

# BKP: An R Package for Beta Kernel Process Modeling

Jiangyan Zhao

School of Statistics

East China Normal University

Kunhai Qing

School of Statistics

East China Normal University

Jin Xu

School of Statistics

and

Key Laboratory of Advanced Theory and Application in Statistics and Data Science - MOE  
East China Normal University

---

## Abstract

We present **BKP**, a user-friendly and extensible R package that implements the Beta Kernel Process (BKP) – a fully nonparametric and computationally efficient framework for modeling spatially varying binomial probabilities. The BKP model combines localized kernel-weighted likelihoods with conjugate beta priors, resulting in closed-form posterior inference without requiring latent variable augmentation or intensive MCMC sampling. The package supports binary and aggregated binomial responses, allows flexible choices of kernel functions and prior specification, and provides loss-based kernel hyperparameter tuning procedures. In addition, BKP extends naturally to the Dirichlet Kernel Process (DKP) for modeling spatially varying categorical or multinomial data. To our knowledge, this is the first publicly available R package for implementing BKP-based methods. We illustrate the use of **BKP** through several synthetic and real-world datasets, highlighting its interpretability, accuracy, and scalability. The package aims to facilitate practical application and future methodological development of kernel-based beta modeling in statistics and machine learning.

*Keywords:* Beta kernel process, Dirichlet kernel process, nonparametric Bayesian modeling, Beta-binomial, Dirichlet-multinomial, R.

---

## 1. Introduction

Estimating a continuous probability function from binary or binomial observations is a fundamental task in modern statistics and machine learning (Hastie *et al.* 2009; Rolland *et al.* 2019; Murphy 2022). Such problems arise in various domains where the goal is to infer a latent probability surface from individual Bernoulli outcomes or aggregated binomial counts over a continuous input space. Representative applications include binary classification (MacKenzie *et al.* 2014; Wen *et al.* 2025), probability calibration (Sung *et al.* 2020; Dimitriadis *et al.* 2022), relative abundance modeling (Martin *et al.* 2020), and longitudinal analysis of patient-

reported outcomes (Najera-Zuloaga *et al.* 2018, 2019).

Classical approaches such as logistic regression and generalized additive models provide computationally efficient tools for binomial regression, but their parametric form often limits their ability to capture complex nonlinear patterns (Hastie *et al.* 2009). Nonparametric models like Gaussian Process (GP) classifiers provide greater modeling flexibility and principled uncertainty quantification (Rasmussen and Williams 2006), yet the non-Gaussian likelihood of binary responses necessitates approximate inference, leading to substantial computational cost and implementation complexity (Nickisch and Rasmussen 2008).

The **BKP** package implements the *Beta Kernel Process* (BKP), a scalable and interpretable nonparametric model for estimating binomial probability surfaces. Developed in R (R Core Team 2025) and available on the Comprehensive R Archive Network (CRAN) at <https://cran.r-project.org/package=BKP>, **BKP** combines localized kernel-weighted likelihoods with conjugate beta priors, yielding closed-form posterior inference without latent variables or numerical approximation. It supports both binary and aggregated binomial data, and delivers full posterior uncertainty quantification while maintaining linear scalability and computational efficiency (Goetschalckx *et al.* 2011; MacKenzie *et al.* 2014; Rolland *et al.* 2019).

Compared to GP-based methods, BKP eliminates the need for sampling or variational inference and provides interpretable, locally updated estimates. This makes it suitable for applications demanding transparency, real-time feedback, or adaptive decision-making. BKP has been extended to the multinomial setting via the *Dirichlet Kernel Process* (DKP), enabling nonparametric modeling of categorical proportions across multiple groups. Despite its practical appeal, the theoretical properties of BKP remain underexplored. Mussi *et al.* (2024) identified the lack of formal guarantees for BKP-based decision processes as an open problem, emphasizing the need for further methodological development.

While GP models are supported by mature software libraries, such as **DiceKriging** (Roustant *et al.* 2012), **GPfit** (MacDonald *et al.* 2015), and **gplite** (Piironen 2022) in R, and **GPy-Torch** (Gardner *et al.* 2018), **GPflow** (Matthews *et al.* 2017), and **GPy** (GPy since 2012) in Python. So far, there are about 100 GP-related packages on CRAN (as seen by running `packagefinder::findPackage("Gaussian process", display = "browser")`). In contrast, software for kernel-based modeling of binomial and multinomial data remains scarce. The **BKP** package fills this gap by offering a natively, unified and extensible framework for nonparametric modeling of binary/binomial, categorical/multinomial data. It provides multiple options for prior specification, kernel functions, and loss functions, enabling flexible model customization. These features make **BKP** applicable to a wide range of problems in biomedicine, ecology, social science, and industrial statistics.

The remainder of the paper is organized as follows. Section 2 introduces the statistical foundation of the BKP model, including prior specification strategies, kernel functions, loss-based hyperparameter tuning, and the DKP extension. Section 3 describes the structure and main functionalities of the **BKP** package, including model fitting, prediction, and simulation for both BKP and DKP. Section 4 presents illustrative examples to demonstrate the use of BKP and DKP in binary classification and compositional modeling tasks. Section 5 demonstrates the BKP workflow using the Loa loa dataset. Section 6 concludes with a discussion of current limitations and potential extensions.

## 2. Statistical Foundation

### 2.1. Beta Kernel Process

Let  $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathcal{X} \subset \mathbb{R}^d$  denote a  $d$ -dimensional input. Suppose the success probability surface  $\pi(\mathbf{x}) \in [0, 1]$  is unknown. At each location  $\mathbf{x}$ , the observed data is modeled as

$$y(\mathbf{x}) \sim \text{Binomial}(m(\mathbf{x}), \pi(\mathbf{x})),$$

where  $y(\mathbf{x})$  is the number of successes out of  $m(\mathbf{x})$  independent trials. The full dataset comprises  $n$  observations  $\mathcal{D}_n = \{(\mathbf{x}_i, y_i, m_i)\}_{i=1}^n$ , where we write  $y_i = y(\mathbf{x}_i)$  and  $m_i = m(\mathbf{x}_i)$  for brevity.

In line with the Bayesian paradigm, assign a Beta prior to the unknown probability function as

$$\pi(\mathbf{x}) \sim \text{Beta}(\alpha_0(\mathbf{x}), \beta_0(\mathbf{x})),$$

where  $\alpha_0(\mathbf{x}) > 0$  and  $\beta_0(\mathbf{x}) > 0$  are spatially varying shape parameters. Details on prior specification are discussed in Section 2.2.

Let  $k : \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$  denote a user-defined kernel function measuring the similarity between input locations. By the kernel-based Bayesian updating strategy (Goetschalckx *et al.* 2011; Rolland *et al.* 2019), the BKP model constructs a closed-form posterior distribution for  $\pi(\mathbf{x})$  as

$$\begin{aligned} \pi(\mathbf{x}) \mid \mathcal{D}_n &\sim \text{Beta}(\alpha_n(\mathbf{x}), \beta_n(\mathbf{x})), \\ \alpha_n(\mathbf{x}) &= \alpha_0(\mathbf{x}) + \sum_{i=1}^n k(\mathbf{x}, \mathbf{x}_i) y_i = \alpha_0(\mathbf{x}) + \mathbf{k}^\top(\mathbf{x}) \mathbf{y}, \\ \beta_n(\mathbf{x}) &= \beta_0(\mathbf{x}) + \sum_{i=1}^n k(\mathbf{x}, \mathbf{x}_i) (m_i - y_i) = \beta_0(\mathbf{x}) + \mathbf{k}^\top(\mathbf{x}) (\mathbf{m} - \mathbf{y}), \end{aligned} \tag{1}$$

where  $\mathbf{k}(\mathbf{x}) = [k(\mathbf{x}, \mathbf{x}_1), \dots, k(\mathbf{x}, \mathbf{x}_n)]^\top$  is the vector of kernel weights and  $\mathbf{y} = (y_1, \dots, y_n)^\top$ .

**Remark 1.** While the BKP model leverages the conjugacy of the Beta-Binomial pair, it differs from the traditional Bayesian paradigm in the sense that the posterior update is induced from a kernel-weighted local likelihood defined by

$$\tilde{L}(\pi(\mathbf{x}); \mathcal{D}_n) \propto \prod_{i=1}^n \{\pi(\mathbf{x})^{y_i} (1 - \pi(\mathbf{x}))^{m_i - y_i}\}^{k(\mathbf{x}, \mathbf{x}_i)} = \pi(\mathbf{x})^{\mathbf{k}^\top(\mathbf{x}) \mathbf{y}} \{1 - \pi(\mathbf{x})\}^{\mathbf{k}^\top(\mathbf{x}) (\mathbf{m} - \mathbf{y})}.$$

This approach mimics the local likelihood method of Fan *et al.* (1998), where data are reweighted based on distance to the target point in the input space. And, the choice of kernel parameters is driven by empirical risk minimization rather than posterior inference. Thus, BKP is best interpreted as a nonparametric, Bayesian-inspired smoothing framework.

**Remark 2.** One prominent advantage of the closed-form updating scheme is its light burden of computational complexity. Fitting the BKP model involves  $\mathcal{O}(n^2)$  operations for computing the kernel matrix, in contrast to the  $\mathcal{O}(n^3)$  operations typically required by Gaussian process regression. While, evaluating the posterior at a new location requires only  $\mathcal{O}(n)$  kernel computations.

Based on the resulting posterior distribution (1), the posterior mean

$$\hat{\pi}_n(\mathbf{x}) = \mathbb{E}[\pi(\mathbf{x}) | \mathcal{D}_n] = \frac{\alpha_n(\mathbf{x})}{\alpha_n(\mathbf{x}) + \beta_n(\mathbf{x})} \quad (2)$$

serves as a smooth estimator of the latent success probability. The corresponding posterior variance

$$\sigma_n^2(\mathbf{x}) = \text{Var}[\pi(\mathbf{x}) | \mathcal{D}_n] = \frac{\hat{\pi}_n(\mathbf{x})\{1 - \hat{\pi}_n(\mathbf{x})\}}{\alpha_n(\mathbf{x}) + \beta_n(\mathbf{x}) + 1} \quad (3)$$

provides a local measure of epistemic uncertainty. These posterior summaries can be used to visualize prediction quality across the input space, particularly highlighting regions with sparse data coverage. See Section 4 for illustrations.

For binary classification, the posterior mean can be thresholded to produce hard predictions through

$$\hat{y}(\mathbf{x}) = \begin{cases} 1 & \text{if } \hat{\pi}_n(\mathbf{x}) > \pi_0, \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

where  $\pi_0 \in (0, 1)$  is a user-specified threshold, typically set to be 0.5.

## 2.2. Prior Specification

We provide three strategies for prior specification as follows.

- **Non-informative prior:** A default and widely used choice is the uniform prior, which sets  $\alpha_0(\mathbf{x}) = \beta_0(\mathbf{x}) \equiv 1$  for all  $\mathbf{x}$ . It is appropriate when no prior knowledge is available.
- **Informative prior with fixed mean:** When prior information about the overall success probability  $p_0 \in (0, 1)$  is available, an informative prior can be constructed by setting

$$\alpha_0(\mathbf{x}) = r_0 p_0, \quad \beta_0(\mathbf{x}) = r_0(1 - p_0),$$

where  $r_0 > 0$  is a scalar precision parameter controlling the strength of the prior. Larger value of  $r_0$  represents greater prior certainty. It contains the previous non-informative prior as a special case with  $r_0 = 2$  and  $p_0 = 0.5$ . In practice, we recommend specifying  $p_0$  according to the empirical overall success proportion  $y_i/m_i$ .

- **Data-adaptive informative prior:** To accommodate spatial variation in the data, BKP supports a locally adaptive prior in which both the prior mean and prior strength vary across the input space. Specifically, at each location  $\mathbf{x}$ , the prior mean  $p(\mathbf{x})$  and prior precision  $r(\mathbf{x})$  are estimated using kernel-weighted averages through

$$p(\mathbf{x}) = \sum_{i=1}^n w_i(\mathbf{x}) \cdot \frac{y_i}{m_i}, \quad r(\mathbf{x}) = r_0 \sum_{i=1}^n k(\mathbf{x}, \mathbf{x}_i),$$

where  $r_0 > 0$  is a global precision parameter and  $w_i(\mathbf{x})$  are normalized kernel weights defined by  $w_i(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}_i) / \sum_{j=1}^n k(\mathbf{x}, \mathbf{x}_j)$ . Then, the prior parameters at  $\mathbf{x}$  are given by

$$\alpha_0(\mathbf{x}) = r(\mathbf{x})p(\mathbf{x}), \quad \beta_0(\mathbf{x}) = r(\mathbf{x})\{1 - p(\mathbf{x})\}.$$

Table 1: Kernel functions implemented in **BKP**.

Kernel Type	Function $k(h)$
Gaussian	$k(h) = \exp(-h^2)$
Matérn $\nu = 5/2$	$k(h) = \left(1 + \sqrt{5}h + \frac{5}{3}h^2\right) \exp(-\sqrt{5}h)$
Matérn $\nu = 3/2$	$k(h) = \left(1 + \sqrt{3}h\right) \exp(-\sqrt{3}h)$

This data-adaptive formulation allows the prior to dynamically respond to the local sampling density. In well-sampled regions, it becomes more concentrated, while in sparse regions, it remains diffuse. Such adaptivity improves calibration in under-sampled areas and reduces overfitting where data are dense, thereby enhancing inference robustness under spatial heterogeneity. This strategy is consistent with the principle of local likelihood modeling ([Fan and Gijbels 1996](#)).

We offer some practical guideline to choose the global precision parameter. In classification when each input location receives a single categorical observation, we recommend choosing a relatively small value, such as  $0.01 \leq r_0 \leq 0.1$ , to prevent the prior domination and keep the posterior inference primarily data-driven. We illustrate this point in Example~1 of Section 4.1. For model fitting with binomial responses at each location, a reasonable choice is to set  $r_0$  to the mean number of trials per location, reflecting the typical information content of the data. In the absence of reliable prior information, a simple default is  $r_0 = 0.1$ , with subsequent adjustment guided by cross-validation or predictive performance over a range of candidate values (e.g.,  $r_0 \in \{0.001, 0.01, 0.1, 1, 5, 10\}$ ) to assess the robustness of predictions and decision boundaries.

### 2.3. Model Selection via Kernel Hyperparameter Tuning

#### *Kernel Functions*

Let  $h(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})$  denote the scaled Euclidean distance:

$$h(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) = \sqrt{\sum_{j=1}^d \left( \frac{x_j - x'_j}{\theta_j} \right)^2},$$

where  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_d)$  are positive kernel hyperparameters governing the relative importance of each input component. Based on this metric, define the kernel function as  $k(\mathbf{x}, \mathbf{x}') = k(h)$ , where the functional form of  $k(\cdot)$  determines the kernel type ([Rasmussen and Williams 2006](#)).

The **BKP** package implements several widely used kernel functions, namely Gaussian (squared-exponential) and Matérn families, summarized in Table 1.

#### *Loss Functions*

Hyperparameter tuning is conducted by minimizing a user-specified loss function evaluated using the leave-one-out cross-validation (LOOCV) method (Montesano and Lopes 2012). For each data point  $\mathbf{x}_i$ , the model is refitted up on the remaining data  $\mathcal{D}_n^{-i}$ , and the posterior mean  $\hat{\pi}_n^{-i}(\mathbf{x}_i; \boldsymbol{\theta})$  is used as the prediction. The LOOCV procedure is known to mitigate overfitting and produce better generalization performance than marginal likelihood-based approaches (Rasmussen and Williams 2006; Vehtari *et al.* 2017).

