell(t_{\lambda,m})\tilde{\phi}_{n,m}(t)\pi(t_m)dt \\ &= \sum_{m=1}^{M} \int_{L_{n,m}} n^{-\beta}\ell(t_{\lambda,m})\nu_{n,m}(u)\pi(t_m)du\end{aligned}$$

where we let $t = t_{\lambda,m} + u/n^{\beta}$ and $L_{n,m} = (-\infty, (z - t_{\lambda,m})n^{\beta}]$. By Lemma 4 and (23) , we have

$$\begin{aligned}\int_{L_{n,m}} n^{-\beta}\ell(t_{\lambda,m})\nu_{n,m}(u)\pi(t_m)du &\lesssim \int_{L_{n,m}} n^{\frac{1}{2}-\beta}\varphi_{11,m}^{-\frac{1}{2}} \exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{u^2}{2\sigma_m^{*2}}\right) du \\ &= \sqrt{\frac{\pi\sigma_m^{*2}}{2}}\left[1 - \operatorname{Erf}\left(\frac{\sqrt{n}(t_{\lambda,m} - z)}{\sqrt{2\sigma_m^{*2}\varphi_{11,m}}}\right)\right] \\ &\lesssim \exp\left(-\frac{n(t_{\lambda,m} - z)^2}{2\sigma_m^{*2}\varphi_{11,m}}\right),\end{aligned}$$

where we use the well known inequality that $1 - \operatorname{Erf}(x) \le \frac{2}{\sqrt{\pi}} e^{-\frac{x^2}{2}}$. Therefore, there holds that for $z \le 0$,

$$\int_{-\infty}^{z} \tilde{h}_n(t)dt \lesssim M \exp\left(-\frac{n(t_{\lambda,1} - z)^2}{4\sigma_m^{*2}\varphi_{11,m}}\right) \to 0. \tag{26}$$

Case (2) $(0 < z \le 1)$. Now the left hand side of (25) becomes

$$\left|\int_{-\infty}^{z} \ell(t)\pi(t)dt - \int_{-\infty}^{z} \tilde{h}_n(t)dt\right| \le \int_{-\infty}^{0} \tilde{h}_n(t)dt + \left|\int_{0}^{z} \ell(t)\pi(t)dt - \int_{0}^{z} \tilde{h}_n(t)dt\right|.$$

Taking $z = 0$ in (26) gives that $\int_{-\infty}^{0} \tilde{h}_n(t)dt \to 0$. We next bound the second term. We divide $[0, 1]$ into M disjoint intervals (up to overlapping endpoints that do not affect estimates of integrals), each of which centering around $t_{\lambda,m}$:

$$[0,1] = \bigcup_{m=1}^{M} I_{n,m}, \quad I_{n,m} = [t_{\lambda,m} - \xi_{n,m-1}, t_{\lambda,m} + \xi_{n,m}], \quad m = 1, \ldots, M,$$

where $\xi_{n,0} = t_{\lambda,1}$, $\xi_{n,m} = (t_{\lambda,m+1} - t_{\lambda,m})/2$ for $m = 1, \ldots, M-1$, and $\xi_{n,M} = 1 - t_{\lambda,M}$. Suppose $z \in I_{n,m_0}$ for some $1 \le m_0 \le M$ and let $I'_{n,m_0} = [t_{\lambda,m_0} - \xi_{n,m_0-1}, z]$. By the triangle inequality, we have

$$\begin{aligned}\left|\int_{0}^{z} \ell(t)\pi(t)dt - \int_{0}^{z} \tilde{h}_n(t)dt\right| &= \left|\left(\sum_{m=1}^{m_0-1}\int_{I_{n,m}} + \int_{I'_{n,m_0}}\right)\left[\ell(t)\pi(t) - \tilde{h}_n(t)\right]dt\right| \\ &\le \sum_{m=1}^{m_0-1}\left|\int_{I_{n,m}}\left[\ell(t)\pi(t) - \ell(t_{\lambda,m})\tilde{\phi}_{n,m}(t)\pi(t_m)\right]dt\right| \end{aligned} \tag{27}$$

$$+\sum_{m=1}^{m_0-1}\int_{[0,z]\setminus I_{n,m}}\ell(t_{\lambda,m})\tilde{\phi}_{n,m}(t)\pi(t_m)dt \tag{28}$$

$$+\left|\int_{I'_{n,m_0}}\left[\ell(t)\pi(t)-\ell(t_{\lambda,m})\tilde{\phi}_{n,m}(t)\pi(t_m)\right]dt\right| \tag{29}$$

$$+\int_{[0,z]\setminus I'_{n,m_0}}\ell(t_{\lambda,m})\tilde{\phi}_{n,m}(t)\pi(t_m)dt. \tag{30}$$

Again, after changing of variable with $t = t_{\lambda,m} + u/n^{\beta}$, each term in (27) becomes

$$\begin{aligned}
&\left|\int_{I_{n,m}}\left[\ell(t)\pi(t)-\ell(t_{\lambda,m})\tilde{\phi}_{n,m}(t)\pi(t_m)\right]dt\right| \\
&=\left|\int_{J_{n,m}}\left[\ell(t_{\lambda,m}+u/n^{\beta})\pi(t_{\lambda,m}+u/n^{\beta})-\ell(t_{\lambda,m})\nu_{n,m}(u)\pi(t_m)\right]n^{-\beta}du\right| \\
&=\left|\int_{J_{n,m}}n^{-\beta}\ell(t_{\lambda,m})\left[Z_{n,m}(u)\pi(t_{\lambda,m}+u/n^{\beta})-\nu_{n,m}(u)\pi(t_m)\right]du\right|,
\end{aligned}$$

where $Z_{n,m}(u) = \ell(t_{\lambda,m}+u/n^{\beta})/\ell(t_{\lambda,m})$ and $J_{n,m} = [-n^{\beta}\xi_{n,m-1}, n^{\beta}\xi_{n,m}]$. Applying the triangle inequality yields an upper bound of the preceding display:

$$\begin{aligned}
&\left|\int_{J_{n,m}}n^{-\beta}\ell(t_{\lambda,m})Z_{n,m}(u)[\pi(t_{\lambda,m}+u/n^{\beta})-\pi(t_m)]du\right| \\
&\quad+\left|\int_{J_{n,m}}n^{-\beta}\ell(t_{\lambda,m})[Z_{n,m}(u)-\nu_{n,m}(u)]\pi(t_m)du\right| \\
&=I_1+I_2.
\end{aligned}$$

By Lemma 4 and Theorem 1, we have

$$\begin{aligned}
I_1 &\lesssim\left|\int_{J_{n,m}}n^{\frac{1}{2}-\beta}\varphi_{11,m}^{-\frac{1}{2}}\exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{\mu_{n,m}^2}{2\sigma_{n,m}^2}\right)\exp\left(n^{1-2\beta}\varphi_{11,m}^{-1}\left(-\frac{(u+\mu_{n,m})^2}{2\sigma_{n,m}^2}+\frac{\mu_{n,m}^2}{2\sigma_{n,m}^2}\right)\right)\right. \\
&\qquad\left.\cdot[\pi(t_{\lambda,m}+u/n^{\beta})-\pi(t_m)]du\right| \\
&\lesssim\left|\int_{J_{n,m}}n^{\frac{1}{2}-\beta}\varphi_{11,m}^{-\frac{1}{2}}\exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{(u+\mu_{n,m})^2}{2\sigma_{n,m}^2}\right)[\pi(t_{\lambda,m}+u/n^{\beta})-\pi(t_m)]du\right| \\
&\lesssim|t_{\lambda,m}-t_m|\int_{J_{n,m}}n^{\frac{1}{2}-\beta}\varphi_{11,m}^{-\frac{1}{2}}\exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{(u+\mu_{n,m})^2}{2\sigma_{n,m}^2}\right)du \\
&\quad+\left|\int_{J_{n,m}}n^{\frac{1}{2}-2\beta}\varphi_{11,m}^{-\frac{1}{2}}u\exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{(u+\mu_{n,m})^2}{2\sigma_{n,m}^2}\right)du\right|
\end{aligned}$$

$= I_{11} + I_{12}$.

