Laplace Approximation for Logistic Gaussian Process Density Estimation and Regression

Jaakko Riihimäki [*] and Aki Vehtari [†]

Abstract. Logistic Gaussian process (LGP) priors provide a flexible alternative for modelling unknown densities. The smoothness properties of the density estimates can be controlled through the prior covariance structure of the LGP, but the challenge is the analytically intractable inference. In this paper, we present approximate Bayesian inference for LGP density estimation in a grid using Laplace's method to integrate over the non-Gaussian posterior distribution of latent function values and to determine the covariance function parameters with type-II maximum a posteriori (MAP) estimation. We demonstrate that Laplace's method with MAP is sufficiently fast for practical interactive visualisation of 1D and 2D densities. Our experiments with simulated and real 1D data sets show that the estimation accuracy is close to a Markov chain Monte Carlo approximation and state-of-the-art hierarchical infinite Gaussian mixture models. We also construct a reduced-rank approximation to speed up the computations for dense 2D grids, and demonstrate density regression with the proposed Laplace approach.

Keywords: Gaussian process, logistic transformation, density estimation, density regression, approximate inference, Laplace's method

1 Introduction

Logistic Gaussian process (LGP) priors provide a flexible alternative for modelling unknown densities (Leonard 1978). With the LGP models densities can be estimated without restricting to any specific parameterized form and the smoothness properties of estimates can be controlled through the prior covariance structure. The challenge with the LGP model is the analytically intractable inference that results from the normalization term required to construct the prior distribution over density functions. This paper focuses on finite dimensional approximations, where the Gaussian process and the integral required to ensure normalization are evaluated in a grid as described by Tokdar (2007). The theoretical properties of LGP are discussed by Tokdar and Ghosh (2007) who establish conditions for posterior consistency of LGP density estimation and by van der Vaart and van Zanten (2009) who consider posterior convergence rates. Tokdar (2007) shows that as the spacing between the grid points decreases, the Kullback–Leibler divergence from an infinite-dimensional model to the finite-dimensional approximation converges to zero. Related considerations about a finite element approach for LGP density estimation are also presented by Griebel and Hegland (2010). Densities are often estimated on bounded intervals, although the estimation can be extended to unbounded intervals by transforming them into bounded intervals as proposed by Tokdar (2007).

[*]BECS, Aalto University, Finland
[†]BECS, Aalto University, Finland, aki.vehtari@aalto.fi

Tokdar (2007) integrated over the latent values with Metropolis–Hastings sampling and over the potential grid point sets with Metropolis–Hastings–Green sampling. The purpose of the latter part is to keep the number of active grid points small and automatically find the most important grid point locations. By replacing the sampling with an analytic approximation, inference can be made faster even if a fixed and finer grid is assumed. In this paper, we consider Laplace's method (LA) to approximate the posterior inference for LGP density estimation in a grid. Our objective is to obtain an accurate and quick approximation that enables practical estimation of densities by focusing on efficient ways to approximate Bayesian inference for LGP density estimation.

The proposed LA approach is related to other approaches in the literature. Given fixed covariance function parameters, the LA approximation for the posterior distribution involves finding the posterior mode, which is equivalent to a penalized maximum likelihood estimator considered by Leonard (1978) and Thorburn (1986). The marginal likelihood can be approximated with LA, which enables a fast gradient-based type-II maximum a posteriori (MAP) estimation of the covariance function parameters, or alternatively, the posterior distribution can be integrated over. Lenk (1991) uses a truncated Karhunen–Loeve representation and derives the moments of the density process, from which it is also possible to obtain a marginal likelihood approximation. Lenk evaluates the marginal likelihood of hyperparameter values in a grid, while we use gradients and quasi-Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization or Markov chain Monte Carlo (MCMC) integration. In Lenk's approach the mean of the density process is guaranteed to be positive, but there is no such guarantee for the whole posterior. Also the spectral representation restricts the selection of covariance functions. Lenk (2003) refines his approach with a Fourier series expansion and MCMC sampling of the hyperparameters.

The computational complexity of the proposed LA approach is dominated by the covariance matrix operations, which scale in a straightforward implementation as $\mathcal{O}(m^3)$, where m is the number of grid points. However, because of the discretization, the computational complexity is independent of the number of observations. Applying the proposed approach with a default grid size 400, 1D and 2D density estimation takes one to two seconds, which facilitates interactive visualization of data with density estimates and violin plots (see, e.g., Figures 2–5 and 8). Additionally, we consider fast Fourier transform (FFT) to speed up the computations with an even grid and stationary covariance functions. To avoid the cubic computational scaling in m with dense 2D grids, we also exploit Kronecker product computations to obtain a reduced-rank approximation of the exact prior covariance structure.

The number of grid points grows exponentially with the data dimension d. Although the Kronecker product approach is suitable for reducing the computation time when $d > 1$, the exponentially increasing number of latent function values makes the proposed approach impractical for more than three or four dimensions. Adams et al. (2009) propose an alternative GP approach called Gaussian process density sampler (GPDS) in which the numerical approximation of the normalizing term in the likelihood is avoided by a conditioning set and an elaborate rejection sampling method. The conditioning set is generated by the algorithm, which automatically places m_s points where they

are needed, making the estimation in higher-dimensional spaces easier. However, the computational complexity of GPDS scales as $\mathcal{O}((n + m_s)^3)$, where n is the number of data points.

The main contribution of this paper is the construction of the quick approximation for the LGP density estimation and regression by combining various ideas from the literature. Using Laplace's method to integrate over the latent values, we avoid the slow mixing of MCMC. We present FFT and reduced-rank based speed-ups tailored for LGP with LA. We demonstrate that LA can be further improved by rejection and importance sampling. We also show that using the type-II MAP estimation for two to three hyperparameters gives good results compared to the integration over the hyperparameters.

In the next section, we review the basics of the logistic Gaussian processes. In Section 3, we present the LA approach for LGP density estimation. We also introduce the additional approximations for speeding up the inference and consider briefly a similar LA approach for LGP density regression. In Section 4, we demonstrate the LA approach with several experiments, and compare it against MCMC and hierarchical infinite Gaussian mixture models by Griffin (2010).

2   Density Estimation with Logistic Gaussian Process

We consider the problem of computing a density estimate $p(\mathbf{x})$ given n independently drawn d-dimensional data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from an unknown distribution in a finite region $\mathcal{V}$ of $\mathbf{R}^d$. In this paper, we focus only on $d \in \{1, 2\}$. To find an estimate p for the unknown density, we can maximize the following log-likelihood functional

$$L(p) = \sum_{i=1}^{n} \log p(\mathbf{x}_i) \tag{1}$$

with the constraints $\int_{\mathcal{V}} p(\mathbf{x})d\mathbf{x} = 1$, and $p(\mathbf{x}) \geq 0$ for $\mathbf{x} \in \mathcal{V}$. The limiting solution leads to a mixture of delta functions located at the observations (Leonard 1978), which is why we need to set prior beliefs about the unknown density to obtain more realistic estimates.

To introduce the constraints of the density being non-negative and that its integral over $\mathcal{V}$ is equal to one, we employ the logistic density transform (Leonard 1978)

$$p(\mathbf{x}) = \frac{\exp(f(\mathbf{x}))}{\int_{\mathcal{V}} \exp(f(\mathbf{s}))d\mathbf{s}}, \tag{2}$$

where f is an unconstrained latent function. To smooth the density estimates, we place a Gaussian process prior for f, which enables us to set the prior assumptions about the smoothness properties of the unknown density p via the covariance structure of the GP prior.

