The Annals of Statistics
2017, Vol. 45, No. 5, 2190–2217
DOI: 10.1214/16-AOS1523

BAYESIAN DETECTION OF IMAGE BOUNDARIES[1]

BY MENG LI AND SUBHASHIS GHOSAL

Duke University and North Carolina State University

Detecting boundary of an image based on noisy observations is a fundamental problem of image processing and image segmentation. For a d-dimensional image ($d = 2, 3, \ldots$), the boundary can often be described by a closed smooth $(d-1)$-dimensional manifold. In this paper, we propose a nonparametric Bayesian approach based on priors indexed by $\mathbb{S}^{d-1}$, the unit sphere in $\mathbb{R}^d$. We derive optimal posterior contraction rates for Gaussian processes or finite random series priors using basis functions such as trigonometric polynomials for 2-dimensional images and spherical harmonics for 3-dimensional images. For 2-dimensional images, we show a rescaled squared exponential Gaussian process on $\mathbb{S}^1$ achieves four goals of guaranteed geometric restriction, (nearly) minimax optimal rate adapting to the smoothness level, convenience for joint inference and computational efficiency. We conduct an extensive study of its reproducing kernel Hilbert space, which may be of interest by its own and can also be used in other contexts. Several new estimates on modified Bessel functions of the first kind are given. Simulations confirm excellent performance and robustness of the proposed method.

1. Introduction. The problem of detecting boundaries of images arises in a variety of areas including epidemiology [44], geology [26], ecology [13], forestry, marine science. A general d-dimensional ($d \geq 2$) image can be described as $(X_i, Y_i)_{i=1}^n$, where $X_i \in T = [0,1]^d$ is the location of the ith observation and Y_i is the corresponding pixel intensity. Let $f(\cdot;\phi)$ be a given regular parametric family of densities with respect to a σ-finite measure ν, indexed by a p-dimensional parameter $\phi \in \Theta$, then we assume that there is a closed region $\Gamma \subset T$ such that

$$Y_i \sim \begin{cases} f(\cdot;\xi) & \text{if } X_i \in \Gamma; \\ f(\cdot;\rho) & \text{if } X_i \in \Gamma^c, \end{cases}$$

where ξ, ρ are distinct but unknown parameters. We assume that both Γ and Γ^c have nonzero Lebesgue measures. The goal here is to recover the boundary $\gamma = \partial\Gamma$ from the noisy image where γ is assumed to be a smooth $(d-1)$-dimensional manifold without boundary, and derive the contraction rate of γ at a given true value γ_0 in terms of the metric defined by the Lebesgue measure of the symmetric difference between the regions enclosed by γ and γ_0. When the boundary itself is of

Received September 2015; revised May 2016.

[1]Research supported in part by NSF Grant DMS-11-06570.

MSC2010 subject classifications. Primary 62G20, 62H35; secondary 62F15, 60G15.

Key words and phrases. Boundary detection, Gaussian process on sphere, image, posterior contraction rate, random series, squared exponential periodic kernel, Bayesian adaptation.

interest such as in image segmentation, we can view the problem as a generalization of the change-point problem in one-dimensional data to images.

A significant part of the literature focuses on the detection of boundary pixels, based on either first-order or second-order derivatives of the underlying intensity function [32], Chapter 6, or Markov random fields [14], resulting in various edge detectors or filters. This approach is especially popular in computer vision [3, 4]. However, the detected boundary pixels are scattered all over the image and do not necessarily lead to a closed region, and hence cannot be directly used for image segmentation. A post-smoothing step can be applied, such as Fourier basis expansion, principal curves [18] or a Bayesian multiscale method proposed by [16]. However the ad-hoc two-step approach makes the theoretical study of convergence intractable. In addition, as pointed out by [2], many applications produce data at irregular spatial locations and do not have natural neighborhoods.

Most existing methods are based on local smoothing techniques [5, 17, 31, 34, 38], which lead to convenient study of theoretical properties benefiting from well-established results. However, local methods suffer when the data is sparse, and thus the global information becomes critical. More importantly, it often leads to *local* (or pointwise) inference such as marginal confidence bands losing the joint information.

A relevant and intensively studied problem is to estimate the underlying intensity function $\mathrm{E}(Y|X)$ with discontinuity at the boundary [10, 17, 29, 33, 35, 36]. These two problems are different for at least two reasons. First, there are many important applications where ξ and ρ affect $f(\cdot)$ not (or not only) in the mean but some other characteristics such as variance [5]. Second, the reconstruction of $\mathrm{E}(Y|X)$ is essentially a curve (or surface) fitting problem with discontinuity and the corresponding asymptotics are mostly on the entire intensity function rather than the boundary itself. Therefore, we may refer the latter as image denoising when boundaries are present, not necessarily guaranteeing the geometric restrictions on the boundary such as closedness and smoothness.

In this paper, we propose a nonparametric Bayesian method tailored to detect the boundary γ_0, which is viewed as a closed smooth $(d-1)$-dimensional manifold without boundary. This paper has three main contributions.

The first main contribution is that the proposed method is, to our best knowledge, the first one in the literature that achieves all the following four goals (i)–(iv) when estimating the boundary:

(i) Guaranteed geometric restrictions on the boundary such as closedness and smoothness.

(ii) Convergence at the (nearly) minimax rate [22, 28], adaptively to the smoothness of the boundary.

(iii) Possibility and convenience of joint inference.

(iv) Computationally efficient algorithm.

To address (i) and (iii), the Bayesian framework has its inherent advantages. For (i), we note that Bayesian methods allow us to put the restrictions on the boundary conveniently via a prior distribution. Specifically, we propose to use a Gaussian process (GP) prior indexed by the unit sphere in $\mathbb{R}^d$, that is, the $(d-1)$-sphere $\mathbb{S}^{d-1} = \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d : x_1^2 + \cdots + x_d^2 = 1\}$, or a random series prior on $\mathbb{S}^{d-1}$. For (iii), Bayesian methods allow for joint inference since we draw samples from the joint posterior distributions, as demonstrated by the numerical results in Section 6. The proposed method achieves the (nearly) minimax optimal rate adapting to the unknown smoothness level based on a random rescaling incorporated by a hierarchical prior [39, 43]. Furthermore, Goal (ii) is achieved for any regular family of noise and general dimensions. In contrast, for instance, the method in [28] is presented only for binary images and does not adapt to the unknown smoothness level. Although the quantification of uncertainty and adaptivity of a method is appealing, Goal (iv) of computational efficiency is important when implementing the procedure. Many adaptive methods are hard to implement since inverses of covariance matrices need to be calculated repeatedly. In the proposed Bayesian approach, an efficient Markov chain Monte Carlo (MCMC) sampling is designed based on the analytical eigendecomposition of the squared exponential periodic (SEP) kernel (see Section 5), for various noise distributions. In addition, we conduct extensive numerical studies to confirm the good performance of the proposed method and indicate that it is robust under model misspecification.

As the second main contribution, we conduct an extensive study on the reproducing kernel Hilbert space (RKHS) of the SEP Gaussian process, which is essential to obtain the optimal rate and adaptation in Goal (ii). For the most important case in applications $d = 2$, by a simple mapping, the squared exponential (SE) Gaussian process on $\mathbb{S}^1$ is equivalent to the SEP Gaussian process on $[0, 1]$ since their RKHS's are isometric (see Lemma 4.1). Recently developed theory of posterior contraction rates implies that nonparametric Bayesian procedures can automatically adapt to the unknown smoothness level using a rescaling factor via a hyperparameter in a stationary Gaussian process on $[0, 1]$ or $[0, 1]^d$ [41, 43]. Rescaled SE Gaussian process is one popular example of this kind. In contrast, the literature lacks results on the rescaling scheme and the resulting properties of the SEP Gaussian process, even though it has been implemented in many applications [27]. It may be due to the apparent similarity shared between the SEP Gaussian process and the SE Gaussian process. However, these two processes have fundamental differences because the rescaling of the argument on $\mathbb{S}^1$ cannot be transformed as a rescaling of the mapped argument on the Euclidean domain. In addition, the spectral measure of the SEP Gaussian process is discrete (see Lemma 4.2) thus lacking the absolute continuity of that of the SE Gaussian process which is critical in establishing many of its properties [43]. As a result, the RKHS of the SEP Gaussian process for different scales do not follow the usual nesting property. We overcome these issues by using the special eigenstructure of the SEP kernel

and intriguing properties of modified Bessel functions of the first kind. Some of the properties of the SE Gaussian process still hold, however, the proofs are remarkably different. Nevertheless, we show that the posterior contraction rate of the boundary by using the SEP Gaussian process is nearly minimax-optimal, which is $n^{-\alpha/(\alpha+1)}$ up to a logarithmic factor, adaptively to the smoothness level α of the boundary. Section 4 establishes a list of properties on the RKHS of the SEP Gaussian process, along with the contraction rate calculation and adaptation.

The third main contribution is that we provide some new estimates on Bessel functions, which are critical when establishing properties on the RKHS of the SEP Gaussian process. Similar to the second main contribution, these new estimates may be of interest by their own and are useful in broader contexts such as function estimation on spheres in addition to the boundary detection problem discussed here.

In addition to establishing key theoretical properties, we also develop an efficient MCMC method for sampling posterior distribution based on a SEP Gaussian process prior using the explicit eigenstructure of the SEP Gaussian process obtained in this paper [taking $O(n)$ time in each MCMC run]. The algorithm is generic, and hence can be used for posterior computation in other curve estimation problems on the circle such as directional data analysis using the SEP Gaussian process prior.

The paper is organized as follows. Section 2 introduces the model and notation. The general results on the posterior contraction rate are given in Section 3, along with examples of priors and posterior rate calculation including a finite random series prior (for $d = 2$ and 3) and the squared exponential Gaussian process prior on $\mathbb{S}^1$ (for $d = 2$). In Section 4, we study the corresponding RKHS of a squared exponential Gaussian process prior on $\mathbb{S}^1$, or equivalently, a squared exponential periodic Gaussian process on $[0, 1]$, heavily relying on the properties of modified Bessel functions of the first kind. Section 5 proposes an efficient Markov Chain Monte Carlo methods for computing the posterior distribution of the boundary using a randomly rescaled Gaussian process prior, for various noise distributions. Section 6 studies the performance of the proposed Bayesian estimator via simulations, under various settings for both binary images and Gaussian noised images. Section 7 contains proofs to all main theorems. Section 8 provides several results on the modified Bessel functions of the first kind. Proofs to all the lemmas and propositions are deferred to the supplementary materials [24].

2. Model and notation. We consider a d-dimensional image $(X_i, Y_i)_{i=1}^n$ for $d = 2, 3, \ldots$, where X_i is the location of the ith observation and Y_i is the image intensity. We consider the locations within a d-dimensional fixed size hypercube, and we specifically use unit hypercube $T = [-1/2, 1/2]^d$ without loss of generality. Depending on the scheme of collecting data, we have the following options for the distribution P_{X_i} of X_i:

- *Completely Random Design.* $X_i \overset{i.i.d.}{\sim} \text{Uniform}(T)$.
- *Jitteredly Random Design.* Let T_i be the ith block when partitioning T into equal-spaced grids. Then X_i is chosen randomly at T_i, that is, $X_i \sim \text{Uniform}(T_i)$ independently.

The region Γ is assumed to be star-shaped with a known reference point $O \in \Gamma$, namely, for any point in Γ the line segment from O to that point is in Γ; see [10] and [22], Chapter 5. If the image is star-shaped, this assumption is mild since in general a reference point can be easily detected by a preliminary estimator of the boundary or it could be directly given in many cases according to some subject matter knowledge. Images of more general shapes can possibly be addressed by an ad-hoc "divide and conquer" strategy. The boundary $\gamma = \partial\Gamma$ is a $(d-1)$-dimensional closed manifold. In view of a converse of the Jordan curve theorem, we represent the closed boundary γ as a function indexed by $\mathbb{S}^{d-1}$, that is, $\gamma : \mathbb{S}^{d-1} \to \mathbb{R}^+ : s \to \gamma(s)$. We further assume that the boundary γ is α-smooth, that is, $\gamma \in \mathbb{C}^\alpha(\mathbb{S}^{d-1})$, where $\mathbb{C}^\alpha(\mathbb{S}^{d-1})$ is the α-Hölder class on $\mathbb{S}^{d-1}$. Specifically, let α_0 be the largest integer strictly smaller than α, then

$$\mathbb{C}^\alpha\big(\mathbb{S}^{d-1}\big) = \big\{ f : \mathbb{S}^{d-1} \to \mathbb{R}^+, \big| f^{(\alpha_0)}(x) - f^{(\alpha_0)}(y) \big| \le L_f \|x - y\|^{\alpha - \alpha_0} \\ \text{for } \forall x, y \in \mathbb{S}^{d-1} \text{ and some } L_f > 0 \big\},$$

where $\|\cdot\|$ is the Euclidean distance. A different definition of smoothness was used by [28] based on the class of sets in [11], which covers cases of unsmooth boundary but with smooth parameterization. Here, we focus on the class of smooth boundary, therefore, it may be more natural to use the definition of $\mathbb{C}^\alpha(\mathbb{S}^{d-1})$ directly. It may be noted that in our setup, the boundary is not affected by reparameterization.

