The Gaussian Process Density Sampler

Ryan Prescott Adams[*]
Cavendish Laboratory
University of Cambridge
Cambridge CB3 0HE, UK
rpa23@cam.ac.uk

Iain Murray
Dept. of Computer Science
University of Toronto
Toronto, Ontario. M5S 3G4
murray@cs.toronto.edu

David J.C. MacKay
Cavendish Laboratory
University of Cambridge
Cambridge CB3 0HE, UK
mackay@mrao.cam.ac.uk

Abstract

We present the Gaussian Process Density Sampler (GPDS), an exchangeable generative model for use in nonparametric Bayesian density estimation. Samples drawn from the GPDS are consistent with exact, independent samples from a fixed density function that is a transformation of a function drawn from a Gaussian process prior. Our formulation allows us to infer an unknown density from data using Markov chain Monte Carlo, which gives samples from the posterior distribution over density functions and from the predictive distribution on data space. We can also infer the hyperparameters of the Gaussian process. We compare this density modeling technique to several existing techniques on a toy problem and a skull-reconstruction task.

1 Introduction

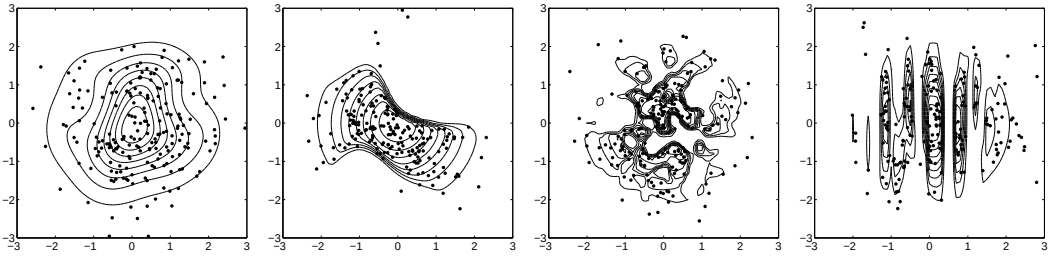
We present the Gaussian Process Density Sampler (GPDS), a generative model for probability density functions, based on a Gaussian process. We are able to draw exact and exchangeable data from a fixed density drawn from the prior. Given data, this generative prior allows us to perform inference of the unnormalized density. We perform this inference by expressing the generative process in terms of a latent history, then constructing a Markov chain Monte Carlo algorithm on that latent history. The central idea of the GPDS is to allow nonparametric Bayesian density estimation where the prior is specified via a Gaussian process covariance function that encodes the intuition that "similar data should have similar probabilities."

One way to perform Bayesian nonparametric density estimation is to use a Dirichlet process to define a distribution over the weights of the components in an infinite mixture model, using a simple parametric form for each component. Alternatively, Neal [1] generalizes the Dirichlet process itself, introducing a spatial component to achieve an exchangeable prior on discrete or continuous density functions with hierarchical characteristics. Another way to define a nonparametric density is to transform a simple latent distribution through a nonlinear map, as in the Density Network [2] and the Gaussian Process Latent Variable Model [3]. Here we use the Gaussian process to define a prior on the density function itself.

2 The prior on densities

We consider densities on an input space $\mathcal{X}$ that we will call the *data space*. In this paper, we assume without loss of generality that $\mathcal{X}$ is the d-dimensional real space $\mathbb{R}^d$. We first construct a Gaussian process prior with the data space $\mathcal{X}$ as its input and the one-dimensional real space $\mathbb{R}$ as its output. The Gaussian process defines a distribution over functions from $\mathcal{X}$ to $\mathbb{R}$. We define a mean function $m(\cdot) : \mathcal{X} \rightarrow \mathbb{R}$ and a positive definite covariance function $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. We

[*] http://www.inference.phy.cam.ac.uk/rpa23/



(a) $\ell_x = 1, \ell_y = 1, \alpha = 1$ (b) $\ell_x = 1, \ell_y = 1, \alpha = 10$ (c) $\ell_x = 0.2, \ell_y = 0.2, \alpha = 5$ (d) $\ell_x = 0.1, \ell_y = 2, \alpha = 5$

Figure 1: Four samples from the GPDS prior are shown, with 200 data samples. The contour lines show the approximate unnormalized densities. In each case the base measure is the zero-mean spherical Gaussian with unit variance. The covariance function was the squared exponential: $K(x, x') = \alpha \exp(-\frac{1}{2}\sum_i \ell_i^{-2}(x_i - x_i')^2)$, with parameters varied as labeled in each subplot. $\Phi(\cdot)$ is the logistic function in these plots.

assume that these functions are together parameterized by a set of hyperparameters θ. Given these two functions and their hyperparameters, for any finite subset of $\mathcal{X}$ with cardinality N there is a multivariate Gaussian distribution on $\mathbb{R}^N$ [4]. We will take the mean function to be zero.