Currently, **BKP** supports two loss functions: the Brier score (i.e., squared error loss) and the log-loss (i.e., negative log-likelihood or cross-entropy).

**Brier Score** Let  $\tilde{\pi}_i = y_i/m_i$  denote the empirical success proportion at input location  $\mathbf{x}_i$ . The LOOCV Brier score is defined as

$$\text{BS}(\boldsymbol{\theta}; \mathcal{D}_n) = \frac{1}{n} \sum_{i=1}^n \left\{ \hat{\pi}_n^{-i}(\mathbf{x}_i; \boldsymbol{\theta}) - \tilde{\pi}_i \right\}^2, \quad (5)$$

which penalizes the squared deviation between the predicted and observed success proportions.

**Log-Loss** The log-loss is widely used for probabilistic classification. Based on the LOOCV predictions, it is defined as

$$\text{LL}(\boldsymbol{\theta}; \mathcal{D}_n) = -\frac{1}{n} \sum_{i=1}^n \left[ \tilde{\pi}_i \log \hat{\pi}_n^{-i}(\mathbf{x}_i; \boldsymbol{\theta}) + (1 - \tilde{\pi}_i) \log \left\{ 1 - \hat{\pi}_n^{-i}(\mathbf{x}_i; \boldsymbol{\theta}) \right\} \right]. \quad (6)$$

It corresponds to the negative log-likelihood of the binomial model evaluated at the LOOCV predictive probabilities.

**Remark 3.** Although both criteria rely on LOOCV estimates, they emphasize different aspects of predictive performance (Gneiting and Raftery 2007; Flores *et al.* 2025). The Brier score directly penalizes squared deviations from empirical proportions, making it more robust to poorly calibrated probabilities and better suited for smooth probability estimation under binomial data. In contrast, the log-loss penalizes overconfident mispredictions more heavily and is highly sensitive to the accuracy of the predicted probabilities. In practice, the choice between the two depends on whether calibration quality or robustness is prioritized in the evaluation (DRatings 2023).

### Hyperparameter Optimization

To improve optimization stability and reduce irregularities near the boundary of the parameter space, we adopt the reparameterization strategy proposed by MacDonald *et al.* (2015). Specifically, we transform the kernel scale parameters via  $\gamma_j = \log_{10}(\theta_j)$  for  $j = 1, \dots, d$ , and denote  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_d)$ . For example, the Gaussian kernel becomes

$$k(\mathbf{x}, \mathbf{x}'; \boldsymbol{\gamma}) = \exp \left\{ - \sum_{j=1}^d \left( \frac{x_j - x'_j}{10^{\gamma_j}} \right)^2 \right\}. \quad (7)$$

Although gradient-based optimizers such as L-BFGS-B (Byrd *et al.* 1995) are computationally efficient, their performance can be sensitive to initialization in non-convex settings. To improve robustness, we adopt a multi-start strategy based on a Latin Hypercube design (LHD) of  $n_0 = 10d$  initial runs within a carefully constructed region (Loeppky *et al.* 2009).

By MacDonald *et al.* (2015), the Gaussian kernel value  $k(\mathbf{x}, \mathbf{x}'; \boldsymbol{\gamma})$  typically lies within the interval  $[0.0067, 0.9999]$ , approximately  $[\exp(-5), \exp(-10^{-4})]$ , to ensure stable numerical behavior. Assuming a space-filling design of size  $n = 10d$  over  $[0, 1]^d$  and isotropic smoothness (i.e.,  $\gamma_j = \gamma_0$  for all  $j$ ), the minimum pairwise distance along each coordinate is roughly  $|x_{ik} - x_{jk}| \approx 0.1$ . This leads to the following search region for  $\boldsymbol{\gamma}$ :

$$\Omega_0 = \left[ \frac{\log_{10} d - \log_{10} 500}{2}, \frac{\log_{10} d + 2}{2} \right]^d. \quad (8)$$

For the mat'rn kernels, we adopt the same search region.

The initial values  $\boldsymbol{\gamma}^{(1)}, \dots, \boldsymbol{\gamma}^{(n_0)} \subset \Omega_0$  are drawn using LHD. For each initial value, solve

$$\hat{\boldsymbol{\gamma}}^{(i)} = \underset{\boldsymbol{\gamma} \in \Omega}{\operatorname{argmin}} L(\boldsymbol{\gamma}; \mathcal{D}_n), \quad i = 1, \dots, n_0,$$

where  $L(\cdot)$  is the chosen loss function from (5) or (6), and  $\Omega = [-3, 3]^d$  is a broader search region than  $\Omega_0$ . The final estimate is chosen as the best local minimum through

$$\hat{\boldsymbol{\gamma}} = \underset{1 \leq i \leq n_0}{\operatorname{argmin}} L(\hat{\boldsymbol{\gamma}}^{(i)}; \mathcal{D}_n).$$

The procedure is summarized below:

1. Generate  $10d$  initial values  $\boldsymbol{\gamma}$  in  $\Omega_0$  using a space-filling LHD.
2. Run the L-BFGS-B algorithm from each initial value.
3. Select the solution that produces the lowest loss function.

**Remark 4.** *The implementation of LOOCV necessitates refitting the model  $n$  times, which can be computationally intensive. However, the closed-form posterior updates in the BKP model mitigate this cost by reducing the computational complexity to  $\mathcal{O}(n^2)$ . This represents a substantial improvement over the typical  $\mathcal{O}(n^3)$  complexity of Gaussian process models with binomial likelihoods, as demonstrated in Example 3.*

## 2.4. Extension to Dirichlet Kernel Process

The BKP model can be naturally extended to handle multi-class responses via the Dirichlet Kernel Process (DKP) on replacing the binomial likelihood with a multinomial model and the Beta prior with a Dirichlet prior (MacKenzie *et al.* 2014).

Let the response at input  $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$  be  $\mathbf{y}(\mathbf{x}) = (y_1(\mathbf{x}), \dots, y_q(\mathbf{x}))$ , where  $y_s(\mathbf{x})$  denotes the count of class  $s$  out of  $m(\mathbf{x}) = \sum_{s=1}^q y_s(\mathbf{x})$  total trials. Assume

$$\mathbf{y}(\mathbf{x}) \sim \text{Multinomial}(m(\mathbf{x}), \boldsymbol{\pi}(\mathbf{x})),$$

with class probabilities  $\boldsymbol{\pi}(\mathbf{x}) = (\pi_1(\mathbf{x}), \dots, \pi_q(\mathbf{x}))$  satisfying  $\sum_{s=1}^q \pi_s(\mathbf{x}) = 1$ .

A Dirichlet prior is imposed on  $\boldsymbol{\pi}(\mathbf{x})$  as

$$\boldsymbol{\pi}(\mathbf{x}) \sim \text{Dirichlet}(\boldsymbol{\alpha}_0(\mathbf{x})),$$

where  $\boldsymbol{\alpha}_0(\mathbf{x}) = (\alpha_{0,1}(\mathbf{x}), \dots, \alpha_{0,q}(\mathbf{x}))$  are prior concentration parameters. Given training data  $\mathcal{D}_n = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ , define the response matrix as  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n]^\top \in \mathbb{R}^{n \times q}$ . Then the kernel-smoothed conjugate posterior becomes

$$\boldsymbol{\pi}(\mathbf{x}) \mid \mathcal{D}_n \sim \text{Dirichlet}(\boldsymbol{\alpha}_n(\mathbf{x})), \quad \text{with } \boldsymbol{\alpha}_n(\mathbf{x}) = \boldsymbol{\alpha}_0(\mathbf{x}) + \mathbf{k}^\top(\mathbf{x})\mathbf{Y}. \quad (9)$$

The posterior means and variances of the class probabilities are

$$\hat{\pi}_{n,s}(\mathbf{x}) = \frac{\alpha_{n,s}(\mathbf{x})}{\sum_{s'=1}^q \alpha_{n,s'}(\mathbf{x})}, \quad \sigma_{n,s}^2(\mathbf{x}) = \frac{\hat{\pi}_{n,s}(\mathbf{x})\{1 - \hat{\pi}_{n,s}(\mathbf{x})\}}{\sum_{s'=1}^q \alpha_{n,s'}(\mathbf{x}) + 1}, \quad s = 1, \dots, q.$$

For classification tasks, labels are assigned by the maximum a posteriori (MAP) decision rule through

$$\hat{y}(\mathbf{x}) = \underset{s \in \{1, \dots, q\}}{\operatorname{argmax}} \hat{\pi}_{n,s}(\mathbf{x}). \quad (10)$$

Prior choices (e.g., non-informative, fixed informative, or locally adaptive) follow similarly from the BKP framework in Section 2.2 by treating component-wise specification of prior class proportions.

Kernel hyperparameters are tuned by minimizing LOOCV-based loss functions, as described in Section 2.3. For multi-class responses, we consider the Brier score and the log-loss, defined respectively as

$$\text{BS}_{\text{multi}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \sum_{s=1}^q \left( \hat{\pi}_{n,s}^{-i}(\mathbf{x}_i; \boldsymbol{\theta}) - \tilde{\pi}_{i,s} \right)^2, \quad (11)$$

$$\text{LL}_{\text{multi}}(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^n \sum_{s=1}^q \tilde{\pi}_{i,s} \log \hat{\pi}_{n,s}^{-i}(\mathbf{x}_i; \boldsymbol{\theta}), \quad (12)$$

where  $\tilde{\pi}_{i,s} = y_{i,s}/m_i$  denotes the empirical class proportion. These multi-class criteria reduce to their binary counterparts when  $q = 2$ . For the log-loss, the expressions coincide exactly, while for the Brier score one obtains  $\text{BS}_{\text{multi}}(\boldsymbol{\theta})|q=2 = 2 \cdot \text{BS}(\boldsymbol{\theta})$ , due to the summation over both classes.

### 3. BKP Package

The **BKP** package provides the implementation of the Beta Kernel Process (BKP) and its extension to the Dirichlet Kernel Process (DKP). The package is designed with usability and extensibility in mind, offering a modular structure with separate functions for model fitting, prediction, simulation, kernel construction, prior specification, and hyperparameter tuning. Users can specify the kernel type and prior form to adapt to various application scenarios. The core functionality of the package is summarized in Table 2.

The package can be obtained from CRAN using `install.packages("BKP")` or the development version from GitHub via `pak::pak("Jiangyan-Zhao/BKP")`.

Table 2: Main functions and S3 methods provided in the **BKP** package.

Description	
<b>Functions</b>	
<code>fit_BKP()</code>	Fit a Beta Kernel Process (BKP) model for binary/binomial data.
<code>fit_DKP()</code>	Fit a Dirichlet Kernel Process (DKP) model for categorical or multinomial data.
<b>S3 Methods</b>	
<code>predict()</code>	Posterior prediction for BKP and DKP models.
<code>simulate()</code>	Posterior simulation for BKP and DKP models.
<code>summary()</code>	Summarize fitted BKP and DKP models.
<code>plot()</code>	Visualization of fitted BKP and DKP models.
<code>print()</code>	Print fitted models, predictions, simulations, and summaries.

### 3.1. BKP Model

We first provide a detailed description of the BKP model. The DKP model, presented in Section 3.2, highlights only the aspects that differ from the BKP implementation.

The function `fit_BKP()` fits the BKP model (1) for binomial or binary response data. Its main arguments are

```
fit_BKP(X, y, m, Xbounds = NULL,
         prior = "noninformative", r0 = 2, p0 = mean(y / m),
         loss = "brier", kernel = "gaussian", n_multi_start = NULL, theta = NULL)
```

The first group specifies the core data inputs: `X` is an  $n \times d$  input matrix, `y` is the vector of observed successes, and `m` is the corresponding vector of total trials. `Xbounds` rescales `X` to the unit hypercube  $[0, 1]^d$  for kernel hyperparameter optimization, and if `NULL` (default), `X` is assumed standardized.

The second group defines the prior for the latent success probability surface. `prior` selects the type ("noninformative", "fixed", or "adaptive"; see Section 2.2), and the prior is constructed using the `get_prior()` function. `r0` and `p0` specify prior precision and mean when applicable.

The third group specifies the loss function and optimization settings. `loss` chooses the criterion for kernel hyperparameter estimation ("brier" (default) or "log\_loss"), which is implemented via the `loss_fun()` function, and `kernel` selects the kernel function ("gaussian" (default), "matern32", or "matern52"). `n_multi_start` sets the number of random initial points for multi-start optimization (default  $10d$ ). Alternatively, `theta` can fix the kernel length scale directly, as a scalar (applied to all dimensions) or a vector of length  $d$ , skipping optimization. Optimization is performed using `multistart()` from the `optimx` package (Nash and Varadhan 2011; Nash 2014). Note that both `get_prior()` and `loss_fun()` are primarily invoked internally by `fit_BKP()`, but exporting them allows advanced users to inspect the construction of priors or customize the evaluation of the loss criterion if desired.

**Remark 5.** *The current version is implemented entirely in R. Computational efficiency can be further improved by implementing additional components in C++ via Rcpp (Eddelbuettel*

and François 2011), and by enabling options that leverage parallel computing architectures, for example, to accelerate multi-start optimization routines used in hyperparameter tuning.

`fit_BKP()` returns an object of class ‘BKP’, containing the estimated model hyperparameters  $\theta$  (`theta_opt`); the minimum achieved loss value (`loss_min`); the prior parameters  $\alpha_0(\mathbf{x})$  and  $\beta_0(\mathbf{x})$  (`alpha_0` and `beta_0`); the fitted posterior parameters  $\alpha_n(\mathbf{x})$  and  $\beta_n(\mathbf{x})$  (`alpha_n` and `beta_n`); and other input arguments.

The resulting ‘BKP’ object can be directly used as the `object` argument in the functions `predict()`, `simulate()`, and `summary()`, or as the `x` argument in the functions `plot()` and `print()`.

For a ‘BKP’ object `BKPmodel`, the function call

```
predict(BKPmodel, Xnew = NULL, CI_level = 0.95, threshold = 0.5, ...)
```

returns, for each  $\mathbf{x}$  in `Xnew` (or in the training data if `Xnew = NULL`), the predictive mean  $\hat{\pi}_n(\mathbf{x})$ , the predictive variance  $\sigma_n^2(\mathbf{x})$ , the lower and upper bounds of the `CI_level` credible interval, and the predictive parameters  $\alpha_n(\mathbf{x})$  and  $\beta_n(\mathbf{x})$ . If `m = 1`, it also returns the predictive binary label (0 or 1), determined by comparing the posterior mean to the specified `threshold` according to the classification rule in (4).

The function call

```
simulate(BKPmodel, nsim = 1, seed = NULL, Xnew = NULL, threshold = NULL, ...)
```

generates random samples from the posterior predictive distribution of a fitted BKP model at the input locations in `Xnew` (or at the training data if `Xnew = NULL`). Specifically, it draws `nsim` samples from the posterior Beta distribution at each row of `Xnew`, representing the distribution of success probabilities implied by the fitted model. If `threshold` is specified and `m = 1`, binary class labels (0 or 1) are generated by thresholding the simulated probabilities. An optional `seed` argument ensures reproducibility. Such posterior samples are useful for downstream decision-making tasks, for example Bayesian optimization via Thompson sampling (Garnett 2023).

The function call

```
summary(BKPmodel, ...)
```

produces a structured summary of a fitted ‘BKP’ object. It reports the number of training observations, input dimensionality, kernel type, prior specification, and the optimized kernel hyperparameters, together with the minimum achieved loss value. In addition, it provides posterior summaries at the training points, including the posterior mean and variance of the success probabilities. This function offers a convenient way to inspect the fitted model configuration and key inferential results.