We assume a zero-mean Gaussian process $g(\mathbf{x}) \sim \mathcal{GP}\left(0, \kappa(\mathbf{x}, \mathbf{x}')\right)$, where the covariance function is denoted with $\kappa(\mathbf{x}, \mathbf{x}')$ for a pair of inputs $\mathbf{x}$ and $\mathbf{x}'$. An example of a widely
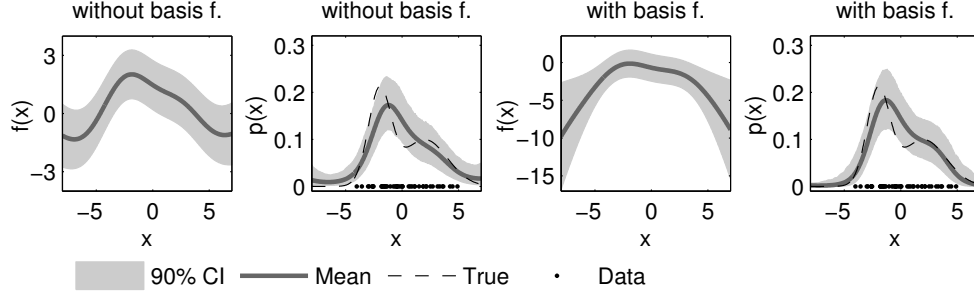
Figure 1: An illustration of the logistic Gaussian process density estimation with and without the basis functions. The first plot visualizes the posterior latent function without the basis functions, and the second plot shows the corresponding density estimate with the logistic density transform (2) given 50 observations from the mixture of two Gaussians: $\frac{1}{2}\mathcal{N}(-2,1) + \frac{1}{2}\mathcal{N}(2,2^2)$. The third plot visualizes the posterior latent function with the second-order polynomials as the basis functions, and the fourth plot shows the corresponding density estimate. The hyperparameter values of the squared exponential covariance function (3) are in this example $\sigma^2 = 1$ and $l_1 = \frac{1}{2}$.

used covariance function is the stationary squared exponential covariance function

$$\kappa(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{1}{2}\sum_{k=1}^{d} l_k^{-2}(x_k - x_k')^2\right), \tag{3}$$

where the hyperparameters $\theta = \{\sigma^2, l_1, \ldots, l_d\}$ govern the smoothness properties of f (Rasmussen and Williams 2006). The length-scale hyperparameters $l_1, \ldots, l_d$ control how fast the correlation decreases in different dimensions, and σ^2 is a magnitude hyperparameter.

For the latent function f in equation 2, we assume the model $f(\mathbf{x}) = g(\mathbf{x}) + \mathbf{h}(\mathbf{x})^T\mathbf{fi}$, where the GP prior is combined with the explicit basis functions $\mathbf{h}(\mathbf{x})$. Regression coefficients are denoted with $\mathbf{fi}$, and by placing a Gaussian prior $\mathbf{fi} \sim \mathcal{N}(\mathbf{b}, B)$ with the mean $\mathbf{b}$ and the covariance B, the parameters $\mathbf{fi}$ can be integrated out from the model, which results in the following GP prior for f (O'Hagan 1978; Rasmussen and Williams 2006):

$$f(\mathbf{x}) \sim \mathcal{GP}\left(\mathbf{h}(\mathbf{x})^T\mathbf{b}, \kappa(\mathbf{x}, \mathbf{x}') + \mathbf{h}(\mathbf{x})^T B \mathbf{h}(\mathbf{x}')\right). \tag{4}$$

We assume the second-order polynomials for the explicit basis functions. We use $\mathbf{h}(\mathbf{x}) = [x_1, x_1^2]^T$ in 1D, and $\mathbf{h}(\mathbf{x}) = [x_1, x_1^2, x_2, x_2^2, x_1 x_2]^T$ in 2D, which leads to a GP prior that can favour density estimates where the tails of the distribution go eventually to zero. The effect of the basis functions is demonstrated in an illustrative 1D example of Figure 1.

We discretize $\mathcal{V}$ into m subregions (or intervals in 1D), and collect the coordinates of the subregions into an $m \times d$ matrix X, where the i'th row denotes the center point of

the i'th subregion. Given X, the GP prior (4) results in the Gaussian distribution over the latent function values

$$p(\mathbf{f}|X,\theta) = \mathcal{N}(\mathbf{f}|H\mathbf{b}, K + HBH^T), \qquad (5)$$

where $\mathbf{f}$ is a column vector of m latent values associated with each subregion. The entries of the $m \times m$ covariance matrix K are determined by the input points X and the covariance function. The matrix H, of size $m \times 2$ in 1D and $m \times 5$ in 2D, contains the values of the fixed basis functions evaluated at X. We assume a weakly informative prior distribution for the regression coefficients by fixing the mean $\mathbf{b} = \mathbf{0}$ (a zero vector of a length 2 in 1D and 5 in 2D), and the covariance $B = 10^2 I$, where I is the identity matrix of a size 2×2 (in 1D) and 5×5 (in 2D). The regression coefficient for the quadratic term x_k^2 should be negative in order to make the tails of a distribution to go towards zero. However, the prior distribution does not force the negativity of the coefficient. In our experiments, the posterior of the coefficient was clearly below zero, but in our implementation it is also possible to force the negativity of the coefficient by rejection sampling (in 1D).

After the discretization, the log-likelihood contribution of an observation belonging to the i'th subregion can be written as

$$L_i = \log\left(\frac{w_i \exp(f_i)}{\sum_{j=1}^{m} w_j \exp(f_j)}\right), \qquad (6)$$

where the latent value f_i is associated with the i'th subregion. Throughout this paper, we assume a regular grid, and therefore the weights $w_1, \ldots, w_m$ have all the same value and can be omitted from (6). The number of observations that fall within the i'th subregion is denoted with y_i and all the count observations with an $m \times 1$ vector $\mathbf{y}$. The overall log-likelihood contribution of the n observations is given by

$$\log p(\mathbf{y}|\mathbf{f}) = \mathbf{y}^T \mathbf{f} - n \log\left(\sum_{j=1}^{m} \exp(f_j)\right). \qquad (7)$$

The prior (5) and the likelihood (7) are combined by Bayes' rule, which results in the conditional posterior distribution of the latent values

$$p(\mathbf{f}|X, \mathbf{y}, \theta) = \frac{1}{Z} p(\mathbf{f}|X, \theta) p(\mathbf{y}|\mathbf{f}), \qquad (8)$$

where $Z = p(\mathbf{y}|X, \theta) = \int p(\mathbf{f}|X, \theta) p(\mathbf{y}|\mathbf{f}) d\mathbf{f}$ is the marginal likelihood. Due to the non-Gaussian likelihood (7), the posterior distribution is also non-Gaussian, and approximate methods are needed to integrate over $\mathbf{f}$.

3 Approximate Inference

In this section, we discuss the implementation issues of Laplace's method for logistic GP density estimation. In Section 3.1, we present an efficient approach for the mode

finding and computation of the marginal likelihood and predictions. Further speed-ups are obtained by using the fast Fourier transform for matrix-vector multiplications and by Kronecker product and reduced-rank approximations suitable for $d > 1$ cases with dense grids (large m). In Section 3.2, we consider the LA approach for logistic GP density regression, and in Section 3.3, we give a brief description of inference with MCMC.