We use θ to denote the triplet (ξ, ρ, γ). Let ϕ_i be the parameters at the ith location, that is, $\phi_i = \xi \mathbb{1}(X_i \in \Gamma) + \rho \mathbb{1}(X_i \in \Gamma^c)$ where $\mathbb{1}(\cdot)$ is the indicator function. The model assumes that $Y|X \sim P_\theta^n$ for some θ, where P_θ^n has density $\prod_{i=1}^n p_{\theta,i}(Y_i) = \prod_{i=1}^n f(Y_i; \phi_i)$ with respect to ν^n. Let

$$d_n^2(\theta, \theta') = \frac{1}{n} \sum_{i=1}^n \int (\sqrt{p_{\theta,i}} - \sqrt{p_{\theta',i}})^2 \, d\nu$$

be the average of the squares of the Hellinger distance for the distributions of the individual observations. Let $K(f,g) = \int f \log(f/g)\, d\nu$, $V(f,g) = \int f |\log(f/g)|^2 \, d\nu$, and $\|\cdot\|_p$ denote the L_p-norm $(1 \le p \le \infty)$. We use $f \lesssim g$ if there is a universal constant C such that $f \lesssim Cg$, and $f \asymp g$ if $f \lesssim g \lesssim f$. For a vector $x \in \mathbb{R}^d$, define $\|x\|_p = \{\sum_i^d |x_i|^p\}^{1/p}$ and $\|x\|_\infty = \max_{1 \le i \le p} |x_i|$. For two sets Γ and Γ', we use $\Gamma \triangle \Gamma'$ for their symmetric difference and $\lambda(\Gamma \triangle \Gamma')$ for its corresponding Lebesgue measure. We also use $\lambda(\gamma, \gamma')$ for $\lambda(\Gamma \triangle \Gamma')$ when $\gamma = \partial\Gamma$ and $\gamma' = \partial\Gamma'$.

3. Posterior convergence. In the following sections, we shall focus on the jitteredly random design; the completely random design is more straightforward and follows the same rate calculation with minor modifications.

3.1. *General theorem.* The likelihood function is given by

$$L(Y|X,\theta) = \prod_{i\in I_1} f(Y_i;\xi) \prod_{i\in I_2} f(Y_i;\rho),$$

where $I_1 = \{i : X_i \in \Gamma\}$ and $I_2 = \{i : X_i \in \Gamma^c\}$. The parameters $(\xi, \rho) \in \Theta^*$, where Θ^* is a subset of $\Theta \times \Theta = \{(\xi, \rho) : \xi \in \Theta, \rho \in \Theta\}$. The set Θ^* is typically given as the full parameter space $\Theta \times \Theta$ with some order restriction between ξ and ρ. For instance, when $f(\cdot)$ is the Bernoulli distribution, then $\Theta^* = \{(\xi, \rho) \in \mathbb{R}^2 : 0 < \rho < \xi < 1\}$ if the inside probability ξ is believed to be larger than the outside probability. We assume that the distribution $f(\cdot)$ has the following regularity conditions:

(A1) For fixed ϕ_0, we have $K(f(\cdot;\phi_0), f(\cdot;\phi)) \lesssim \|\phi - \phi_0\|^2$ and $V(f(\cdot;\phi_0), f(\cdot;\phi)) \lesssim \|\phi - \phi_0\|^2$ as $\|\phi - \phi_0\|^2 \to 0$;

(A2) There exist constants $C_0, b_0 > 0$ such that for ϕ_1, ϕ_2 with $\|\phi_1\|, \|\phi_2\| \le M$, we have $h^2(f(\cdot;\phi_1), f(\cdot;\phi_2)) \le C_0(1 + M^{b_0})\|\phi - \phi_0\|^2$, where $h(f(\cdot;\phi), f(\cdot;\phi'))$ is the Hellinger distance between the two densities $f(\cdot;\phi)$ and $f(\cdot;\phi')$.

Assumptions (A1) and (A2) relate divergences and distances between two distributions to the Euclidean distance between the corresponding parameters. Most common distributions where the parameters are bounded away from the boundary of their supports satisfy these two assumptions, particularly including all the distribution families discussed in the paper.

The observations Y_i's are conditionally independent given parameters. In the following sections, we let θ_0 denote the true value of the parameter vector $(\xi_0, \rho_0, \gamma_0)$ generating the data, and the corresponding region with boundary γ_0 is denoted by Γ_0.

We shall denote the prior on θ by Π. By a slight abuse of notation, we denote the priors on (ξ, ρ) and γ also by Π. We next present the abstract forms of the required prior distributions in order to satisfy the minimax-optimal posterior contraction rate later on. The prior on (ξ, ρ) is independent with the prior on γ and satisfies that:

(B1) $\Pi(\xi, \rho)$ has a positive and continuous density on Θ^*;

(B2) Sub-polynomial tails: there are some constants $t_1, t_2 > 0$ such that for any $M > 0$, we have $\Pi(\xi : \xi \notin [-M, M]^p) \le t_1 M^{-t_2}$ and $\Pi(\rho : \rho \notin [-M, M]^p) \le t_1 M^{-t_2}$.

As inference on γ is of primary interest, (ξ, ρ) are considered as two nuisance parameters. When γ is modeled nonparametrically, the contraction rate for θ is primarily influenced by γ. The following condition is critical to relate $d_n(\theta, \theta')$ to $\lambda(\gamma, \gamma')$, which will lead to the contraction rate for γ.

(C) For given $(\xi_0, \rho_0) \in \Theta^*$, there exists a positive constant $c_{0,n}$ such that for arbitrary $(\xi, \rho) \in \Theta^*$, $h(f(\cdot;\xi_0), f(\cdot;\rho)) + h(f(\cdot;\rho_0), f(\cdot;\xi)) \geq c_{0,n}$.

Above we allow the constant $c_{0,n}$, which is usually $h(\xi_0, \rho_0)$, to depend on n if we consider a sequence of true values (ξ_0, ρ_0). Assumption (C) can be interpreted as quantifying the separation of the inside and outside densities in terms of the Hellinger distance. The separation becoming smaller with n indicates the increasing level of difficulty of the problem. Assumption (C) holds for most commonly used distribution families $\{f(\cdot;\phi); \phi \in \Theta\}$ when Θ^* considers the order restriction between ξ and ρ. Examples include but are not limited to:

- One-parameter family such as Bernoulli, Poisson, exponential distributions and $\Theta^* = \Theta^2 \cap \{(\xi, \rho) : \rho < \xi\}$, or $\Theta^* = \Theta^2 \cap \{(\xi, \rho) : \rho > \xi\}$.
- Two-parameter family such as Gaussian distributions, and $\Theta^* = \Theta^2 \cap \{((\mu_1, \sigma_1), (\mu_2, \sigma_2)) : \mu_1 > \mu_2, \sigma_1 = \sigma_2\}$, or $\Theta^* = \Theta^2 \cap \{((\mu_1, \sigma_1), (\mu_2, \sigma_2)) : \mu_1 > \mu_2, \sigma_1 > \sigma_2\}$, or $\Theta^* = \Theta^2 \cap \{((\mu_1, \sigma_1), (\mu_2, \sigma_2)) : \mu_1 = \mu_2, \sigma_1 > \sigma_2\}$.

The assertions above can be verified by noting that, by keeping one argument fixed, the Hellinger distance increase in the other argument in each direction as that moves away from the fixed value in terms of the Euclidean distance. In practice, the order restriction is often naturally obtained depending on the concrete problem. For instance, in brain oncology, a tumor often has higher intensity values than its surroundings in a positron emission tomography scan, while for astronomical applications objects of interest emit light and will be brighter. In this paper, we use the abstract condition (C) to provide a general framework for various relevant applications.

Throughout this paper, we shall use $h(\phi, \phi')$ to abbreviate $h(f(\cdot;\phi), f(\cdot;\phi'))$. The following general theorem gives a posterior contraction rate for parameters θ and γ.

THEOREM 3.1. *Let a sequence $\varepsilon_n \to 0$ be such that $n\varepsilon_n^2/\log n$ is bounded away from 0. Under Conditions* (A1), (A2), (B1), (B2), *if there exists Borel measurable subsets $\Sigma_n \subset \mathbb{C}^\alpha(\mathbb{S}^{d-1})$ with $\sigma_n = \sup\{\|\gamma\|_\infty : \gamma \in \Sigma_n\}$ such that*

$$-\log \Pi\big(\gamma : \lambda(\Gamma_0 \triangle \Gamma) \leq \varepsilon_n^2\big) \lesssim n\varepsilon_n^2, \tag{3.1}$$

$$-\log \Pi\big(\gamma \in \Sigma_n^c\big) \gtrsim n\varepsilon_n^2, \tag{3.2}$$

$$\log N\big(\varepsilon_n^2/\sigma_n^{d-1}, \Sigma_n, \|\cdot\|_\infty\big) \lesssim n\varepsilon_n^2, \tag{3.3}$$

then for the entire parameter $\theta = (\xi, \rho, \gamma)$, we have that for every $M_n \to \infty$,

$$\mathrm{P}_{\theta_0}^{(n)} \Pi\big(\theta : d_n(\theta, \theta_0) \geq M_n\varepsilon_n | X^{(n)}, Y^{(n)}\big) \to 0. \tag{3.4}$$

Further, if also Condition (C) *holds, then for the boundary γ, we have that for every $M_n \to \infty$,*

$$\mathrm{P}_{\theta_0}^{(n)} \Pi\big(\gamma : \lambda(\gamma, \gamma_0) \geq M_n\varepsilon_n^2/c_{0,n}^2 | X^{(n)}, Y^{(n)}\big) \to 0. \tag{3.5}$$

Equation (3.5) claims that if the rate for θ is ε_n, then the boundary γ has the rate $\varepsilon_n^2/c_{0,n}^2$ in terms of the discrepancy metric $\lambda(\cdot,\cdot)$ and can be faster than $n^{-1/2}$ which is an interesting aspect of a boundary detection problem. The condition that $n\varepsilon_n^2/\log n$ is bounded away from 0 is not restrictive since its sufficient condition $\varepsilon_n \gtrsim n^{-c}$ for some $c < 1/2$ is expected for nonparametric problems. It ensures that the parametric components with priors of sub-polynomial tails are not influential for the posterior contraction rate compared with the nonparametric part.

REMARK 3.2. It follows immediately that $\|\xi_0 - \xi\| \lesssim \varepsilon_n$ and $\|\rho_0 - \rho\| \lesssim \varepsilon_n$. These two parameters are not of our interest and are actually estimable at $n^{-1/2}$ rate. To see this, we assume the separation $c_{0,n}^2$ decays at a rate such that $\varepsilon_n^2/c_{0,n}^2 \to 0$ (particularly $c_{0,n} = c_0$ satisfies this), one may use a two-step semiparametric procedure: first, estimate the boundary curve consistently using Theorem 3.1, remove a small section of pixels neighboring the estimated boundary and then estimate (ξ, ρ) based on observations at the remaining pixels. The $n^{-1/2}$-rate possibly also holds for the original posterior of (ξ, ρ) and will follow if a semiparametric Bernstein—von Mises theorem can be established using the rate in Theorem 3.1 as a preliminary rate; see [6].

In the next two subsections, we consider two general classes of priors suitable for applications of Theorem 3.1.

3.2. *Rate calculation using finite random series priors.* The boundary γ_0 is a function on $\mathbb{S}^{d-1}$, which can be regarded also as a function on $[0,1]^{d-1}$ with periodicity restrictions. We construct a sequence of finite dimensional approximation for functions in $\mathbb{C}^\alpha([0,1]^{d-1})$ with periodicity restriction by linear combinations of the first J elements of a collection of fixed basis functions. Let $\boldsymbol{\eta} = \boldsymbol{\eta}_J = (\eta_1, \ldots, \eta_J)^T$ be the vector formed by the first J basis functions, and $\boldsymbol{\beta}_{0,J}^T\boldsymbol{\eta}$ be a linear approximation to γ_0 with $\|\boldsymbol{\beta}_{0,J}\|_\infty < \infty$. We assume that the basis functions satisfy the following condition:

(D) $\max_{1\le j\le J} \|\eta_j\|_\infty \le t_3 J^{t_4}$ for some constants $t_3, t_4 \ge 0$.

Priors. We use a random series prior for γ induced from $\boldsymbol{\beta}^T\boldsymbol{\eta}$ through the number of basis functions J and the corresponding coefficients $\boldsymbol{\beta}$ given J. For the simplicity of notation, we use $\boldsymbol{\beta}$ ($\boldsymbol{\beta}_0$) for $\boldsymbol{\beta}_J$ ($\boldsymbol{\beta}_{0,J}$) when J is explicit from the context. Let Π stand for the probability mass function (p.m.f.) of J and also for the prior of $\boldsymbol{\beta}$ satisfying the following conditions:

(E1) $-\log\Pi(J > j) \gtrsim j\log j$, and $-\log\Pi(J = j) \lesssim j\log j$.