Probability density functions must be everywhere nonnegative and must integrate to unity. We define a map from a function $g(x) : \mathcal{X} \to \mathbb{R}$, $x \in \mathcal{X}$, to a proper density $f(x)$ via

$$f(x) = \frac{1}{\mathcal{Z}_\pi[\boldsymbol{g}]}\, \Phi(g(x))\, \pi(x) \tag{1}$$

where $\pi(x)$ is an arbitrary base probability measure on $\mathcal{X}$, and $\Phi(\cdot) : \mathbb{R} \to (0, 1)$ is a nonnegative function with upper bound 1. We take $\Phi(\cdot)$ to be a sigmoid, e.g. the logistic function or cumulative normal distribution function. We use the bold notation $\boldsymbol{g}$ to refer to the function $g(x)$ compactly as a vector of (infinite) length, versus its value at a particular x. The normalization constant is a functional of $g(x)$:

$$\mathcal{Z}_\pi[\boldsymbol{g}] = \int \mathrm{d}x'\ \Phi(g(x'))\, \pi(x'). \tag{2}$$

Through the map defined by Equation 1, a Gaussian process prior becomes a prior distribution over normalized probability density functions on $\mathcal{X}$. Figure 2 shows several sample densities from this prior, along with sample data.

3 Generating exact samples from the prior

We can use rejection sampling to generate samples from a common density drawn from the the prior described in Section 2. A rejection sampler requires a proposal density that provides an upper bound for the unnormalized density of interest. In this case, the proposal density is $\pi(x)$ and the unnormalized density of interest is $\Phi(g(x))\pi(x)$.

If $g(x)$ were known, rejection sampling would proceed as follows: First generate proposals $\{\tilde{x}_q\}$ from the base measure $\pi(x)$. The proposal $\tilde{x}_q$ would be accepted if a variate r_q drawn uniformly from $(0, 1)$ was less than $\Phi(g(\tilde{x}_q))$. These samples would be exact in the sense that they were not biased by the starting state of a finite Markov chain. However, in the GPDS, $g(x)$ is not known: it is a random function drawn from a Gaussian process prior. We can nevertheless use rejection sampling by "discovering" $g(x)$ as we proceed at just the places we need to know it, by sampling from the prior distribution of the latent function. As it is necessary only to know $g(x)$ at the $\{x_q\}$ to accept or reject these proposals, the samples are still exact. This retrospective sampling trick has been used in a variety of other MCMC algorithms for infinite-dimensional models [5, 6]. The generative procedure is shown graphically in Figure 2.

In practice, we generate the samples sequentially, as in Algorithm 1, so that we may be assured of having as many accepted samples as we require. In each loop, a proposal is drawn from the base measure $\pi(x)$ and the function $g(x)$ is sampled from the Gaussian process at this proposed coordinate, conditional on all the function values already sampled. We will call these data the *conditioning set* for the function $g(x)$ and will denote the conditioning inputs $\boldsymbol{X}$ and the conditioning

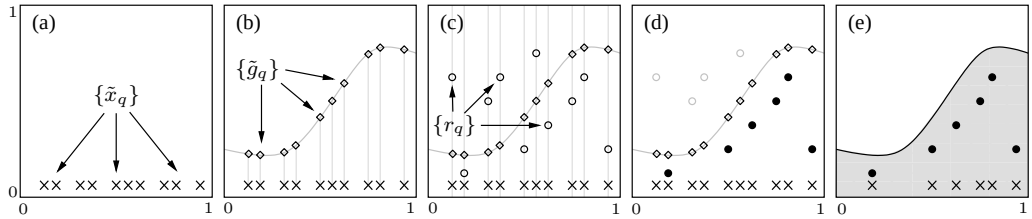


Figure 2: These figures show the procedure for generating samples from a single density drawn from the GP-based prior. (a): Draw Q samples $\{\tilde{x}_q\}^Q$ from the base measure $\pi(x)$, which in this case is uniform on $[0, 1]$. (b): Sample the function $g(x)$ at the randomly chosen locations, generating the set $\{\tilde{g}_q = g(\tilde{x}_q)\}^Q$. The squashed function $\Phi(g(x))$ is shown. (c): Draw a set of variates $\{r_q\}^Q$ uniformly beneath the bound in the vertical coordinate. (d): Accept only the points whose uniform draws are beneath the squashed function value, i.e. $r_q < \Phi(\tilde{g}_q)$. (e): The accepted points $(\tilde{x}_q, r_q)$ are uniformly drawn from the shaded area beneath the curve and the marginal distribution of the accepted $\tilde{x}_q$ is proportional to $\Phi(g(x))\pi(x)$.

function values $\boldsymbol{G}$. After the function is sampled, a uniform variate is drawn from beneath the bound and compared to the Φ-squashed function at the proposal location.