3.1   Inference with the Laplace Approximation

Our approach resembles the Laplace approximation for GP classification (Williams and Barber 1998; Rasmussen and Williams 2006) and GP point process intensity estimation (Cunningham et al. 2008), but the implementation is different because in LGP each term in the likelihood (7) depends on all the latent values $\mathbf{f}$.

The Laplace approximation is based on a second-order Taylor expansion for $\log p(\mathbf{f}|X, \mathbf{y}, \theta)$ around the posterior mode $\hat{\mathbf{f}}$, which results in the Gaussian approximation

$$q(\mathbf{f}|X, \mathbf{y}, \theta) = \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}}, \Sigma), \tag{9}$$

where $\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} p(\mathbf{f}|X, \mathbf{y}, \theta)$. The covariance matrix is given by

$$\Sigma = (C^{-1} + W)^{-1}, \tag{10}$$

where $C = K + HBH^T$ and $W = -\nabla\nabla_{\mathbf{f}} \log p(\mathbf{y}|\mathbf{f})|_{\mathbf{f}=\hat{\mathbf{f}}}$. The likelihood (7) leads to a full matrix W with the following structure[1]

$$W = n(\mathrm{diag}(\mathbf{u}) - \mathbf{u}\mathbf{u}^T), \tag{11}$$

where the non-negative entries of the vector $\mathbf{u}$ are given by

$$u_i = \frac{\exp(f_i)}{\sum_{j=1}^m \exp(f_j)}. \tag{12}$$

In the implementation, forming the full matrix W can be avoided by using the vector $\mathbf{u}$ and the structure (11). Similarly to multiclass classification with the softmax likelihood (Williams and Barber 1998; Rasmussen and Williams 2006), W is positive semidefinite, and because C is positive definite, $p(\mathbf{f}|X, \mathbf{y}, \theta)$ has a unique maximum.

Newton's Method for Finding the Mode

We use Newton's method for finding the mode $\hat{\mathbf{f}}$. At each iteration, we need to compute

$$\mathbf{f}^{\mathrm{new}} = (C^{-1} + W)^{-1}\mathbf{v}, \tag{13}$$

---

[1]We use the following notation: $\mathrm{diag}(\mathbf{w})$ with a vector argument means a diagonal matrix with the elements of the vector $\mathbf{w}$ on its diagonal, and $\mathrm{diag}(W)$ with a matrix argument means a column vector consisting of the diagonal elements of the matrix W.

where $\mathbf{v} = W\mathbf{f} + \nabla_{\mathbf{f}} \log p(\mathbf{y}|\mathbf{f})$. To increase the numerical stability of the computations, we use the factorization $W = RR^T$, where

$$R = \sqrt{n}((\text{diag}(\mathbf{u}))^{\frac{1}{2}} - \mathbf{u}\mathbf{u}^T(\text{diag}(\mathbf{u}))^{-\frac{1}{2}}). \tag{14}$$

Instead of a direct implementation of equation 13, we can apply the matrix inversion lemma (see, e.g., Harville 1997) to write the Newton step in a numerically preferable way as

$$\mathbf{f}^{\text{new}} = C(I_m - R(I_m + R^T C R)^{-1} R^T C)\mathbf{v}. \tag{15}$$

The inversion of equation 15 can be computed by solving $\mathbf{z}$ from the linear system $(I_m + R^T C R)\mathbf{z} = R^T C\mathbf{v}$ with the conjugate gradient method. As discussed, for example, by Cunningham et al. (2008), a stationary covariance function and evenly spaced grid points lead to a Toeplitz covariance matrix which can be embedded in a larger circulant matrix enabling efficient computations by using the fast Fourier transform. With a Toeplitz covariance matrix K, we can achieve a small speed-up with larger grid sizes in the evaluation of the Newton step (15) because all the multiplications of K and any vector can be done efficiently in the frequency domain. By using FFT, these matrix-vector multiplications become convolution operations for which we are required only to form a single row of the embedded circulant matrix instead of the full matrix K. The rest of the matrix-vector multiplications in equation 15 are fast because the matrix H has only two (in 1D) or five (in 2D) columns, and instead of forming the full matrix R, we can use the vector $\mathbf{u}$ and exploit the structure of R from equation 14.

The Approximate Marginal Likelihood and Predictions

Approximating the integration over $\mathbf{f}$ with the LA approximation enables fast gradient-based type-II MAP estimation for choosing the values for the covariance function hyper-parameters. After finding $\hat{\mathbf{f}}$, the approximate log marginal likelihood can be evaluated as

$$\log p(\mathbf{y}|X, \theta) \approx \log q(\mathbf{y}|X, \theta) = -\frac{1}{2}\hat{\mathbf{f}}^T C^{-1}\hat{\mathbf{f}} + \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2}\log|I_m + R^T C R| \tag{16}$$

(see, e.g., Rasmussen and Williams 2006). The first and second terms of equation 16 are fast to compute by using the results from Newton's algorithm, but the determinant term is more difficult to evaluate.

Cunningham et al. (2008) show that for the point process intensity estimation with the GP priors, the evaluation of a corresponding determinant term can be done efficiently by exploiting a low-rank structure of the observation model. In density estimation W has rank $m - 1$ due to the normalization over $\mathcal{V}$, and a similar low-rank representation as in intensity estimation cannot be used for W. Therefore, we form the full $m \times m$ matrix and compute the Cholesky decomposition that scales as $\mathcal{O}(m^3)$. Although this is computationally burdensome for large m, it is required only once per each evaluation of the marginal likelihood after the convergence of Newton's algorithm. Typically, in

1D cases with the number of intervals m being less than one thousand, the LA approach is sufficiently fast for practical purposes (in our implementation the default size for m is 400). However, the computations can become restrictive in 2D with large m. In such cases, we consider a reduced-rank approximation for the prior covariance K to find a faster way to compute the determinant term and the approximate posterior covariance, as will be discussed in Section 3.1. The gradients of the log marginal likelihood with respect to the hyperparameters θ can be computed by evaluating explicit and implicit derivatives similarly as shown by Rasmussen and Williams (2006).

We place a prior distribution $p(\theta)$ for the hyperparameters to improve the identifiability of the ratio of the magnitude and the length-scale parameters of the covariance function. We assume a weakly informative half Student-t distribution, as recommended for hierarchical models by Gelman (2006), with one degree of freedom and a variance that is equal to ten in 1D and a thousand in 2D for σ (magnitude). For $l_1, \ldots, l_d$ (length-scales), we assume otherwise the same prior but with a variance that is equal to one. The MAP estimate for the hyperparameters is found by maximizing the approximate marginal posterior $p(\theta|\mathbf{y}, X) \propto q(\mathbf{y}|X, \theta)p(\theta)$, in which we use the BFGS quasi-Newton optimization. In addition to the MAP estimate, we also approximate the integration over the hyperparameters with the central composite design (LA-CCD-LGP) scheme, similarly as proposed by Rue et al. (2009).