To facilitate user interaction, the core ‘BKP’ object and the outputs of `predict()`, `simulate()`, and `summary()` are all implemented as S3 class objects. Specifically, the ‘BKP’ object itself is equipped with a customized `print()` method for concise display, while the outputs of `predict()`, `simulate()`, and `summary()` are returned as dedicated classes (i.e., ‘`predict_BKP`’,

‘simulate\_BKP’, ‘summary\_BKP’). Assigning these classes not only provides customized `print()` methods for human-readable summaries, but also supports future extensibility and backward-compatible addition of new S3 methods, such as `plot()` or specialized extraction functions.

The `plot()` method provides graphical visualization of a fitted ‘BKP’ object, adapting to the input dimensionality. For one-dimensional inputs ( $d = 1$ ), it displays the posterior mean of the predicted success probability as a line plot, with a shaded region representing the 95% credible interval and the observed proportions ( $\tilde{\pi}_i = y_i/m_i$ ) overlaid as points; for classification tasks, an optional curve of the maximum posterior probability can be added to illustrate decision confidence. For two-dimensional inputs ( $d = 2$ ), the method generates contour plots over a 2D prediction grid, by default producing four summary surfaces: the predictive mean, the predictive variance, and the 2.5% and 97.5% quantiles of the posterior distribution, providing a comprehensive view of predicted probabilities and associated uncertainty; if the argument `only_mean = TRUE` is specified, only the contour plot of the predictive mean surface is displayed. For high-dimensional inputs ( $d > 2$ ), users must specify which dimensions to plot via the `dims` argument.

In addition to the core methods in Table 2, the **BKP** package provides auxiliary methods for inspecting fitted models, including `fitted()` to extract fitted values (i.e., predictive means), `parameter()` to retrieve model parameters ( $\alpha_n(\mathbf{x})$  and  $\beta_n(\mathbf{x})$ ), and `quantile()` to obtain posterior quantiles, with the quantile levels specified by the argument `probs`, from a fitted ‘BKP’ object. These functions complement the main workflow and facilitate exploration of model outputs and posterior summaries.

### 3.2. DKP Model

The function `fit_DKP()` implements the DKP model~(9), extending the BKP framework to multinomial response data. The response  $\mathbf{Y}$  should be an  $n \times q$  matrix of observed counts, where each row represents an observation and each column a category. The `m` argument is not required, as the total count is implicitly defined by the row sums of  $\mathbf{Y}$ . When using a fixed prior (`prior = "fixed"`), the user must provide a global prior mean vector  $\mathbf{p}^0$  of length  $q$ , which defaults to the empirical class proportions `colMeans(Y / rowSums(Y))`.

The function returns an object of class ‘DKP’, which contains elements analogous to those in a ‘BKP’ object, including optimized kernel hyperparameters, the selected kernel function, the loss specification, and prior information. Distinctively, the DKP posterior parameters are vectors  $\boldsymbol{\alpha}_n(\mathbf{x})$  associated with each input  $\mathbf{x}$ , representing the smoothed Dirichlet parameters across categories. Other elements, such as the original and normalized input matrices, normalization bounds, and observed counts, are also included to support downstream prediction, simulation, and visualization.

Prediction, simulation, and visualization methods extend directly from the class ‘BKP’. The `predict()` method returns per-class predictive means, variances, and credible intervals; if all row sums of  $\mathbf{Y}$  are 1, it additionally provides predicted class labels following the MAP classification rule in (10). The `simulate()` method draws samples from the posterior predictive Dirichlet distribution and can optionally generate sampled class labels via Thompson sampling. For one-dimensional inputs ( $d = 1$ ), `plot()` displays one curve per class with credible intervals, while for two-dimensional inputs ( $d = 2$ ), it produces a panel of contour plots illustrating the predictive mean or other posterior summaries for each class.

## 4. Examples using BKP

### 4.1. BKP Model

This subsection presents detailed illustrative examples based on the BKP model. Examples for the DKP model are provided in Section 4.2.

**Example 1** Let  $x \in [-2, 2]$ , and suppose the true Bernoulli probability function is given by

$$\pi_1(x) = \frac{1}{1 + e^{-3x}}, \quad (13)$$

which is referred as the function `true_pi_fun` in the code below. We aim to fit the BKP model based on seven input locations that are uniformly distributed over  $[-2, 2]$ , with each location associated with a binomial observation having a maximum trial count of 100. The input locations are generated using the `lhs()` function from the R package `tgp` (Gramacy and Taddy 2010). The following R code illustrates how to simulate the data and fit the BKP model using the `fit_BKP()` function.

```
R> n <- 7
R> Xbounds <- matrix(c(-2,2), nrow = 1)
R> X <- lhs(n = n, rect = Xbounds)
R> true_pi <- true_pi_fun(X)
R> m <- sample(100, n, replace = TRUE)
R> y <- rbinom(n, size = m, prob = true_pi)
R> BKP_model_1D_1 <- fit_BKP(X, y, m, Xbounds = Xbounds)
```

The estimates of the parameters of the fitted BKP model can be displayed using the `print()` function:

```
R> print(BKP_model_1D_1)
```

```
Beta Kernel Process (BKP) Model

Number of observations (n): 7
Input dimensionality (d): 1
Kernel type: gaussian
Optimized kernel parameters: 0.1748
Minimum achieved loss: 0.01165
Loss function: brier
Prior type: noninformative
```

The `BKP_model_1D_1` object can be used for visualization and prediction over a grid of input values via the `plot()` and `simulate()` methods:

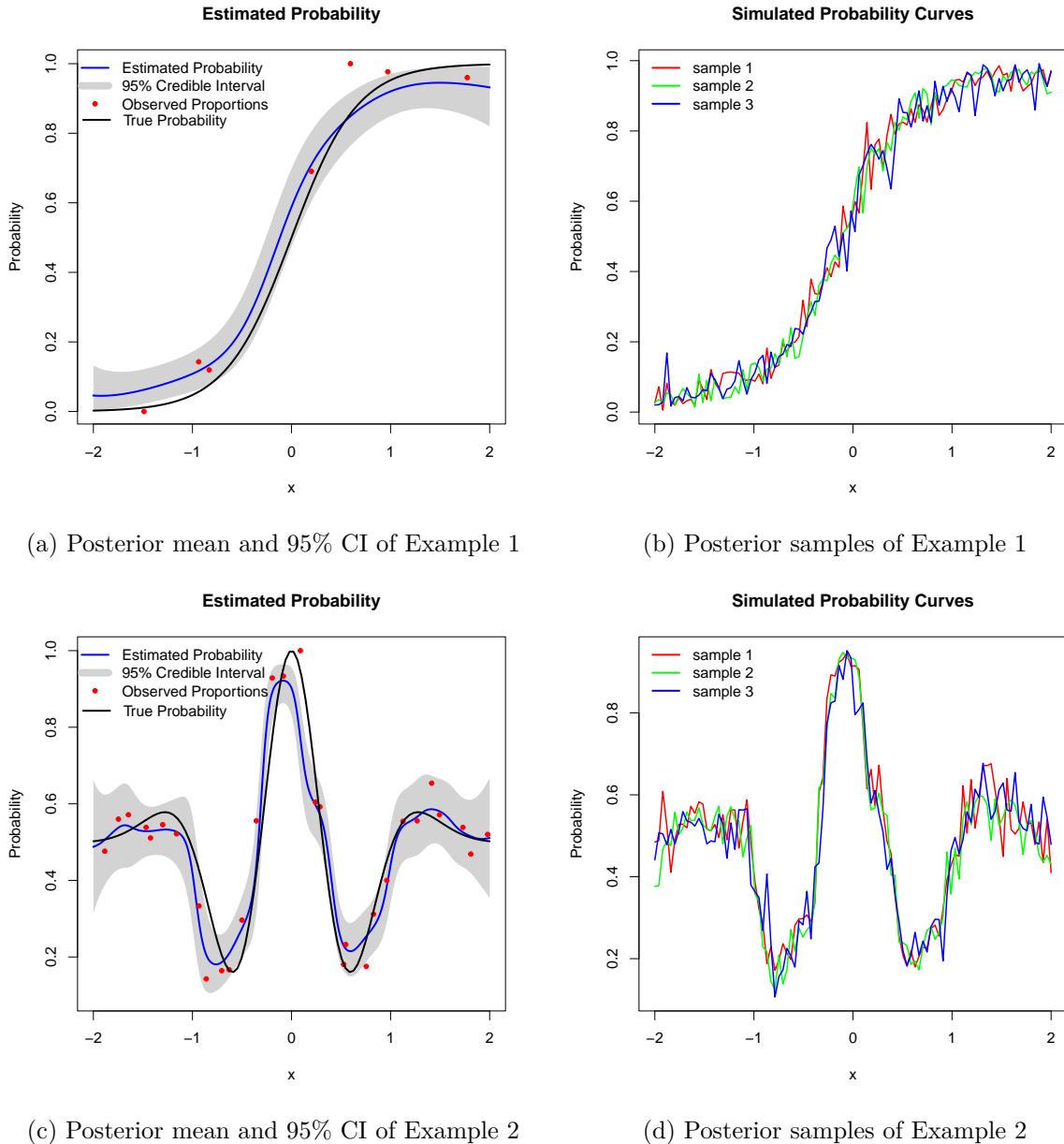


Figure 1: Posterior inference and simulation results from the fitted BKP models

```
R> plot(BKP_model_1D_1)
R> Xnew = matrix(seq(-2, 2, length = 100), ncol = 1)
R> sim <- simulate(BKP_model_1D_1, nsim = 3, Xnew = Xnew)
```

Figure 1 presents two views of the model output. Panel~(a) shows the posterior mean estimate of the probability function  $\pi(x)$  (blue), along with a 95% credible interval (gray band), observed proportions (red dots), and the true underlying probability function (black). Panel~(b) presents three posterior sample curves generated using the `simulate()` method, illustrating the variability of the estimated probability surface.

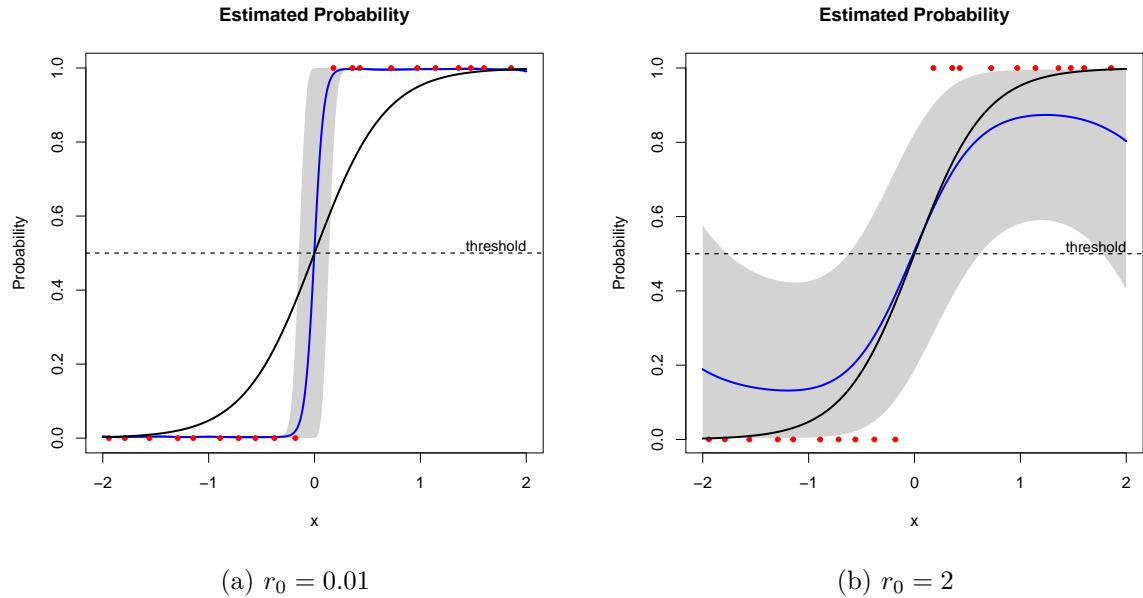


Figure 2: Posterior mean and 95% CI of Example 1 (classification task) under  $r_0 = 0.01$  and 2

We continue with the same probability function of this example to show the impact of the global precision parameters,  $r_0$ , in classification tasks. However, the sample size is changed from 7 to 20 for classification. The following R code generates the label of response and fit the BKP model for classification with  $r_0$  being 0.01 and 2.

```
R> # Fit BKP model with r0 = 0.01
R> BKP_model_1D_1_class_1 <- fit_BKP(
+   X, y, m, Xbounds = Xbounds,
+   prior = "fixed", r0 = 0.01, loss = "log_loss")
R> # Fit BKP model with r0 = 2
R> BKP_model_1D_1_class_2 <- fit_BKP(
+   X, y, m, Xbounds = Xbounds,
+   prior = "fixed", r0 = 2, loss = "log_loss")
```

Figure 2a shows a sigmoidal curve with steep slope around zero (in solid line) and a narrow 95% credible interval (in grey band) under  $r_0 = 0.01$ , indicating a decisive and confident classification boundary. In contrast, Figure 2b displays a sine-shape curve with a much wider credible interval, reflecting greater uncertainty and less effective separation. This demonstrates the preference of a small value of  $r_0$  value for classification task.

**Example 2** The first example is essentially a generalized linear model with a smooth logit link, and thus poses limited modeling complexity. To demonstrate the capability of the BKP model in handling more challenging classification structures, we consider a second example with a highly nonlinear underlying probability surface. Define the true Bernoulli probability

as

$$\pi_2(x) = \frac{1}{2} \left[ 1 + e^{-x^2} \cos \left( 10 \frac{1 - e^{-x}}{1 + e^{-x}} \right) \right], \quad (14)$$

where  $x \in [-2, 2]$  (Goetschalckx *et al.* 2011). This example involves rapid local oscillations and strong nonlinearity, making it substantially more difficult to fit than Example~1. Here, we increase the number of locations to 30.

```
R> n <- 30
R> Xbounds <- matrix(c(-2,2), nrow = 1)
R> X <- lhs(n = n, rect = Xbounds)
R> true_pi <- true_pi_fun(X)
R> m <- sample(100, n, replace = TRUE)
R> y <- rbinom(n, size = m, prob = true_pi)
R> BKP_model_1D_2 <- fit_BKP(X, y, m, Xbounds = Xbounds)
```

The results are shown in panels~(c) and (d) of Figure 1. Panel~(c) demonstrates that the BKP model accurately recovers the highly nonlinear pattern, while panel~(d) illustrates posterior variability through three representative realizations of  $\pi(x)$ .

**Example 3** We now consider a two-dimensional test function to further illustrate the modeling capabilities of the **BKP** package. Let  $\mathbf{x} \in [0, 1]^2$ , and define the latent surface using a re-scaled version of the Goldstein–Price function (Picheny *et al.* 2013):

$$\begin{aligned} f(\mathbf{x}) &= \frac{\log[\{1 + a(\mathbf{x})\}\{30 + b(\mathbf{x})\} - 8.6928]}{2.4269}, \quad \text{with} \\ a(\mathbf{x}) &= (4x_1 + 4x_2 - 3)^2 \times \\ &\quad \{75 - 56(x_1 + x_2) + 3(4x_1 - 2)^2 + 6(4x_1 - 2)(4x_2 - 2) + 3(4x_2 - 2)^2\}, \\ b(\mathbf{x}) &= (8x_1 - 12x_2 + 2)^2 \times \\ &\quad \{-14 - 128x_1 + 12(4x_1 - 2)^2 + 192x_2 - 36(4x_1 - 2)(4x_2 - 2) + 27(4x_2 - 2)^2\}. \end{aligned}$$

The true Bernoulli probability surface is then defined by

$$\pi_3(\mathbf{x}) = \Phi\{f(\mathbf{x})\}, \quad (15)$$

where  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal distribution. This formulation produces a smooth yet highly non-linear response surface, providing a challenging test scenario for probabilistic modeling.