(E2) $-\log\Pi(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 \le \varepsilon | J) \lesssim J\log(1/\varepsilon)$, and $\Pi(\boldsymbol{\beta} \notin [-M, M]^J | J) \le J\exp\{-CM^2\}$ for some constant C.

For instance, a Poisson prior on J and a multivariate normal for $\boldsymbol{\beta}$ meet the required conditions.

Finite random series priors with a random number of coefficients form a very tractable flexible class of priors offering alternative to Gaussian process priors. Their properties including asymptotic behavior of posterior distributions have been thoroughly studied by [1] and [39].

We derive conditions to obtain the posterior contraction rate as follows.

THEOREM 3.3. *Let ε_n be a sequence such that $\varepsilon_n \to 0$ and $n\varepsilon_n^2/\log n$ is bounded away from 0, and $J_n \le n$ be a sequence such that $J_n \to \infty$. Under conditions* (A1), (A2), (B1), (B2), (C), (D), (E1) *and* (E2), *if ε_n, J_n satisfy*

$$\|\gamma_0 - \boldsymbol{\beta}_{0,J_n}^T \boldsymbol{\eta}\|_\infty \le \varepsilon_n^2/2, \qquad c_1 n\varepsilon_n^2 \le J_n \log J_n \le J_n \log n \le c_2 n\varepsilon_n^2,$$

for some constants $c_1 > 0$ and $c_2 > 0$, then the posterior contraction rate for γ in terms of the distance $\lambda(\cdot,\cdot)$ is $\varepsilon_n^2/c_{0,n}^2$.

REMARK 3.4. The optimal value of J_n, say $J_{n,\alpha}$ typically depends on the degree of smoothness α. We can use a fixed value $J = J_n$ when α is given. The posterior distribution can be easily computed, for example, by a Metropolis–Hastings algorithm. If α is unknown, one will need to put a prior on J and reversible-jump MCMC may be needed for computation.

EXAMPLE 3.5 (Trigonometric polynomials). For the case $d = 2$ (2D image), we use trigonometric polynomials $\{1, \cos 2\pi j\omega, \sin 2\pi j\omega, \ldots : \omega \in [0,1]\}$ as the basis. If γ_0 is α-smooth, we have $\|\gamma_0 - \boldsymbol{\beta}_{0,j}^T \boldsymbol{\eta}\|_\infty \lesssim j^{-\alpha}$ for some appropriate choice of $\boldsymbol{\beta}_{0,j}$ (cf. [19]). Therefore, according to Theorem 3.3, we can obtain the rate ε_n by equating $J_n^{-\alpha} \asymp \varepsilon_n^2$ and $J_n \log J_n \asymp n\varepsilon_n^2$, which gives the following rate ε_n and the corresponding J_n:

$$J_n \asymp n^{1/(\alpha+1)} (\log n)^{-1/(\alpha+1)}, \qquad \varepsilon_n^2 \asymp n^{-\alpha/(\alpha+1)} (\log n)^{\alpha/(\alpha+1)}.$$

EXAMPLE 3.6 (Spherical harmonics). For 3D images ($d = 3$), periodic functions on the sphere can be expanded in the spherical harmonic basis functions. Spherical harmonics are eigenfunctions of the Laplacian on the sphere. It satisfies condition (D) and more technical details and the analytical expressions of spherical harmonics can be found in [40], Chapter 2, while MATLAB implementation is available in [12]. Let K_n be the degree of the spherical harmonics, then the number of basis functions are $J_n = K_n^2$. The approximation error for spherical harmonics is $J_n^{-\alpha/2}$ [9], Theorem 4.4.2. Therefore, we can obtain the posterior contraction rate by equating $J_n^{-\alpha/2} \asymp \varepsilon_n^2$ and $J_n \log J_n \asymp n\varepsilon_n^2$, which gives

$$J_n \asymp n^{2/(\alpha+2)} (\log n)^{-2/(\alpha+2)}, \qquad \varepsilon_n^2 \asymp n^{-\alpha/(\alpha+2)} (\log n)^{\alpha/(\alpha+2)}.$$

3.3. *Rescaled squared exponential Gaussian process prior on* $\mathbb{S}^1$. We use a rescaled squared exponential Gaussian process (GP) to induce priors on γ when $d = 2$. Specifically, let W be a GP with the squared exponential kernel function $K(t, t') = \exp(-\|t - t'\|^2)$, where $t, t' \in \mathbb{S}^1$ and $\|\cdot\|$ is the Euclidean distance. Let $W^a = (W_{at}, t \in \mathbb{S}^1)$ be the scaled GP with scale $a > 0$, whose covariance kernel becomes $K_a(t, t') = \exp(-a^2\|t - t'\|^2)$. The rescaling factor a acts as a smoothing parameter and allows us to control smoothness of a sample path from the prior distribution.

When $d = 2$, it is natural to use the map $Q : [0, 1] \to \mathbb{S}^1, \omega \to (\cos 2\pi\omega, \sin 2\pi\omega)$ as in [27], then by Lemma 4.1, the squared exponential kernel $K_a(\cdot, \cdot)$ on $\mathbb{S}^1$ has the equivalent RKHS as of the kernel $G_a(t_1, t_2)$ on $[0, 1]$ defined by

$$\begin{aligned} G_a(t_1, t_2) &= \exp\bigl(-a^2\bigl\{(\cos 2\pi t_1 - \cos 2\pi t_2)^2 + (\sin 2\pi t_1 - \sin 2\pi t_2)^2\bigr\}\bigr) \\ &= \exp\bigl\{-4a^2 \sin^2(\pi t_1 - \pi t_2)\bigr\}. \end{aligned}$$

We call $G_a(\cdot, \cdot)$ on the unit interval as *squared exponential periodic* (SEP) kernel. Theorem 3.7 gives the posterior contraction rate if a rescaled SEP Gaussian process is used as the prior.

THEOREM 3.7. *Let Conditions* (A1), (A2), (B1), (B2) *and* (C) *hold*:

(i) *Deterministic rescaling*: *If the smoothness level* α *is known, and we choose* $a = a_n = n^{1/(\alpha+1)}(\log n)^{-2/(\alpha+1)}$, *then the posterior contraction rate in Theorem* 3.1 *is determined by* $\varepsilon_n^2 = n^{-\alpha/(\alpha+1)}(\log n)^{2\alpha/(\alpha+1)}$.

(ii) *Random rescaling*: *If the rescaling factor a follows a gamma prior, then the contraction rate in Theorem* 3.1 *is determined by* $\varepsilon_n^2 = n^{-\alpha/(\alpha+1)}(\log n)^{2\alpha/(\alpha+1)}$ *for any* $\alpha > 0$.

Thus, when the underlying smoothness level α is unknown, the SEP Gaussian process prior can adapt to α in a hierarchical Bayesian approach by assigning the rescaling parameter an appropriate prior such as a gamma distribution [43].

The proof to Theorem 3.7 relies on an extensive study of the RKHS of the rescaled SEP Gaussian process (see Section 4). We also obtain the eigenstructure of the SEP Gaussian process analytically, leading to efficient MCMC method for posterior sampling for various distribution families (see Section 5).

REMARK 3.8. The rates obtained in Examples 3.5, 3.6 and Theorem 3.7 are optimal in the minimax sense up to a logarithmic factor; see [22], Chapter 7. By a uniform strengthening of Theorem 3 of [15], the conclusion can be strengthened to uniform in $(\xi_0, \rho_0, \gamma_0) \in \Theta_0$ if on Θ_0 Assumptions (A1) and (C) hold uniformly and $\|\gamma_0\|_\infty \le C_0$ for a universal constant $C_0 > 0$.

3.4. *Rescaled Gaussian process with the heat kernel.* For a Gaussian process prior, various kernels may be used in addition to the squared exponential kernel, for example, the heat kernel Gaussian processes studied by [7]. Without introducing the technical details on heat kernel theory, a Gaussian process with the heat kernel on $\mathbb{S}^{d-1}$ can be represented as

$$W^T(x) = \sum_{k=0}^{\infty} \sum_{l=1}^{N_k(d)} e^{-\lambda_k T/2} Z_{k,l} e_{k,l}(x),$$

for any $x \in \mathbb{S}^{d-1}$, where $Z_{k,l}$ are independent standard normal variables indexed by (k,l), and $\{e_{k,l}(\cdot), 1 \le l \le N_k(d)\}$ is an orthonormal basis of the space of spherical harmonics of order k whose eigenvalue is λ_k and dimension is $N_k(d) = (2k + d - 2)\binom{d+k-3}{k}/(d-2)$. Starting from eigendecomposition of the heat kernel, the rescaling is applied directly to the eigenvalues λ_k. Note that this is a different rescaling strategy compared to the SEP Gaussian process where the rescaling is applied to the distances between two points. If we randomly rescale the process by letting T follow a gamma prior, then similar results as in Theorem 3.7 hold (possibly with a different logarithmic factor) based on the study of RKHS of $W^T(\cdot)$ available in [7] which are parallel to Lemma 4.4 to Lemma 4.8, following the argument in proving Theorem 3.7.

4. RKHS of SEP Gaussian processes. The RKHS of a GP plays a critical role in calculating the posterior contraction rate. There has been an extensive study of the RKHS of a GP indexed by $[0,1]^{d-1}$ (e.g., [41, 43]). A GP indexed by $\mathbb{S}^{d-1}$ can be naturally related to a GP indexed by $[0,1]^{d-1}$ by a surjection $Q : [0,1]^{d-1} \to \mathbb{S}^{d-1}$ (e.g., using the spherical coordinate system). Define the following kernels on $[0,1]^{d-1}$: $G(s_1, s_2) = K(Qs_1, Qs_2)$ for any $s_1, s_2 \in [0,1]^{d-1}$. Let $\mathbb{H}$ be the RKHS of the GP defined by the kernel K, equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and the RKHS norm $\| \cdot \|_{\mathbb{H}}$. For the GP with covariance kernel G, we denote the RKHS, its inner product and norm by $\mathbb{H}'$, $\langle \cdot, \cdot \rangle_{\mathbb{H}'}$ and $\| \cdot \|_{\mathbb{H}'}$, respectively. Then the following lemma shows that the two RKHSs related by the map Q are isomorphic.

LEMMA 4.1. $(\mathbb{H}', \| \cdot \|_{\mathbb{H}'})$ *and* $(\mathbb{H}, \| \cdot \|_{\mathbb{H}})$ *are isometric*; $(\mathbb{H}', \| \cdot \|_{\infty})$ *and* $(\mathbb{H}, \| \cdot \|_{\infty})$ *are also isometric*.

However, if K is the squared exponential kernel on $\mathbb{S}^{d-1}$, the kernel $G(\cdot, \cdot)$ is no longer a squared exponential kernel on $[0,1]^{d-1}$. More importantly, it is not even stationary for general $d > 2$. The case $d = 2$ is an exception, for which the RKHS can be studied via an explicit treatment such as analytical eigendecompositions of its equivalent kernel on the unit interval. We next focus on the case $d = 2$, and study the RKHS of a GP $W^a = \{W_t^a : t \in [0,1]\}$ with the SEP kernel $G_a(\cdot, \cdot)$ which was used in Section 3.3.

The SEP kernel $G_a(\cdot,\cdot)$ is stationary since $G_a(t_1,t_2)=\phi_a(t_1-t_2)$, where $\phi_a(t)=\exp\{-4a^2\sin^2(\pi t)\}$. The following result gives the explicit form of the spectral measure μ_a of the process W_t^a. Let δ_x be the Kronecker delta function and $I_n(x)$ be the modified Bessel function of the first kind with order n and argument x where $n\in\mathbb{Z}$ and $x\in\mathbb{R}$.

LEMMA 4.2. *We have $\phi_a(t)=\int e^{-its}\,d\mu_a(s)$, where μ_a is a symmetric and finite measure and given by $\mu_a=\sum_{n=-\infty}^{\infty}e^{-2a^2}I_n(2a^2)\delta_{2\pi n}$.*

In addition, the Karhunen–Loève expansion of the covariance kernel is $G_a(t,t')=\sum_{k=1}^{\infty}v_k(a)\psi_k(t)\psi_k(t')$, where the eigenvalues are given by

$$v_1(a)=e^{-2a^2}I_0\big(2a^2\big),\qquad v_{2j}(a)=v_{2j+1}(a)=e^{-2a^2}I_j\big(2a^2\big),\qquad j\geq 1,$$

with eigenfunctions $\psi_j(t)$, $j=1,2,\ldots$ given by the Fourier basis functions $\{1,\cos 2\pi t,\sin 2\pi t,\ldots\}$ in that order.

The measure μ_a is the so-called spectral measure of W^a. In the existing literature [41, 43], posterior contraction properties of rescaled GP on $[0,1]^{d-1}$ were studied relying on the absolute continuity of the spectral measure and the scaling relation $\mu_a(B)=\mu_1(aB)$. However, Lemma 4.2 shows that the spectral measure of an SEP Gaussian process is discrete and the simple relationship $\mu_a(B)=\mu_1(aB)$ does not hold. We instead heavily use properties of modified Bessel functions to study the RKHS of an SEP Gaussian process.