To compute the joint posterior predictive distribution, we marginalize over the latent values by Monte Carlo using 8000 latent samples drawn from the multivariate normal posterior predictive distribution. This sampling is needed only once after the MAP estimate for the hyperparameters has been found, and time consumed in this step is negligible compared to the other parts of the computations. In many cases, we have observed that the posterior weights of the quadratic terms x_k^2 of the basis function are automatically small and thus the effect of basis functions is not strong for densities not well represented by them. We consider an optional rejection sampling step based on finite differences to force the tails to go to zero (in 1D), if necessary (not used if the density is defined to be bounded).

To improve the Gaussian approximation of the posterior distribution, we also consider an additional importance sampling step (LA-IS-LGP). Following Geweke (1989) we use the multivariate split Gaussian density as an approximation for the exact posterior. The multivariate split Gaussian is based on the posterior mode and covariance, but the density is scaled along the principal component axes (in the positive and negative direction separately) adaptively to match to the skewness of the true distribution (see also Villani and Larsson 2006). To further improve the performance the discontinuous split Gaussian used by Geweke (1989) was replaced with a continuous version. We observed that scaling is not needed along the principal component axis corresponding to the smallest eigenvalues. To speed up the computation, we scale only along the first 50 principal component axes. The importance sampling step took approximately an additional 0.3 seconds, whereas direct sampling from the Gaussian posterior distribution (to compute the predictions) took about 0.05 seconds. Due to this small additional computational demand, we added the importance sampling correction also in the experiments. In all the experiments the estimated effective number of samples (Kong et al. 1994) was rea-

sonable, but we also included a sanity check to the code and give a warning and use a soft thresholding of the importance weights if the estimated effective number of samples is less than 200. As a comparison, using scaled Metropolis–Hastings sampling to obtain 500 effectively independent samples from the latent posterior given fixed hyperparameters took 24 minutes (similar inefficiency in mixing was observed with other MCMC methods usually used for GPs).

The Reduced-Rank Approximation for 2D Density Estimation

To speed up the inference with 2D grids when m is large, we propose a reduced-rank approximation that can be formed efficiently by exploiting Kronecker product computations. For separable covariance functions, the covariance structure of evenly spaced grid points can be presented in a Kronecker product form $K = K_1 \otimes K_2$, where K_k is an $m_k \times m_k$ matrix representing the covariances between the latent values associated with m_k grid points in the k'th dimension. For Kronecker products, many matrix operations scale efficiently: for example, the determinant $|K_1 \otimes K_2|$ can be computed by using the determinants of the smaller matrices K_1 and K_2 as $|K_1|^{m_2}|K_2|^{m_1}$ (Harville 1997). To compute the approximate marginal likelihood (16), we need to evaluate the determinant term of a form $|I_m + R^T(K_1 \otimes K_2 + HBH^T)R|$. Unfortunately, the Kronecker product covariance structure does not preserve due to the multiplication and summation operations, leading to the unfavourable $\mathcal{O}(m^3)$ scaling. However, we can exploit the Kronecker product $K_1 \otimes K_2$ to obtain the eigendecomposition of K efficiently, and then form a reduced-rank approximation for K by using only the largest eigenvalues with the corresponding eigenvectors. The idea of using the eigendecomposition to construct a reduced-rank approximation for the covariance matrix has been previously mentioned by Rasmussen and Williams (2006, Chapter 8). By denoting an eigenvalue of K_1 with r_1 and an eigenvector of K_1 with $\mathbf{v}_1$, and similarly, an eigenvalue of K_2 with r_2 and an eigenvector of K_2 with $\mathbf{v}_2$, then, $r_1 r_2$ is an eigenvalue of $K_1 \otimes K_2$ and $\mathbf{v}_1 \otimes \mathbf{v}_2$ an eigenvector of $K_1 \otimes K_2$ corresponding to the eigenvalue $r_1 r_2$ (Harville 1997). Thus, instead of computing the eigendecomposition of K, which is an $\mathcal{O}(m^3)$ operation, we can compute the eigendecompositions of K_1 and K_2, which scales as $\mathcal{O}(m_1^3 + m_2^3)$, and form a desired number of eigenvalues and eigenvectors with the Kronecker product computations to obtain the reduced-rank approximation for K.

We approximate the exact prior covariance with

$$K \approx VSV^T + \Lambda, \tag{17}$$

where S is a diagonal matrix of size $s \times s$ having the s largest eigenvalues on its diagonal and V is an $m \times s$ matrix consisting of the corresponding eigenvectors. Similarly to the fully independent conditional (FIC) approximation (Snelson and Ghahramani 2006), we use the exact full-rank diagonal by setting $\Lambda = \mathrm{diag}(\mathrm{diag}(K) - \mathrm{diag}(VSV^T))$ in equation 17 to obtain a more accurate approximation.

With the approximate prior (17), the mode can be found using Newton's method in a similar way as described in Section 3.1. To evaluate the marginal likelihood approximation, we need to compute efficiently the determinant in equation 16. The determinant

term can be written as $n|A||\mathbf{1}^T D^{1/2} A^{-1} D^{1/2} \mathbf{1}|$, where $\mathbf{1}$ is an $m \times 1$ column vector of ones, $D = n(\text{diag}(\mathbf{u}))$ is a diagonal matrix, and $A = I_m + D^{1/2}(\Lambda + \tilde{V}\tilde{S}\tilde{V}^T)D^{1/2}$. We have defined $\tilde{V} = \begin{bmatrix} V & H \end{bmatrix}$, which is of size $m \times (s+5)$, and $\tilde{S} = \begin{bmatrix} S & 0 \\ 0 & B \end{bmatrix}$ of size $(s+5) \times (s+5)$. To avoid forming any $m \times m$ matrix, $|A|$ can be evaluated by applying the matrix determinant lemma (see, e.g., Harville 1997) and A^{-1} by applying the matrix inversion lemma. With the prior (17), we can also compute the approximate gradients of the marginal likelihood with respect to hyperparameters without resorting to any $\mathcal{O}(m^3)$ matrix operations.

For large m, we use the same fixed grid for the observations and predictions. After we have found the MAP estimate for the hyperparameters, we need to draw samples from $q(\mathbf{f}|X, \mathbf{y}, \theta)$ to marginalize over the latent values. The posterior covariance is approximated by

$$\Sigma \approx \tilde{C} - \tilde{C}(\tilde{C} + W^{-1})^{-1}\tilde{C}, \tag{18}$$

where $\tilde{C} = \Lambda + \tilde{V}\tilde{S}\tilde{V}^T$ is the approximate prior covariance with the explicit basis functions evaluated at the grid points. We can rewrite $(\tilde{C} + W^{-1})^{-1}$ in equation 18 as

$$(\tilde{C} + W^{-1})^{-1} = E - E\mathbf{1}(\mathbf{1}^T E\mathbf{1})^{-1}\mathbf{1}^T E, \tag{19}$$

where

$$E = D^{1/2}(Z - ZD^{1/2}\tilde{V}\tilde{S}^{1/2}(LL^T)^{-1}\tilde{S}^{1/2}\tilde{V}^T D^{1/2}Z)D^{1/2}. \tag{20}$$

In equation 20 we have denoted $Z = (I_m + D\Lambda)^{-1}$ which is diagonal and therefore fast to evaluate. The matrix L in equation 20 is an $(s+5) \times (s+5)$ lower triangular matrix, and it can be computed using the Cholesky decomposition

$$L = \text{chol}(I_{s+5} + \tilde{S}^{1/2}\tilde{V}^T D^{1/2}ZD^{1/2}\tilde{V}\tilde{S}^{1/2}), \tag{21}$$

which scales as $\mathcal{O}((s+5)^3)$. When drawing samples from the Gaussian approximation $q(\mathbf{f}|X, \mathbf{y}, \theta)$ given the covariance matrix (18), the structure of $\tilde{C}$ can be exploited to avoid forming any $m \times m$ matrix.