In view of (15), (16), (23) and Lemma 2, it follows that

$$I_{11} \lesssim |t_{\lambda,m} - t_m| \cdot \sqrt{2\pi\sigma_{n,m}^2} \lesssim |t_{\lambda,m} - t_m| \lesssim \lambda^{r_1},$$

and

$$\begin{aligned}
I_{12} &= n^{-\beta} \Bigg| \int_{J'_{n,m}} n^{\frac{1}{2}-\beta} \varphi_{11,m}^{-\frac{1}{2}} u \exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{u^2}{2\sigma_{n,m}^2}\right) du \\
&\qquad - \mu_{n,m} \int_{J'_{n,m}} n^{\frac{1}{2}-\beta} \varphi_{11,m}^{-\frac{1}{2}} \exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{u^2}{2\sigma_{n,m}^2}\right) du \Bigg| \\
&\lesssim n^{-\beta} \Bigg[2\int_0^{n^\beta(\xi_{n,m-1}\vee\xi_{n,m})} n^{\frac{1}{2}-\beta} \varphi_{11,m}^{-\frac{1}{2}} u \exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{u^2}{2\sigma_{n,m}^2}\right) du \\
&\qquad + |\mu_{n,m}| \int_{J_{n,m}} n^{\frac{1}{2}-\beta} \varphi_{11,m}^{-\frac{1}{2}} \exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{(u+\mu_{n,m})^2}{2\sigma_{n,m}^2}\right) du \Bigg] \\
&\lesssim n^{-\beta} \left[\sigma_{n,m}^2 \varphi_{11,m}^{\frac{1}{2}} \left(1 - \exp\left(-\frac{n(\xi_{n,m-1}\vee\xi_{n,m})^2}{2\sigma_{n,m}^2\varphi_{11,m}}\right)\right) n^{-\frac{1}{2}+\beta} + \mu_{n,m}\sqrt{2\pi\sigma_{n,m}^2} \right] \\
&\lesssim \sqrt{\frac{\varphi_{11,m}}{n}} \wedge n^{-\beta},
\end{aligned}$$

where $J'_{n,m} = [-n^\beta \xi_{n,m-1} + \mu_{n,m}, n^\beta \xi_{n,m} + \mu_{n,m}]$. Hence, $I_1 \lesssim \lambda^{r_1} \wedge \sqrt{\frac{\varphi_{11,m}}{n}} \wedge n^{-\beta} \to 0$ under Assumption B2. By Lemma 4 and Theorem 1, we have

$$\begin{aligned}
I_2 &\lesssim \Bigg| \int_{J_{n,m}} n^{\frac{1}{2}-\beta} \varphi_{11,m}^{-\frac{1}{2}} \exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{\mu_{n,m}^2}{2\sigma_{n,m}^2}\right) \Bigg[\exp\left(n^{1-2\beta}\varphi_{11,m}^{-1}\left(-\frac{(u+\mu_{n,m})^2}{2\sigma_{n,m}^2} + \frac{\mu_{n,m}^2}{2\sigma_{n,m}^2}\right)\right) \\
&\qquad - \exp\left(n^{1-2\beta}\varphi_{11,m}^{-1}\left(-\frac{u^2}{2\sigma_m^{*2}} + \frac{\mu_{n,m}^2}{2\sigma_{n,m}^2}\right)\right) \pi(t_m) \Bigg] du \Bigg| \\
&= \left| \int_{J_{n,m}} n^{\frac{1}{2}-\beta} \varphi_{11,m}^{-\frac{1}{2}} \left[\exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{(u+\mu_{n,m})^2}{2\sigma_{n,m}^2}\right) - \exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{u^2}{2\sigma_m^{*2}}\right) \right] du \right| \\
&\le \left| \int_{J_{n,m}} n^{\frac{1}{2}-\beta} \varphi_{11,m}^{-\frac{1}{2}} \left[\exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{(u+\mu_{n,m})^2}{2\sigma_{n,m}^2}\right) - \exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{u^2}{2\sigma_{n,m}^2}\right) \right] du \right| \\
&\quad + \left| \int_{J_{n,m}} n^{\frac{1}{2}-\beta} \varphi_{11,m}^{-\frac{1}{2}} \left[\exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{u^2}{2\sigma_{n,m}^2}\right) - \exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{u^2}{2\sigma_m^{*2}}\right) \right] du \right| \\
&= I_{21} + I_{22}.
\end{aligned}$$

Without loss of generality we assume $\mu_{n,m} \geq 0$. Then, combining (15), (16) and(23) gives that

$$\begin{aligned}
I_{21} = & \left| \sqrt{\frac{\pi\sigma_{n,m}^2}{2}} \left[\mathrm{Erf}\left(\frac{\xi_{n,m-1} n^{\frac{1}{2}} - \mu_{n,m} n^{\frac{1}{2}-\beta}}{\sqrt{2\sigma_{n,m}^2 \varphi_{11,m}}} \right) + \mathrm{Erf}\left(\frac{\xi_{n,m} n^{\frac{1}{2}} + \mu_{n,m} n^{\frac{1}{2}-\beta}}{\sqrt{2\sigma_{n,m}^2 \varphi_{11,m}}} \right) \right] \right. \\
& \left. - \sqrt{\frac{\pi\sigma_{n,m}^2}{2}} \left[\mathrm{Erf}\left(\frac{\xi_{n,m-1} n^{\frac{1}{2}}}{\sqrt{2\sigma_{n,m}^2 \varphi_{11,m}}} \right) + \mathrm{Erf}\left(\frac{\xi_{n,m} n^{\frac{1}{2}}}{\sqrt{2\sigma_{n,m}^2 \varphi_{11,m}}} \right) \right] \right| \\
\lesssim & \left| \mathrm{Erf}\left(\frac{\xi_{n,m-1} n^{\frac{1}{2}} - \mu_{n,m} n^{\frac{1}{2}-\beta}}{\sqrt{2\sigma_{n,m}^2 \varphi_{11,m}}} \right) - \mathrm{Erf}\left(\frac{\xi_{n,m-1} n^{\frac{1}{2}}}{\sqrt{2\sigma_{n,m}^2 \varphi_{11,m}}} \right) \right| \\
& + \left| \mathrm{Erf}\left(\frac{\xi_{n,m} n^{\frac{1}{2}} + \mu_{n,m} n^{\frac{1}{2}-\beta}}{\sqrt{2\sigma_{n,m}^2 \varphi_{11,m}}} \right) - \mathrm{Erf}\left(\frac{\xi_{n,m} n^{\frac{1}{2}}}{\sqrt{2\sigma_{n,m}^2 \varphi_{11,m}}} \right) \right| \\
\leq & \mu_{n,m} n^{\frac{1}{2}-\beta} \cdot \exp\left(-\frac{\xi_{n,m-1}^2 n}{2\sigma_{n,m}^2 \varphi_{11,m}} \right) + \mu_{n,m} n^{\frac{1}{2}-\beta} \cdot \exp\left(-\frac{\xi_{n,m}^2 n}{2\sigma_{n,m}^2 \varphi_{11,m}} \right) \\
\lesssim & e^{-cn\varphi_{11,m}^{-1}}
\end{aligned}$$

for some $c > 0$. In view of (16), (23) and Theorem 1, we obtain

$$\begin{aligned}
I_{22} = & \left| \sqrt{\frac{\pi\sigma_{n,m}^2}{2}} \left[\mathrm{Erf}\left(\frac{\xi_{n,m-1} n^{\frac{1}{2}}}{\sqrt{2\sigma_{n,m}^2 \varphi_{11,m}}} \right) + \mathrm{Erf}\left(\frac{\xi_{n,m} n^{\frac{1}{2}}}{\sqrt{2\sigma_{n,m}^2 \varphi_{11,m}}} \right) \right] \right. \\
& \left. - \sqrt{\frac{\pi\sigma_m^{*2}}{2}} \left[\mathrm{Erf}\left(\frac{\xi_{n,m-1} n^{\frac{1}{2}}}{\sqrt{2\sigma_m^{*2} \varphi_{11,m}}} \right) + \mathrm{Erf}\left(\frac{\xi_{n,m} n^{\frac{1}{2}}}{\sqrt{2\sigma_m^{*2} \varphi_{11,m}}} \right) \right] \right| \lesssim |\sigma_m^{*2} - \sigma_{n,m}^2| \lesssim \lambda^r.
\end{aligned}$$

Therefore, $I_2 \to 0$.