Note that the discrete measure μ_a has subexponential tails since

$$\int e^{|s|}\mu_a(ds)=\sum_{n=-\infty}^{\infty}e^{2\pi|n|}e^{-2a^2}I_n\big(2a^2\big)\leq 2e^{-2a^2}\sum_{n=0}^{\infty}e^{2\pi n}I_n\big(2a^2\big)$$

which is bounded by $2e^{a^2(e^{2\pi}+e^{-2\pi}-2)}<\infty$ [Proposition 8.1(a)]. The following Lemma 4.3 describes the RKHS $\mathbb{H}^a$ of the rescaled process W^a, as real parts of a closed set containing complex valued functions.

LEMMA 4.3. *The RKHS $\mathbb{H}^a$ of the process W^a is the set of real parts of all functions in*

$$\Bigg\{h:[0,1]\to\mathbb{C},\,h(t)=\sum_{n=-\infty}^{\infty}e^{-it2\pi n}b_{n,a}e^{-2a^2}I_n\big(2a^2\big),$$
$$b_{n,a}\in\mathbb{C},\,\sum_{n=-\infty}^{\infty}|b_{n,a}|^2e^{-2a^2}I_n\big(2a^2\big)<\infty\Bigg\},$$

and it is equipped with the squared norm

$$\|h\|_{\mathbb{H}^a}^2=\sum_{n=-\infty}^{\infty}|b_{n,a}|^2e^{-2a^2}I_n\big(2a^2\big)=\sum_{n=-\infty}^{\infty}\frac{1}{e^{-2a^2}I_n(2a^2)}\left|\int_0^1 h(t)e^{it2\pi n}\,dt\right|^2.$$

We then study the approximation property of $\mathbb{H}^a$ to an arbitrary smooth function $w \in \mathbb{C}^\alpha[0,1]$. Unlike the approach approximating w by a convolution of w_0 with a smooth function as used in [41, 43], we use a finite Fourier approximation to w.

LEMMA 4.4. *For any function $w \in \mathbb{C}^\alpha[0,1]$, there exist constants C_w and D_w depending only on w such that $\inf\{\|h\|_{\mathbb{H}^a}^2 : \|w-h\|_\infty \le C_w a^{-\alpha}\} \le D_w a$, as $a \to \infty$.*

Lemma 4.5 obtains an entropy estimate using Proposition 8.3 on modified Bessel functions.

LEMMA 4.5. *Let $\mathbb{H}_1^a$ be the unit ball of the RKHS of the process $W^a = (W_t^a : 0 \le t \le 1)$, then we have $\log N(\varepsilon, \mathbb{H}_1^a, \|\cdot\|_\infty) \lesssim \max(a,1) \cdot \{\log(1/\varepsilon)\}^2$.*

As a corollary of Lemma 4.5, using the connection between the entropy of the unit ball of the RKHS and the small ball probability [23, 25], we have the following estimate of the small ball probability.

LEMMA 4.6 (Lemma 4.6 in [43]). *For any $a_0 > 0$, there exits constants C and ε_0 that depend only on a_0 such that, for $a \ge a_0$ and $\varepsilon \le \varepsilon_0$,*

$$-\log \mathrm{P}\Big(\sup_{0\le t\le 1} |W_t^a| \le \varepsilon\Big) \le Ca\left(\log \frac{a}{\varepsilon}\right)^2.$$

The proof of Theorem 3.7(ii) needs a nesting property of the RKHS of W^a for different values of a. Lemma 4.7 in [43] proved that $\sqrt{a}\mathbb{H}_1^a \subset \sqrt{b}\mathbb{H}_1^b$ if $a \le b$ for a squared exponential GP indexed by $[0,1]^{d-1}$. For the SEP Gaussian process prior, this does not hold but can be modified up to a global constant.

LEMMA 4.7. *If $a \le b$, then $\sqrt{a}\mathbb{H}_1^a \subset \sqrt{cb}\mathbb{H}_1^b$ for a universal constant c.*

When $a \downarrow 0$, sample paths of W^a tend to concentrate near a constant value by the following lemma. This property is crucial in controlling the variation of sample paths for small a.

LEMMA 4.8. *For $h \in \mathbb{H}_1^a$, we have $|h(0)| \le 1$ and $|h(t) - h(0)| \le 2\sqrt{2\pi a t}$ for every $t \in [0,1]$.*

5. Sampling algorithms. We assume that the origin (center of the image) is inside the boundary, and thus use it as the reference point to represent the observed image in a polar coordinate system as $(\boldsymbol{\omega}, \boldsymbol{r}; \boldsymbol{Y})$, where $(\boldsymbol{\omega}, \boldsymbol{r}) = \{(\omega_i, r_i)\}_{i=1}^n$ are the locations using polar coordinates and $\boldsymbol{Y} = \{Y_i\}_{i=1}^n$ are the image intensities. Let γ be a closed curve, and $\boldsymbol{\gamma}$ be values of γ evaluated at each $\boldsymbol{\omega}$.

For most kernels, the eigenfunctions and eigenvalues are challenging to obtain although there are several exceptions ([37], Chapter 4.3). Therefore, a randomly rescaled GP prior may be infeasible in practice since the numerical inversion of a covariance matrix is often needed when no analytical forms are available. However, thanks to the analytical eigendecomposition the SEP kernel in Lemma 4.2, we can implement this theoretically appealing prior in an computationally efficient way. If a curve $\gamma(\omega) \sim \mathrm{GP}(\mu(\omega), G_a(\cdot,\cdot)/\tau)$, we then have the equivalent representation $\gamma(\omega) = \mu(\omega) + \sum_{k=1}^{\infty} z_k \psi_k(\omega)$, where $z_k \sim N(0, v_k(a)/\tau)$ independently.

The modified Bessel function of the first kind used in $v_k(a)$'s is a library function in most software, such as `besselI` in `R` language. Figure 1(a) shows that eigenvalues decay very fast when $a = 1, 10$. When a increases, the smoothness level of the kernel decreases. In practice, we typically do not use values as large as 100 since then the kernel becomes very close to the identity matrix, and thus the resulting prior paths become very rough. The fast decay rate of $v_k(a)$ for fixed a [Proposition 8.1(c3)] guarantees that some suitable finite order truncation of the Karhunen–Loève expansion is able to approximate the kernel function well. If we use $L = 2J + 1$ basis functions, then the truncated process is given by $\gamma(\omega) = \sum_{k=1}^{L} z_k \psi_k(\omega) + \mu(\omega)$. Let $\mathrm{PVE}_a = \sum_{k=1}^{L} v_k(a) / \sum_{k=1}^{\infty} v_k(a)$ be the percentage of variance explained by the first L basis functions, where the denom-

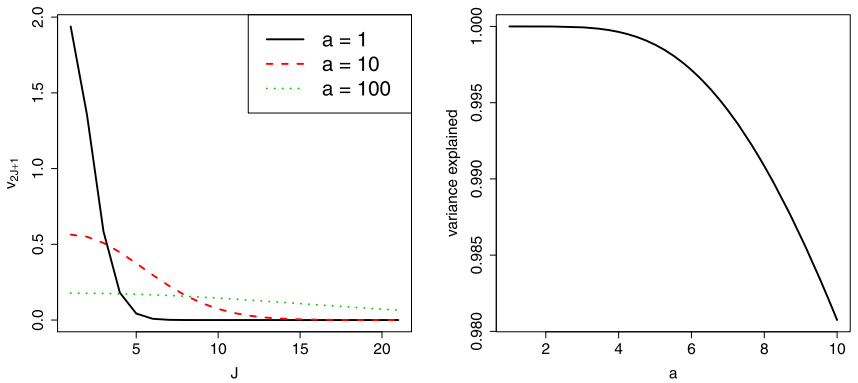


FIG. 1. *Decay rate of the eigenvalues of the squared exponential kernel. Figure* (a) *plots the values of* $v_{2J+1}(a)$ *at* $J = 0, \ldots, 20$ *when* $a = 1, 10, 100$. *Figure* (b) *plots the percentage of the variance explained* (*i.e.*, PVE_a) *if using the first* 21 ($J = 10$) *basis functions at different values of* a.

inator $\sum_{k=1}^{\infty} v_k(a) = e^{-2a^2} \sum_{k=-\infty}^{\infty} I_k(2a^2) = 1$ according to the definition of v_k in Lemma 4.2 and the properties of modified Bessel functions of the first kind obtained in Proposition 8.1. Figure 1(b) shows that with $J = 10$, we are able to explain at least 98% of all the variability for a reasonable range of a's from 0 to 10.

Let $\boldsymbol{\Psi}$ be the n by L matrix with the kth column comprising of the evaluations of $\psi_k(\cdot)$ at the components of $\boldsymbol{\omega}$, and $\boldsymbol{\mu}$ comprising of the evaluations of $\mu(\cdot)$ at the components of $\boldsymbol{\omega}$. Then the Gaussian process prior for the boundary curve can be expressed as

$$\boldsymbol{\gamma} = \boldsymbol{\Psi} z + \boldsymbol{\mu}; \qquad z \sim N(0, \boldsymbol{\Sigma}_a/\tau),$$

where $\boldsymbol{\Sigma}_a = \operatorname{diag}(v_1(a), \ldots, v_L(a))$. We use the following priors for the hyper-parameters involved in the covariance kernel: $\tau \sim \text{Gamma}(\alpha_\tau = 500, \beta_\tau = 1)$ and $a \sim \text{Gamma}(\alpha_a = 2, \beta_a = 1)$. For the mean $\mu(\cdot)$, we use a constant 0.1. Note that here we also can use the empirical Bayes approach to estimate the prior mean by any ordinary one-dimensional change-point method with an extra step of smoothing. However, our numerical investigation shows that our method is robust with respect to the specification of $\mu(\cdot)$.

The priors for (ξ, ρ) depend on the error distributions. We also need to use the order information between the parameters to keep the two regions distinguishable. We use OIB, OIN, OIG for ordered independent beta, normal and gamma distributions, respectively. If not specified explicitly, the parameters are assumed to be in a decreasing order. It is easy to see that this convention is for simplicity of notation, and any order between the two region parameters are allowed in practice. Below we give the conjugate priors for (ξ, ρ) for some commonly used noise distributions:

- Binary images: the parameters are the probabilities $(\pi_1, \pi_2) \sim \text{OIB}(\alpha_1, \beta_1, \alpha_1, \beta_1)$;
- Gaussian noise: the parameters are the mean and standard deviation $(\mu_1, \sigma_1, \mu_2, \sigma_2)$ with prior distributions $(\mu_1, \mu_2) \sim \text{OIN}(\mu_0, \sigma_0^2, \mu_0, \sigma_0^2)$ and $(\sigma_1^{-2}, \sigma_2^{-2}) \sim \text{OIG}(\alpha_2, \beta_2, \alpha_2, \beta_2)$.
- Poisson noise: the parameters are the rates $(\lambda_1, \lambda_2) \sim \text{OIG}(\alpha_3, \beta_3, \alpha_3, \beta_3)$;
- Exponential noise: the parameters are the rates $(\lambda_1, \lambda_2) \sim \text{OIG}(\alpha_4, \beta_4, \alpha_4, \beta_4)$.

In fact, any error distributions with conjugacy properties conditionally on the boundary can be directly used. We specify the hyper-parameters such that the corresponding prior distributions are spread out. For example, in the simulation, we use $\alpha_1 = \beta_1 = 0$ for binary images; we use $\mu_0 = \bar{y}, \sigma_0 = 10^3$ and $\alpha_2 = \beta_2 = 10^{-2}$ for Gaussian noise.

We use the slice sampling technique [30] within the Gibbs sampler to draw samples from the posterior distribution for (z, ξ, ρ, τ, a). Below is a detailed description of the sampling algorithms for binary images:

1. Initialize the parameters to be $z = \mathbf{0}$, $\tau = 500$ and $a = 1$. The parameters $(\xi, \rho) = (\pi_1, \pi_2)$ are initialized by the maximum likelihood estimates (MLE) given the boundary to be $\mu(\cdot)$.

2. $z|(\pi_1, \pi_2, \tau, a, \boldsymbol{Y})$: the conditional posterior density of z (in a logarithmic scale and up to an additive constant) is equal to

$$\sum_{i\in I_1} \log f(Y_i; \pi_1) + \sum_{i\in I_2} \log f(Y_i; \pi_2) - \frac{\tau z^T \boldsymbol{\Sigma}_a^{-1} z}{2}$$

$$= N_1 \log \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} + n_1 \log \frac{1-\pi_1}{1-\pi_2} - \frac{\tau z^T \boldsymbol{\Sigma}_a^{-1} z}{2},$$

where $n_1 = \sum_i \mathbb{1}(r_i < \boldsymbol{\gamma}_i)$ and $N_1 = \sum_i \mathbb{1}(r_i < \boldsymbol{\gamma}_i) Y_i$. We use slice sampling of one-coordinate-at-a-time for this step.