With the reduced-rank approximation based on the eigendecomposition, we avoid choosing (or optimizing) the locations of inducing inputs, which is required in many sparse approximations, such as, fully independent conditional (FIC) sparse approaches (e.g. Snelson and Ghahramani 2006; Quiñonero-Candela and Rasmussen 2005). With the possibility to choose the locations of inducing inputs, the reduced-rank approximation would be more expressive although choosing automatically the locations of inputs can be challenging if the correlation structure of a GP prior is wanted to be preserved using a smaller number of inducing inputs. In addition, the optimization of the locations of inducing inputs is not trivial and can lead to overfitting (see, e.g., the discussion and visualizations of different correlation structures by Vanhatalo et al. 2010). The problem with the reduced-rank approximation based on Kronecker products, however, is that the covariance function must be separable with respect to the inputs, which leads to a restricted class of possible covariance functions. In this paper, we have tested the reduced-rank approximation with the squared exponential covariance function.

3.2 Density Regression with the Laplace Approximation

Logistic Gaussian processes are also suitable for estimating conditional densities $p(t|\mathbf{x})$, where t is a target variable (see, e.g., Tokdar et al. 2010). We discretize both the covariate and target space in a finite region to model the conditional densities with the logistic GP. In this paper, we focus on modelling the conditional density of t given a univariate predictor x, which leads to a 2D grid similarly as in the case of 2D density estimation. We denote the number of intervals in the covariate space with m_x and the number of intervals in target space with m_t. To estimate the conditional densities, we write the log-likelihood contribution of all the observations as

$$\log p(\mathbf{y}|\mathbf{f}) = \sum_{i=1}^{m_x} \left(\mathbf{y}_i^T \mathbf{f}_i - n_i \log \left(\sum_{j=1}^{m_t} \exp(f_{i,j}) \right) \right), \tag{22}$$

where $f_{i,j}$ is the j'th element of $\mathbf{f}_i$. The $m_t \times 1$ vector $\mathbf{f}_i$ contains all the latent values associated with each subregion conditioned to the covariate x_i, that is, the latent values associated with the i'th slice of the grid. Similarly, $\mathbf{y}_i$ is a vector of length m_t and consists of the number of observations in each subregion of the i'th slice of the grid associated with x_i. The vector $\mathbf{f}$ contains all the latent values of the grid and $\mathbf{y}$ all the count observations. To approximate the resulting non-Gaussian posterior distribution, we use Laplace's method. The likelihood (22) results in that W in equation 10 is a block-diagonal matrix. Similarly as in Section 3.1, the i'th block of W can be factorized into $R_i R_i^T$, where

$$R_i = \sqrt{n_i}((\operatorname{diag}(\mathbf{u}_i))^{\frac{1}{2}} - \mathbf{u}_i \mathbf{u}_i^T (\operatorname{diag}(\mathbf{u}_i))^{-\frac{1}{2}}). \tag{23}$$

The total number of observations in the i'th slice of the grid is denoted with n_i, and the vector $\mathbf{u}_i$ is formed as in equation 12, but by considering only the latent values $\mathbf{f}_i$. Because the LA approach for the LGP density regression follows closely to the derivation presented in Section 3.1, the implementation issues are omitted here. The density regression becomes computationally challenging with dense grids and when applied to a larger number of covariate dimensions d, and therefore, computational speed-ups are required, but we do not consider these in this paper.

3.3 Markov Chain Monte Carlo

MCMC sampling enables approximating the integration over the posterior distribution without limiting to a Gaussian form approximation. MCMC can be extremely slow but in the limit of a long run it provides an exact result by which the accuracy of the LA approach can be measured.

We approximate the posterior distribution by sampling alternatively from the conditional posterior of the latent values $p(\mathbf{f}|X, \mathbf{y}, \theta)$ by using scaled Metropolis–Hastings sampling (Neal 1998) and from the conditional posterior of the hyperparameters $p(\theta|\mathbf{f}, X, \mathbf{y})$ by using the no-U-turn sampler (NUTS) (Hoffmann and Gelman 2013).

In the experiments, this combination was more efficient than other MCMC methods usually used for GPs.

A fixed length of MCMC chains was determined by estimating the convergence and effective sample size for different data sets. The convergences of MCMC chains were diagnosed with visual inspection and the potential scale reduction factor (Brooks and Gelman 1998). The effective sample size was estimated with Geyer's initial monotone sequence estimator (Geyer 1992). As the sampling from the conditional posterior of the latent values was significantly faster, each meta-iteration consisted of a hundred sampling steps for $p(\mathbf{f}|X, \mathbf{y}, \theta)$ and one step for $p(\theta|\mathbf{f}, X, \mathbf{y})$. The sampling was initialised by using the hyperparameter values at the mode of the LA approximated marginal posterior and the initial latent values were sampled from the Gaussian approximation given the initial hyperparameter values. The sampling consisted of 5100 meta-iterations of which 100 iterations were removed as a burn-in. The effective sample sizes with different data sets and random number sequences were about 50–100, which gives sufficient accuracy for the posterior mean of the density estimate. For a grid size of 400, the computation time was about 2200 seconds.

4   Experiments

In this section, we examine the performance of the Laplace approximation for the logistic Gaussian process (LA-LGP) with several simulated and real data sets. We compare LA-LGP to the MCMC approximation (MCMC-LGP) and to Griffin's (2010) Dirichlet process mixture of Gaussians with common component variance (CCV-mixture) and different component variance (DCV-mixture). The CCV model assumes that all mixture components have equal variances. The DCV model allows different variances for the components. Computation for Dirichlet process mixture models is done with Gibbs sampling. We compare LA-LGP only to advanced Bayesian kernel methods (Griffin 2010), since Tokdar (2007) and Adams (2009) have already shown that logistic GP works better than simple kernel methods (such as the Parzen method). Griffin showed that CCV- and DCV-mixture models performed equally or better than other default priors for mixture models, which is why the other priors are excluded in our comparisons. We do not compare the performance to other MCMC based LGP approaches by Tokdar (2007) and Adams (2009) as the prior is the same and if using the same grid the difference would only be in the implemantation speed and convergence speed of the different MCMC methods. LA-LGP and MCMC-LGP were implemented using the GPstuff toolbox[2] (Vanhatalo et al. 2013), and CCV/DCV-mixtures were computed using Griffin's code[3].