To construct the training data, we generate a LHD of size 100 over  $[0, 1]^2$ . Each location is associated with a binomial observation whose number of trials is randomly drawn from  $\{1, \dots, 100\}$ . The following R code demonstrates the data simulation and model fitting using the `fit_BKP()` function:

```
R> n <- 100
R> Xbounds <- matrix(c(0, 0, 1, 1), nrow = 2)
R> X <- lhs(n = n, rect = Xbounds)
```

```
R> true_pi <- true_pi_fun(X)
R> m <- sample(100, n, replace = TRUE)
R> y <- rbinom(n, size = m, prob = true_pi)
R> BKP_model_2D <- fit_BKP(X, y, m, Xbounds=Xbounds)
R> print(BKP_model_2D)
```

### Beta Kernel Process (BKP) Model

```
Number of observations (n): 100
Input dimensionality (d): 2
Kernel type: gaussian
Optimized kernel parameters: 0.1112, 0.0680
Minimum achieved loss: 0.01041
Loss function: brier
Prior type: noninformative
```

Figure 3 is obtained by the following code, which is the heatmaps with contour lines for the true probability and the estimated.

```
R> plot(BKP_model_2D)
```

The results of model prediction are as follows:

```
R> Xnew <- lhs(n = 10, rect = Xbounds)
R> predict(BKP_model_2D, Xnew)

Prediction results on new data (Xnew).
Total number of prediction points: 10

Preview of predictions for new data (first 6 of 10 points):
  x1     x2   mean variance 2.5% quantile 97.5% quantile
0.8249 0.0605 0.7106   0.0012      0.6394      0.7770
0.3784 0.5314 0.3324   0.0009      0.2754      0.3919
0.0566 0.1437 0.5519   0.0032      0.4409      0.6603
0.5402 0.8851 0.7994   0.0033      0.6756      0.8993
0.7031 0.7113 0.3282   0.0009      0.2707      0.3883
0.2709 0.6661 0.6178   0.0013      0.5460      0.6871
...
...
```

We continue with Example 3 to demonstrate the scalability of the BKP model in comparison to the logistic Gaussian process (LGP) model (Rasmussen and Williams 2006). Let the sample size range from 200 to 5000. Two scenarios for hyperparameter determination are considered, namely, i) one scenario with fixed value of `theta = 1`, and ii) the other scenario with hyperparameters optimized (through `n_multi_start = 1` or `n_multi_start = NULL`).

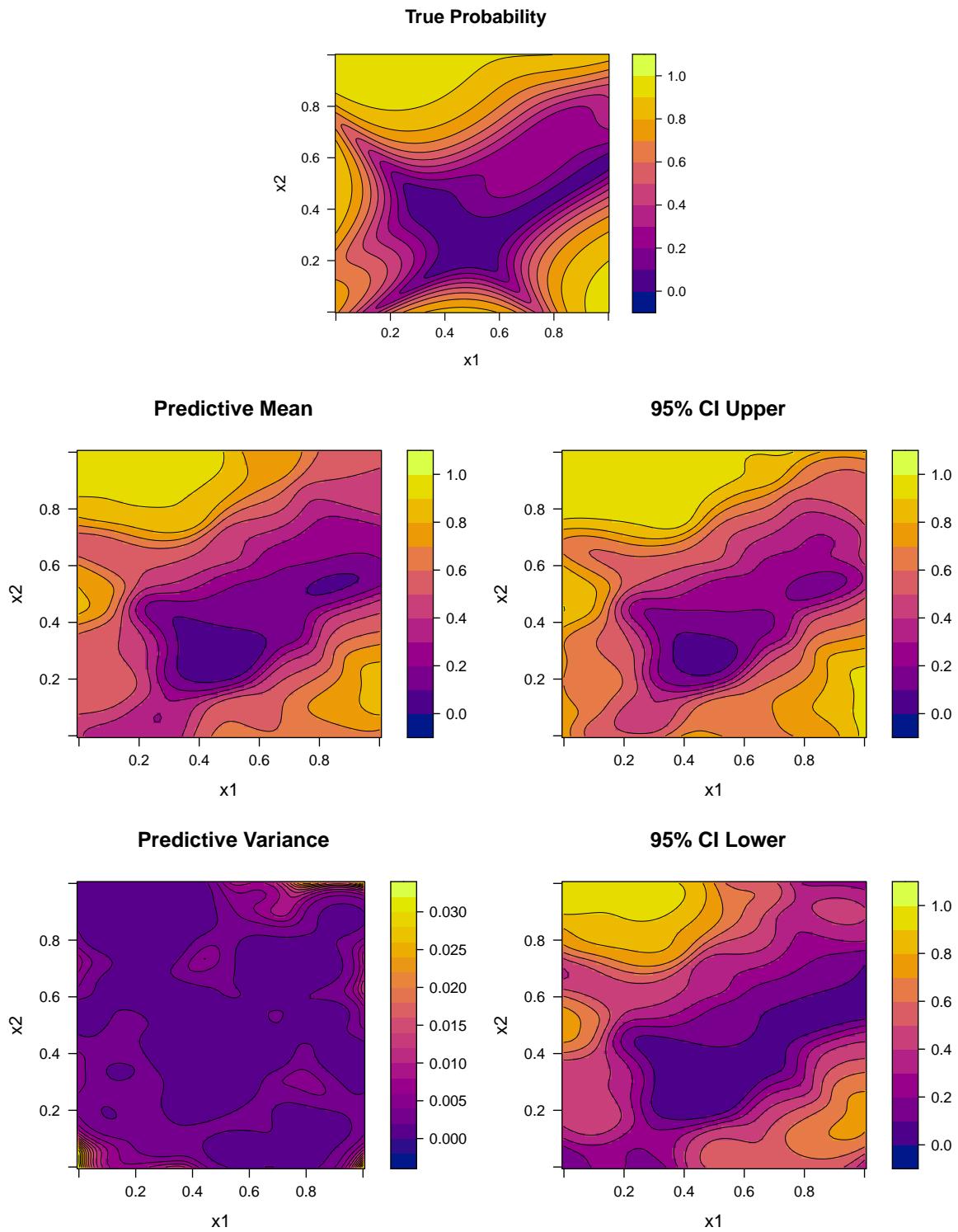


Figure 3: Comparison of the true probability surface and BKP-based posterior summaries of Example 3

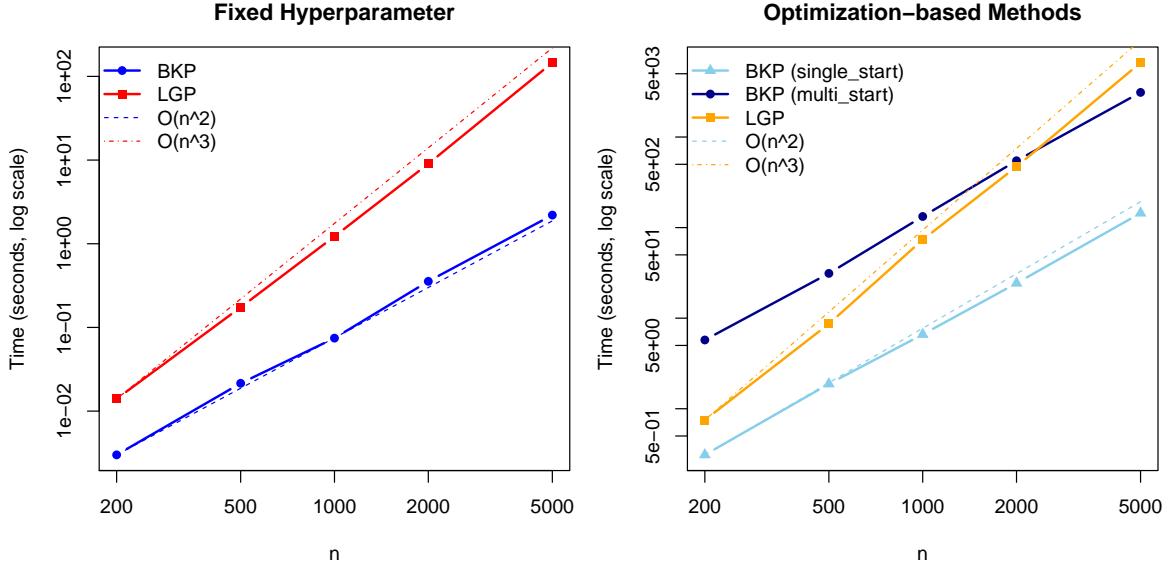


Figure 4: Comparison of computation times (in log scale) between BKP and LGP methods: (a) fixed hyperparameter; (b) optimization-based methods

The LGP model was implemented using the `gp_fit()` and `gp_optim()` functions from the `gplite` package (Piironen 2022). Notably, the `gplite` package employs a multi-start optimization strategy that is only triggered upon optimization failure; therefore, the additional computational cost of multiple restarts was not included in our timing measurements.

All experiments were conducted on a workstation equipped with an Intel(R) Xeon(R) W-2235 CPU @ 3.80GHz (12 cores) and 16 GB RAM. The computation times reported here are averages over 20 independent repetitions to mitigate the effects of variability due to random initialization and computational fluctuations.

The average computation times for both models under these scenarios are presented in Figure 4. The observed computational costs align well with the theoretical complexity predictions, confirming the expected  $\mathcal{O}(n^2)$  scalability for BKP and  $\mathcal{O}(n^3)$  for the LGP model. Additionally, the computational cost of the multi-start optimization approach for BKP is roughly proportional to the number of restarts, which was set to  $10d$ . For the current setting, this corresponds to approximately 20 times the cost of the single-start method, reflecting the increased complexity incurred by multiple initializations. This trade-off between computational expense and potentially improved optimization robustness should be carefully considered when selecting an appropriate strategy.

**Example 4** We next consider a binary classification task using the *Two Spirals* dataset (Chalup and Wiklundt 2007), a well-known benchmark consisting of two intertwined spirals in a bounded two-dimensional input space. This dataset is particularly challenging due to the complex, non-linearly separable class structure.

We generate  $n = 250$  observations using the `mlbench.spirals()` function from the R package `mlbench` (Leisch and Dimitriadou 2024), with two complete rotations and additive Gaussian noise of standard deviation `sd = 0.05`. The inputs  $\mathbf{x}$  are constrained to the domain

$[-1.7, 1.7]^2$ , and the binary class labels are encoded as 0 and 1. We fit the BKP model using a fixed prior specification with  $r_0 = 0.1$  and  $p_0 = 0.5$ .

```
R> n <- 250
R> n_train <- 200
R> n_test <- 50
R> data <- mlbench.spirals(n, cycles = 2, sd = 0.05)
R> X_train <- data$x[1:n_train, ]
R> y_train <- as.numeric(data$classes[1:n_train]) - 1 # Convert to 0/1 for BKP
R> X_test <- data$x[(n_train + 1):n, ]
R> y_test <- as.numeric(data$classes[(n_train + 1):n]) - 1
R> m <- rep(1, n_train)
R> Xbounds <- rbind(c(-1.7, 1.7), c(-1.7, 1.7))
R> BKP_model_Class <- fit_BKP(
+   X_train, y_train, m, Xbounds = Xbounds,
+   prior = "fixed", r0 = 0.1, loss = "log_loss")
R> prediction <- predict(BKP_model_Class, X_test)
```

Figure 5 presents the predictive mean and variance surfaces of the class probabilities. The upper left panel shows that the BKP model accurately captures the intricate spiral structure, with smoothly varying predicted probabilities that delineate the nonlinear decision boundary. The upper right panel displays the associated predictive uncertainty, which is highest near the boundary regions between the spirals. Circles and crosses indicate training observations from the two respective classes. These results illustrate the capacity of BKP to flexibly model complex classification boundaries while providing coherent uncertainty quantification.

We compare the prediction performance with the LGP model implemented by R package **gplite**.

```
R> gp <- gp_init(cf = cf_sexp(), lik = lik_bernoulli())
R> gp <- gp_optim(gp, X_train, y_train, method = method_full(),
+                   approx = approx_ep(), verbose = FALSE)
R> prediction_gp <- gp_pred(gp, as.matrix(X_test), transform = TRUE)
```

The lower panel of Figure 5 indicates that the LGP model encounters difficulty in capturing the intricate geometry of the Two Spirals dataset. Although it roughly follows the spiral structure, its predictive mean exhibits a more irregular and less sharply defined decision boundary compared to the BKP model. This pattern suggests that the LGP model may be more prone to overfitting or less capable of generalizing effectively. The corresponding predictive variance plots reinforce this interpretation when viewed separately in regions with and without training data. In regions where data are available, the LGP model exhibits elevated variance (brighter regions) across a broader portion of the input space indicating greater uncertainty in areas critical for classification. In contrast, the BKP model maintains lower variance concentrated more closely around the decision boundary, reflecting higher confident predictions. Interestingly, in regions without training data, the LGP model's uncertainty is not necessarily the highest; instead, it sometimes shows comparatively lower variance than the BKP model. This suggests that the LGP model may over-smooth or fail to express appropriate uncertainty

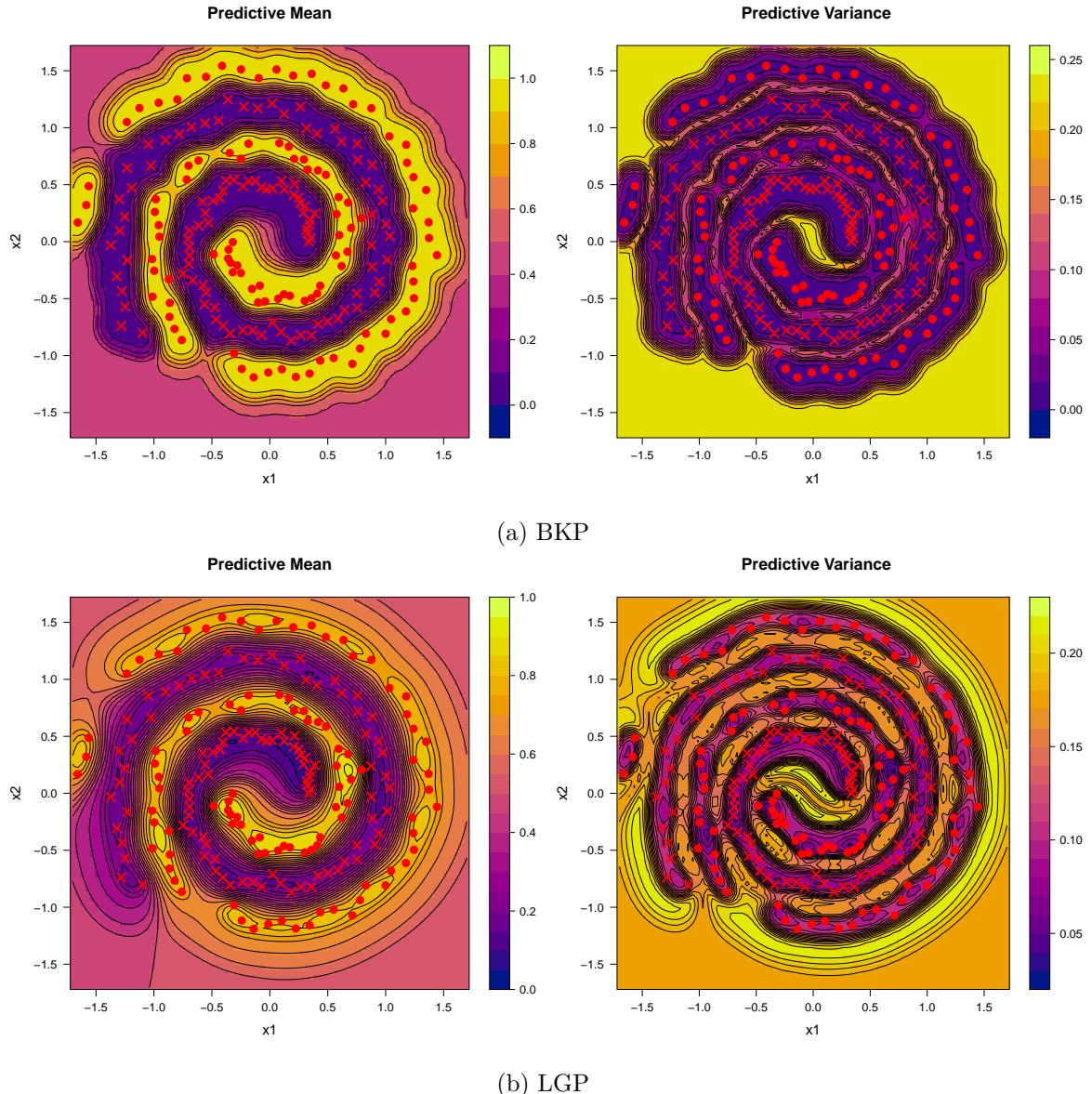


Figure 5: Predictions of the two spirals by BKP and LGP models in terms of posterior predictive mean (left panel) and predictive variance (right panel) of class probabilities, where color shading represents the estimated success probability, circles and crosses indicate training observations from the two classes

far from observed data, whereas BKP better captures uncertainty behavior in unobserved regions. The receiver operating characteristic (ROC) curves and the corresponding area under the curve (AUC) values in Figure 6 further corroborate this conclusion: BKP achieves an AUC of 0.926, compared to 0.889 for LGP, reflecting superior discriminatory performance and a better balance between sensitivity and specificity.