The sequential procedure is exchangeable, which means that the probability of the data is identical under reordering. First, the base measure draws are i.i.d.. Second, conditioned on the proposals from the base measure, the Gaussian process is a simple multivariate Gaussian distribution, which is exchangeable in its components. Finally, conditioned on the draw from the Gaussian process, the acceptance/rejection steps are independent Bernoulli samples, and the overall procedure is exchangeable. This property is important because it ensures that the sequential procedure generates data from the same distribution as the simultaneous procedure described above. More broadly, exchangeable priors are useful in Bayesian modeling because we may consider the data conditionally independent, given the latent density.

Algorithm 1 Generate P exact samples from the prior

Purpose: Draw P exact samples from a common density on $\mathcal{X}$ drawn from the prior in Equation 1
Inputs: GP hyperparameters θ, number of samples to generate P

Initialize empty conditioning sets for the Gaussian process: $\boldsymbol{X} = \emptyset$ and $\boldsymbol{G} = \emptyset$
repeat
  Draw a proposal from the base measure: $\tilde{x} \sim \pi(x)$
  Sample the function from the Gaussian process at $\tilde{x}$: $\tilde{g} \sim \mathcal{GP}(g \,|\, \boldsymbol{X}, \boldsymbol{G}, \tilde{x}, \theta)$
  Draw a uniform variate on $[0, 1]$: $r \sim \mathcal{U}(0, 1)$
  if $r < \Phi(\tilde{g})$ (Acceptance rule) **then**
    Accept $\tilde{x}$
  else
    Reject $\tilde{x}$
  end if
  Add $\tilde{x}$ and $\tilde{g}$ to the conditioning sets: $\boldsymbol{X} = \boldsymbol{X} \cup \tilde{x}$ and $\boldsymbol{G} = \boldsymbol{G} \cup \tilde{g}$
until P samples have been accepted

4 Inference

We have N data $\mathcal{D} = \{x_n\}_{n=1}^N$ which we model as having been drawn independently from an unknown density $f(x)$. We use the GPDS prior from Section 2 to specify our beliefs about $f(x)$, and we wish to generate samples from the posterior distribution over the latent function $g(x)$ corresponding to the unknown density. We may also wish to generate samples from the predictive distribution or perform hierarchical inference of the prior hyperparameters.

By using the GPDS prior to model the data, we are asserting that the data can be explained as the result of the procedure described in Section 3. We do not, however, know what rejections were made en route to accepting the observed data. These rejections are critical to defining the latent function $g(x)$. One might think of defining a density as analogous to putting up a tent: pinning the canvas down with pegs is just as important as putting up poles. In density modeling, defining regions with little probability mass is just as important as defining the areas with significant mass.

Although the rejections are not known, the generative procedure provides a probabilistic model that allows us to traverse the posterior distribution over possible *latent histories* that resulted in the data. If we define a Markov chain whose equilibrium distribution is the posterior distribution over latent histories, then we may simulate plausible explanations of every step taken to arrive at the data. Such samples capture all the information available about the unknown density, and with them we may ask additional questions about $g(x)$ or run the generative procedure further to draw predictive samples. This approach is related to that described by Murray [7], who performed inference on an exactly-coalesced Markov chain [8], and by Beskos et al. [5].

We model the data as having been generated exactly as in Algorithm 1, with $P = N$, i.e. run until exactly N proposals were accepted. The state space of the Markov chain on latent histories in the GPDS consists of: 1) the values of the latent function $g(x)$ at the data, denoted $\mathcal{G}_N = \{g_n\}_{n=1}^N$, 2) the number of rejections M, 3) the locations of the M rejected proposals, denoted $\mathcal{M} = \{x_m\}_{m=1}^M$, and 4) the values of the latent function $g(x)$ at the M rejected proposals, denoted $\mathcal{G}_M = \{g_m = g(x_m)\}_{m=1}^M$. We will address hyperparameter inference in Section 4.3.