The squared exponential covariance function was employed in all the experiments with LGP. We also tested Matérn, exponential, rational quadratic and additive combinations, but the results did not improve considerably with these different covariance functions. To ensure that the same prior is suitable for different scales, we normalized the grid to

---

[2]http://mloss.org/software/view/451/ or http://becs.aalto.fi/en/research/bayes/gpstuff/
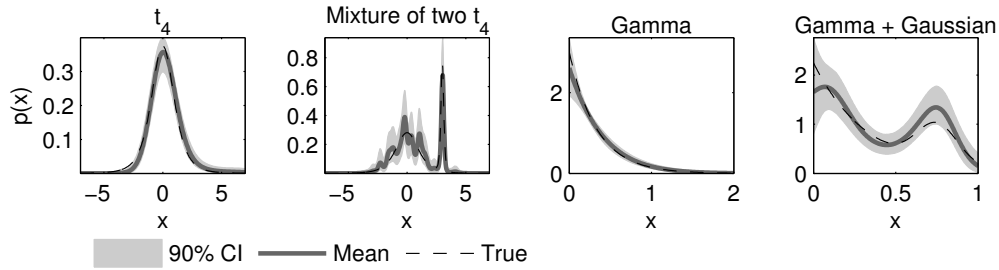[3]http://www.kent.ac.uk/ims/personal/jeg28/BDEcode.zip

Figure 2: Example results of density estimation and related uncertainties for four different simulated data sets with $n = 100$.

have a zero mean and unit variance in all the experiments. If not otherwise mentioned we used LGP with a grid size 400.

The rest of this section is divided into five parts. We compare the performances of LA-LGP, MCMC-LGP and CCV/DCV-mixtures with simulated 1D data in Section 4.1, and with real 1D data in Section 4.2. In Section 4.3, we test how different grid sizes and the number of data points affect density estimates. We illustrate density estimation with simulated 2D data in Section 4.4, and finally, we demonstrate density regression with one simulated data in Section 4.5.

4.1 Simulated 1D Data

Figure 2 shows the density estimates and the corresponding 95% credible intervals for single random realisations from simulated data sets with $n = 100$. The simulated data sets are:

- t_4: Student-$t_4 (0, 1)$

- Mixture of two t_4: $\frac{3}{4} t_4 (0, 1) + \frac{1}{4} t_4 \left(3, \frac{1}{8^2}\right)$

- Gamma: $\mathrm{Gamma}\left(1, \frac{1}{3}\right)$

- Truncated Gamma+Gaussian $x \in (0, 1)$: $\frac{3}{4} \mathrm{Gamma}\left(1, \frac{1}{3}\right) + \frac{1}{4} \mathcal{N}\left(\frac{3}{4}, \frac{1}{8^2}\right)$.

Student's t_4-distribution was chosen as a simple unimodal distribution with thicker tails than Gaussian. The mixture of two t_4-distributions was chosen as a more challenging case, where there are two separate modes with different widths. In the second plot of Figure 2, it can be seen that the short length-scale required to model the narrow peak makes the density estimate bumpy also elsewhere. Gamma was chosen as a simple distribution with a mode on the boundary, and a truncated Gamma+Gaussian, $x \in (0, 1)$, as a more difficult case with one mode on the boundary and another mode in the middle of the distribution. Truncated Gamma+Gaussian was also used by Tokdar (2007) and Adams (2009).
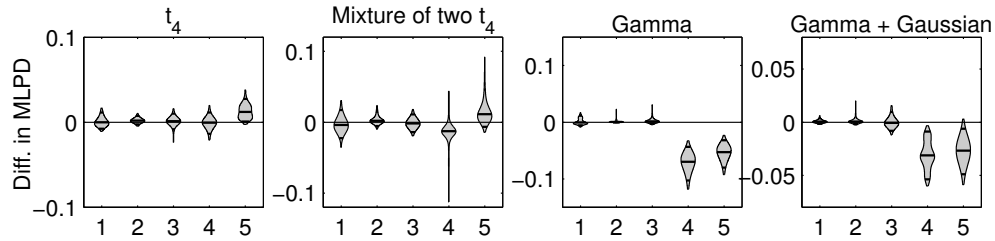
Figure 3: The pairwise comparison of LA-LGP to (1) LA-LGP (no IS), (2) LA-CCD-LGP, (3) MCMC-LGP, (4) CCV-mixture and (5) DCV-mixture. The plot shows the distribution of differences in the mean log-predictive density (MLPD) along with the median and 95% lines. The values above zero indicate that a method is performing better than LA-LGP. The MLPDs were computed using the true density as the reference. The violin plots were produced using LA-LGP from the results of 100 independent random repetitions.

Figure 3 shows the pairwise comparison of LA-LGP to LA-LGP without importance sampling (LA-LGP, no IS), LA-LGP with the CCD integration (LA-CCD-LGP), MCMC-LGP, CCV-mixture and DCV-mixture. We made 100 realisations of $n = 100$ random samples from the true density and computed for each method the mean log-predictive densities (MLPD) over the true distribution. Distributions of the differences between MLPDs are plotted with violin plots generated using LA-LGP along with the median and 95% lines. There are no statistically significant differences between the methods for the first three data sets. For the last data set the Gaussian mixture models do not work well for the data with the mode on the boundary. The violin plots in Figures 3 and 5 also illustrate another practical application of density estimation with the LA approach which facilitates interactive visualisation.

4.2   Real 1D Data

Figure 4 shows the density estimates and the corresponding 95% credible intervals for the following real data sets:

- Galaxy $n = 82$
- Enzyme $n = 245$
- Log acidity $n = 155$
- Sodium lithium $n = 190$,

which were studied by Griffin (2010). The density estimates are visually similar to the results by Griffin (2010, Figure 5).

Figure 5 shows the pairwise comparison of LA-LGP to LA-LGP (no IS), LA-CCD-LGP, MCMC-LGP, CCV-mixture and DCV-mixture with the real data sets. For each
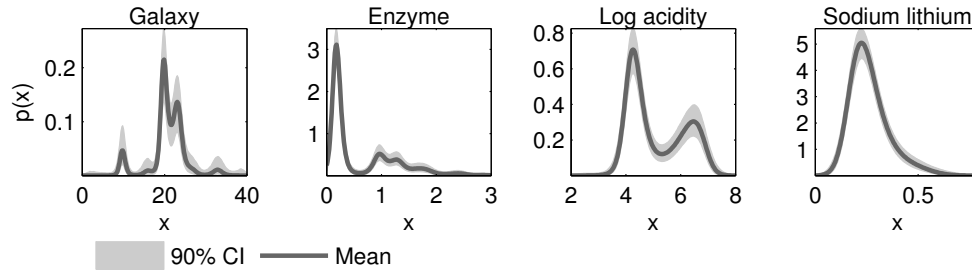
Figure 4: Example results of density estimation and related uncertainties for four different real data sets with $n = \{82, 245, 155, 190\}$.

method, we computed leave-one-out cross-validation mean log-predictive densities (CV-MLPD). Samples from the distributions of the differences between CV-MLPDs were obtained using the Bayesian bootstrap (Rubin 1981; Vehtari and Lampinen 2002) and violin plots were generated using LA-LGP.

There is no substantial difference in the performances of LA-LGP and MCMC-LGP across the data sets. Importance sampling improves the performance of the Laplace approximation for the Galaxy data set, but there is no improvement for the other data sets. CCD and MCMC improved the performance only for the Enzyme data set. CCV-mixture and DCV-mixture perform similar to LA-LGP for all the other data sets except for the Sodium lithium data set for which they have slightly worse performance.

Figure 6 shows the effect of the rejection sampling and the importance sampling in the density estimation of the Galaxy data set. The rejection sampling helps to make the tails decreasing. The importance sampling makes the estimate to be lower on the areas with no observations and respectively higher at the modes. When looking at MCMC posterior samples, the Galaxy data had most skewed marginal posterior distributions of the latent values, explaining the benefit of the importance sampling.