Similarly, by changing of variable and Lemma 4, each term in (28) becomes

$$\begin{aligned}
I_3 &= \int_{[0,z]\setminus I_{n,m}} \ell(t_{\lambda,m}) \tilde{\phi}_{n,m}(t) \pi(t_m) dt \\
&= \int_{K_{n,m}} n^{-\beta} \ell(t_{\lambda,m}) \nu_{n,m}(u) \pi(t_m) du \\
&\lesssim \int_{K_{n,m}} n^{\frac{1}{2}-\beta} \varphi_{11,m}^{-\frac{1}{2}} \exp\left(-n^{1-2\beta} \varphi_{11,m}^{-1} \frac{\mu_{n,m}^2}{2\sigma_{n,m}^2} \right) \exp\left(n^{1-2\beta} \varphi_{11,m}^{-1} \left(-\frac{u^2}{2\sigma_m^{*2}} + \frac{\mu_{n,m}^2}{2\sigma_{n,m}^2} \right) \right) du \\
&= \int_{K_{n,m}} n^{\frac{1}{2}-\beta} \varphi_{11,m}^{-\frac{1}{2}} \exp\left(-n^{1-2\beta} \varphi_{11,m}^{-1} \frac{u^2}{2\sigma_m^{*2}} \right) du,
\end{aligned}$$

where $K_{n,m} = [-n^\beta t_{\lambda,m}, -n^\beta \xi_{n,m-1}] \cup [n^\beta \xi_{n,m}, n^\beta (z - t_{\lambda,m})]$. It then follows that

$$I_3 \lesssim n^\beta \cdot n^{\frac{1}{2}-\beta} \varphi_{11,m}^{-\frac{1}{2}} \exp\left(-n^{1-2\beta} \varphi_{11,m}^{-1} \frac{n^{2\beta}(z - t_{\lambda,m})^2}{2\sigma_m^{*2}} \right) \to 0.$$

Following the same arguments, we can show that (29) and (30) converge to zero. This proves (25) for $0 < z \leq 1$.

Case (3) ($z > 1$). From Case (2) we can see that

$$\left| \int_{-\infty}^{1} \ell(t)\pi(t)dt - \int_{-\infty}^{1} \tilde{h}_n(t)dt \right| \to 0.$$

Note that $\ell(t) = 0$ in $\mathbb{R}\backslash[0,1]$. Using similar arguments as in Case (1), it holds that $\int_1^z \tilde{h}_n(t)dt \to 0$, proving (25) for $z > 1$.

Step 3: We normalize $\tilde{h}_n(t)$ to a density

$$h_n(t) = \frac{\tilde{h}_n(t)}{\int_{\mathbb{R}} \tilde{h}_n(t)dt}.$$

Note that (25) implies

$$\left| \left(\int_{\mathbb{R}} \ell(t)\pi(t)dt \right)^{-1} - \left(\int_{\mathbb{R}} \tilde{h}_n(t)dt \right)^{-1} \right| \to 0. \tag{31}$$

Hence, for any $z \in \mathbb{R}$, we have

$$\begin{aligned} \left| \int_{-\infty}^{z} \pi_n(t \mid X, \boldsymbol{y})dt - \int_{-\infty}^{z} h_n(t)du \right| &\leq \left(\int_{\mathbb{R}} \ell(t)\pi(t)dt \right)^{-1} \left| \int_{-\infty}^{z} \ell(t)\pi(t)dt - \int_{-\infty}^{z} \tilde{h}_n(t)dt \right| \\ &\quad + \int_{-\infty}^{z} \left| \left(\int_{\mathbb{R}} \ell(t)\pi(t)dt \right)^{-1} - \left(\int_{\mathbb{R}} \tilde{h}_n(t)dt \right)^{-1} \right| \tilde{h}_n(t)dt \\ &\to 0, \end{aligned} \tag{32}$$

where the last line follows from (25) and (31).

Rewrite $\tilde{h}_n(t)$ defined in (24) to

$$\tilde{h}_n(t) = \sum_{m=1}^{M} \pi(t_m)\ell(t_{\lambda,m}) \exp\left(n^{1-2\beta}\varphi_{11,m}^{-1}\frac{\mu_{n,m}^2}{2\sigma_{n,m}^2}\right)\sqrt{2\pi n^{-1}\varphi_{11,m}\sigma_m^{*2}}\phi(t \mid t_{\lambda,m}, n^{-1}\varphi_{11,m}\sigma_m^{*2}),$$

which is a linear combination of $\phi(t \mid t_{\lambda,m}, n^{-1}\varphi_{11,m}\sigma_m^{*2})$. Hence, the density function after normalization is

$$h_n(t) = \sum_{m=1}^{M} \tilde{\pi}_{n,m}\phi(t \mid t_{\lambda,m}, n^{-1}\varphi_{11,m}\sigma_m^{*2}),$$

with weights

$$\begin{aligned}
\tilde{\pi}_{n,m} &= \frac{\pi(t_m)\ell(t_{\lambda,m}) \exp\left(n^{1-2\beta}\varphi_{11,m}^{-1}\frac{\mu_{n,m}^2}{2\sigma_{n,m}^2}\right)\sqrt{2\pi n^{-1}\varphi_{11,m}\sigma_m^{*2}}}{\sum_{m=1}^{M}\pi(t_m)\ell(t_{\lambda,m}) \exp\left(n^{1-2\beta}\varphi_{11,m}^{-1}\frac{\mu_{n,m}^2}{2\sigma_{n,m}^2}\right)\sqrt{2\pi n^{-1}\varphi_{11,m}\sigma_m^{*2}}} \\
&= \frac{\pi(t_m)\sigma_m^* n^{-\frac{1}{2}}\varphi_{11,m}^{\frac{1}{2}}\ell(t_{\lambda,m}) / \exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{\mu_{n,m}^2}{2\sigma_{n,m}^2}\right)}{\sum_{m=1}^{M}\pi(t_m)\sigma_m^* n^{-\frac{1}{2}}\varphi_{11,m}^{\frac{1}{2}}\ell(t_{\lambda,m}) / \exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{\mu_{n,m}^2}{2\sigma_{n,m}^2}\right)} \\
&= \frac{\pi(t_m)C|f_0''(t_m)|^{-1} + c_{n,m}}{\sum_{m=1}^{M}\pi(t_m)C|f_0''(t_m)|^{-1} + c_{n,m}}
\end{aligned}$$

where the existence of sequences $c_{n,m} = O(n^{\frac{1}{2}-2\beta}(\log n)^{-1-a})$ is guaranteed by Lemma 4. Hence, we arrive at

$$\tilde{\pi}_{n,m} = \frac{\pi(t_m)|f_0''(t_m)|^{-1}}{\sum_{m=1}^{M}\pi(t_m)|f_0''(t_m)|^{-1}} + c'_{n,m} =: \pi_m + c'_{n,m},$$

for some $c'_{n,m} = O(n^{\frac{1}{2}-2\beta}(\log n)^{-1-a})$. It then holds that

$$\begin{aligned}
&\left|\int_{-\infty}^{z} h_n(t)dt - \int_{-\infty}^{z}\sum_{m=1}^{M}\pi_m\phi(t \mid t_{\lambda,m}, n^{-1}\varphi_{11,m}\sigma_m^{*2})dt\right| \\
&\leq \sum_{m=1}^{M}\int_{-\infty}^{z}|\tilde{\pi}_{n,m} - \pi_m|\phi(t \mid t_{\lambda,m}, n^{-1}\varphi_{11,m}\sigma_m^{*2})dt \to 0. \qquad (33)
\end{aligned}$$

Combining (32) and (33), we obtain that for any z,

$$\mathbb{E}_{\mathbb{P}_0}\left|\int_{-\infty}^{z}\pi_n(t \mid X, \boldsymbol{y})dt - \int_{-\infty}^{z}\sum_{m=1}^{M}\pi_m\phi(t \mid t_{\lambda,m}, n^{-1}\varphi_{11,m}\sigma_m^{*2})dt\right|\mathbf{1}_{A_n} \to 0.$$

This together with $\mathbb{E}_{\mathbb{P}_0}(\mathbf{1}_{A_n^c}) = \mathbb{P}_0(A_n^c) \le n^{-10}$ gives that

$$\left| \Pi_n(t \le z \mid X, \boldsymbol{y}) - \sum_{m=1}^{M} \pi_m \Phi(z \mid t_{\lambda,m}, n^{-1}\varphi_{11,m}\sigma_m^{*2}) \right| \to 0$$

for any $z \in \mathbb{R}$ in $\mathbb{P}_0$-probability. This completes the proof.