We perform Gibbs-like sampling of the latent history by alternating between modification of the number of rejections M and block updating of the rejection locations $\mathcal{M}$ and latent function values $\mathcal{G}_M$ and $\mathcal{G}_N$. We will maintain an explicit ordering of the latent rejections for reasons of clarity, although this is not necessary due to exchangeability. We will also assume that $\Phi(\cdot)$ is the logistic function, i.e. $\Phi(z) = (1 + \exp\{-z\})^{-1}$.

4.1 Modifying the number of latent rejections

We propose a new number of latent rejections $\hat{M}$ by drawing it from a proposal distribution $q(\hat{M} \leftarrow M)$. If $\hat{M}$ is greater than M, we must also propose new rejections to add to the latent state. We take advantage of the exchangeability of the process to generate the new rejections: we imagine these proposals were made *after* the last observed datum was accepted, and our proposal is to call them rejections and move them *before* the last datum. If $\hat{M}$ is less than M, we do the opposite by proposing to move some rejections to after the last acceptance.

When proposing additional rejections, we must also propose times for them among the current latent history. There are $\binom{\hat{M}+N-1}{\hat{M}-M}$ such ways to insert these additional rejections into the existing latent history, such that the sampler terminates after the Nth acceptance. When removing rejections, we must choose which ones to place after the data, and there are $\binom{M}{M-\hat{M}}$ possible sets. Upon simplification, the proposal ratios for both addition and removal of rejections are identical:

$$\frac{\overbrace{q(M \leftarrow \hat{M})\binom{\hat{M}+N-1}{\hat{M}-M}}^{\hat{M}>M}}{q(\hat{M} \leftarrow M)\binom{\hat{M}}{\hat{M}-M}} = \frac{\overbrace{q(M \leftarrow \hat{M})\binom{M}{M-\hat{M}}}^{\hat{M}<M}}{q(\hat{M} \leftarrow M)\binom{M+N-1}{M-\hat{M}}} = \frac{q(M \leftarrow \hat{M})M!(\hat{M}+N-1)!}{q(\hat{M} \leftarrow M)\hat{M}!(M+N-1)!}.$$

When inserting rejections, we propose the locations of the additional proposals, denoted $\mathcal{M}^+$, and the corresponding values of the latent function, denoted $\mathcal{G}_M^+$. We generate $\mathcal{M}^+$ by making $\hat{M} - M$ independent draws from the base measure. We draw $\mathcal{G}_M^+$ jointly from the Gaussian process prior, conditioned on all of the current latent state, i.e. $(\mathcal{M},\ \mathcal{G}_M,\ \mathcal{D},\ \mathcal{G}_N)$. The joint probability of this state is

$$p(\mathcal{D}, \mathcal{M}, \mathcal{M}^+, \mathcal{G}_N, \mathcal{G}_M, \mathcal{G}_M^+) = \left[\prod_{n=1}^{N} \pi(x_n)\Phi(g_n)\right]\left[\prod_{m=1}^{M} \pi(x_m)(1-\Phi(g_m))\right]\left[\prod_{m=M+1}^{\hat{M}} \pi(x_m)\right] \times \mathcal{GP}(\mathcal{G}_M, \mathcal{G}_N, \mathcal{G}_M^+ \,|\, \mathcal{D}, \mathcal{M}, \mathcal{M}^+). \quad (3)$$

The joint in Equation 3 expresses the probability of all the base measure draws, the values of the function draws from the Gaussian process, and the acceptance or rejection probabilities of the proposals *excluding* the newly generated points. When we make an insertion proposal, exchangeability allows us to shuffle the ordering without changing the probability; the only change is that now we must account for labeling the new points as rejections. In the acceptance ratio, all terms except for the "labeling probability" cancel. The reverse proposal is similar, however we denote the removed

proposal locations as $\mathcal{M}^-$ and the corresponding function values as $\mathcal{G}_M^-$. The overall acceptance ratios for insertions or removals are

$$a = \begin{cases} \frac{q(M \leftarrow \hat{M})}{q(\hat{M} \leftarrow M)} \frac{M!}{\hat{M}!} \frac{(\hat{M}+N-1)!}{(M+N-1)!} \prod_{g \in \mathcal{G}_M^+} (1 - \Phi(g)) & \text{if } \hat{M} > M \\ \frac{q(M \leftarrow \hat{M})}{q(\hat{M} \leftarrow M)} \frac{M!}{\hat{M}!} \frac{(\hat{M}+N-1)!}{(M+N-1)!} \prod_{g \in \mathcal{G}_M^-} (1 - \Phi(g))^{-1} & \text{if } \hat{M} < M. \end{cases} \tag{4}$$