4.3   The Effect of the Number of Data and Grid Points

Figure 7 illustrates the effect of the number of grid and data points to the accuracy of LA-LGP and MCMC-LGP. The number of grid points was 400 when the number of data points was varied, and the number of data points was 100 when the number of grid points was varied. For each combination, we made 100 realisations of random samples from the mixture of two t_4-distributions and the truncated Gamma+Gaussian distribution. For both data sets the KL divergence approaches to zero when the number of data points increase. It seems that about 50–100 grid points are sufficient for the mixture of two t_4-distributions, whereas the truncated Gamma+Gaussian distribution is less sensitive to the number of grid points. There is no practical difference in the performances of LA-LGP and MCMC-LGP.

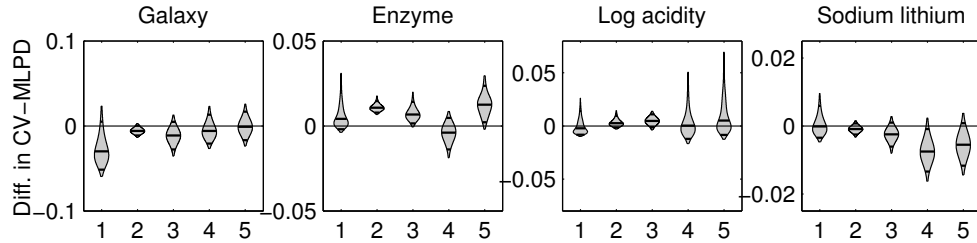We measured computation times for density estimation with different grid sizes using

Figure 5: The pairwise comparison of LA-LGP to (1) LA-LGP (no IS), (2) LA-CCD-LGP, (3) MCMC-LGP, (4) CCV-mixture and (5) DCV-mixture. The plot shows the distribution of differences in the cross-validation mean log-predictive density (CV-MLPD) along with the median and 95% lines. The values above zero indicate that a method is performing better than LA-LGP. CV-MLPDs were computed using leave-one-out cross-validation. Violin plots were produced with LA-LGP given 1000 Bayesian bootstrap draws from the distribution of CV-MLPD.
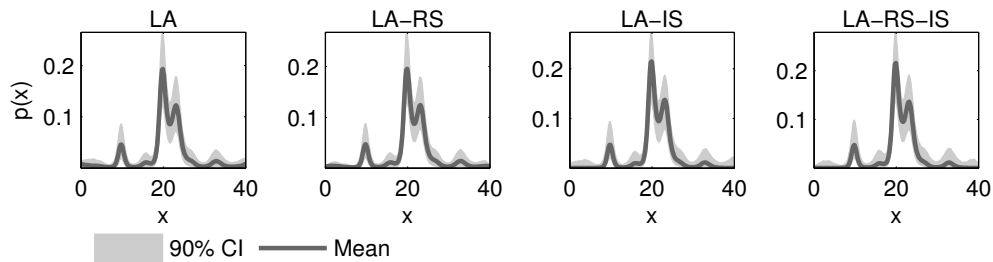
Figure 6: Density estimates for the Galaxy data set with different sampling options. The plots show the estimate with the Laplace approximation (LA), with additional rejection sampling (LA-RS) or importance sampling (LA-IS), or with both rejection and importance sampling (LA-RS-IS). See the text for an explanation.
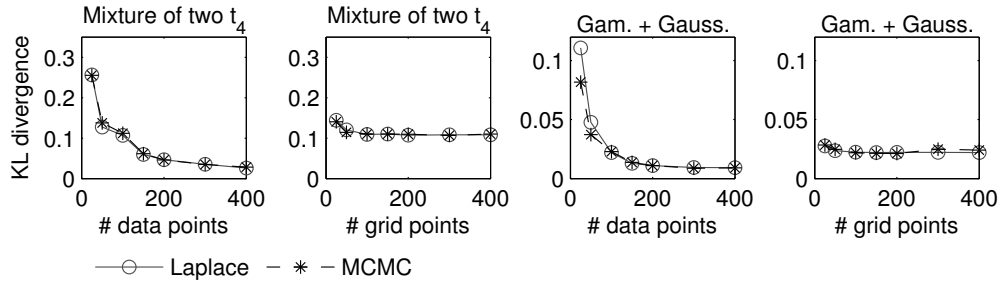
Figure 7: The effect of the number of data and grid points to the accuracy of density estimation with two simulated data sets. The number of grid points was 400 when the number of data points was varied, and the number of data points was 100 when the number of grid points was varied. For both data sets, as the number of data points increases, the KL divergence decreases to near zero. About 50–100 grid points seem to be sufficient for the mixture of two t_4-distributions, whereas the truncated Gamma+Gaussian distribution is less sensitive to the number of grid points. There is no practical difference in the performances of LA-LGP and MCMC-LGP.

100 random samples from the Student-t_4 distribution. LA-LGP with the grid sizes $(50, 100, 200, 400, 900)$ took about $(0.1, 0.2, 0.3, 0.9, 4.5)$ seconds using four cores of Intel(R) Xeon(R) 2.67GHz. If the matrix-vector multiplications in Newton's algorithm were made with FFT, the times were about $(0.3, 0.3, 0.4, 0.8, 3.4)$ seconds. MCMC-LGP with the same grid sizes took approximately $(350, 460, 640, 2200, 8100)$ seconds. Using Griffin's code with the default options, CCV-mixture took about 200 seconds and DCV-mixture about 800 seconds. Note that these time comparisons are only approximate, and the computation times depend considerably on the specific implementation and on the chosen convergence criteria.

4.4 Simulated and Real 2D Data

The columns 1–4 of Figure 8 show the following four simulated 2D distributions:

- Student-t_8: $t_8 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & .7 \\ .7 & 1 \end{bmatrix} \right)$

- Mixture of two Gaussians: $\frac{1}{2} \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) + \frac{1}{2} \mathcal{N} \left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \right)$

- Banana: $x_1 \sim \mathcal{N} \left(0, 10^2 \right), \quad x_2 \sim \mathcal{N} \left(\frac{1}{50} x_1^2 - \frac{1}{5}, 1 \right)$

- Ring: $\frac{1}{2\pi} \int_{-\pi}^{\pi} \mathcal{N} \left(\frac{3}{2} \begin{bmatrix} \cos \varphi \\ \sin \varphi \end{bmatrix}, \begin{bmatrix} 0.2^2 & 0 \\ 0 & 0.2^2 \end{bmatrix} \right) d\varphi.$

Given a random sample with $n = 100$ from each distribution, we compute the 2D density estimates with LA-LGP. The mean estimates are illustrated in the lower row of Figure
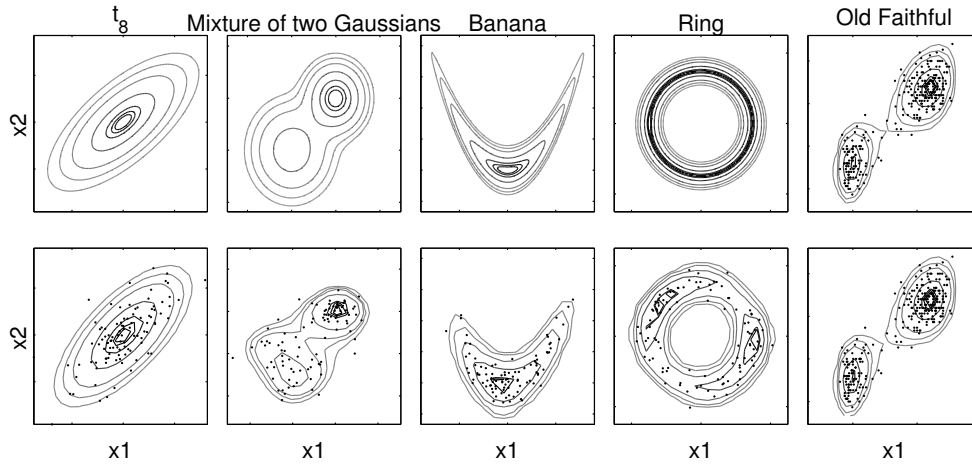
Figure 8: Example results of 2D density estimation for four different simulated data sets (Columns 1–4, $n = 100$ observations) and for one real data set (Column 5, $n = 272$ observations). The upper row shows the contour plots of the true densities for the simulated data sets and the MCMC-LGP result for the Old Faithful data. The lower row shows the contour plots of the estimated densities inferred with LA-LGP.