3. $\tau|(z, a) \sim \text{Gamma}(\alpha_\tau^*, \beta_\tau^*)$, where $\alpha_\tau^* = \alpha_\tau + L/2$ and $\beta_\tau^* = \beta_\tau + z^T \boldsymbol{\Sigma}_a^{-1} z/2$;

4. $(\pi_1, \pi_2)|(z, \boldsymbol{Y}) \sim \text{OIB}(\alpha_1 + N_1, \beta_1 + n_1 - N_1, \alpha_1 + N_2, \beta_1 + n_2 - N_2)$, where N_2 is the count of 1's outside γ and n_2 is the number of observations outside γ.

5. $a|z, \tau$: use slice sampling by noting that the conditional posterior density of a (in a logarithmic scale and up to an additive constant) is

$$-\log|\boldsymbol{\Sigma}_a|/2 - \tau z^T \boldsymbol{\Sigma}_a^{-1} z/2 + (\alpha_a - 1)\log a - \beta_a$$

$$= -\sum_{k=1}^{L} \log v_k(a)/2 - \sum_{k=1}^{L} \tau z_k^2 / \{2 v_k(a)\} + (\alpha_a - 1)\log a - \beta_a.$$

The above algorithm is generic beyond binary images. For other noise distributions, the update of τ and a are the same. The update of z and (ξ, ρ) in Step 2 and Step 4 will be changed using the corresponding priors and conjugacy properties. For example, for Gaussian noise, the parameters (ξ, ρ) are $(\mu_1, \sigma_1, \mu_2, \sigma_2)$, and the conditional posterior density (in the logarithmic scale and up to an additive constant) used in Step 2 is changed to

$$-n_1(\log\sigma_1 - \log\sigma_2) - \sum_{i\in I_1} \frac{(y_i - \mu_1)^2}{2\sigma_1^2} - \sum_{i\in I_2} \frac{(y_i - \mu_2)^2}{2\sigma_2^2} - \frac{\tau z^T \boldsymbol{\Sigma}_a^{-1} z}{2}.$$

For Step 4, the conjugacy step is changed to

$$(\mu_1, \mu_2)|(z, \sigma_1, \sigma_2, \boldsymbol{Y}) \sim \text{OIN}, \qquad (\sigma_1^{-2}, \sigma_2^{-2})|(z, \mu_1, \mu_2, \boldsymbol{Y}) \sim \text{OIG}.$$

Similarly, it is straightforward to apply this algorithm to images where the noise follows Poisson, exponential or another distribution family with ordered conjugate prior.

6. Simulations.

6.1. *Numerical results for binary images.* We use jitteredly random design for locations $(\boldsymbol{\omega}, \boldsymbol{r})$ and three cases for boundary curves:

- Case B1. Ellipse given by $r(\omega) = b_1 b_2 / \sqrt{(b_2 \cos\omega)^2 + (b_1 \sin\omega)^2}$, where $b_1 \geq b_2$ and ω is the angular coordinate measured from the major axis. We set $b_1 = 0.35$ and $b_2 = 0.25$.
- Case B2. Ellipse with shift and rotation: centered at (0.1, 0.1) and rotated by 60° counterclockwise. We use this setting to investigate the influence of the specification of the reference point.
- Case B3. Regular triangle centered at the origin with the height to be 0.5. We use this setting to investigate the performance of our method when the true boundary is not smooth at some points.

We keep using $\pi_2 = 0.2$ and vary the values of π_1 to be $(0.5, 0.25)$. For each combination of (π_1, π_2), the observed image is $m \times m$ where $m = 100, 500$ (therefore the total number of observations is $n = m^2$). The MCMC procedure is iterated 5000 times after 1000 steps burn-in period. For the estimates, we calculate the Lebesgue error (area of mismatched regions) between the estimates and the true boundary. For the proposed Bayesian approach, we use the posterior mean as the estimate and construct a variable-width uniform credible band. Specifically, let $\{\gamma_i(\omega)\}_{1000}^{5000}$ be the posterior samples and $(\widehat{\gamma}(\omega), \widehat{s}(\omega))$ be the posterior mean and standard deviation functions derived from $\{\gamma_i(\omega)\}$. For each MCMC run, we calculate the distance $u_i = \|(\gamma_i - \widehat{\gamma})/s\|_\infty = \sup_\omega\{|\gamma_i(\omega) - \widehat{\gamma}(\omega)|/\widehat{s}(\omega)\}$ and obtain the 95th percentile of all the u_i's, denoted as L_0. Then a 95% uniform credible band is given by $[\widehat{\gamma}(\omega) - L_0\widehat{s}(\omega), \widehat{\gamma}(\omega) + L_0\widehat{s}(\omega)]$.

We compare the proposed approach with a maximum contrast estimator (MCE) which first detects boundary pixels followed by a post-smoothing via a penalized Fourier regression. In the 1-dimensional case, the MCE selects the location which maximizes the differences of the parameter estimates at the two sides. This is similar to many pixel boundary detection algorithms discussed in [32]. In images, for a selected number of angles (say 1000 equal-spaced angles from 0 to 2π), we choose the neighboring bands around each angle and apply MCE to obtain the estimated radius and then smooth those estimates via a penalized Fourier regression. Note that unlike the proposed Bayesian approach, a joint confidence band is not conveniently obtainable for the method of MCE, due to its use of a two-step procedure.

As indicated by Table 1, the proposed Bayesian method has Lebesgue errors typically less than 2.5%. In addition, the proposed method outperforms the benchmark method MCE significantly. We also observe that the MCE method is highly affected by the number of basis functions; in contrast, the proposed Bayesian method adapts to the smoothness level automatically. The comparison between Case B1 and Case B2 shows that the specification of the reference point will not

TABLE 1
Lebesgue errors ($\times 10^{-2}$) *of the methods based on* 100 *simulations. The standard errors are reported below in the parentheses*

	$m = 100$, $(\pi_1, \pi_2) = (0.50, 0.20)$			$m = 100$, $(\pi_1, \pi_2) = (0.25, 0.20)$		
	Case B1	**Case B2**	**Case B3**	**Case B1**	**Case B2**	**Case B3**
Bayesian method	0.64	0.67	2.26	0.71	0.8	2.36
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
MCE with 5 bases	6.57	6.58	6.03	6.39	10.09	7.03
	(0.25)	(0.21)	(0.07)	(0.19)	(0.20)	(0.11)
MCE with 31 bases	8.75	7.84	5.96	9.19	11.8	7.86
	(0.18)	(0.19)	(0.10)	(0.16)	(0.19)	(0.14)

influence the performance of our methods since the differences are not significant compared with the standard error.

Figure 2 confirms the superior performance of the proposed method compared with the smoothed MCE method with 5 and 31 basis functions when the true boundary curve is an ellipse, an ellipse with shift and rotation and a triangle. Even in the case of $\pi_1 = 0.25$ where the contrast at two sides of the boundary is small, the proposed method is still able to capture the boundary when $m = 500$. This observation is consistent with the result derived from the infill asymptotics when the number of data points increase within images of fixed size. In addition, we also obtain joint credible bands using the samples drawn from the joint posterior distribution. Figure 3 plots the trace plots of (a, π_1, π_2) and the histogram of a to illustrate the mixing and convergence of the posterior samples for Case B1 when $m = 500$ and the true parameters $(\pi_1, \pi_2) = (0.25, 0.20)$. Similar plots are obtained for other scenarios but are not presented here.

6.2. *Numerical results for Gaussian noised images*. For Gaussian noised images, we keep using an ellipse with shift and rotation as the true boundary curve (i.e., Case B2). We consider the following scenarios where the two standard deviations are all given by $(\sigma_1, \sigma_2) = (1.5, 1)$ and the observed image is 100×100:

- Case G1. $\mu_1 = 4$, $\mu_2 = 1$, that is, distributions in the two regions differ in both of the first two moments;
- Case G2. $\mu_1 = \mu_2 = 1$, that is, distributions in the two regions only differ in the standard deviation;
- Case G3. (μ_1, μ_2) are functions of the location. Let r^I be the smallest radius inside the boundary, and r^O the largest radius outside the boundary. We use $\mu(i)$ for the mean of Y_i and let $\mu(i) = r_i - r^I + 0.2$ if it is inside, while $\mu(i) = r_i + r^O$ if outside. Therefore, the mean values vary at each location but have a gap of 0.2 between the two regions.

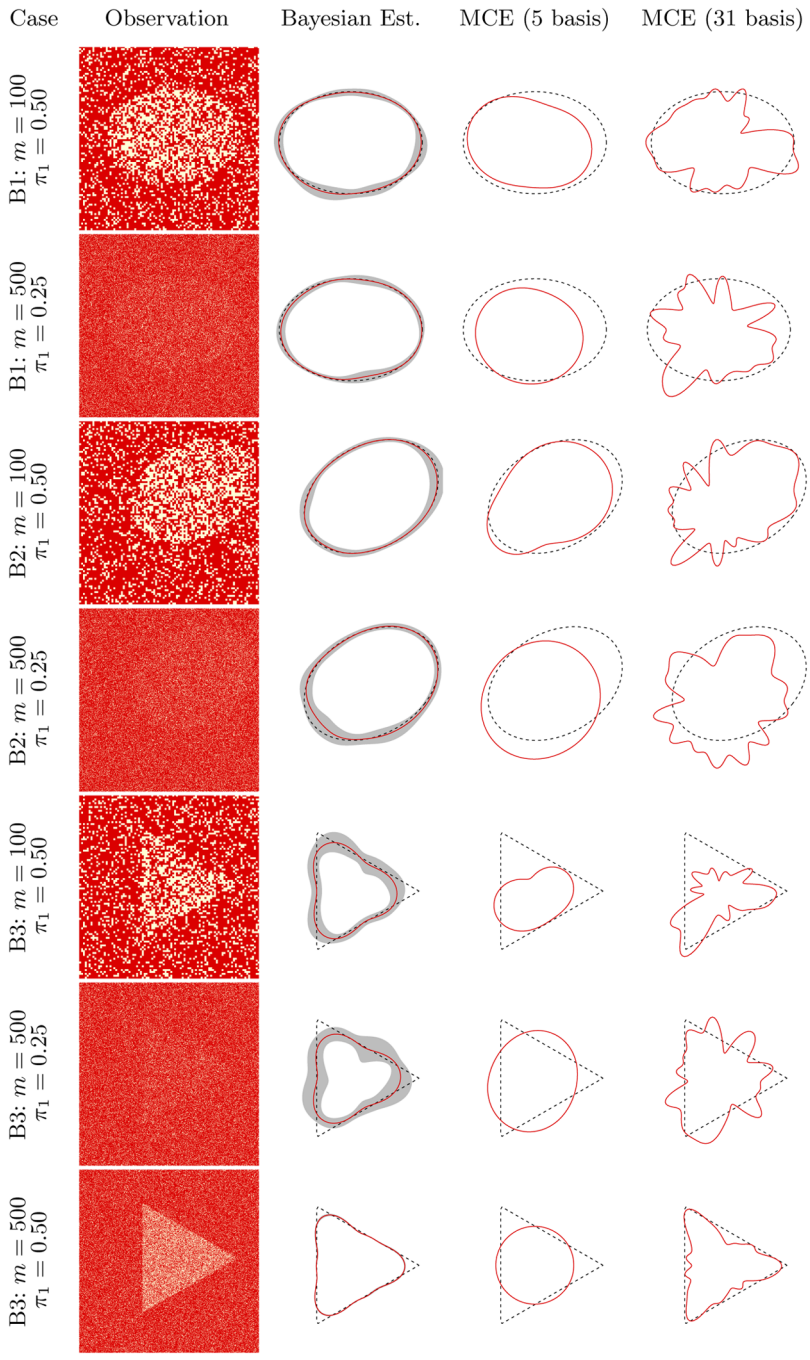
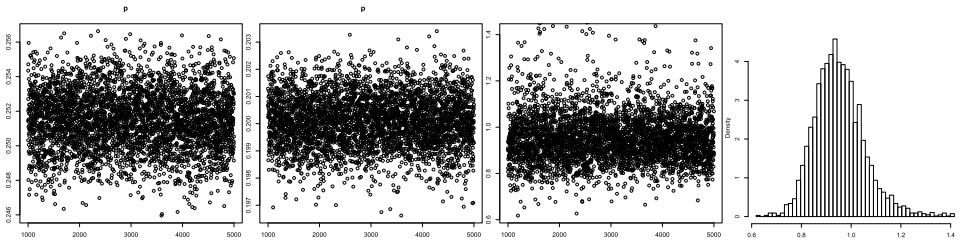


FIG. 2. *Performance on binary images* (*Column* 1: *observations*) *in various cases. Columns* 2–4 *plot the estimate* (*solid line in red*) *against the true boundary* (*dotted line in black*). *A* 95% *uniform credible band* (*in gray*) *is provided for the Bayesian estimate* (*Column* 2).