4.2 Modifying rejection locations and function values

Given the number of latent rejections M, we propose modifying their locations $\mathcal{M}$, their latent function values $\mathcal{G}_M$, and the values of the latent function at the data $\mathcal{G}_N$. We will denote these proposals as $\hat{\mathcal{M}} = \{\hat{x}_m\}_{m=1}^M$, $\hat{\mathcal{G}}_M = \{\hat{g}_m = \hat{g}(\hat{x}_m)\}_{m=1}^M$ and $\hat{\mathcal{G}}_N = \{\hat{g}_n = \hat{g}(x_n)\}_{n=1}^N$, respectively. We make simple perturbative proposals of $\mathcal{M}$ via a proposal density $q(\hat{\mathcal{M}} \leftarrow \mathcal{M})$. For the latent function values, however, perturbative proposals will be poor, as the Gaussian process typically defines a narrow mass. To avoid this, we propose modifications to the latent function that leave the prior invariant.

We make joint proposals of $\hat{\mathcal{M}}$, $\hat{\mathcal{G}}_M$ and $\hat{\mathcal{G}}_N$ in three steps. First, we draw new rejection locations from $q(\hat{\mathcal{M}} \leftarrow \mathcal{M})$. Second, we draw a set of M intermediate function values from the Gaussian process at $\hat{\mathcal{M}}$, conditioned on the current rejection locations and their function values, as well as the function values at the data. Third, we propose new function values at $\hat{\mathcal{M}}$ and the data $\mathcal{D}$ via an underrelaxation proposal of the form

$$\hat{g}(x) = \alpha \, g(x) + \sqrt{1 - \alpha^2} \, h(x)$$

where $h(x)$ is a sample from the Gaussian process prior and α is in $[0, 1)$. This is a variant of the overrelaxed MCMC method discussed by Neal [9]. This procedure leaves the Gaussian process prior invariant, but makes conservative proposals if α is near one. After making a proposal, we accept or reject via the ratio of the joint distributions:

$$a = \frac{q(\mathcal{M} \leftarrow \hat{\mathcal{M}}) \left[\prod_{m=1}^M \pi(\hat{x}_m)(1 - \Phi(\hat{g}_m))\right] \left[\prod_{n=1}^N \Phi(\hat{g}_n)\right]}{q(\hat{\mathcal{M}} \leftarrow \mathcal{M}) \left[\prod_{m=1}^M \pi(x_m)(1 - \Phi(g_m))\right] \left[\prod_{n=1}^N \Phi(g_n)\right]}.$$

4.3 Hyperparameter inference

Given a sample from the posterior on the latent history, we can also perform a Metropolis–Hasting step in the space of hyperparameters. Parameters θ, governing the covariance function and mean function of the Gaussian process provide common examples of hyperparameters, but we might also introduce parameters ϕ that control the behavior of the base measure $\pi(x)$. We denote the proposal distributions for these parameters as $q(\hat{\theta} \leftarrow \theta)$ and $q(\hat{\phi} \leftarrow \phi)$, respectively. With priors $p(\theta)$ and $p(\phi)$, the acceptance ratio for a Metropolis–Hastings step is

$$a = \frac{q(\theta \leftarrow \hat{\theta}) \; q(\phi \leftarrow \hat{\phi}) \; p(\hat{\theta}) \; p(\hat{\phi}) \; \mathcal{N}(\{\mathcal{G}_M, \mathcal{G}_N\} \,|\, \mathcal{M}, \mathcal{D}, \hat{\theta})}{q(\hat{\theta} \leftarrow \theta) \; q(\hat{\phi} \leftarrow \phi) \; p(\theta) \; p(\phi) \; \mathcal{N}(\{\mathcal{G}_M, \mathcal{G}_N\} \,|\, \mathcal{M}, \mathcal{D}, \theta)} \left[\prod_{m=1}^M \frac{\pi(x_m \,|\, \hat{\phi})}{\pi(x_m \,|\, \phi)}\right] \left[\prod_{n=1}^N \frac{\pi(x_n \,|\, \hat{\phi})}{\pi(x_n \,|\, \phi)}\right].$$

4.4 Prediction

The predictive distribution is the one that arises on the space $\mathcal{X}$ when the posterior on the latent function $g(x)$ (and perhaps hyperparameters) is integrated out. It is the expected distribution of the next datum, given the ones we have seen and taking into account our uncertainty. In the GPDS we sample from the predictive distribution by running the generative process of Section 3, initialized to the current latent history sample from the Metropolis–Hastings procedure described above.