A.5.2 Proof of (ii)

Denote $\zeta_0 = t_1$, $\zeta_m = \frac{1}{2}(t_{m+1} - t_m)$, $m = 1, \ldots, M-1$, $\zeta_M = 1 - t_M$. Then,

$$[0,1] = \bigcup_{m=1}^{M} I_m, \quad I_m = [t_m - \zeta_{m-1}, t_m + \zeta_m], \quad m = 1, \ldots, M.$$

We first bound the unnormalized difference

$$\left| \int_{-\infty}^{z} \ell(t)\mathbf{1}_{I_m}(t)\pi(t)dt - \int_{-\infty}^{z} \ell(t_{\lambda,m})\tilde{\phi}_{n,m}(t)\pi(t_m)dt \right| \tag{34}$$

under A_n by considering three cases for z: (1) $z \le t_m - \zeta_{m-1}$, (2) $t_m - \zeta_{m-1} < z < t_m + \zeta_m$, and (3) $z \ge t_m + \zeta_m$.

Case 1 ($z \le t_m - \zeta_{m-1}$). Since $z \notin I_m$, (34) becomes

$$\begin{aligned}
\int_{-\infty}^{z} \ell(t_{\lambda,m})\tilde{\phi}_{n,m}(t)\pi(t_m)dt &= \int_{-\infty}^{(z-t_{\lambda,m})n^{\beta}} \ell(t_{\lambda,m})\nu_{n,m}(u)\pi(t_m)du \\
&\lesssim \int_{-\infty}^{(z-t_{\lambda,m})n^{\beta}} n^{\frac{1}{2}-\beta}\varphi_{11,m}^{-\frac{1}{2}} \exp\left(-n^{1-2\beta}\varphi_{11,m}^{-1}\frac{u^2}{2\sigma_m^{*2}}\right) du \\
&= \sqrt{\frac{\pi\sigma_m^{*2}}{2}}\left[1 - \operatorname{Erf}\left(\frac{\sqrt{n}(t_{\lambda,m} - z)}{\sqrt{2\sigma_m^{*2}\varphi_{11,m}}}\right)\right] \\
&\lesssim \exp\left(-\frac{n(t_{\lambda,m} - t_m - \zeta_{m-1})^2}{2\sigma_m^{*2}\varphi_{11,m}}\right).
\end{aligned}$$

Case 2 ($t_m - \zeta_{m-1} < z < t_m + \zeta_m$). In this case, we consider

$$\left| \int_{t_m - \zeta_{m-1}}^{z} \left[\ell(t)\pi(t) - \ell(t_{\lambda,m})\tilde{\phi}_{n,m}(t)\pi(t_m)\right] dt \right|$$

$$
\begin{aligned}
&= \left| \int_{H_{n,m}} \left[\ell(t_{\lambda,m} + u/n^{\beta}) \pi(t_{\lambda,m} + u/n^{\beta}) - \ell(t_{\lambda,m}) \nu_{n,m}(u) \pi(t_m) \right] n^{-\beta} du \right| \\
&= \left| \int_{H_{n,m}} n^{-\beta} \ell(t_{\lambda,m}) \left[Z_{n,m}(u) \pi(t_{\lambda,m} + u/n^{\beta}) - \nu_{n,m}(u) \pi(t_m) \right] du \right| \\
&\leq \left| \int_{H_{n,m}} n^{-\beta} \ell(t_{\lambda,m}) Z_{n,m}(u) [\pi(t_{\lambda,m} + u/n^{\beta}) - \pi(t_m)] du \right| \\
&\quad + \left| \int_{H_{n,m}} n^{-\beta} \ell(t_{\lambda,m}) [Z_{n,m}(u) - \nu_{n,m}(u)] \pi(t_m) du \right| \\
&\quad = I_1' + I_2',
\end{aligned}
$$

where $H_{n,m} = [(t_m - t_{\lambda,m} - \zeta_{m-1})n^{\beta}, (z - t_{\lambda,m})n^{\beta}]$. Following similar arguments as used in the proof of Part (i), it can be shown that $I_1' \lesssim \lambda^{r_1}$ and

$$
I_2' \lesssim \mu_{n,m} n^{\frac{1}{2}-\beta} \cdot \exp\left(-\frac{(z - t_{\lambda,m})^2 n}{2\sigma_{n,m}^2 \varphi_{11,m}} \right).
$$

Case 3 ($z \geq t_m + \zeta_m$). Again, $z \notin I_m$ and (34) becomes

$$
\begin{aligned}
\int_{t_m+\zeta_m}^{z} \ell(t_{\lambda,m}) \tilde{\phi}_{n,m}(t) \pi(t_m) dt &= \int_{(t_m - t_{\lambda,m} + \zeta_m)n^{\beta}}^{(z - t_{\lambda,m})n^{\beta}} \ell(t_{\lambda,m}) \nu_{n,m}(u) \pi(t_m) du \\
&\lesssim \int_{(t_m - t_{\lambda,m} + \zeta_m)n^{\beta}}^{\infty} n^{\frac{1}{2}-\beta} \varphi_{11,m}^{-\frac{1}{2}} \exp\left(-n^{1-2\beta} \varphi_{11,m}^{-1} \frac{u^2}{2\sigma_m^{*2}} \right) du \\
&= \sqrt{\frac{\pi \sigma_m^{*2}}{2}} \left[1 - \operatorname{Erf}\left(\frac{\sqrt{n}(t_m - t_{\lambda,m} + \zeta_m)}{\sqrt{2\sigma_m^{*2} \varphi_{11,m}}} \right) \right] \\
&\lesssim \exp\left(-\frac{n(t_m - t_{\lambda,m} + \zeta_m)^2}{2\sigma_m^{*2} \varphi_{11,m}} \right).
\end{aligned}
$$

Combining the three cases, we obtain that under A_n,

$$
\left| \int_{-\infty}^{z} \ell(t) \mathbf{1}_{I_m}(t) \pi(t) dt - \int_{-\infty}^{z} \ell(t_{\lambda,m}) \tilde{\phi}_{n,m}(t) \pi(t_m) dt \right| \lesssim \lambda^{r_1} \vee n^{\frac{1}{2}-\beta} \cdot \exp\left(-\frac{(z - t_{\lambda,m})^2 n}{2\sigma_{n,m}^2 \varphi_{11,m}} \right).
$$

Let $\Pi_{n,m}(\cdot \mid X, \boldsymbol{y})$ be the posterior of $t\mathbf{1}_{I_m}$. Following the same arguments as in part (i) again, we can show that

$$
\left| \Pi_{n,m}(t' \leq z \mid X, \boldsymbol{y}) - \Phi(z \mid t_{\lambda,m}, n^{-1} \varphi_{11,m} \sigma_m^{*2}) \right| \lesssim \lambda^{r_1} \vee n^{\frac{1}{2}-\beta} \cdot \exp\left(-\frac{(z - t_{\lambda,m})^2 n}{2\sigma_{n,m}^2 \varphi_{11,m}} \right)
$$

in $\mathbb{P}_0$-probability.

By Lemma 2, we have $|\hat{t}_m - b_n - t_{\lambda,m}| \lesssim n^{-\beta}\sqrt{\log n} \vee n^{-\beta}\log n = n^{-\beta}\log n$. Thus,

$$\left|\Pi_{n,m}(t' \le z \mid X, \boldsymbol{y}) - \Phi(z \mid \hat{t}_m - b_n, n^{-1}\varphi_{11,m}\sigma_m^{*2})\right| \lesssim \lambda^{r_1} \vee n^{\frac{1}{2}-\beta} \cdot \exp\left(-\frac{(z - t_{\lambda,m})^2 n}{2\sigma_{n,m}^2 \varphi_{11,m}}\right) \vee n^{-\beta}\log n.$$

Now we consider the posterior of $\sqrt{\frac{n}{\varphi_{11,m}}}(t\mathbf{1}_{I_m}(t) - \hat{t}_m + b_n)$. By changing of variable, it follows that

$$\begin{aligned}
&\left|\Pi'_{n,m}(t' \le z \mid X, \boldsymbol{y}) - \Phi(z \mid 0, \sigma_m^{*2})\right| \\
&= \left|\Pi_{n,m}\left(t' \le \sqrt{\frac{\varphi_{11,m}}{n}}z + \hat{t}_m - b_n \mid X, \boldsymbol{y}\right) - \Phi\left(\sqrt{\frac{\varphi_{11,m}}{n}}z + \hat{t}_m - b_n | \hat{t}_m - b_n, n^{-1}\varphi_{11,m}\sigma_m^{*2}\right)\right| \\
&\lesssim n^{\frac{1}{2}-\beta} \cdot \exp\left\{-\frac{(\sqrt{\frac{\varphi_{11,m}}{n}}z + \hat{t}_m - t_{\lambda,m} - b_n)^2 n}{2\sigma_{n,1}^2 \varphi_{11,m}}\right\} \\
&\lesssim n^{\frac{1}{2}-\beta} \cdot \exp\left\{-\left(\sqrt{\frac{\varphi_{11,m}}{n}}z \vee (\hat{t}_m - t_{\lambda,m}) \vee b_n\right)^2 \frac{n}{2\sigma_{n,m}^2 \varphi_{11,m}}\right\} \\
&\lesssim n^{\frac{1}{2}-\beta} \cdot \exp\left\{-\frac{b_n^2 n}{2\sigma_{n,m}^2 \varphi_{11,m}}\right\} \\
&\lesssim n^{\frac{1}{2}-\beta} \cdot \exp\left\{-\frac{n^{1-2\beta}(\log n)^2}{2\sigma_{n,m}^2 \varphi_{11,m}}\right\} \to 0.
\end{aligned}$$

This completes the proof.