(a) Trace plot of π_1 (b) Trace plot of π_2 (c) Trace plot of a (d) Histogram of a

FIG. 3. *Trace plots and histograms of posterior samples of* (π_1, π_2, a) *for Case* B1 *when* $m = 500$ *and* $\pi_1 = 0.25$.

- Case G4. We use mixture normal distribution $0.6\mathrm{N}(2, \sigma_1^2) + 0.4\mathrm{N}(1, \sigma_2^2)$ for the inside distribution; the outside distribution is still Gaussian with mean $\mu_2 = 1$.

Cases G3 and G4 allow us to investigate the performance of the proposed method when the distribution $f(\cdot)$ in the model is misspecified. For comparison, we use a 1-dimensional change-point detection algorithm [8, 21] via the `R` package `changepoint` [20]. For the post-smoothing step, we use a penalized Fourier regression with 5 and 31 basis functions (method CP5 and CP31 in Table 2). Here, we use the estimates of CP5 as the mean in the Gaussian process prior. Table 2 shows that the proposed method has good performance for all the four cases. The method of CP5 and CP10 produce small errors in Case G1, but suffer a lot from the other three cases. It shows that the change-point method highly depends on the distinction between the means (Case G2), and also it loses its way when the model is misspecified. In fact, for Cases G2, G3 and G4, the CP5 and CP31 methods lead to a curve almost containing the whole frame of the image. The proposed Bayesian approach which models the boundary directly, seems to be not affected even when the model is substantially misspecified (Case G3). Figure 4 shows the noisy observation and our estimation from 1 replication for all the four cases. We can see the impressive performance of the proposed method. It also shows that the

TABLE 2

Performance of the methods for Gaussian noised images based on 100 *simulations. The Lebesgue error* $(\times 10^{-2})$ *between the estimated boundary the true boundary is presented. The maximum standard errors of each column are reported in the last row*

	Case G1	**Case G2**	**Case G3**	**Case G4**
Bayesian Method	0.11	0.99	0.69	0.99
CP5	2.90	62.91	62.2	61.12
CP31	1.99	64.00	63.26	62.10
SE	0.01	0.26	0.19	0.27

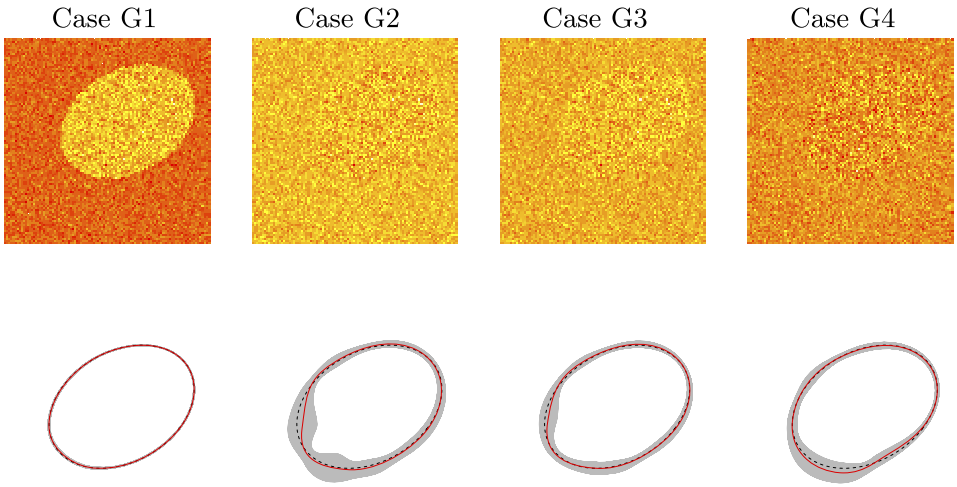


FIG. 4. *Proposed Bayesian estimates for Gaussian noised images with elliptic boundary. The first row plots the noisy observations, while the second row is the corresponding estimate (solid line in red) against the true boundary (dotted line in black), with a 95% uniform credible band (in gray).*

contrast between the two regions are visible for Cases G3 and G4, and the proposed method is capable of capturing the boundary even though the distributions are misspecified.

7. Proofs to main theorems.

PROOF OF THEOREM 3.1. *Step 1: Prior concentration.* Let

$$B_n^*(\theta_0,\varepsilon)=\left\{\theta:\frac{1}{n}\sum_{i=1}^n K_i(\theta_0,\theta)\le\varepsilon^2,\frac{1}{n}\sum_{i=1}^n V_i(\theta_0,\theta)\le\varepsilon^2\right\},$$

where $K_i(\theta_0,\theta)=K(P_{\theta_0,i},P_{\theta,i})$ and $V_i(\theta_0,\theta)=V(P_{\theta_0,i},P_{\theta,i})$. When $\|\xi-\xi_0\|\le\varepsilon^2$ and $\|\rho-\rho_0\|\le\varepsilon^2$ for some small ε, it follows that

$$\begin{aligned}K_i(\theta_0,\theta)&=K(\xi_0,\xi)P(X_i\in\Gamma_0\cap\Gamma)+K(\rho_0,\rho)P\big(X_i\in\Gamma_0^c\cap\Gamma^c\big)\\&\quad+K(\xi_0,\rho)P\big(X_i\in\Gamma_0\cap\Gamma^c\big)+K(\rho_0,\xi)P\big(X_i\in\Gamma_0^c\cap\Gamma\big)\\&\lesssim\|\xi_0-\xi\|^2+\|\rho_0-\rho\|^2+P\big(X_i\in\Gamma_0^c\cap\Gamma\big)+P\big(X_i\in\Gamma_0\cap\Gamma^c\big)\\&=\|\xi_0-\xi\|^2+\|\rho_0-\rho\|^2+n\lambda\big[(\Gamma_0\triangle\Gamma)\cap T_i\big],\end{aligned}\tag{7.1}$$

according to the Assumption (A). Consequently, the average Kullback–Leibler divergence:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n K_i(\theta_0,\theta)&\lesssim\|\xi_0-\xi\|^2+\|\rho_0-\rho\|^2+\frac{1}{n}n\lambda\big[(\Gamma_0\triangle\Gamma)\cap(\cup T_i)\big]\\&=\|\xi_0-\xi\|^2+\|\rho_0-\rho\|^2+\lambda(\Gamma_0\triangle\Gamma).\end{aligned}\tag{7.2}$$

Similarly, the second moment V_i of the log-likelihood ratio is bounded in the same way, that is, $V_i(\theta_0,\theta) \lesssim \|\xi_0-\xi\|^2 + \|\rho_0-\rho\|^2 + \lambda(\Gamma_0 \triangle \Gamma)$, which leads to

$$B_n^*(\theta_0,\varepsilon) \supset \{(\xi,\rho,\gamma) : \|\xi_0-\xi\|^2 \le \varepsilon^2/3, \|\rho_0-\rho\|^2 \le \varepsilon^2/3, \lambda(\Gamma_0 \triangle \Gamma) \le \varepsilon^2/3\}.$$

Therefore, we have $\Pi(B_n^*(\theta_0,\varepsilon)) \gtrsim \Pi(\|\xi_0-\xi\|^2 \le \varepsilon^2/3, \|\rho_0-\rho\|^2 \le \varepsilon^2/3) \times \Pi(\lambda(\Gamma_0 \triangle \Gamma) \le \varepsilon^2/3)$, or equivalently,

$$\begin{aligned}-\log \Pi\big(B_n^*(\theta_0,\varepsilon_n)\big) \lesssim &-\log \Pi\big(\|\xi_0-\xi\|^2 \le \varepsilon^2/3, \|\rho_0-\rho\|^2 \le \varepsilon^2/3\big)\\ &-\log \Pi\big(\gamma : \lambda(\Gamma_0 \triangle \Gamma) \le \varepsilon^2/3\big).\end{aligned}$$

By Assumption (B1), the prior density of (ξ,ρ) is bounded below in a neighborhood of (ξ_0,ρ_0), indicating that $\Pi(\|\xi_0-\xi\|^2 \le \varepsilon^2/3, \|\rho_0-\rho\|^2 \le \varepsilon^2/3) \gtrsim \varepsilon^{2p}$, and thus $-\log \Pi(\|\xi_0-\xi\|^2 \le \varepsilon^2/3, \|\rho_0-\rho\|^2 \le \varepsilon^2/3) \lesssim \log(1/\varepsilon^2)$.

Let ε_n be a sequence such that $\varepsilon_n \to 0$ and $n\varepsilon_n^2/\log n$ bounded away from 0, and hence $\log(1/\varepsilon_n^2) \lesssim \log n \lesssim n\varepsilon_n^2$. In order to ensure that $-\log \Pi(B_n^*(\theta_0,\varepsilon_n)) \lesssim n\varepsilon_n^2$, it suffices that $-\log \Pi(\gamma : \lambda(\Gamma_0 \triangle \Gamma) \le \varepsilon_n^2) \lesssim n\varepsilon_n^2$ in equation (3.1).

Step 2: Sieves. For each prior, we shall define a sieve Σ_n for γ, and consider $\Theta_n = [-c_n,c_n]^p \times [-c_n,c_n]^p \times \Sigma_n$ as the sieve for θ. Because

$$\Pi(\Theta_n^c) \le \Pi\big(\xi : \xi \notin [-c_n,c_n]^p\big) + \Pi\big(\rho : \rho \notin [-c_n,c_n]^p\big) + \Pi(\gamma : \gamma \notin \Sigma_n),$$

in order to ensure that the sieve contains most of the prior mass, it is sufficient to show $-\log \Pi(\Sigma_n^c) \gtrsim n\varepsilon_n^2$ as in equation (3.2) provided that $-\log \Pi(\xi : \xi \notin [-c_n,c_n]^p) \gtrsim n\varepsilon_n^2$ and $-\log \Pi(\rho : \rho \notin [-c_n,c_n]^p) \gtrsim n\varepsilon_n^2$. For the later two conditions, we let $c_n = e^{n\varepsilon_n^2}$. Then $-\log \Pi(\xi : \xi \notin [-c_n,c_n]^p) \gtrsim -\log c_n^{-t_2}$ by Assumption (B2), which is $t_2 \cdot n\varepsilon_n^2 \gtrsim n\varepsilon_n^2$; similarly, we have $-\log \Pi(\xi : \xi \notin [-c_n,c_n]^p) \gtrsim n\varepsilon_n^2$.

Step 3: Entropy bounds. Let $\sigma_n = \sup_{\gamma \in \Sigma_n} \|\gamma\|_\infty$, for $\gamma, \gamma' \in \Sigma_n$, we then have

$$\begin{aligned}\lambda(\gamma,\gamma') &= \int_{\mathbb{S}^{d-1}} \left| \int_{\gamma(\boldsymbol{\omega})}^{\gamma'(\boldsymbol{\omega})} r^{d-1}\, dr \right| d\boldsymbol{\omega} \le \sigma_n^{d-1} \|\gamma'-\gamma\|_\infty \int_{\mathbb{S}^{d-1}} d\boldsymbol{\omega}\\ &\lesssim \sigma_n^{d-1} \|\gamma-\gamma'\|_\infty.\end{aligned}$$

Like in equation (7.2), the average squared Hellinger distance d_n^2 has the following bound when $\|\xi-\xi'\| \le \varepsilon$ and $\|\rho-\rho'\| \le \varepsilon$ for some small ε and $\max(\|\xi\|, \|\xi'\|, \|\rho\|, \|\rho'\|) \le M$:

$$\begin{aligned}d_n^2(\theta,\theta') &= \frac{1}{n}\sum_{i=1}^n \int h^2(\phi_i,\phi_i')\, dP_{X_i} \lesssim h^2(\xi,\xi') + h^2(\rho,\rho') + \lambda(\Gamma \triangle \Gamma')\\ &\lesssim (1+M^{b_0})\big(\|\xi-\xi'\|^2 + \|\rho-\rho'\|^2\big) + \sigma_n^{d-1}\|\gamma-\gamma'\|_\infty\end{aligned}$$

by Condition (A2). Therefore, the entropy $\log N(\varepsilon_n,\Theta_n,d_n)$ is bounded by

$$\begin{aligned}&2\log N\big(\varepsilon_n^2/(1+c_n)^{b_0}, [-c_n,c_n]^p, \|\cdot\|\big) + \log N\big(\varepsilon_n^2/\sigma_n^{d-1}, \Sigma_n, \|\cdot\|_\infty\big)\\ &\lesssim \log\frac{c_n}{\varepsilon_n^2} + \log N\big(\varepsilon_n^2/\sigma_n^{d-1}, \Sigma_n, \|\cdot\|_\infty\big).\end{aligned}$$

Recall that $c_n = e^{n\varepsilon_n^2}$. Therefore, $\log(c_n/\varepsilon_n^2) \le n\varepsilon_n^2 + \log(1/\varepsilon_n^2) \lesssim n\varepsilon_n^2$. Hence, in order to ensure $\log N(\varepsilon_n, \Theta_n, d_n) \lesssim n\varepsilon_n^2$, it is sufficient to verify that $\log N(\varepsilon_n^2/\sigma_n^{d-1}, \Sigma_n, \|\cdot\|_\infty) \lesssim n\varepsilon_n^2$ which is equation (3.3).