#### 4.2. DKP Model

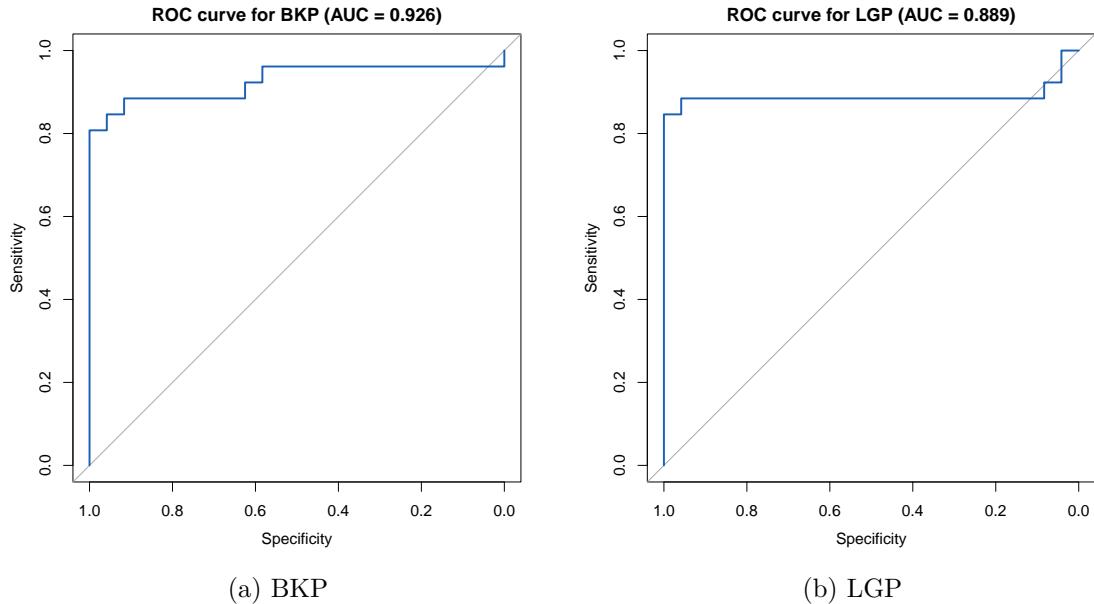


Figure 6: ROC curves and their respective AUCs for BKP and LGP models

**Example 5** Consider a one-dimensional three-class classification problem. The input is defined on the interval  $x \in [-2, 2]$ , and the true class probability vector is given by

$$\boldsymbol{\pi}(x) = \left[ \frac{\pi_1(x)}{2}, \frac{\pi_2(x)}{2}, 1 - \frac{\pi_1(x)}{2} - \frac{\pi_2(x)}{2} \right]^\top,$$

where  $\pi_1(x)$  and  $\pi_2(x)$  are smooth functions defined in (13) and (14), respectively.

We generate  $n = 30$  input locations using Latin hypercube sampling over the interval  $[-2, 2]$ . At each location, the response is a multinomial vector with probability  $\boldsymbol{\pi}(x)$  and a random total count sampled from  $\{1, \dots, 150\}$ . The DPK model is then fitted using the `fit_DPK()` function and visualized with the `plot()` method:

```
R> n <- 30
R> Xbounds <- matrix(c(-2, 2), nrow = 1)
R> X <- lhs(n = n, rect = Xbounds)
R> true_pi <- true_pi_fun(X)
R> m <- sample(150, n, replace = TRUE)
R> Y <- t(sapply(1:n, function(i) rmultinom(1, size = m[i], prob = true_pi[i, ])))
R> DPK_model_1D <- fit_DPK(X, Y, Xbounds = Xbounds)
R> plot(DPK_model_1D)
```

Figure 7 shows that the DPK model accurately recovers the true class probability functions and provides meaningful uncertainty quantification. The predictive mean curves align with the true underlying structure, and the shaded bands reflect posterior uncertainty.

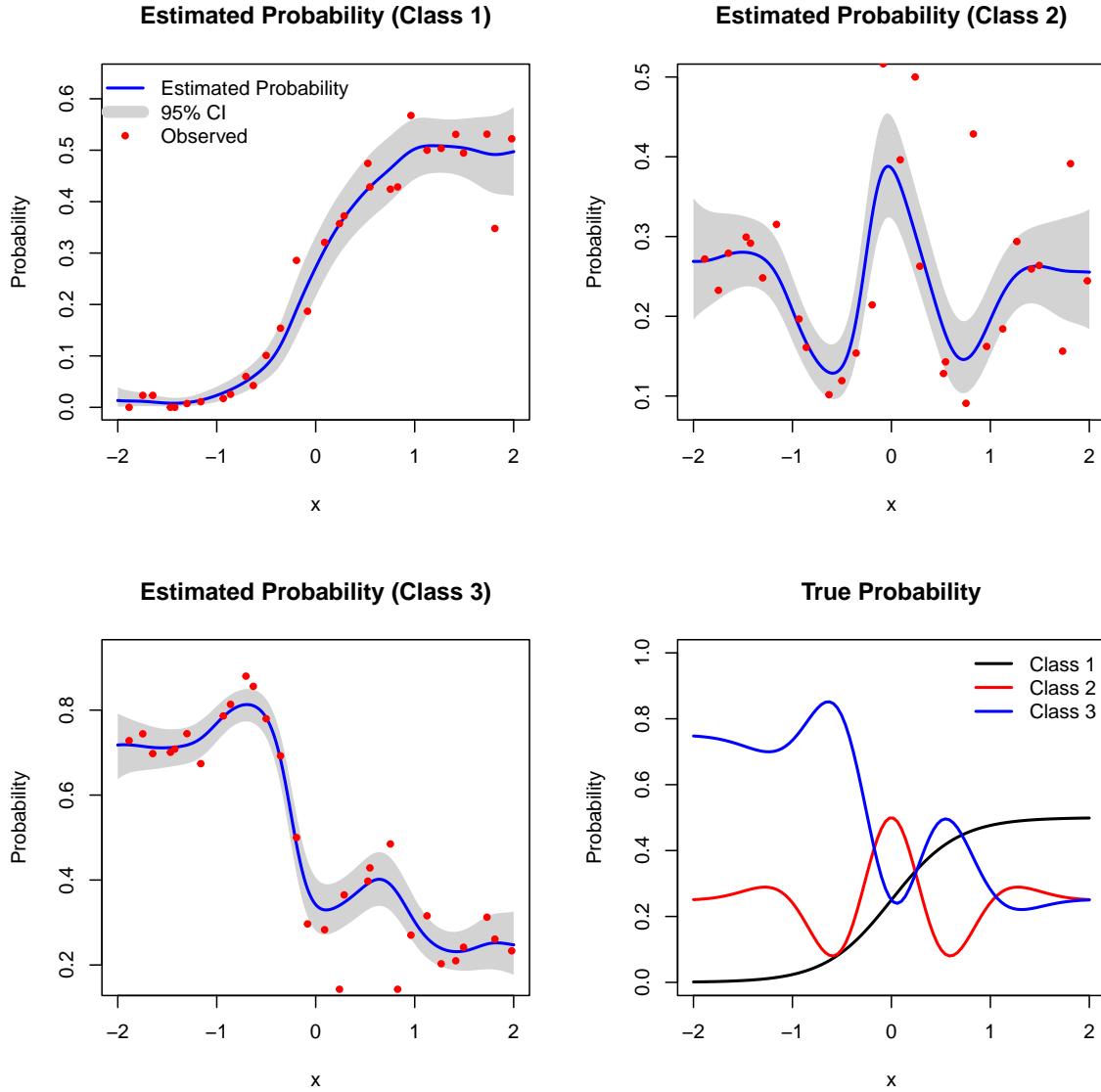


Figure 7: Posterior inference from the fitted DKP model for a one-dimensional three-class problem

**Example 6** Consider a two-dimensional three-class classification problem. Let  $\mathbf{x} = [x_1, x_2]^\top \in [0, 1]^2$ , and define the true class probability function as

$$\boldsymbol{\pi}(\mathbf{x}) = \left[ \frac{\pi_3(\mathbf{x})}{2}, \frac{\pi_4(\mathbf{x})}{2}, 1 - \frac{\pi_3(\mathbf{x})}{2} - \frac{\pi_4(\mathbf{x})}{2} \right]^\top,$$

where  $\pi_3(\mathbf{x})$  is defined in (15), and

$$\pi_4(\mathbf{x}) = \sin(\pi x_1) \sin(\pi x_2).$$

We generate  $n = 100$  input points using Latin hypercube sampling over  $[0, 1]^2$ . At each location, the response is a multinomial vector with probability  $\boldsymbol{\pi}(\mathbf{x})$  and a total count randomly

sampled from  $\{1, \dots, 150\}$ . We fit the DKP model using `fit_DKP()` and visualize the results with the `plot()` method:

```
R> n <- 100
R> Xbounds <- matrix(c(0, 0, 1, 1), nrow = 2)
R> X <- lhs(n = n, rect = Xbounds)
R> true_pi <- true_pi_fun(X)
R> m <- sample(150, n, replace = TRUE)
R> Y <- t(sapply(1:n, function(i) rmultinom(1, size = m[i], prob = true_pi[i, ])))
R> DKP_model_2D <- fit_DKP(X, Y, Xbounds=Xbounds)
```

Figures~8–10 display the ground truth surfaces and corresponding posterior inference for each of the three classes. In each figure, the top panel shows the true class probability surface, while the bottom row presents the posterior mean, variance, and uncertainty bounds produced by the DKP model.

**Example 7** Consider another three-class classification task based on the well-known *Iris* dataset available in R. This classic benchmark dataset contains measurements of three iris species—*setosa*, *versicolor*, and *virginica*—each described by four features: sepal length, sepal width, petal length, and petal width. Due to the overlap in feature space, particularly between *versicolor* and *virginica*, the class boundaries are not linearly separable, making this dataset a standard testbed for evaluating multi-class classification algorithms. For visualization purposes, we restrict our analysis to the first two features: sepal length and sepal width. We fit the DKP model using a fixed prior specification with `r0 = 0.1` and `p0 = rep(1/3, 3)`, assuming equal prior probability for each class.

```
R> data(iris)
R> X <- as.matrix(iris[, 1:2])
R> Xbounds <- rbind(c(4.2, 8), c(1.9, 4.5))
R> labels <- iris$Species
R> Y <- model.matrix(~ labels - 1) # expand factors to a set of dummy variables
R> train_indices <- sample(1:nrow(iris), 0.7 * nrow(iris))
R> X_train <- X[train_indices, ]
R> Y_train <- Y[train_indices, ]
R> DKP_model_Class <- fit_DKP(
+   X_train, Y_train, Xbounds = Xbounds, loss = "log_loss",
+   prior = "fixed", r0 = 0.1, p0 = rep(1/3, 3))
R> X_test <- X[-train_indices, ]
R> Y_test <- Y[-train_indices, ]
R> dkp_pred_probs <- predict(DKP_model_Class, X_test)$mean
```

Figure 11 provides a visual summary of the fitted DKP model. The upper left panel displays the MAP classification boundaries, which clearly separate *setosa* from the other two species. In contrast, the boundary between *versicolor* and *virginica* is more intricate, reflecting the known overlap between these two species in sepal measurements.

To further illustrate the model’s prediction uncertainty, the upper right panel shows a contour map of the maximum predicted probability  $\max_j \pi_j(\mathbf{x})$  across the input space. This diagnostic

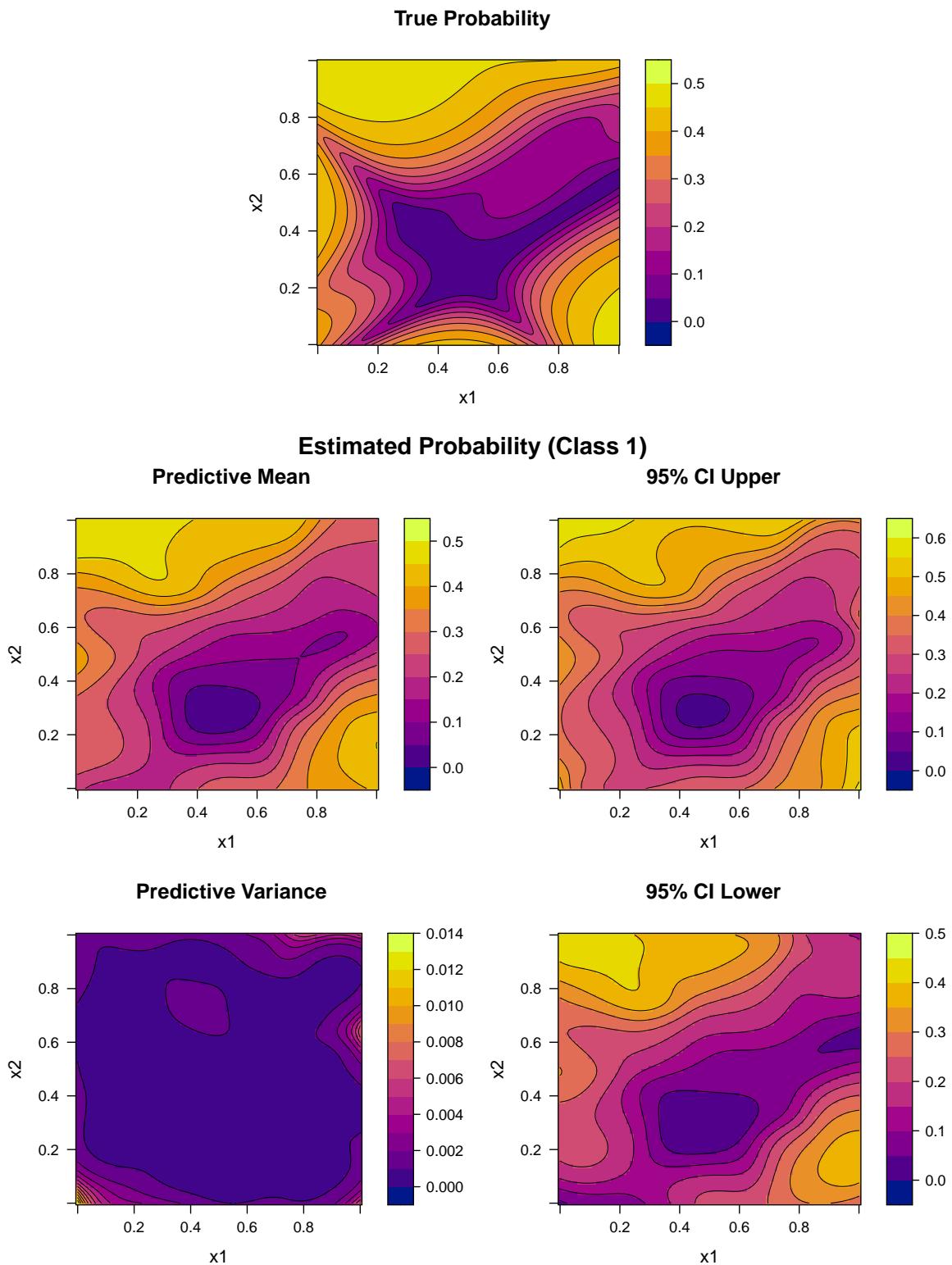


Figure 8: Class 1: true distribution (top) and DKP model posterior summaries (bottom)