A.6 Proof of Theorem 3

A.6.1 Proof of (i)

Let $F_n(z) = \Pi_n(t \le z \mid X, \boldsymbol{y})$ and $G_n(z) = \sum_{m=1}^M \pi_m \Phi(z \mid t_m, n^{-1}\varphi_{11,m}\sigma_m^{*2})$. Note that $G_n(\cdot)$ is a deterministic function, and its derivative $G'_n(\cdot)$ is the density function of a Gaussian mixture. The variance of each component distribution in G_n goes to zero in view of Assumption B2 and conditions in Theorem 2. For sufficiently large n, using the analytical expression of $G'_n(\cdot)$ and elementary calculus, we can show that $G'_n(\cdot)$ has at least M local modes, denoted by $t_{m,G}$, such that $t_{m,G} \to t_m$. On the other hand, $G'_n(\cdot)$ cannot have more

than M local modes in view of Corollary 2.4 in Carreira-Perpinán and Williams (2003); hence, $\{t_{m,1}, \ldots, t_{M,G}\}$ are the only local modes of $G_n'(\cdot)$. For large enough n and each m, we consider an interval $(t_{m,G} - \delta_m, t_{m,G} + \delta_m)$ for some $\delta_m > 0$ such that $G_n''(z) > 0$ when $z \in (t_{m,G} - \delta_m, t_{m,G})$ and $G_n''(z) < 0$ when $z \in (t_{m,G}, t_{m,G} + \delta_m)$.

By Theorem 2 (i), we have $|F_n(z) - G_n(z)| \to 0$ for any $z \in \mathbb{R}$ in $\mathbb{P}_0$-probability. The following arguments and conclusions in Step 1–4 hold with $\mathbb{P}_0$-probability tending to 1 because of this convergence in $\mathbb{P}_0$-probability.

Step 1: We first show that there exists a $t_{m,F}$ in the neighborhood of $t_{m,G}$ such that $F_n''(t_{m,F}) = 0$, for $m = 1, \ldots, M$. Suppose $F_n''(z) \neq 0$ for any $z \in (t_{m,G} - \delta_m, t_{m,G} + \delta_m)$, Without loss of generality we assume $F_n''(z) > 0$ when $z \in (t_{m,G} - \delta_m, t_{m,G} + \delta_m)$. Since $G_n(z)$ is concave on $(t_{m,G}, t_{m,G} + \delta_m)$,

$$G_n(t_{m,G} + \delta_m/2) > (G_n(t_{m,G}) + G_n(t_{m,G} + \delta_m))/2 + \epsilon, \tag{35}$$

for some $\epsilon > 0$. Since $F_n(z)$ is convex on $(t_{m,G}, t_{m,G} + \delta_m)$,

$$F_n(t_{m,G} + \delta_m/2) < (F_n(t_{m,G}) + F_n(t_{m,G} + \delta_m))/2.$$

For sufficiently large n, it holds that with $\mathbb{P}_0$-probability tending to 1 $|F_n(z) - G_n(z)| < \epsilon/2$ for $z = t_{m,G}, t_{m,G} + \delta_m/2, t_{m,G} + \delta_m$. Therefore,

$$G_n(t_{m,G} + \delta_m/2) > (F_n(t_{m,G}) + F_n(t_{m,G} + \delta_m))/2 + \epsilon/2 > F_n(t_{m,G} + \delta_m/2) + \epsilon/2,$$

which is a contradiction. This proves that there exists $t_{m,F} \in (t_{m,G} - \delta_m, t_{m,G} + \delta_m)$ such that $F_n''(t_{m,F}) = 0$.

Step 2: We show that $t_{m,F} \to t_m$ in $\mathbb{P}_0$-probability. Suppose there exists $\delta > 0$ such that $|t_{m,G} - t_{m,F}| > \delta$ for any sufficiently large n. Without loss of generality we assume $t_{m,G} < t_{m,F}$ and $F_n''(z) < 0$ when $z \in (t_{m,F}, t_{m,G} + \delta_m)$ and $F_n''(z) > 0$ when $z \in (t_{m,G} - \delta_m, t_{m,F})$. Thus, G_n is concave on $(t_{m,G}, t_{m,F})$ while F_n is convex on $(t_{m,G}, t_{m,F})$. This is a contradiction using

the same argument in Step 1. Combining this with $t_{m,G} \to t_m$ shows the convergence of $t_{m,F}$.

Step 3: In this step, we show that $t_{m,F}$ must be a local mode of $F_n'(z)$. Suppose that $F_n''(z) > 0$ when $z \in (t_{m,F}, t_{m,G} + \delta_m)$ and $F_n''(z) < 0$ when $z \in (t_{m,G} - \delta_m, t_{m,F})$, yielding

$$F_n(t_{m,F} + \delta_m/2) < (F_n(t_{m,F}) + F_n(t_{m,F} + \delta_m))/2.$$

For sufficiently large n, it holds with $\mathbb{P}_0$-probability tending to 1 that $|F_n(z) - G_n(z)| < \epsilon/4$ for $x = t_{m,G}, t_{m,G} + \delta_m/2, t_{m,G} + \delta_m$. Invoking (35),

$$G_n(t_{m,G} + \delta_m/2) > (F_n(t_{m,G}) + F_n(t_{m,G} + \delta_m))/2 + 3\epsilon/4.$$

For sufficiently large n, it holds with $\mathbb{P}_0$-probability tending to 1 that $|F_n(z_1) - F_n(z_2)| < \epsilon/4$ for $z_1 = t_{m,G}$, $z_2 = t_{m_F}$, $z_1 = t_{m,G} + \delta_m/2$, $z_2 = t_{m,F} + \delta_m/2$ and $z_1 = t_{m,G} + \delta_m$, $z_2 = t_{m,F} + \delta_m$. Therefore,

$$G_n(t_{m,G} + \delta_m/2) > (F_n(t_{m,F}) + F_n(t_{m,F} + \delta_m))/2 + \epsilon/2 > F_n(t_{m,F} + \delta_m/2) + \epsilon/2.$$

However,

$$G_n(t_{m,G} + \delta_m/2) < F_n(t_{m,G} + \delta_m/2) + \epsilon/4 < F_n(t_{m,F} + \delta_m/2) + \epsilon/2,$$

which is a contradiction. This completes Step 3.

Step 4: In the last step, we show that the number of local modes of $F_n'(z)$ is exactly M. We have proven that $F_n'(\cdot)$ has at least M local modes. Suppose that there exists $t_{m',F} \in (0, 1)$ and $\delta_{m'} > 0$ such that $t_{m',F}$ is a local mode of $F_n'(z)$ and $G_n''(z) \neq 0$ for $z \in (t_{m',F} - \delta_{m'}, t_{m',F} + \delta_{m'})$ for any sufficiently large n. Without loss of generality assume $G_n''(z) > 0$ for $z \in (t_{m',F} - \delta_{m'}, t_{m',F} + \delta_{m'})$. Thus, on $(t_{m',F} - \delta_{m'}, t_{m',F} + \delta_{m'})$, $G_n(z)$ is convex while $F_n(z)$ is concave. By similar arguments used in Step 1, we can obtain a contradiction. Hence, the number of local modes of $F_n'(\cdot)$ is exactly M.

This completes the proof.