Then equation (3.4) follows by applying Theorem 4 of [15].

Equation (3.5) will follow if we show that $d_n(\theta, \theta_0) \le \varepsilon_n$ implies $\lambda(\Gamma \triangle \Gamma_0) \lesssim \varepsilon_n^2/c_{0,n}^2$. As argued in the derivation of (7.1) and (7.2), $d_n^2(\theta, \theta_0)$ is given by

$$\frac{1}{n}\sum_{i=1}^{n}\int h^2(\phi, \phi_0)\, dP_{X_i} = h^2(\xi_0, \xi)\lambda(\Gamma_0 \cap \Gamma) + h^2(\rho_0, \rho)\lambda\big(\Gamma_0^c \cap \Gamma^c\big) + h^2(\xi_0, \rho)\lambda\big(\Gamma_0 \cap \Gamma^c\big) + h^2(\rho_0, \xi)\lambda\big(\Gamma_0^c \cap \Gamma\big).$$

The above expression is larger than each of the following three expressions:

$$\begin{aligned}
&h^2(\xi_0, \xi)\lambda(\Gamma_0 \cap \Gamma) + h^2(\rho_0, \xi)\lambda\big(\Gamma_0^c \cap \Gamma\big) \\
&\quad \ge \frac{(h(\xi_0, \xi) + h(\rho_0, \xi))^2}{2} \cdot \big(\lambda(\Gamma_0 \cap \Gamma) \wedge \lambda\big(\Gamma_0^c \cap \Gamma\big)\big), \\
&h^2(\rho_0, \rho)\lambda\big(\Gamma_0^c \cap \Gamma^c\big) + h^2(\xi_0, \rho)\lambda\big(\Gamma_0 \cap \Gamma^c\big) \\
&\quad \ge \frac{(h(\rho_0, \rho) + h(\xi_0, \rho))^2}{2} \cdot \big(\lambda\big(\Gamma_0^c \cap \Gamma^c\big) \wedge \lambda\big(\Gamma_0 \cap \Gamma^c\big)\big), \\
&h^2(\xi_0, \rho)\lambda\big(\Gamma_0 \cap \Gamma^c\big) + h^2(\rho_0, \xi)\lambda\big(\Gamma_0^c \cap \Gamma\big) \\
&\quad \ge \frac{(h(\xi_0, \rho) + h(\rho_0, \xi))^2}{2} \cdot \big(\lambda\big(\Gamma_0 \cap \Gamma^c\big) \wedge \lambda\big(\Gamma_0^c \cap \Gamma\big)\big).
\end{aligned}$$

We further have $h(\xi_0, \xi) + h(\rho_0, \xi) \ge h(\xi_0, \rho_0)$ and $h(\xi_0, \rho) + h(\rho_0, \rho) \ge h(\xi_0, \rho_0)$, by the triangle inequality, and $h(\xi_0, \rho) + h(\rho_0, \xi) \ge c_{0,n} > 0$ by Condition (C). Combining with the last three displays respectively, we have

$$\lambda(\Gamma_0 \cap \Gamma) \wedge \lambda\big(\Gamma_0^c \cap \Gamma\big) \lesssim \varepsilon_n^2/c_{0,n}^2, \tag{7.3}$$

$$\lambda\big(\Gamma_0^c \cap \Gamma^c\big) \wedge \lambda\big(\Gamma_0 \cap \Gamma^c\big) \lesssim \varepsilon_n^2/c_{0,n}^2, \tag{7.4}$$

$$\lambda\big(\Gamma_0 \cap \Gamma^c\big) \wedge \lambda\big(\Gamma_0^c \cap \Gamma\big) \lesssim \varepsilon_n^2/c_{0,n}^2, \tag{7.5}$$

whenever $d_n^2(\theta_0, \theta) \le \varepsilon_n^2$. By adding (7.3) and (7.4) to (7.5), we derive

$$\lambda(\Gamma_0) \wedge \lambda\big(\Gamma_0^c \cap \Gamma\big) \lesssim \varepsilon_n^2/c_{0,n}^2, \qquad \lambda\big(\Gamma_0^c\big) \wedge \lambda\big(\Gamma_0 \cap \Gamma^c\big) \lesssim \varepsilon_n^2/c_{0,n}^2. \tag{7.6}$$

Since Γ_0 is fixed with $\lambda(\Gamma_0) > 0$ and $\lambda(\Gamma_0^c) > 0$ by the assumption, (7.6) implies that $\lambda(\Gamma_0^c \cap \Gamma) \lesssim \varepsilon_n^2/c_{0,n}^2$, and $\lambda(\Gamma_0 \cap \Gamma^c) \lesssim \varepsilon_n^2/c_{0,n}^2$. Consequently, $\lambda(\Gamma_0 \triangle \Gamma) = \lambda(\Gamma_0^c \cap \Gamma) + \lambda(\Gamma_0 \cap \Gamma^c) \lesssim \varepsilon_n^2/c_{0,n}^2$, which completes the proof. $\square$

PROOF OF THEOREM 3.3. We verify equations (3.1), (3.2) and (3.3) in Theorem 3.1. Since $\|\gamma_0 - \boldsymbol{\beta}_{0,J_n}^T \boldsymbol{\eta}\|_\infty \le \varepsilon_n^2/2$, we have

$$\begin{aligned}
&\Pi\{\gamma : \gamma = \boldsymbol{\beta}^T\boldsymbol{\eta}, \|\gamma - \gamma_0\|_\infty \le \varepsilon_n^2\} \\
&\quad \ge \Pi(J = J_n)\Pi\big(\|\boldsymbol{\beta}^T\boldsymbol{\eta} - \boldsymbol{\beta}_0^T\boldsymbol{\eta}\|_\infty \le \varepsilon_n^2/2 | J = J_n\big) \\
&\quad \ge \Pi(J = J_n)\Pi\big(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 \le t_3^{-1} J^{-t_4}\varepsilon_n^2/2 | J = J_n\big),
\end{aligned}$$

where the last step is because $\|\boldsymbol{\beta}_{1,J}^T\boldsymbol{\eta} - \boldsymbol{\beta}_{2,J}^T\boldsymbol{\eta}\|_\infty \le \|\boldsymbol{\beta}_{1,J} - \boldsymbol{\beta}_{2,J}\|_1 \times \max_{1\le j\le J}\|\eta_j\|_\infty$ which is bounded by $t_3 J^{t_4}\|\boldsymbol{\beta}_{1,J} - \boldsymbol{\beta}_{2,J}\|_1$ according to the triangle inequality and Assumption (D). Therefore, equation (3.1) holds by noting that $-\log\Pi\{\gamma = \boldsymbol{\beta}^T\boldsymbol{\eta} : \|\gamma - \gamma_0\|_\infty \le \varepsilon_n^2\} \le -\log\Pi(J = J_n) - \log\Pi(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 \le t_3^{-1}J^{-t_4}\varepsilon_n^2/2 | J = J_n) \lesssim J_n\log J_n + J_n\log(J_n/\varepsilon_n) \lesssim J_n\log n \lesssim n\varepsilon_n^2$.

Considering the sieve $\Sigma_n = \{\gamma : \gamma = \boldsymbol{\beta}^T\boldsymbol{\eta}, \boldsymbol{\beta} \in \mathbb{R}^j, j \le J_n, \|\boldsymbol{\beta}\|_\infty \le \sqrt{n/C}\}$, the estimate of the prior mass of the complement of the sieve is given by $\Pi(\gamma : \gamma \notin \Sigma_n) \le \Pi(J > J_n) + J_n e^{-n}$ (see equation (2.10) in [39]). For any $a, b > 0$, we have $\log(a+b) \le \log(2(a\vee b))$, leading to $-\log(a+b) \ge -\log 2 + (-\log a)\wedge(-\log b)$. Noting that $-\log\Pi(J > J_n) \gtrsim J_n\log J_n$, and $-\log(J_n e^{-n}) = n - \log J_n \ge n - \log n \asymp n \gtrsim J_n\log J_n$ (because $J_n\log J_n \lesssim n\varepsilon_n^2 \lesssim n$), we then obtain $-\log\Pi(\gamma : \gamma \notin \Sigma_n) \gtrsim J_n\log J_n \gtrsim n\varepsilon_n^2$ verifying equation (3.2).

For the entropy calculation, we first notice that for any $\gamma \in \Sigma_n$, we have $\|\gamma\|_\infty = \|\boldsymbol{\beta}^T\boldsymbol{\eta}\|_\infty \le \max_{1\le j\le J_n}\|\eta_j\|_\infty\|\boldsymbol{\beta}\|_1 \lesssim J_n^{t_4}\sqrt{n} \lesssim n^{t_4+1/2}$, which is an upper bound for σ_n. We estimate the packing number $D(\varepsilon_n^2/\sigma_n^{d-1}, \Sigma_n, \|\cdot\|_\infty)$ by

$$\sum_{j=1}^{J_n} D\big(\varepsilon_n^2/\sigma_n^{d-1}, \{\boldsymbol{\beta} \in \mathbb{R}^j, \|\boldsymbol{\beta}\|_\infty \le \sqrt{n/C}\}, \|\cdot\|_\infty\big) \le \sum_{j=1}^{J_n}\left(1 + \frac{\sqrt{n/C}\sigma_n^{d-1}}{\varepsilon_n^2}\right)^j,$$

which is further bounded by $J_n(1 + \sqrt{n/C}\sigma_n^{d-1}/\varepsilon_n^2)^{J_n}$. Equation (3.3) follows since $\log N(\varepsilon_n^2/\sigma_n^{d-1}, \Sigma_n, \|\cdot\|_\infty) \le \log D(\varepsilon_n^2/\sigma_n^{d-1}, \Sigma_n, \|\cdot\|_\infty) \lesssim \log J_n + J_n\log n + J_n\log(1/\varepsilon_n^2) \lesssim J_n\log n \lesssim n\varepsilon_n^2$. □

PROOF OF THEOREM 3.7. We first obtain the contraction rate for deterministic rescaling when the smoothness level α is known.

Let $\mathbb{B}$ be $\mathbb{C}^\alpha(\mathbb{S}^1)$ equipped with the $\|\cdot\|_\infty$ norm. Let $\phi_{\gamma_0}^a(\varepsilon) = \inf_{\gamma\in\mathbb{H}^a:\|\gamma-\gamma_0\|_\infty\le\varepsilon}\frac{1}{2}\|\gamma\|_{\mathbb{H}^a}^2 - \log\mathrm{P}(\|W^a\|_\infty \le \varepsilon)$ stand for the concentration function at γ_0. Note that $\phi_0^a(\varepsilon) = -\log\mathrm{P}(\|W^a\|_\infty \le \varepsilon)$.

The selection of sieves and entropy calculation is similar to Theorem 2.1 in [42] with ε_n replaced by ε_n^2 and adjustment because of the involvement of σ_n later on. Define the sieve as $\Sigma_n = (M_n\mathbb{H}_1^a + \frac{1}{4}\varepsilon_n^2 M_n^{-1}\mathbb{B}_1)$, where $\mathbb{H}_1^a$ and $\mathbb{B}_1$ are the unit balls of $\mathbb{H}^a$ and $\mathbb{B}$, respectively. Let $M_n = -2\Phi^{-1}(\exp(-Cn\varepsilon_n^2))$ for a large constant $C > 1$. By Borell's inequality, we can bound $\Pi(\Sigma_n^c) \le \Pi(\gamma \notin M_n\mathbb{H}_1^a + \varepsilon_n^2\mathbb{B}_1) \lesssim \exp(-Cn\varepsilon_n^2)$, provided that $\phi_0^a(\frac{1}{4}\varepsilon_n^2 M_n^{-1}) \le n\varepsilon_n^2$.