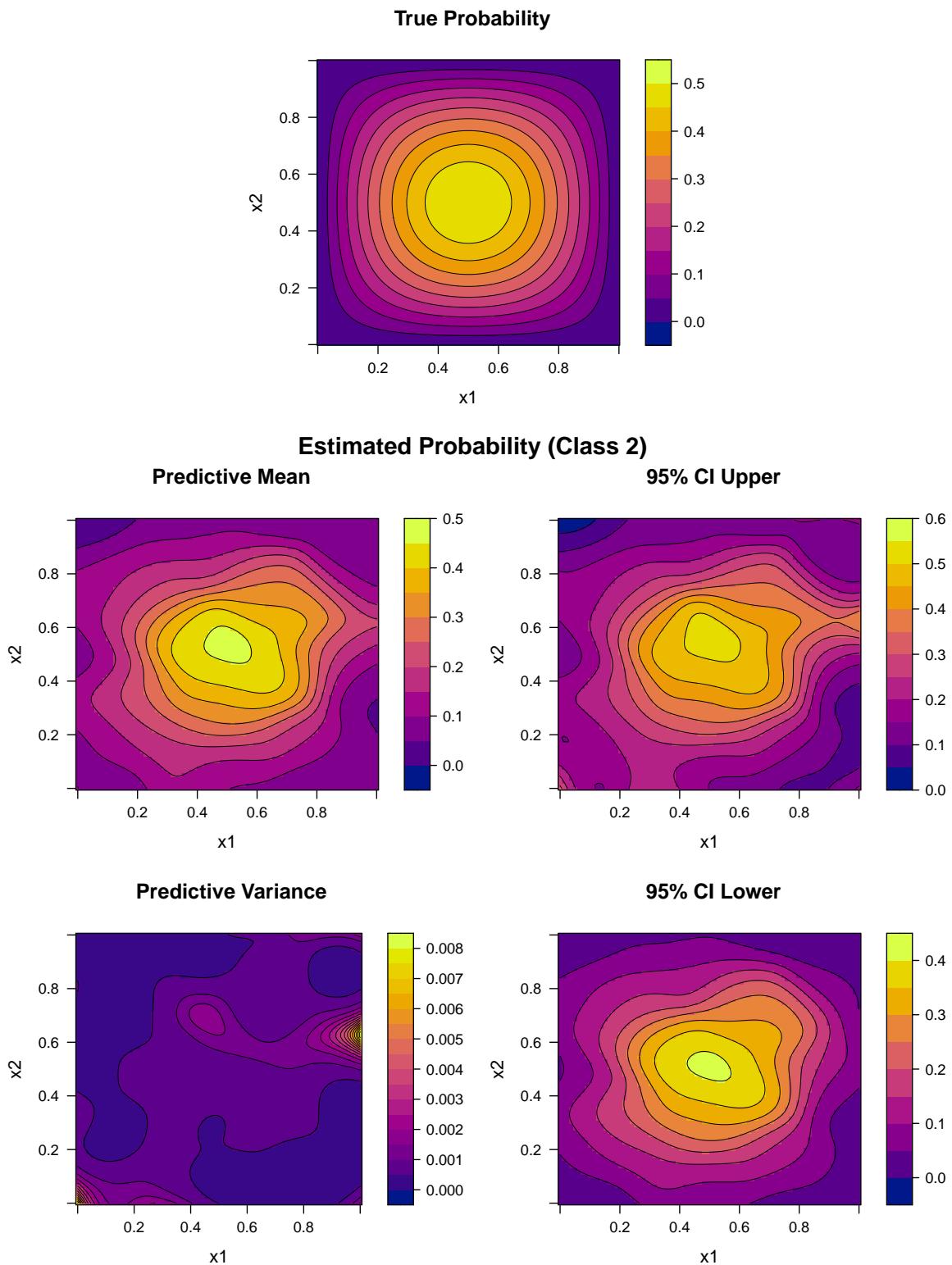


Figure 9: Class 2: true distribution (top) and DKP model posterior summaries (bottom)

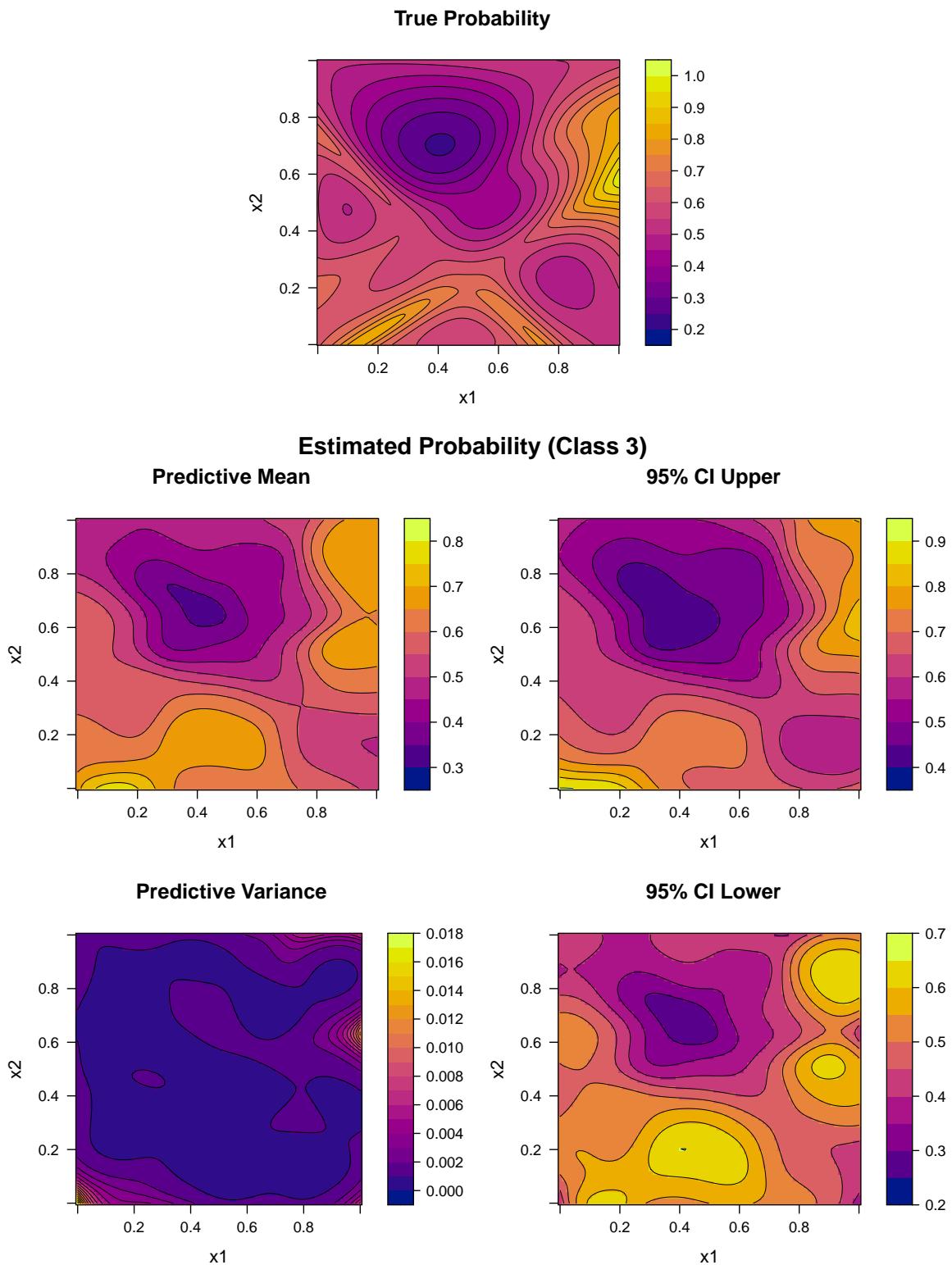


Figure 10: Class 3: true distribution (top) and DKP model posterior summaries (bottom)

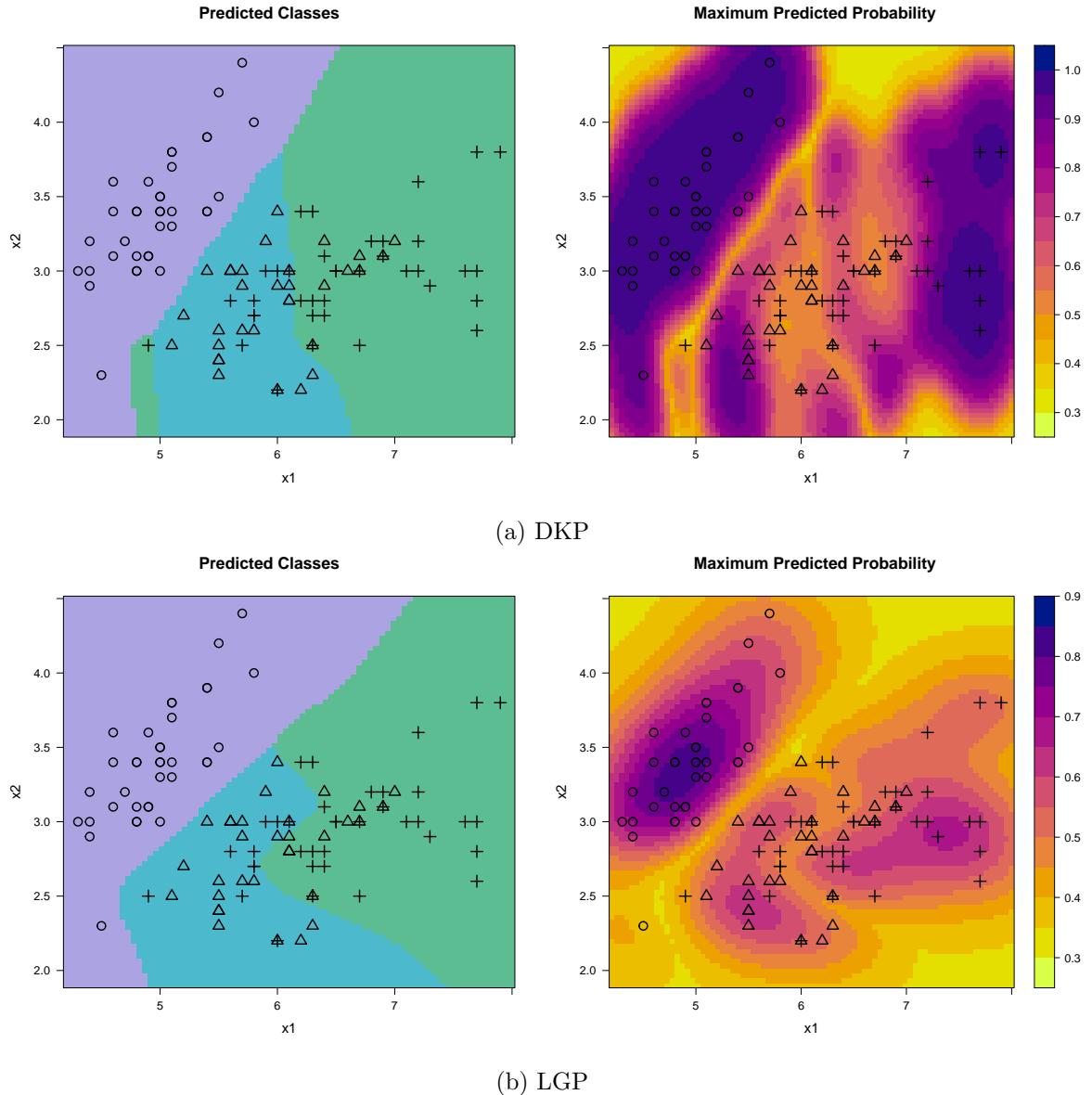


Figure 11: Classification and uncertainty visualization after applying the DGP model on the *Iris* dataset. The left panel shows the predicted classification regions based on the MAP decision rule, where each color corresponds to one of the three iris species. The training data points are overlaid using shape-coded markers. The right panel visualizes the model's predictive uncertainty using the maximum predicted class probability. Lower values (light-colored regions) indicate higher classification uncertainty, which usually occurs near the class boundaries.

highlights regions where the classifier is most confident (values near 1) versus uncertain (values close to 1/3), which effectively identifies decision boundaries and ambiguous areas. Such visualization aids in understanding the reliability of classification decisions and the structure of the learned decision surfaces.

Again, we compare the performance with LGP model, which is implemented by R package **kernlab** (Karatzoglou *et al.* 2004). We adopt the one-vs-rest (OvR) approach for comparison, which works by training a separate binary classifier for each class. Each classifier is tasked with distinguishing its assigned class from all other classes combined.

```
R> iris_data <- data.frame(
+   Sepal.Length = iris$Sepal.Length,
+   Sepal.Width = iris$Sepal.Width,
+   Species = iris$Species
+ )
R> iris_train <- iris_data[train_indices, ]
R> iris_test <- iris_data[-train_indices, ]
R> gausspr_model <- gausspr(Species ~ ., data = iris_train,
+                               kernel = "rbfdot", kpar = "automatic")
R> lgp_pred_probs <- predict(gausspr_model, newdata = iris_test,
+                             type = "probabilities")
```

It is seen that the decision boundaries in the 'Predicted Classes' plot (lower left panel) appear less smooth, particularly in the regions between the two clusters of 'triangle' and 'plus' data points. Additionally, the 'Maximum Predicted Probability' plot (lower right panel) shows a more fragmented and less uniform confidence landscape. There are pockets of high confidence, but also areas of lower confidence (yellows and greens) scattered over the region, even within what appear to be the core regions of the classes. This indicates that DKP model is more accurate in prediction and more consistent in uncertainty quantification than the LGP model. The corresponding multiclass ROC curves with macro-average AUC (0.936 vs 0.927) are presented in Figure 12.

## 5. Working Example: Modeling Loa loa Infection

This section presents the workflow of the **BKP** package based on a real data analysis, including data preprocessing, model fitting, prediction, and visualization. %As an illustrative application, we analyze the prevalence of Loa loa parasite infection in North Cameroon using the **Loaloa** dataset (Diggle *et al.* 2007).

### 5.1. Data Preparation

The **Loaloa** dataset, available in the **mdhglm** package (Lee *et al.* 2018), contains 197 observations collected in North Cameroon between 1991 and 2001.