A.6.2 Proof of (ii)

By Taylor expansion of $\widehat{\mu}_{f'}$, we obtain

$$\widehat{\mu}_{f'}(t) = \widehat{\mu}_{f'}(t_m) + (t - t_m)\widehat{\mu}'_{f'}(\xi)$$

for some ξ between t and t_m. Since $\hat{t}_m$ is a local extremum of $\hat{\mu}_f$, there holds $\widehat{\mu}_{f'}(\hat{t}_m) = 0$. Substituting $t = \hat{t}_m$ into the expansion above yields

$$\widehat{\mu}_{f'}(t_m) + (\hat{t}_m - t_m)\widehat{\mu}'_{f'}(\xi) = 0.$$

Lemma 1 and Assumption C ensure that $\widehat{\mu}'_{f'}(x) \xrightarrow{p} f_0''(x)$, and Lemma 2 implies that $\hat{t}_m \xrightarrow{p} t_m$. Therefore, $\widehat{\mu}'_{f'}(\xi) \xrightarrow{p} f_0''(t_m)$, and thus $\widehat{\mu}'_{f'}(\xi)$ is bounded away from zero and infinity in view of Assumption A3. It thus follows that

$$\hat{t}_m - t_m = -\frac{\widehat{\mu}_{f'}(t_m)}{\widehat{\mu}'_{f'}(\xi)}.$$

Let $\Delta_n(\cdot) = K_{10}(\cdot, X)[K(X,X) + n\lambda \boldsymbol{I}_n]^{-1} f_0(X)$. Conditioning on X, it holds that

$$\begin{aligned}\widehat{\mu}_{f'}(t_m) &= K_{10}(t_m, X)[K(X,X) + n\lambda \boldsymbol{I}_n]^{-1}\boldsymbol{y} \\ &\sim N\left(\Delta_n(t_m), \sigma^2 K_{10}(t_m, X)[K(X,X) + n\lambda \boldsymbol{I}_n]^{-2} K_{10}(t_m, X)^T\right).\end{aligned}$$

Hence,

$$\left.\frac{\widehat{\mu}_{f'}(t_m) - \Delta_n(t_m)}{\sigma\sqrt{K_{10}(t_m, X)[K(X,X) + n\lambda \boldsymbol{I}_n]^{-2} K_{10}(t_m, X)^T}}\right| X \sim N(0,1),$$

which implies that

$$\frac{\widehat{\mu}_{f'}(t_m) - \Delta_n(t_m)}{\sigma\sqrt{K_{10}(t_m, X)[K(X,X) + n\lambda \boldsymbol{I}_n]^{-2} K_{10}(t_m, X)^T}} \sim N(0,1).$$

By Slutsky's theorem, we obtain

$$\sqrt{\frac{1}{K_{10}(t_m,X)[K(X,X)+n\lambda \boldsymbol{I}_n]^{-2}K_{10}(t_m,X)^T}}\left[\hat{t}_m - t_m + \frac{\Delta_n(t_m)}{\widehat{\mu}'_{f'}(t_m)}\right] \xrightarrow{d} N\left(0, \sigma^2 f_0''(t_m)^{-2}\right).$$

Note that

$$\begin{aligned}
&\frac{K_{10}(\hat{t}_m,X)[K(X,X)+n\lambda \boldsymbol{I}_n]^{-2}K_{10}(\hat{t}_m,X)^T}{K_{10}(t_m,X)[K(X,X)+n\lambda \boldsymbol{I}_n]^{-2}K_{10}(t_m,X)^T} - 1 \\
&= \frac{K_{10}(\hat{t}_m,X)[K(X,X)+n\lambda \boldsymbol{I}_n]^{-2}(K_{10}(\hat{t}_m,X)^T - K_{10}(t_m,X))}{K_{10}(t_m,X)[K(X,X)+n\lambda \boldsymbol{I}_n]^{-2}K_{10}(t_m,X)^T} \\
&\quad + \frac{(K_{10}(\hat{t}_m,X)^T - K_{10}(t_m,X))[K(X,X)+n\lambda \boldsymbol{I}_n]^{-2}K_{10}(t_m,X))}{K_{10}(t_m,X)[K(X,X)+n\lambda \boldsymbol{I}_n]^{-2}K_{10}(t_m,X)^T}.
\end{aligned}$$

Consider the eigendecomposition of $K(X,X) = Q_n \Lambda_n Q_n^T$, where $\Lambda_n = \mathrm{diag}(u_1, \ldots, u_n)$ and $Q_n^T = Q_n^{-1}$. Denote $(p_1, \ldots, p_n) = K_{10}(t_m, X)Q_n$, likewise $(q_1, \ldots, q_n) = K_{10}(\hat{t}_m, X)Q_n$. Then

$$\begin{aligned}
K_{10}(t_m,X)[K(X,X)+n\lambda \boldsymbol{I}_n]^{-2}K_{10}(t_m,X)^T &= K_{10}(t_m,X)Q_n\Lambda_n^{-2}Q_n^T K_{10}(t_m,X) \\
&= \sum_{i=1}^{\infty} \frac{p_i^2}{(u_i+n\lambda)^2}.
\end{aligned}$$

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&K_{10}(\hat{t}_m,X)[K(X,X)+n\lambda \boldsymbol{I}_n]^{-2}(K_{10}(\hat{t}_m,X)^T - K_{10}(t_m,X)) \\
&= \sum_{i=1}^{\infty} \frac{q_i(q_i-p_i)}{(u_i+n\lambda)^2} \le \sqrt{\sum_{i=1}^{\infty} \frac{q_i^2}{(u_i+n\lambda)^2} \sum_{i=1}^{\infty} \frac{(q_i-p_i)^2}{(u_i+n\lambda)^2}}.
\end{aligned}$$

Since $K_{10}(\hat{t}_m, X_i) - K_{10}(t_m, X_i) \xrightarrow{p} 0$ uniformly for $1 \le i \le n$, we have

$$\sum_{i=1}^{\infty} \frac{q_i^2}{(u_i+n\lambda)^2} \Big/ \sum_{i=1}^{\infty} \frac{p_i^2}{(u_i+n\lambda)^2} \xrightarrow{p} 1$$

and

$$\sum_{i=1}^{\infty} \frac{(q_i - p_i)^2}{(u_i + n\lambda)^2} \Big/ \sum_{i=1}^{\infty} \frac{p_i^2}{(u_i + n\lambda)^2} \xrightarrow{p} 0.$$

Hence, it follows that

$$\frac{K_{10}(\hat{t}_m, X)[K(X,X) + n\lambda \boldsymbol{I}_n]^{-2}(K_{10}(\hat{t}_m, X)^T - K_{10}(t_m, X))}{K_{10}(t_m, X)[K(X,X) + n\lambda \boldsymbol{I}_n]^{-2} K_{10}(t_m, X)^T} \xrightarrow{p} 0.$$

Similarly, it can be shown that

$$\frac{(K_{10}(\hat{t}_m, X)^T - K_{10}(t_m, X))[K(X,X) + n\lambda \boldsymbol{I}_n]^{-2} K_{10}(t_m, X))}{K_{10}(t_m, X)[K(X,X) + n\lambda \boldsymbol{I}_n]^{-2} K_{10}(t_m, X)^T} \xrightarrow{p} 0.$$

Therefore,

$$\frac{K_{10}(\hat{t}_m, X)[K(X,X) + n\lambda \boldsymbol{I}_n]^{-2} K_{10}(\hat{t}_m, X)^T}{K_{10}(t_m, X)[K(X,X) + n\lambda \boldsymbol{I}_n]^{-2} K_{10}(t_m, X)^T} \xrightarrow{p} 1.$$

Recall that $\widehat{\mu}'_{f'}(\hat{t}_m) \xrightarrow{p} f_0''(t_m)$. Therefore, by Slutsky's theorem again, we arrive at

$$\frac{\sigma |\widehat{\mu}'_{f'}(\hat{t}_m)|}{\sqrt{K_{10}(\hat{t}_m, X)[K(X,X) + n\lambda \boldsymbol{I}_n]^{-2} K_{10}(\hat{t}_m, X)^T}} \left[\hat{t}_m - t_m + \frac{\Delta_n(t_m)}{\widehat{\mu}'_{f'}(t_m)} \right] \xrightarrow{d} N(0,1).$$

Hence, an asymptotic $1 - \alpha$ confidence interval of $t_m + \Delta_n(t_m)/f_0''(t_m)$ is

$$\hat{t}_m \pm z_{\alpha/2} \frac{\sigma \sqrt{K_{10}(\hat{t}_m, X)[K(X,X) + n\lambda \boldsymbol{I}_n]^{-2} K_{10}(\hat{t}_m, X)^T}}{|\widehat{\mu}'_{f'}(\hat{t}_m)|}.$$

This completes the proof.