For the entropy calculation, first observe that $M_n^2 \lesssim n\varepsilon_n^2$ since $|\Phi^{-1}(u)| \le \sqrt{2\log(1/u)}$ for $u \in (0, 1/2)$. We further notice that $\mathbb{H}_1^a \subset \mathbb{B}_1$ since by Lemma 4.3, for any function $h \in \mathbb{H}_1^a$, we have $\|h\|_\infty^2 \le \{\sum_{n=-\infty}^\infty |b_{n,a}| e^{-2a^2} I_n(2a^2)\}^2 \le \sum_{n=-\infty}^\infty |b_{n,a}|^2 e^{-2a^2} I_n(2a^2) \cdot \sum_{n=-\infty}^\infty e^{-2a^2} I_n(2a^2) = \|h\|_{\mathbb{H}^a}^2 \cdot 1$ [the last step uses Proposition 8.1(a) by letting $z = 1$ and $x = a^2$]. Therefore, the sieve Σ_n is a subset of $(M_n + \varepsilon_n^2 M_n^{-1}/4)\mathbb{B}_1$, and thus $\sigma_n = \sup\{\|\gamma\|_\infty : \gamma \in \Sigma_n\} \le M_n + \varepsilon_n^2 M_n^{-1}/4 \le 2M_n$ for sufficiently large n. By construction of Σ_n, a $\frac{1}{4}\varepsilon_n^2 M_n^{-2}$-net for $\mathbb{H}_1^a$ is a $\frac{1}{2}\varepsilon_2^2 M_n^{-1}$-net for Σ_n. Therefore, by Lemma 4.5, we have

$$\begin{aligned}\log N\big(\varepsilon_n^2/\sigma_n, \Sigma_n, \|\cdot\|_\infty\big) &\le \log N\big(\varepsilon_n^2/\{4M_n^2\}, \mathbb{H}_1^a, \|\cdot\|_\infty\big)\\ &\lesssim a\big(\log M_n^2/\{2\varepsilon_n^2\}\big)^2 \lesssim a(\log n)^2.\end{aligned}$$

To evaluate the prior concentration probability, we proceed as follows. Let $\Pi^a(\cdot)$ be a SEP Gaussian process with the rescaling factor a. By the approximation property of $\mathbb{H}^a$ in Lemma 4.4, there exists $h_0 \in \mathbb{H}^a$ such that $\|h_0 - \gamma_0\|_\infty \lesssim a^{-\alpha}$ and $\|h_0\|_{\mathbb{H}^a}^2 \lesssim a$. Therefore, if $a^{-\alpha} \le \varepsilon_n^2/2$, then

$$\Pi^a\big(\gamma : \|\gamma - \gamma_0\|_\infty \le \varepsilon_n^2\big) \ge \Pi^a\big(\gamma : \|\gamma - h_0\|_\infty \le \varepsilon_n^2/2\big) \ge \exp\big\{-\phi_0^a\big(\varepsilon_n^2/4\big)\big\},$$

leading to $-\log \Pi^a(\gamma : \|\gamma - \gamma_0\|_\infty \le \varepsilon_n^2) \lesssim \phi_0^a(\varepsilon_n^2/4)$.

Note that $\phi_0^a(\varepsilon) \lesssim a(\log(a/\varepsilon))^2$ (Lemma 4.6). To satisfy the conditions in Theorem 3.1, we choose $a = a_n$ depending on the sample size such that $\varepsilon_n^2 \asymp a_n^{-\alpha}$, and $a_n(\log n)^2 \asymp n\varepsilon_n^2$. Then the posterior contraction rate is obtained as $\varepsilon_n^2 = n^{-\alpha/(\alpha+1)}(\log n)^{-2\alpha/(\alpha+1)}$, with $a_n = n^{1/(\alpha+1)}(\log n)^{-2/(\alpha+1)}$.

Now consider the random rescaling for unknown smoothness α. The established properties of the RKHS of W^a from Lemma 4.3 to Lemma 4.8 are parallel to the case when a GP is indexed by $[0, 1]^{d-1}$ with a stationary kernel, therefore, we can directly follow the argument in the proof of Theorem 3.1 in [43]. There is need for a slight modification since the nesting property given in Lemma 4.7 has a universal constant c, but this does not affect the asymptotic rate. The posterior contraction rate ε_n^2 is thus obtained. □

8. Modified Bessel function of the first kind. Modified Bessel functions of the first kind are solutions of the modified Bessel equation [45]. Throughout the paper, we consider integer orders and positive argument, that is, $I_n(x)$ with $n \in \mathbb{Z}$ and $x > 0$. We first introduce some basic properties of $I_n(x)$ in Proposition 8.1, for easy reference.

PROPOSITION 8.1. *For the modified Bessel functions $I_n(2x)$, we have*:

(a) *Generating functions. For $x \in \mathbb{R}$,*

$$G(x, z) =: e^{x(z+1/z)} = \sum_{n=-\infty}^{\infty} I_n(2x) z^n, \qquad z \in \mathbb{C}, z \ne 0.$$

(b) *Symmetry about the order*: $I_n(2x) = I_{-n}(2x)$ *for* $x \in \mathbb{R}$ *and* $n \in \mathbb{Z}$.
(c) *For* $n \geq 0$ *and fixed* $x > 0$, *the following properties hold*:

(c1) *Series representation*: $I_n(2x) = x^n \sum_{j=0}^{\infty} x^{2j}/(j!(n+j)!)$.
(c2) $I_n(2x)$ *is positive and strictly decreasing in* n.
(c3) $I_n(2x) \leq I_0(2x)(2x)^n/n!$.

The estimate below is obtained when $x \to \infty$ with n being fixed or $n \to \infty$ in such a way that $nx^{-1/2}$ tends to a finite nonnegative number.

PROPOSITION 8.2. *Let* $n \in \mathbb{Z}$ *and* $nx^{-1/2} \to c$ *for some constant* $c \geq 0$ *as* $x \to \infty$. *Then* $\sqrt{x}e^{-x} I_n(x) \to (2\pi)^{-1/2} e^{-c^2/2}$ *as* $x \to \infty$.

PROPOSITION 8.3. *For any* $x \geq 0$ *and* $j = 0, 1, 2, \ldots$, *we have*

$$\sum_{n=-\infty}^{\infty} e^{-2x} I_n(2x) n^{2j} \leq \frac{(4j)!}{(2j)!} \max\bigl(x^j, 1\bigr).$$

PROPOSITION 8.4. *The function* $f_n(x) = \sqrt{x}e^{-x} I_n(x)$ *is increasing in* x *when* $x \in B_n$, *where* $B_n = [0, n + 1/2]$ *if* $n = 0, 1$ *and* $B_n = [0, \infty)$ *if* $n \geq 2$.

PROPOSITION 8.5. *For any* $x \geq 0$, *we have* $\sum_{n=-\infty}^{\infty} e^{-2x} I_n(2x) n^2 = 2x$.

Acknowledgments. We thank Professor Aad van der Vaart for many helpful discussions and pointing out important references. We thank the Editor, the Associate Editor and three anonymous referees for constructive comments that helped to improve the paper. The work was conducted when the first author was a graduate student at North Carolina State University.

SUPPLEMENTARY MATERIAL

Supplement to "Bayesian detection of image boundaries" (DOI: 10.1214/16-AOS1523SUPP; .pdf). The supplementary file contains proofs to all lemmas and propositions in the paper.

REFERENCES

[1] ARBEL, J., GAYRAUD, G. and ROUSSEAU, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scand. J. Stat.* **40** 549–570. MR3091697
[2] BANERJEE, S. and GELFAND, A. E. (2006). Bayesian wombling: Curvilinear gradient assessment under spatial process models. *J. Amer. Statist. Assoc.* **101** 1487–1501. MR2279474
[3] BASU, M. (2002). Gaussian-based edge-detection methods-a survey. *IEEE Trans. Syst. Man Cybern., Part C Appl. Rev.* **32** 252–260.
[4] BHARDWAJ, S. and MITTAL, A. (2012). A survey on various edge detector techniques. *Proc. Technol.* **4** 220–226.

[5] CARLSTEIN, E. and KRISHNAMOORTHY, C. (1992). Boundary estimation. *J. Amer. Statist. Assoc.* **87** 430–438. MR1173808
[6] CASTILLO, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields* **152** 53–99. MR2875753
[7] CASTILLO, I., KERKYACHARIAN, G. and PICARD, D. (2014). Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields* **158** 665–710. MR3176362
[8] CHEN, J. and GUPTA, A. K. (2012). *Parametric Statistical Change Point Analysis*: *With Applications to Genetics*, *Medicine*, *and Finance*, 2nd ed. Birkhäuser/Springer, New York. MR3025631
[9] DAI, F. and XU, Y. (2013). *Approximation Theory and Harmonic Analysis on Spheres and Balls*. Springer, New York. MR3060033
[10] DONOHO, D. L. (1999). Wedgelets: Nearly minimax estimation of edges. *Ann. Statist.* **27** 859–897. MR1724034
[11] DUDLEY, R. M. (1974). Metric entropy of some classes of sets with differentiable boundaries. *J. Approx. Theory* **10** 227–236. MR0358168
[12] ENNIS, D. (2005). Spherical Harmonics. http://www.mathworks.com/matlabcentral/fileexchange/8638-spherical-harmonics. MATLAB Central File Exchange.
[13] FITZPATRICK, M. C., PREISSER, E. L., PORTER, A., ELKINTON, J., WALLER, L. A., CARLIN, B. P. and ELLISON, A. M. (2010). Ecological boundary detection using Bayesian areal wombling. *Ecology* **91** 3448–3455.
[14] GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.
[15] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274
[16] GU, K., PATI, D. and DUNSON, D. B. (2014). Bayesian multiscale modeling of closed curves in point clouds. *J. Amer. Statist. Assoc.* **109** 1481–1494. MR3293605
[17] HALL, P., PENG, L. and RAU, C. (2001). Local likelihood tracking of fault lines and boundaries. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 569–582. MR1858403
[18] HASTIE, T. and STUETZLE, W. (1989). Principal curves. *J. Amer. Statist. Assoc.* **84** 502–516. MR1010339
[19] JACKSON, D. (1994). *The Theory of Approximation*. *American Mathematical Society Colloquium Publications* **11**. Amer. Math. Soc., Providence, RI. Reprint of the 1930 original. MR1451140
[20] KILLICK, R. and ECKLEY, I. A. (2011). Changepoint: An R package for changepoint analysis. R Package Version 0.6, http://CRAN.R-project.Org/package=changepoint.
[21] KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418
[22] KOROSTELËV, A. P. and TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction*. *Lecture Notes in Statistics* **82**. Springer, New York. MR1226450
[23] KUELBS, J. and LI, W. V. (1993). Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal.* **116** 133–157. MR1237989
[24] LI, M. and GHOSAL, S. (2017). Supplement to "Bayesian detection of image boundaries." DOI:10.1214/16-AOS1523SUPP.
[25] LI, W. V. and LINDE, W. (1999). Approximation, metric entropy and small ball estimates for Gaussian measures. *Ann. Probab.* **27** 1556–1578. MR1733160
[26] LU, H. and CARLIN, B. P. (2005). Bayesian areal wombling for geographical boundary analysis. *Geogr. Anal.* **37** 265–285.
[27] MACKAY, D. J. (1998). Introduction to Gaussian processes. In *NATO ASI Series F Computer and Systems Sciences* **168** 133–166.
[28] MAMMEN, E. and TSYBAKOV, A. B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.* **23** 502–524. MR1332579

[29] MÜLLER, H.-G. and SONG, K. S. (1994). Maximin estimation of multidimensional boundaries. *J. Multivariate Anal.* **50** 265–281. MR1293046

[30] NEAL, R. M. (2003). Slice sampling. *Ann. Statist.* **31** 705–767. With discussions and a rejoinder by the author. MR1994729

[31] POLZEHL, J. and SPOKOINY, V. (2003). Image denoising: Pointwise adaptive approach. *Ann. Statist.* **31** 30–57. MR1962499

[32] QIU, P. (2005). *Image Processing and Jump Regression Analysis*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. MR2111430

[33] QIU, P. (2007). Jump surface estimation, edge detection, and image restoration. *J. Amer. Statist. Assoc.* **102** 745–756. MR2370864

[34] QIU, P. and SUN, J. (2007). Local smoothing image segmentation for spotted microarray images. *J. Amer. Statist. Assoc.* **102** 1129–1144. MR2412538

[35] QIU, P. and SUN, J. (2009). Using conventional edge detectors and postsmoothing for segmentation of spotted microarray images. *J. Comput. Graph. Statist.* **18** 147–164. MR2649642

[36] QIU, P. and YANDELL, B. (1997). Jump detection in regression surfaces. *J. Comput. Graph. Statist.* **6** 332–354. MR1466871

[37] RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA. MR2514435

[38] RUDEMO, M. and STRYHN, H. (1994). Approximating the distribution of maximum likelihood contour estimators in two-region images. *Scand. J. Stat.* **21** 41–55. MR1267042

[39] SHEN, W. and GHOSAL, S. (2015). Adaptive Bayesian procedures using random series priors. *Scand. J. Stat.* **42** 1194–1213. MR3426318

[40] TERRAS, A. (2013). *Harmonic Analysis on Symmetric Spaces—Euclidean Space, the Sphere, and the Poincaré Upper Half-Plane*, 2nd ed. Springer, New York. MR3100414

[41] VAN DER VAART, A. and VAN ZANTEN, H. (2007). Bayesian inference with rescaled Gaussian process priors. *Electron. J. Stat.* **1** 433–448. MR2357712

[42] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663

[43] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442

[44] WALLER, L. A. and GOTWAY, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. MR2075123

[45] WATSON, G. N. (1995). *A Treatise on the Theory of Bessel Functions*. Cambridge Univ. Press, Cambridge. Reprint of the second (1944) edition. MR1349110

DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
DURHAM, NORTH CAROLINA 27708-0251
USA
E-MAIL: ml371@stat.duke.edu

DEPARTMENT OF STATISTICS
NORTH CAROLINA STATE UNIVERSITY
RALEIGH, NORTH CAROLINA 27695-8203
USA
E-MAIL: sghosal@stat.ncsu.edu