For each surveyed location, the dataset records geographical coordinates (**longitude**, **latitude**), the number of infected individuals (**npos**), the total number of surveyed individuals (**ntot**), and additional covariates such as altitude and the maximum normalized difference vegetation index (NDVI).

In this analysis, we focus on geographical coordinates as predictors (**X**), the number of infected individuals as the response (**y**), and the total number surveyed as the trial counts (**m**). The dataset is randomly divided into a training set (70%) and a testing set (30%):

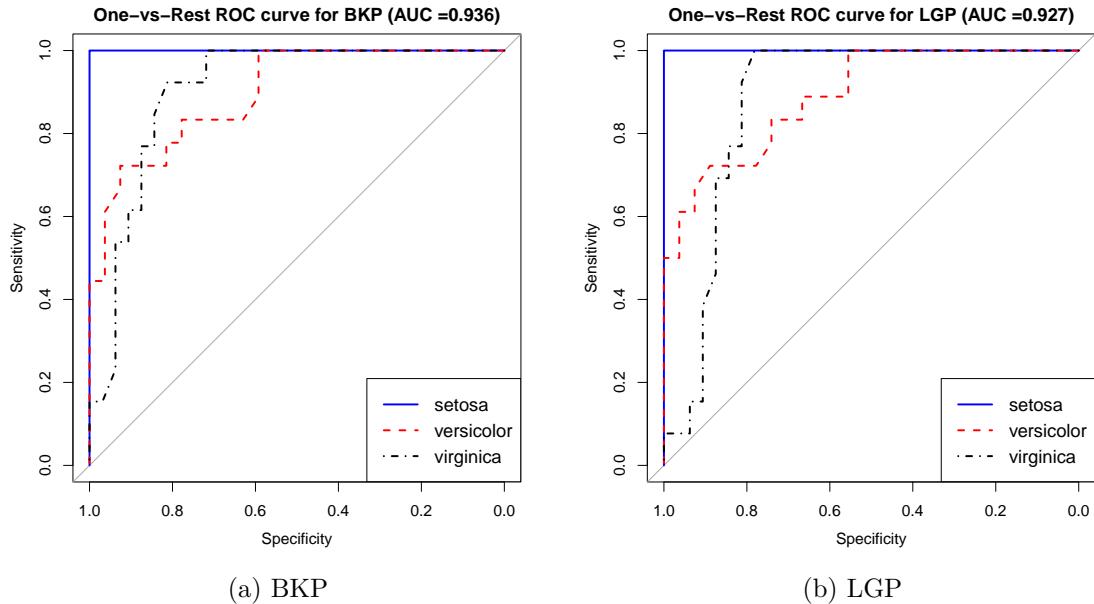


Figure 12: One-vs-Rest ROC curves and their respective macro-average AUCs for BKP and LGP models

```
R> data("Loaloa")
R> # Extract input variables (X), response variable (y), and trial counts (m)
R> X <- as.matrix(Loaloa[, 1:2])
R> y <- Loaloa$npos
R> m <- Loaloa$ntot
R> # Randomly split into training (70%) and testing (30%) sets
R> train_idx <- sample(1:nrow(Loaloa), 0.7 * nrow(Loaloa))
R> X_train <- X[train_idx, ]
R> y_train <- y[train_idx]
R> m_train <- m[train_idx]
R> X_test <- X[-train_idx, ]
R> y_test <- y[-train_idx]
R> m_test <- m[-train_idx]
```

To investigate the spatial heterogeneity of infection prevalence, Figure 13 presents the observed infection proportion ( $y/m$ ) at each survey location. The symbol size is proportional to the sample size, reflecting the relative precision of the observed proportions, while the color gradient encodes the infection rate ranging from 0.0 to 0.5. The map further distinguishes between the training set (70% of locations) and the testing set (30% of locations), which are used for model fitting and predictive validation, respectively. Distinct spatial patterns emerge, with clusters of elevated infection prevalence observed between 10°E and 12°E longitude and another group located between 14°E and 15°E longitude.

## 5.2. Model fitting and visualization

We fit the BKP model to the training data using the `fit_BKP()` function. The model was

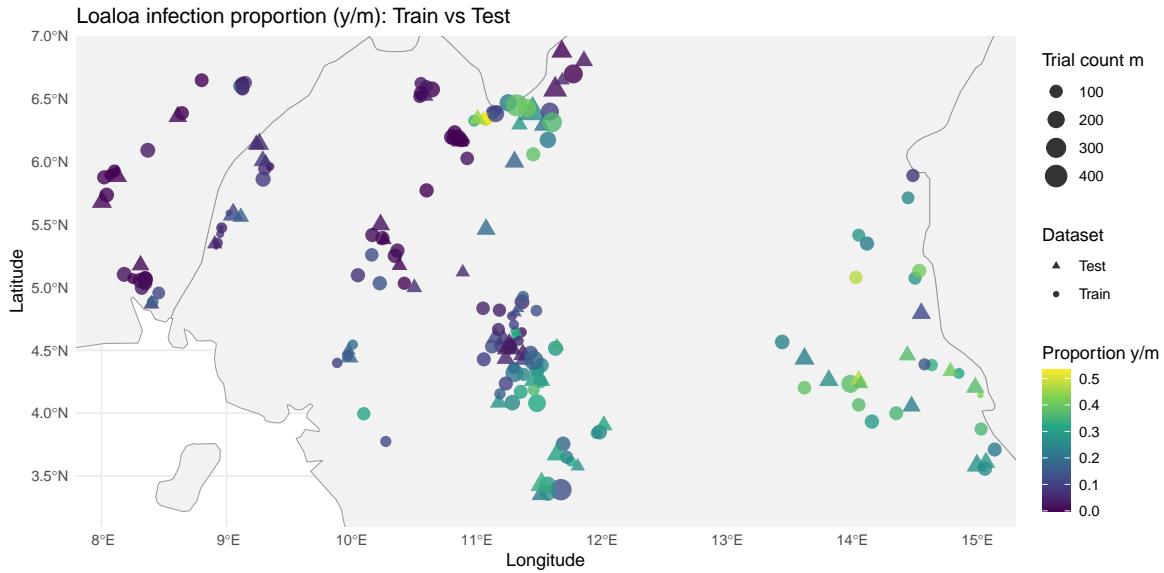


Figure 13: Observed infection proportions ( $y/m$ ) across North Cameroon. Symbol size is proportional to the trial count, color indicates the observed infection rate, and shape distinguishes between training (circle, 70% of locations) and testing (triangle, 30% of locations) sets.

trained with the Brier score as the loss function, and an adaptive prior was adopted to stabilize estimation in regions with small sample sizes. The global precision parameter  $r_0$  is set to the mean number of trials per location, representing a moderately informative prior that provides smoothing while still allowing the data to contribute substantially to posterior inference.

```
R> Xbounds <- matrix(c(7.8, 15.3, 3.1, 7.0), ncol = 2, byrow = TRUE)
R> Loaloa_bkp_model <- fit_BKP(
+   X_train, y_train, m_train, Xbounds, loss = "brier",
+   prior = "adaptive", r0 = mean(m_train))
R> summary(Loaloa_bkp_model)
```

#### Beta Kernel Process (BKP) Model

```
Number of observations (n): 137
Input dimensionality (d): 2
Kernel type: gaussian
Optimized kernel parameters: 0.0076, 0.0460
Minimum achieved loss: 0.00780
Loss function: brier
Prior type: adaptive
r0: 135.533
```

Posterior predictive summary (training points):

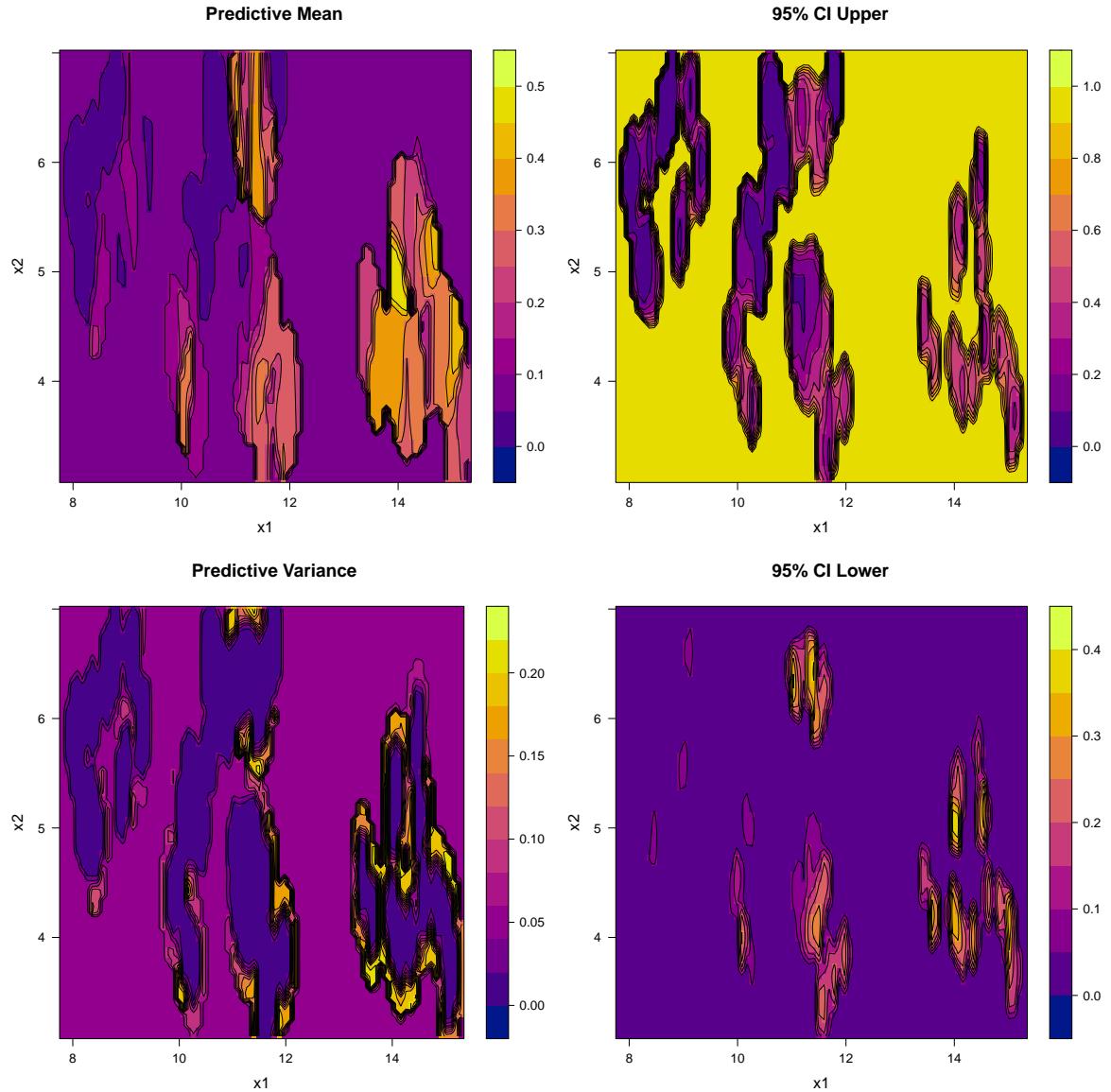


Figure 14: Posterior summaries of the BKP model fitted to the Loa loa dataset, including the predictive mean, variance, and 95% credible interval surfaces.

	Mean	Median	SD	Min	Max
Posterior means	0.1508	0.1220	0.1227	0	0.4619
Posterior variances	0.0002	0.0002	0.0003	0	0.0013

```
R> plot(Loaloa_bkp_model)
```

Figure 14 displays the posterior predictive mean, variance, and 95% credible interval surfaces, providing a comprehensive view of the model's predictive uncertainty. The posterior mean surface highlights clear spatial variation in Loa loa infection risk, with elevated prevalence predicted in the central-eastern part of the study region. Notably, the predictive mean map

from the BKP model effectively captures the major high-prevalence clusters, correctly identifying concentrated yellow and orange regions that correspond to the high-proportion areas observed in Figure 13. The variance surface shows that predictive uncertainty is generally low in areas with dense observations, but increases substantially in peripheral regions where data are sparse. The 95% credible interval surfaces further illustrate this pattern: the upper bound suggests that high infection rates cannot be ruled out in some regions, while the lower bound remains close to zero across much of the area. Overall, the BKP model not only captures spatial heterogeneity in infection prevalence but also accurately reflects the spatial distribution and concentration of Loa loa infections while quantifying the associated predictive uncertainty.

For comparison, we also fit a logistic Gaussian process (LGP) model using the `gp_fit()` function from the `gplite` package (Piironen 2022).

```
R> Loaloa_gp_model <- gp_init(cf = cf_sexp(), lik = lik_binomial())
R> Loaloa_gp_model <- gp_optim(
+   Loaloa_gp_model, X_train, y_train, trials = m_train,
+   method = method_full(), approx = approx_ep(), verbose = FALSE)
```

Figure 15 presents the predictive results of the LGP model. Compared with the BKP model, several important differences emerge.

First, the BKP predictive mean surface is smoother and spatially more coherent, identifying continuous regions of high and low infection prevalence. In contrast, the LGP predictive mean appears more fragmented, with multiple isolated pockets of elevated infection rates. Second, the predictive variance differs remarkably: the BKP variance surface is relatively uniform, indicating a consistent level of uncertainty across the study region, whereas the LGP variance surface exhibits highly localized areas of elevated uncertainty (dark purple regions), suggesting less stable predictions in those locations. Overall, the BKP model provides a more robust and spatially coherent characterization of Loa loa infection prevalence compared to the LGP model.

### 5.3. Predictive performance on test data

Finally, we evaluate predictive accuracy using predictions from the fitted BKP and LGP models, with performance measured by the Brier score.

```
R> predict_on_test_bkp <- predict(Loaloa_bkp_model, Xnew = X_test)
R> predict_on_test_gp <- gp_pred(Loaloa_gp_model, xnew = X_test,
+                                 transform = TRUE, var = TRUE)
R> # Calculate the empirical success rate for the test data
R> pi_tilde_test <- y_test / m_test
R> # Calculate the Mean Squared Error (Brier Score)
R> mse_bkp <- mean((predict_on_test_bkp$mean - pi_tilde_test)^2)
R> mse_gp <- mean((predict_on_test_gp$mean - pi_tilde_test)^2)
```

The results indicate that the BKP model achieves a lower Brier score (0.006) compared to the LGP model (0.019), demonstrating its superior predictive accuracy.