It may also be desirable to estimate the actual value of the predictive density. We use the method of Chib and Jeliazkov [10], and observe by detailed balance of a Metropolis–Hastings move:

$$p(x \,|\, \boldsymbol{g}, \theta, \phi)\pi(x') \min\left(1, \frac{\Phi(g(x'))}{\Phi(g(x))}\right) = p(x' \,|\, \boldsymbol{g}, \theta, \phi)\pi(x) \min\left(1, \frac{\Phi(g(x))}{\Phi(g(x'))}\right).$$

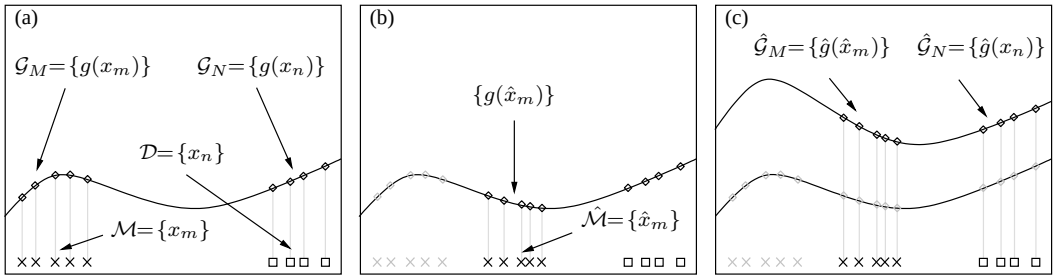


Figure 3: These figures show the sequence of proposing new rejection locations, new function values at those locations, and new function values at the data. (a): The current state, with rejections labeled $\mathcal{M} = \{x_m\}$ on the left, along with the values of the latent function $\mathcal{G}_M = \{g_m\}$. On the right side are the data $\mathcal{D} = \{x_n\}$ and the corresponding values of the latent function $\mathcal{G}_N = \{g_n\}$. (b): New rejections $\hat{\mathcal{M}} = \{\hat{x}_m\}$ are proposed via $q(\hat{\mathcal{M}} \leftarrow \mathcal{M})$, and the latent function is sampled at these points. (c): The latent function is perturbed at the new rejection locations and at the data via an underrelaxed proposal.

We find the expectation of each side under the posterior of g and the hyperparameters θ and ϕ:

$$\int \mathrm{d}\theta \int \mathrm{d}\phi \; p(\theta, \phi \,|\, \mathcal{D}) \int \mathrm{d}\boldsymbol{g} \; p(\boldsymbol{g} \,|\, \theta, \mathcal{D}) \int \mathrm{d}x' \; p(x \,|\, \boldsymbol{g}, \theta, \phi) \pi(x') \min\left(1, \frac{\Phi(g(x'))}{\Phi(g(x))}\right)$$
$$= \int \mathrm{d}\theta \int \mathrm{d}\phi \; p(\theta, \phi \,|\, \mathcal{D}) \int \mathrm{d}\boldsymbol{g} \; p(\boldsymbol{g} \,|\, \theta, \mathcal{D}) \int \mathrm{d}x' \; p(x' \,|\, \boldsymbol{g}, \theta, \phi) \pi(x) \min\left(1, \frac{\Phi(g(x))}{\Phi(g(x'))}\right).$$

This gives an expression for the predictive density:

$$p(x \,|\, \mathcal{D}) = \frac{\int \mathrm{d}\theta \int \mathrm{d}\phi \int \mathrm{d}\boldsymbol{g} \int \mathrm{d}x' \; p(\theta, \phi, \boldsymbol{g}, x' \,|\, \mathcal{D}) \; \pi(x) \; \min\left(1, \frac{\Phi(g(x))}{\Phi(g(x'))}\right)}{\int \mathrm{d}\theta \int \mathrm{d}\phi \int \mathrm{d}\boldsymbol{g} \int \mathrm{d}x' \; p(\theta, \phi, \boldsymbol{g} \,|\, x, \mathcal{D}) \; \pi(x') \; \min\left(1, \frac{\Phi(g(x'))}{\Phi(g(x))}\right)} \tag{5}$$

Both the numerator and the denominator in Equation 5 are expectations that can be estimated by averaging over the output from the GPDS Metropolis–Hasting sampler. The denominator requires sampling from the posterior distribution with the data augmented by x.