A.7 Proof of Theorem 4

For any $j, l \leq s$, we have

$$|K_{jl}^{\alpha,s}(x, x')| = \left| \sum_{i=1}^{\infty} \mu_i \psi_i^{(j)}(x) \psi_i^{(l)}(x') \right| \lesssim \sum_{i=1}^{\infty} i^{-2\alpha + j + l},$$

which is finite when $\alpha > \frac{j+l+1}{2}$. Thus, Assumption B1 holds when $s \geq 4$ and $\alpha > 9/2$. According to Lemma 11 in Liu and Li (2023), when $\alpha > \frac{j+l+1}{2}$, we have

$$\sup_{x\in\mathcal{X}} |\varphi_{jl}(x)| = \sup_{x\in\mathcal{X}} \left| \sum_{i=1}^{\infty} \frac{\mu_i}{\lambda+\mu_i} \psi^{(j)}(x)\psi^{(l)}(x) \right| \lesssim \sum_{i=1}^{\infty} \frac{\mu_i i^{j+l}}{\mu_i+\lambda} \asymp \lambda^{-\frac{j+l+1}{2\alpha}}.$$

Hence, Assumption B2 is satisfied when $\alpha > 3$. In view of Lemma 1, Lemma 11 and Lemma 13 in Liu and Li (2023), when $\alpha > k + 1/2$, we have

$$\|f_\lambda^{(k)} - f_0^{(k)}\|_\infty \lesssim \lambda^{\frac{1}{2}-\frac{k}{2\alpha}}.$$

This verifies Assumption C with $r_1 = \frac{\alpha-1}{2\alpha}$, $r_2 = \frac{\alpha-2}{2\alpha}$ when $\alpha > 5/2$. Finally, by Assumption E we have $\varphi_{11}(x) = \sum_{i=1}^{\infty} \frac{\mu_i}{\lambda+\mu_i} \psi_i'(x)^2 \to \infty$ as $\lambda \to 0$. Thus, a sufficient condition for the boundedness of $n^{\frac{3}{2}-4\beta}\varphi_{11,m}^{-2}$ in Theorem 1 and 2 is $n^{\frac{3}{2}-4\beta} = O(1)$, which implies that $\beta \geq \frac{3}{8}$. This completes the proof.

A.8 Proof of Lemma 3

Let $f_0 = \sum_{i=1}^{\infty} f_i\psi_i$. Then, for any $k \leq s$,

$$f_\lambda^{(k)} - f_0^{(k)} = -\sum_{i=1}^{\infty} \frac{\lambda}{\lambda+\mu_i} f_i \psi_i^{(k)}.$$

Hence,

$$\|f_\lambda^{(k)} - f_0^{(k)}\|_\infty \leq \sum_{i=1}^{\infty} \frac{\lambda}{\lambda+\mu_i} |f_i| \cdot i^k \lesssim \lambda^{\frac{1}{2}-\frac{k}{2e\gamma}} \sum_{i=1}^{\infty} \frac{\lambda^{\frac{1}{2}+\frac{k}{2e\gamma}} \cdot e^{-\gamma i} i^k}{\lambda + e^{-2\gamma i}} e^{\gamma i} |f_i|.$$

Note that $i^k \leq e^{\frac{ki}{e}}$, then by Young's inequality for products, we have

$$\lambda^{\frac{1}{2}+\frac{k}{2e\gamma}} \cdot e^{-\gamma i} i^k \leq \lambda^{\frac{1}{2}+\frac{k}{2e\gamma}} \cdot e^{(\frac{k}{e}-\gamma)i} \leq \left(\frac{1}{2} + \frac{k}{2e\gamma}\right)\lambda + \left(\frac{1}{2} - \frac{k}{2e\gamma}\right) e^{-2\gamma i} \leq \lambda + e^{-2\gamma i}.$$

Therefore, $\|f_\lambda^{(k)} - f_0^{(k)}\|_\infty \lesssim \lambda^{\frac{1}{2}-\frac{k}{2e\gamma}} \sum_{i=1}^{\infty} e^{\gamma i} |f_i| \lesssim \lambda^{\frac{1}{2}-\frac{k}{2e\gamma}}$. This completes the proof.

A.9 Proof of Theorem 5

It is easy to see that $K_{\gamma,s} \in C^8(\mathcal{X}, \mathcal{X})$ for any $\gamma > 0$ and $s \geq 4$; thus Assumption B1 is satisfied.

Note that

$$\sup_{x \in \mathcal{X}} |\varphi_{jl}(x)| = \left| \sum_{i=1}^{\infty} \frac{\mu_i}{\lambda + \mu_i} \psi_i^{(j)}(x) \psi_i^{(l)}(x) \right| \lesssim \sum_{i=1}^{\infty} \frac{e^{-2\gamma i} i^{j+l}}{\lambda + e^{-2\gamma i}} \leq \frac{e^{-2\gamma i} e^{\frac{(j+l)i}{e}}}{\lambda + e^{-2\gamma i}}.$$

By Young's inequality for products, when $2e\gamma > j + l$ we have

$$\lambda^{\frac{j+l}{2e\gamma}} \cdot e^{-2\gamma i + \frac{(j+l)i}{e}} \leq \left(1 - \frac{j+l}{2e\gamma}\right) \lambda + \left(\frac{j+l}{2e\gamma}\right) e^{-2\gamma i} \leq \lambda + e^{-2\gamma i}.$$

Hence, $\varphi_{jl} \lesssim \lambda^{-\frac{j+l}{2e\gamma}}$ for $2e\gamma > j+l$ and Assumption B2 holds for $\gamma > \frac{5}{2e}$. In view of Lemma 3, we have Assumption C satisfied with $r_1 = r_2 = \frac{e\gamma - 2}{2e\gamma}$ and $\gamma > \frac{2}{e}$.

Finally, since $\lambda = o(1)$, we have $\varphi_{11,m}^{-2} = o(1)$ under Assumption E. Thus, a sufficient condition for the boundedness of $n^{\frac{3}{2}-4\beta} \varphi_{11,m}^{-2}$ is $\beta \geq \frac{3}{8}$. This completes the proof.

A.10 Proof of Lemma 4

The likelihood function (5) gives

$$\begin{aligned} n^{-\frac{1}{2}} \varphi_{11,m}^{\frac{1}{2}} \ell(t_{\lambda,m}) &= \frac{C}{\sqrt{n \varphi_{11,m}^{-1} \widehat{\sigma}_{f'}^2(t_{\lambda,m})}} \exp\left(-\frac{\widehat{\mu}_{f'}(t_{\lambda,m})^2}{2\widehat{\sigma}_{f'}^2(t_{\lambda,m})} \right) \\ &= \frac{C}{\sqrt{n \varphi_{11,m}^{-1} \widehat{\sigma}_{f'}^2(t_{\lambda,m})}} \exp\left(-n^{1-2\beta} \varphi_{11,m}^{-1} \frac{\mu_{n,m}^2}{2\sigma_{n,m}^2} \right), \end{aligned}$$

where $\mu_{n,m}$ and $\sigma_{n,m}^2$ are defined in Theorem 1. Note that

$$\begin{aligned} \left| \frac{1}{\sqrt{n \varphi_{11,m}^{-1} \widehat{\sigma}_{f'}^2(t_{\lambda,m})}} - \frac{1}{|f_0''(t_m)| \sigma_m^*} \right| &\asymp \left| |f_0''(t_m)| \sigma_m^* - \sqrt{n \varphi_{11,m}^{-1} \widehat{\sigma}_{f'}^2(t_{\lambda,m})} \right| \\ &\asymp |f_0''(t_m)^2 \sigma_m^{*2} - n \varphi_{11,m}^{-1} \widehat{\sigma}_{f'}^2(t_{\lambda,m})|. \end{aligned}$$

Substituting $f_0''(t_m)^2\sigma_m^{*2} = \sigma^2$ into the right side yields

$$\left| f_0''(t_m)^2\sigma_m^{*2} - n\varphi_{11,m}^{-1}\widehat{\sigma}_{f'}^2(t_{\lambda,m}) \right| = \left| n\varphi_{11,m}^{-1}\widehat{\sigma}_{f'}^2(t_{\lambda,m}) - \sigma^2 \right| \lesssim n^{\frac{1}{2}-2\beta}(\log n)^{-1-a}.$$

This completes the proof.

B Additional simulation results

B.1 Effect of noise standard derivation and credible level

We carried out additional experiments to investigate the effect of noise standard derivation and credible levels $1-\alpha$. We used the same regression function shown in the paper and generated more noisy data by increasing the noise standard deviation σ from 0.1 to 0.2. As expected, results worsen, particularly for smaller sample sizes. This is because the GP tends to produce more wiggly curves. For example, looking at the percentages of correctly estimating M for $\alpha = .05$, calculated over 100 replicated datasets, we observed the following results: for $n = 100$ we obtained 19% and 85% for Beta (1,1) and Beta(2,3), respectively, versus 47% and 86% of Figure 3 in the paper; for $n = 500$ we obtained 52% and 94% for Beta (1,1) and Beta(2,3), respectively, versus 95% and 99% for $\sigma = 0.1$. We notice that, as already shown in the main simulation, the $Beta(2, 3)$ prior and larger sample sizes help identifying the correct number of local extrema.