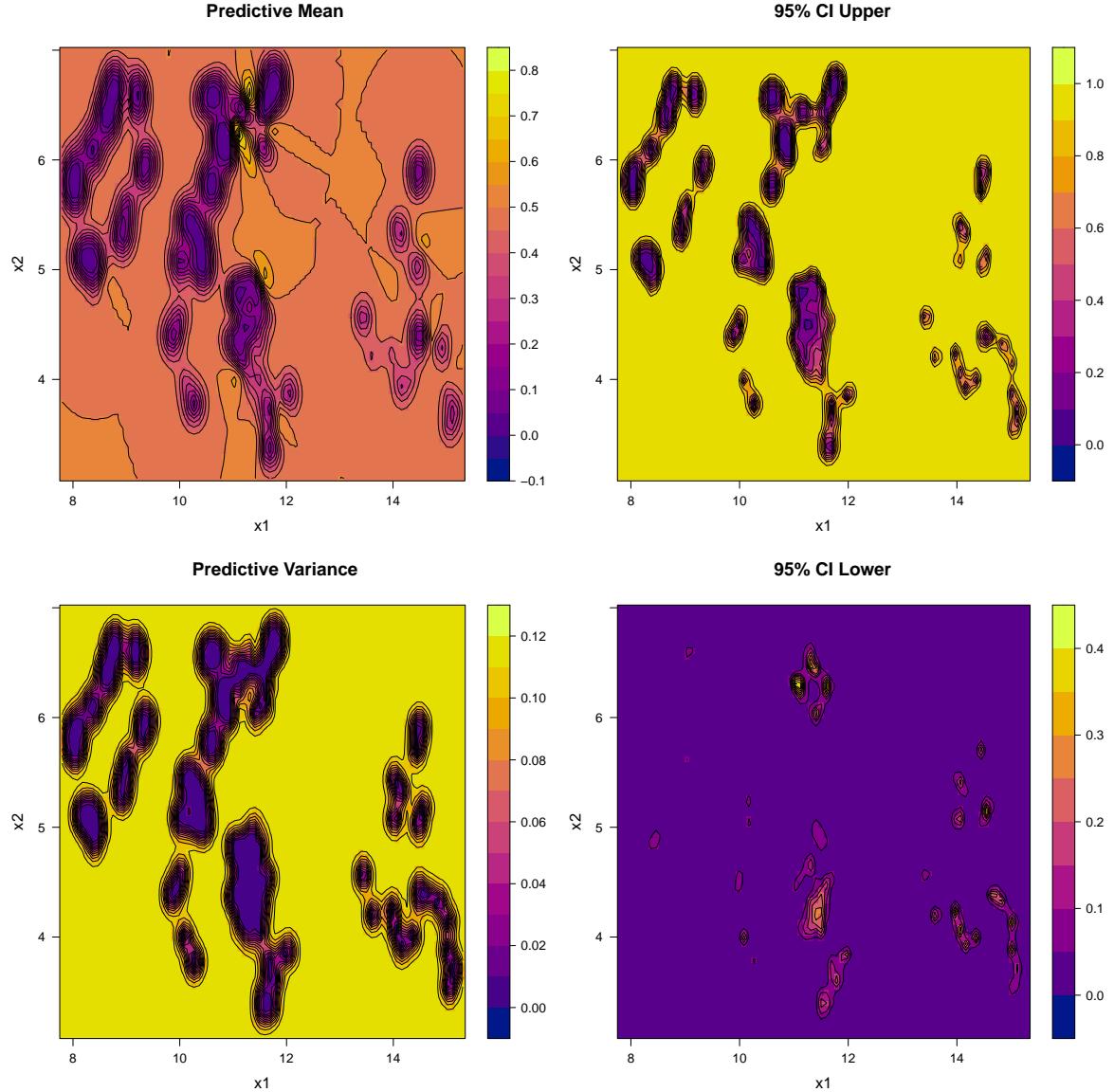


Figure 15: Posterior summaries of the LGP model fitted to the Loa loa dataset, showing the predictive mean and variance surfaces for comparison with the BKP results.

This shows that the BKP predictions are closer to the observed Loa loa infection rates, and its probabilistic forecasts are therefore more reliable.

## 6. Summary and discussion

This article has presented **BKP**, a user-friendly R package that implements the Beta Kernel Process (BKP) – a scalable, interpretable, and fully Bayesian nonparametric framework for modeling spatially varying binomial probabilities. The package provides a flexible and modular interface for fitting BKP models to both binary and aggregated binomial data, and

extends naturally to the Dirichlet Kernel Process (DKP) for handling multinomial and compositional responses with multiple categories. To our knowledge, **BKP** is the first publicly available software for implementing BKP methodology, thereby filling an important gap in the toolkit for spatially varying binomial and multinomial modeling.

Future development directions include extending the BKP framework to support more complex data structures, such as multivariate responses, functional data, time series, and combinations of qualitative and quantitative covariates. Another promising avenue is to generalize the BKP methodology under alternative likelihoods beyond the binomial family. For example, negative binomial likelihoods are particularly suitable for over-dispersed count data, where the variance exceeds the mean, a common phenomenon in ecological surveys, RNA-seq gene expression counts, and epidemiological incidence data. Geometric likelihoods, as a special case of the negative binomial, naturally model the number of trials until the first success and are useful in reliability analysis, survival studies, and modeling waiting times in event processes. <https://www.datacamp.com/tutorial/negative-binomial-distribution> These methodological and computational extensions would substantially broaden the applicability of BKP across applied statistics, biostatistics, and machine learning.

At last, we welcome contributions from the community and invite developers to participate in the ongoing maintenance and extension of the package by submitting pull requests via GitHub at <https://github.com/Jiangyan-Zhao/BKP/pulls>.

## References

- Byrd RH, Lu P, Nocedal J, Zhu C (1995). “A Limited Memory Algorithm for Bound Constrained Optimization.” *SIAM Journal on Scientific Computing*, **16**(5), 1190–1208. [doi:10.1137/0916069](https://doi.org/10.1137/0916069).
- Chalup SK, Wiklendt L (2007). “Variations of the two-spiral task.” *Connection Science*, **19**(2), 183–199. [doi:10.1080/09540090701398017](https://doi.org/10.1080/09540090701398017).
- Diggle PJ, Thomson MC, Christensen OF, Rowlingson B, Obsomer V, Gardon J, Wanji S, Takougang I, Enyong P, Kamgno J, Remme JH, Boussinesq M, Molyneux DH (2007). “Spatial modelling and the prediction of Loa loa risk: decision making under uncertainty.” *Annals of Tropical Medicine & Parasitology*, **101**(6), 499–509. [doi:10.1179/136485913X13789813917463](https://doi.org/10.1179/136485913X13789813917463).
- Dimitriadis T, Dümbgen L, Henzi A, Puke M, Ziegel J (2022). “Honest calibration assessment for binary outcome predictions.” *Biometrika*, **110**(3), 663–680. [doi:10.1093/biomet/asac068](https://doi.org/10.1093/biomet/asac068).
- DRatings (2023). “Log Loss vs. Brier Score.” Accessed on August 02, 2025, URL <https://www.dratings.com/log-loss-vs-brier-score/>.
- Eddelbuettel D, François R (2011). “Rcpp: Seamless R and C++ Integration.” *Journal of Statistical Software*, **40**(8), 1–18. [doi:10.18637/jss.v040.i08](https://doi.org/10.18637/jss.v040.i08).
- Fan J, Farmen M, Gijbels I (1998). “Local Maximum Likelihood Estimation and Inference.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **60**(3), 591–608. [doi:10.1111/1467-9868.00142](https://doi.org/10.1111/1467-9868.00142).

- Fan J, Gijbels I (1996). *Local Polynomial Modelling and Its Applications*, volume 66 of *Monographs on Statistics and Applied Probability*. 1st edition. Routledge, Boca Raton, Florida. doi:[10.1201/9780203748725](https://doi.org/10.1201/9780203748725).
- Flores G, Schiff A, Smith AH, Fukuyama JA, Wilson AC (2025). “A Consequentialist Critique of Binary Classification Evaluation Practices.” URL <https://arxiv.org/abs/2504.04528>.
- Gardner J, Pleiss G, Weinberger KQ, Bindel D, Wilson AG (2018). “GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration.” In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc. URL [https://proceedings.neurips.cc/paper\\_files/paper/2018/file/27e8e17134dd7083b050476733207ea1-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2018/file/27e8e17134dd7083b050476733207ea1-Paper.pdf).
- Garnett R (2023). *Bayesian Optimization*. Cambridge University Press, Cambridge, United Kingdom. doi:[10.1017/9781108348973](https://doi.org/10.1017/9781108348973). <https://bayesoptbook.com/>.
- Gneiting T, Raftery AE (2007). “Strictly Proper Scoring Rules, Prediction, and Estimation.” *Journal of the American Statistical Association*, **102**(477), 359–378. doi:[10.1198/016214506000001437](https://doi.org/10.1198/016214506000001437).
- Goetschalckx R, Poupart P, Hoey J (2011). “Continuous Correlated Beta Processes.” In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI’11, p. 1269–1274. AAAI Press. URL <https://dl.acm.org/doi/10.5555/2283516.2283608>.
- GPy (since 2012). “GPy: A Gaussian process framework in python.” <http://github.com/SheffieldML/GPy>.
- Gramacy RB, Taddy M (2010). “Categorical Inputs, Sensitivity Analysis, Optimization and Importance Tempering with tgp Version 2, an R Package for Treed Gaussian Process Models.” *Journal of Statistical Software*, **33**(6), 1–48. URL <https://www.jstatsoft.org/v33/i06/>.
- Hastie T, Tibshirani R, Friedman J (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer New York, New York, NY. doi:[10.1007/978-0-387-84858-7](https://doi.org/10.1007/978-0-387-84858-7).
- Karatzoglou A, Smola A, Hornik K, Zeileis A (2004). “kernlab – An S4 Package for Kernel Methods in R.” *Journal of Statistical Software*, **11**(9), 1–20. doi:[10.18637/jss.v011.i09](https://doi.org/10.18637/jss.v011.i09).
- Lee Y, Molas M, Noh M (2018). *mdhglm: Multivariate Double Hierarchical Generalized Linear Models*. doi:[10.32614/CRAN.package.mdhglm](https://doi.org/10.32614/CRAN.package.mdhglm). R package version 1.8, URL <https://CRAN.R-project.org/package=mdhglm>.
- Leisch F, Dimitriadou E (2024). *mlbench: Machine Learning Benchmark Problems*. R package version 2.1-6, URL <https://CRAN.R-project.org/package=mlbench>.
- Loeppky JL, Sacks J, Welch WJ (2009). “Choosing the Sample Size of a Computer Experiment: A Practical Guide.” *Technometrics*, **51**(4), 366–376. doi:[10.1198/TECH.2009.08040](https://doi.org/10.1198/TECH.2009.08040).

- MacDonald B, Ranjan P, Chipman H (2015). “GPfit: An R Package for Fitting a Gaussian Process Model to Deterministic Simulator Outputs.” *Journal of Statistical Software*, **64**(12), 1–23. [doi:10.18637/jss.v064.i12](https://doi.org/10.18637/jss.v064.i12).
- MacKenzie CA, Trafalis TB, Barker K (2014). “A Bayesian Beta Kernel Model for Binary Classification and Online Learning Problems.” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **7**(6), 434–449. [doi:10.1002/sam.11241](https://doi.org/10.1002/sam.11241).
- Martin BD, Witten D, Willis AD (2020). “Modeling microbial abundances and dysbiosis with beta-binomial regression.” *The Annals of Applied Statistics*, **14**(1), 94–115. [doi:10.1214/19-AOAS1283](https://doi.org/10.1214/19-AOAS1283).
- Matthews AGdG, van der Wilk M, Nickson T, Fujii K, Boukouvalas A, León-Villagrá P, Ghahramani Z, Hensman J (2017). “GPflow: A Gaussian process library using TensorFlow.” *Journal of Machine Learning Research*, **18**(40), 1–6. URL <http://jmlr.org/papers/v18/16-537.html>.
- Montesano L, Lopes M (2012). “Active learning of visual descriptors for grasping using non-parametric smoothed beta distributions.” *Robotics and Autonomous Systems*, **60**(3), 452–462. [doi:10.1016/j.robot.2011.07.013](https://doi.org/10.1016/j.robot.2011.07.013).
- Murphy KP (2022). *Probabilistic Machine Learning: An introduction*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, Massachusetts. URL <http://probml.github.io/book1>.
- Mussi M, Drago S, Metelli AM (2024). “Open Problem: Tight Bounds for Bernoulli Rewards in Kernelized Multi-Armed Bandits.” In *Workshop on Aligning Reinforcement Learning Experimentalists and Theorists at the International Conference on Machine Learning (ICML)*. ArXiv: 2407.06321, URL <https://arxiv.org/abs/2407.06321>.
- Najera-Zuloaga J, Lee DJ, Arostegui I (2018). “Comparison of beta-binomial regression model approaches to analyze health-related quality of life data.” *Statistical Methods in Medical Research*, **27**(10), 2989–009. [doi:10.1177/0962280217690413](https://doi.org/10.1177/0962280217690413).
- Najera-Zuloaga J, Lee DJ, Arostegui I (2019). “A beta-binomial mixed-effects model approach for analysing longitudinal discrete and bounded outcomes.” *Biometrical Journal*, **61**(3), 600–615. [doi:10.1002/bimj.201700251](https://doi.org/10.1002/bimj.201700251).
- Nash JC (2014). “On Best Practice Optimization Methods in R.” *Journal of Statistical Software*, **60**(2), 1–14. [doi:10.18637/jss.v060.i02](https://doi.org/10.18637/jss.v060.i02).
- Nash JC, Varadhan R (2011). “Unifying Optimization Algorithms to Aid Software System Users: optimx for R.” *Journal of Statistical Software*, **43**(9), 1–14. [doi:10.18637/jss.v043.i09](https://doi.org/10.18637/jss.v043.i09).
- Nickisch H, Rasmussen CE (2008). “Approximations for Binary Gaussian Process Classification.” *Journal of Machine Learning Research*, **9**(67), 2035–2078. URL <http://jmlr.org/papers/v9/nickisch08a.html>.
- Picheny V, Wagner T, Ginsbourger D (2013). “A Benchmark of Kriging-based Infill Criteria for Noisy Optimization.” *Structural and Multidisciplinary Optimization*, **48**(3), 607–626. [doi:10.1007/s00158-013-0919-4](https://doi.org/10.1007/s00158-013-0919-4).

- Piironen J (2022). *gplite: General Purpose Gaussian Process Modelling.* doi:10.32614/[CRAN.package.gplite](https://CRAN.R-project.org/package=gplite). R package version 0.13.0, URL <https://CRAN.R-project.org/package=gplite>.
- R Core Team (2025). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Rasmussen CE, Williams CK (2006). *Gaussian Processes for Machine Learning.* the MIT Press, Cambridge, MA. URL <https://gaussianprocess.org/gpml/>.
- Rolland P, Kavis A, Singla A, Cevher V (2019). “Efficient learning of smooth probability functions from Bernoulli tests with guarantees.” In K Chaudhuri, R Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5459–5467. PMLR. URL <https://proceedings.mlr.press/v97/rolland19a.html>.
- Roustant O, Ginsbourger D, Deville Y (2012). “DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization.” *Journal of Statistical Software*, **51**(1), 1–55. doi:10.18637/jss.v051.i01.
- Sung CL, Hung Y, Rittase W, Zhu C, Wu CFJ (2020). “Calibration for Computer Experiments With Binary Responses and Application to Cell Adhesion Study.” *Journal of the American Statistical Association*, **115**(532), 1664–1674. doi:10.1080/01621459.2019.1699419.
- Vehtari A, Gelman A, Gabry J (2017). “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC.” *Statistics and Computing*, **27**(5), 1413–1432. doi:10.1007/s11222-016-9696-4.
- Wen H, Betken A, Hang H (2025). “Optimal Learning of Kernel Logistic Regression for Complex Classification Scenarios.” In *The Thirteenth International Conference on Learning Representations*. URL <https://openreview.net/forum?id=WlhVRh2rQ0>.

### Affiliation:

Jiangyan Zhao  
 School of Statistics  
 East China Normal University  
 3663 North Zhongshan Road,  
 Shanghai 200062, China  
 E-mail: [jyzhao@sfs.ecnu.edu.cn](mailto:jyzhao@sfs.ecnu.edu.cn)

Kunhai Qing  
 School of Statistics  
 East China Normal University  
 3663 North Zhongshan Road,  
 Shanghai 200062, China  
 E-mail: [51254404017@stu.ecnu.edu.cn](mailto:51254404017@stu.ecnu.edu.cn)

Jin Xu

School of Statistics

*and*

Key Laboratory of Advanced Theory and Application in Statistics and Data Science - MOE

East China Normal University

3663 North Zhongshan Road,

Shanghai 200062, China

E-mail: [jxu@stat.ecnu.edu.cn](mailto:jxu@stat.ecnu.edu.cn)