5 Results

We examined the GPDS prior and the latent history inference procedure on a toy data set and on a skull reconstruction task. We compared the approach described in this paper to a kernel density estimate (Parzen windows), an infinite mixture of Gaussians (iMoG), and Dirichlet diffusion trees (DFT). The kernel density estimator used a spherical Gaussian with the bandwidth set via ten-fold cross validation. Neal's Flexible Bayesian Modeling (FBM) Software [1] was used for the implementation of both iMoG and DFT.

The toy data problem consisted of 100 uniform draws from a two-dimensional ring with radius 1.5, and zero-mean Gaussian noise added with $\sigma = 0.2$. The test data were 50 additional samples, and comparison used mean log probability of the test set. Each of the three Bayesian methods improved on the Parzen window estimate by two or more nats, with the DFT approach being the most successful. A bar plot of these results is shown in Figure 5.

We also compared the methods on a real-data task. We modeled the the joint density of ten measurements of linear distances between anatomical landmarks on 228 rhesus macaque (*Macaca mulatta*) skulls. These linear distances were generated from three-dimensional coordinate data of anatomical landmarks taken by a single observer from dried skulls using a digitizer [11]. Linear distances are commonly used in morphological studies as they are invariant under rotation and translation of the objects being compared [12]. Figure 5 shows a computed tomography (CT) scan reconstruction of a macaque skull, along with the ten linear distances used. Each skull was measured three times in different trials, and these were modeled separately. 200 randomly-selected skulls were used as a training set and 28 were used as a test set. To be as fair as possible, the data was logarithmically transformed and whitened as a preprocessing step, to have zero sample mean and spherical sample covariance. Each of the Bayesian approaches outperformed the Parzen window technique in mean log probability of the test set, with comparable results for each. This result is not surprising, as flexible nonparametric Bayesian models should have roughly similar expressive capabilities. These results are shown in Figure 5.

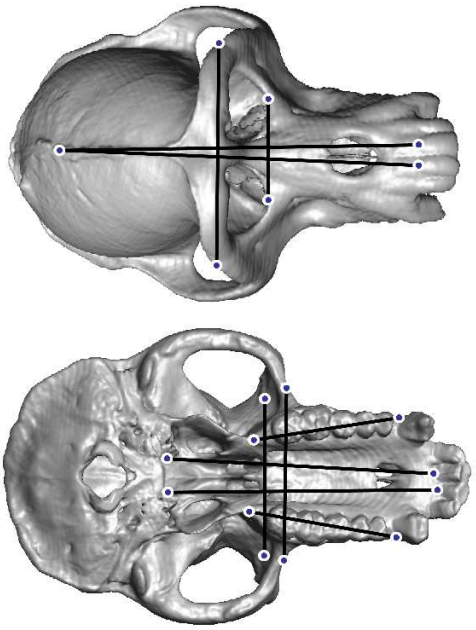


Figure 4: The macaque skull data are linear distances calculated between three-dimensional coordinates of anatomical landmarks. These are superior and inferior views of a computed tomography (CT) scan of a male macaque skull, with the ten linear distances superimposed. The anatomical landmarks are based on biological relevance and repeatability across individuals.

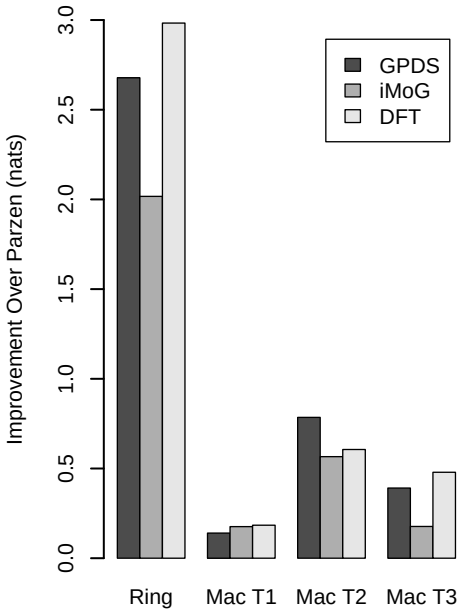


Figure 5: This bar plot shows the improvement of the GPDS, infinite mixture of Gaussians (iMoG), and Dirichlet diffusion trees (DFT) in mean log probability (base e) of the test set over cross-validated Parzen windows on the toy ring data and the macaque data. The baseline log probability of the Parzen method for the ring data was -2.253 and for the macaque data was -15.443, -15.742, and -15.254 for each of three trials.

6 Discussion

Valid MCMC algorithms for fully Bayesian kernel regression methods are well-established. This work introduces the first such prior that enables tractable density estimation, complementing alternatives such as Dirichlet Diffusion Trees [1] and infinite mixture models.