Next, we used this additional simulation study to investigate the performance of HPDR for different values of α. Results for sample sizes $n = 100$, $n = 500$ and $n = 1000$ and the two Beta priors are reported in the two tables below. For each combination of prior and sample size, we generated 100 simulated datasets.

$Beta(1,1)$	$\alpha = 0.001$	$\alpha = 0.005$	$\alpha = 0.01$	$\alpha = 0.03$	$\alpha = 0.05$	$\alpha = 0.1$
$n = 100$	56%	53%	51%	18%	19%	35%
$n = 500$	25%	32%	34%	43%	52%	60%
$n = 1000$	27%	35%	44%	57%	76%	84%

Table 4: $Beta(1, 1)$. Percentages of correctly estimated number of t's. The results are calculated on 100 simulated data.

$Beta(2,3)$	$\alpha = 0.001$	$\alpha = 0.005$	$\alpha = 0.01$	$\alpha = 0.03$	$\alpha = 0.05$	$\alpha = 0.1$
$n = 100$	87%	88%	88%	86%	85%	77%
$n = 500$	95%	96%	95%	95%	94%	89%
$n = 1000$	95%	95%	96%	94%	94%	95%

Table 5: $Beta(2,3)$. Percentages of correctly estimated number of t's. The results are calculated on 100 simulated data.

In this additional study, we observed that increasing values of α did not necessarily correspond to larger estimated numbers of local extrema. This is because situations like the one shown in Figure 7 can occur. Therefore, larger or smaller α values do not necessarily imply more or fewer separated HPDR segments. Overall, results confirm the fairly robust estimation performance of the Beta(2,3) prior in estimating M.

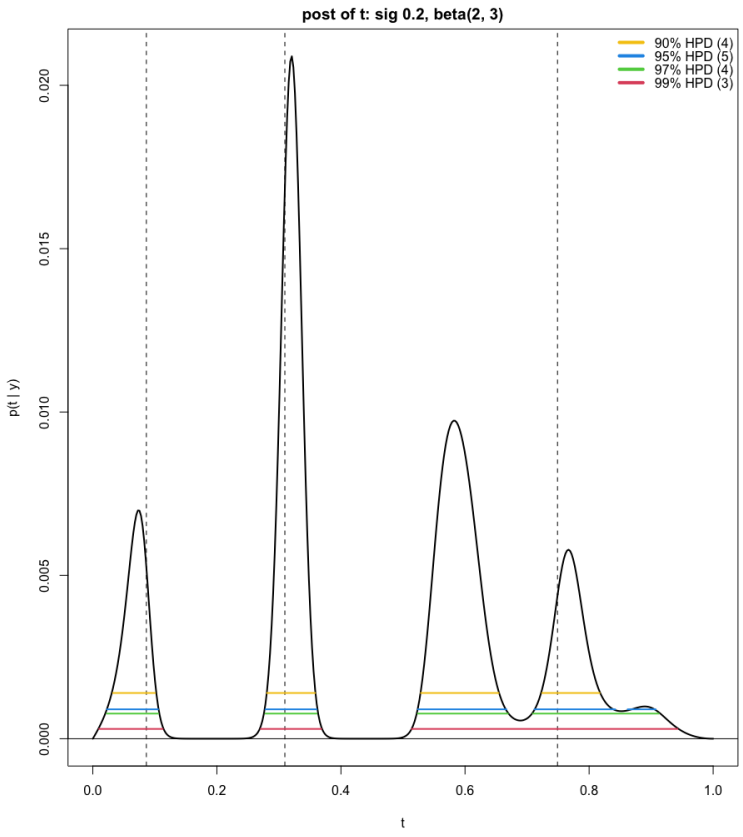


Figure 7: Effect of α on the estimated number of local extrema. The posterior density function is based on one simulated dataset with $n = 100$.

B.2 Highly fluctuated regression function with large M

Upon suggestion from one of the reviewers, we performed a new simulation using the regression function $\sin(k\pi x)$ for $x \in [0, 1]$, and assessed how the estimated number of local extrema converges to the true M. We considered $k = 10$ and $k = 100$ with varying n; with this regression function, the true number of local extrema is $M = k$. Other simulation configurations mirrored the main paper's setup, including the noise standard deviation, observed x values, and the number of replications. The proposed method is implemented using the same settings as in the simulation study in the main paper, unless otherwise stated.

We observe that when $k = 10$, our method is able to correctly estimate M 77% of the time even with sample size as small as 30. This percentage increases steadily to (93%, 99%, 100%) as n increases to $(200, 300, 500)$, respectively.

When $k = 100$, M is correctly estimated only 4% of the time when $n = 300$ (compared to 99% when $k = 10$), indicating the challenge of large $k = 100$. We have looked into this challenging scenario and found that for this highly fluctuated function, even simpler tasks such as function estimation become challenging. For example, the model struggles to distinguish between a highly fluctuated function and a flat function when $n = 300$, which is not surprising as indicated in the top plot of Figure 8. This has prompted us to find an effective strategy for this challenging function in which we incorporate the shape of the function into guided hyperparameter tuning. If we have prior knowledge that there are many local extrema, we can confine the hyperparameter searching space, ruling out some basins of the marginal likelihood that do not result in the regression shape being interested. For example, setting the upper bound when searching for (h, λ) to (0.1, 0.0001) as opposed to (10000, 10000) used in our default implementation, leads to the results reported in Table 6, which show a substantially improved estimation of M. For example, the proposed method can estimate the correct value of M with $n = 300$ in all 100 simulations. The posterior distribution of t in one simulation when $k = 100$ is shown in Figure 8. In this simulation, which is typical across 100 replications, our method correctly identifies the number and location of 100 local extrema. We acknowledge that prior information on the shape of the

unknown function might not always be available.

	70-79	80-89	90-99	100	> 100
$n = 200$	18	76	6	0	0
$n = 225$	0	1	14	85	0
$n = 250$	0	0	0	99	1
$n = 300$	0	0	0	100	0

Table 6: Frequency of $\hat{M}$ falling in each interval when $k = 100$. Results are based on 100 repeated simulations.

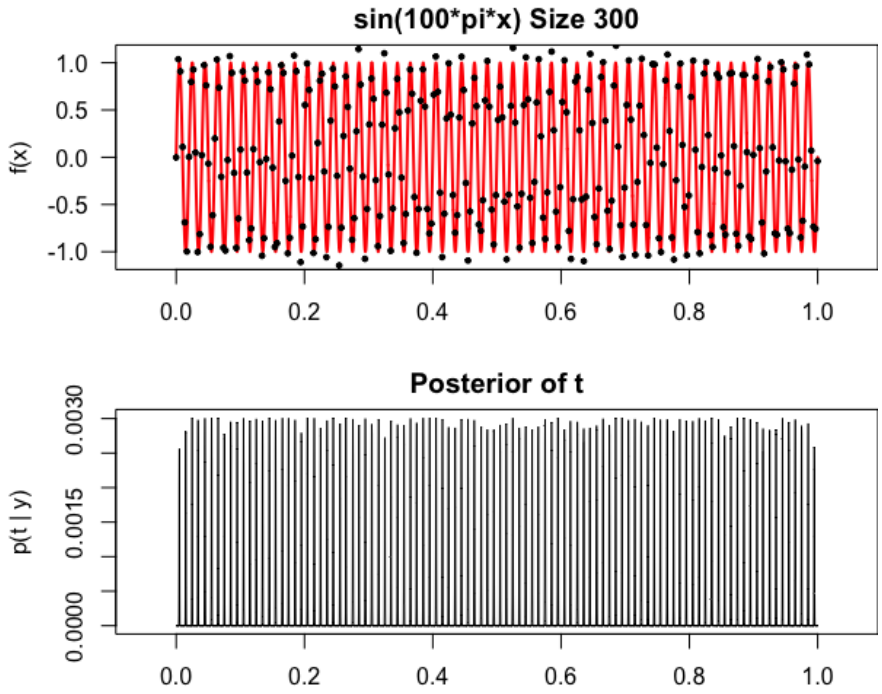


Figure 8: Data (top) and the posterior density of t (bottom) when $f(x) = \sin(100\pi x)$ (red curve in the top plot). Results are based on one simulated dataset with sample size $n = 300$.