Although the GPDS has similar motivation to the logistic Gaussian process [13, 14, 15, 16], it differs significantly in its applicability and practicality. All known treatments of the logistic GP require a finite-dimensional proxy distribution. This proxy distribution is necessary both for tractability of inference and for estimation of the normalization constant. Due to the complexity constraints of both the basis-function approach of Lenk [15] and the lattice-based approach of [16], these have only been implemented on single-dimensional toy problems. The GPDS construction we have presented here not only avoids numerical estimation of the normalization constant, but allows infinite-dimensional inference both in theory and in practice.

6.1 Computational complexity

The inference method for the GPDS prior is "practical" in the sense that it can be implemented without approximations, but it has potentially-steep computational costs. To compare two latent histories in a Metropolis–Hastings step we must evaluate the marginal likelihood of the Gaussian process. This requires a matrix decomposition whose cost is $O((N+M)^3)$. The model explicitly allows M to be any nonnegative integer and so this cost is unbounded. The *expected* cost of an M–H step is determined by the expected number of rejections M. For a given $g(x)$, the expected M is $N(\mathcal{Z}_\pi[g]^{-1} - 1)$. This expression is derived from the observation that $\pi(x)$ provides an upper bound on the function $\Phi(g(x))\pi(x)$ and the ratio of acceptances to rejections is determined by the proportion of the mass of $\pi(x)$ contained by $\Phi(g(x))\pi(x)$.

We are optimistic that more sophisticated Markov chain Monte Carlo techniques may realize constant-factor performance gains over the basic Metropolis–Hasting scheme presented here, without compromising the correctness of the equilibrium distribution. Sparse approaches to Gaussian process regression that improve the asymptotically cubic behavior may also be relevant to the GPDS, but it is unclear that these will be an improvement over other approximate GP-based schemes for density modeling.

6.2 Alternative inference methods

In developing inference methods for the GPDS prior, we have also explored the use of *exchange sampling* [17, 7]. Exchange sampling is an MCMC technique explicitly developed for the situation where there is an intractable normalization constant that prevents exact likelihood evaluation, but exact samples may be generated for any particular parameter setting. Undirected graphical models such as the Ising and Potts models provide common examples of cases where exchange sampling is applicable via coupling from the past [8]. Using the exact sampling procedure of Section 3, it is applicable to the GPDS as well. Exchange sampling for the GPDS, however, requires more evaluations of the function $g(x)$ than the latent history approach. In practice the latent history approach of Section 4 does perform better.

Acknowledgements

The authors wish to thank Radford Neal and Zoubin Ghahramani for valuable comments. Ryan Adams' research is supported by the Gates Cambridge Trust. Iain Murray's research is supported by the government of Canada. The authors thank the Caribbean Primate Research Center, the University of Puerto Rico, Medical Sciences Campus, Laboratory of Primate Morphology and Genetics, and the National Institutes of Health (Grant RR03640 to CPRC) for support.

References

[1] R. M. Neal. Defining priors for distributions using Dirichlet diffusion trees. Technical Report 0104, Department of Statistics, University of Toronto, 2001.

[2] D. J. C. MacKay. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research, Section A*, 354(1):73–80, 1995.

[3] N. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

[4] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

[5] A. Beskos, O. Papaspiliopoulos, G. O. Roberts, and P. Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B*, 68:333–382, 2006.

[6] O. Papaspiliopoulos and G. O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

[7] I. Murray. *Advances in Markov chain Monte Carlo methods*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, London, 2007.

[8] J. G. Propp and D. B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(1&2):223–252, 1996.

[9] R. M. Neal. Supressing random walks in Markov chain Monte Carlo using ordered overrelaxation, 1998.

[10] S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

[11] K. E. Willmore, C. P. Klingenberg, and B. Hallgrimsson. The relationship between fluctuating asymmetry and environmental variance in rhesus macaque skulls. *Evolution*, 59(4):898–909, 2005.

[12] S. R. Lele and J. T. Richtsmeier. *An invariant approach to statistical analysis of shapes*. Chapman and Hall/CRC Press, London, 2001.

[13] T. Leonard. Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society, Series B*, 40(2):113–146, 1978.

[14] D. Thorburn. A Bayesian approach to density estimation. *Biometrika*, 73(1):65–75, 1986.

[15] P. J. Lenk. Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, 78(3):531–543, 1991.

[16] S. T. Tokdar and J. K. Ghosh. Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, 137:34–42, 2007.

[17] I. Murray, Z. Ghahramani, and D. J.C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 359–366, 2006.