8. The fifth column of Figure 8 shows two density estimates for the Old Faithful data ($n = 272$). The upper plot shows the estimate with MCMC-LGP and the lower plot shows the estimate with LA-LGP. The density estimation for LA-LGP with a 20×20 grid took about two seconds and for MCMC-LGP about 29 minutes.

We tested the reduced-rank approximation with two 2D data sets. Given $n = 100$ observations from the Student-t_8 and the mixture of two Gaussians distributions, we measured the computation times and the KL divergences from the true density to the estimated density with different grid sizes. Figure 9 shows the elapsed times and the KL divergences as a function of the grid sizes for LA-LGP with the exact prior covariance (Full) and with the reduced-rank prior covariance (Kron) of equation 17. We formed the reduced-rank approximation by excluding all the eigenvalues smaller than 10^{-6}, or taking at most 50% of all the eigenvalues. The differences between the exact and the reduced-rank prior are small in the KL sense, but for the grid sizes larger than 30×30, LA-LGP with the full prior covariance matrix becomes computationally more expensive than with the reduced-rank approximation.

4.5   Density Regression with Simulated Data

The estimation of a conditional density in a grid with the LA approach is essentially similar to 1D density modelling with LA-LGP. Therefore, we consider density regression with only one simulated data set in this paper.
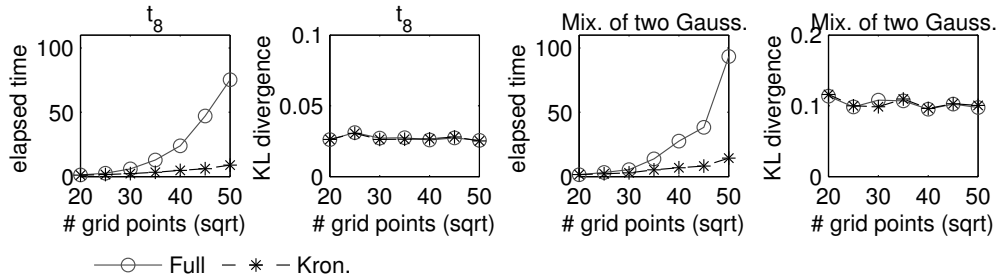
Figure 9: The comparison of LA-LGP with the full prior covariance (Full) and with the reduced-rank approximation (Kron.) of equation 17 for two 2D data sets. The performances are similar in the KL sense, but for grid sizes larger than 30×30, the computation times are larger with the full prior covariance matrix.

We demonstrate density regression with a simulated case studied by Kundu and Dunson (2011). The first plot of Figure 10 shows the simulated density regression scheme, where the predictors z_i $(i = 1, \dots, 50)$ are generated from the following trimodal density: $\frac{9}{20}\mathcal{N}(-\frac{6}{5}, (\frac{3}{5})^2) + \frac{9}{20}\mathcal{N}(\frac{6}{5}, (\frac{3}{5})^2) + \frac{1}{10}\mathcal{N}(0, (\frac{1}{4})^2)$. The generating model for a univariate target is $t_i = \lambda \exp\left(-\frac{e^{z_i}}{1+e^{z_i}}\right) + \frac{e^{z_i}}{1+e^{z_i}}\epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$. We fixed $\lambda = 3$ and $\sigma_\epsilon = 1$. The second and third plots of Figure 10 show estimates with the Laplace approximation (LA-DR) and the MCMC sampling (MCMC-DR) given a single realisation of $n = 50$ samples. Density regression in a 20×20 grid with LA-DR took about three seconds and with MCMC 1600 seconds. Finally, to show the differences between the estimates obtained with LA-DR and LA-LGP, we illustrate a conditional density estimate computed directly from a 2D density estimate obtained with LA-LGP (the fourth plot of Figure 10).

5 Discussion

In this paper, we have proposed to use Laplace's method for fast logistic Gaussian process density estimation. The empirical results with 1D data sets indicate that the accuracy of the proposed LA approach with type-II MAP estimation is close to the state-of-the-art MCMC methods for density estimation. The logistic Gaussian process with Laplace's method also avoids the sampling and convergence assessment problems related to the highly multimodal posterior distributions of the mixture models.

Density estimation with LA-LGP and 400 grid points takes one to two seconds, which in interactive use is a reasonable waiting time. For a finer grid or more dimensions, more grid points could be placed, but this requires additional approximations. In this paper, we have considered a reduced-rank approximation for LA-LGP that avoids the infamous cubic scaling of the basic Gaussian process computation. In addition, we have demonstrated the suitability of the Laplace approach for estimating conditional
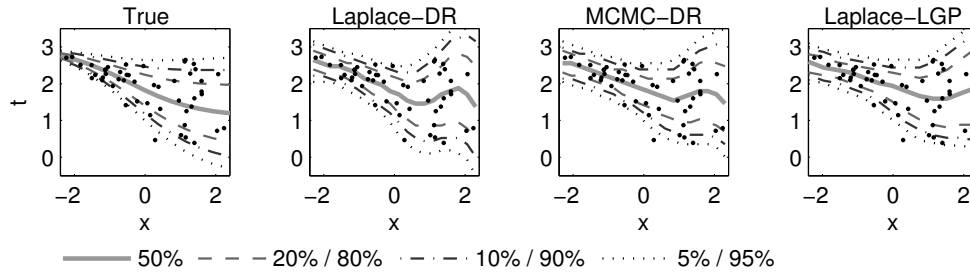
Figure 10: An illustration of density regression with one simulated data set. Dots represent a single realisation of $n = 50$ samples from the simulated density. The first plot shows the percentiles of the true conditional density. The second and third plots show the estimates with the Laplace approximation (LA-DR) and with the MCMC sampling (MCMC-DR). The fourth plot shows the conditional density estimate computed from a 2D density estimate with LA-LGP.

densities with one predictor variable.

Instead of Laplace's method, other deterministic approximations could be applied to speed up the posterior inference compared to MCMC. Variational approximations, including variational bounding and factorized variational approximations have been considered for GP models. Although these are close to LA in speed, and can be more accurate than LA with suitable hyperparameter settings, they can also have problems in estimating the hyperparameters (see, e.g., the comparisons between various Gaussian approximations in GP classification problems by Nickisch and Rasmussen 2008). Expectation propagation (EP) has been shown to perform better than the Laplace or variational approximations for binary classification (Nickisch and Rasmussen 2008) and for Student-t regression (Jylänki et al. 2011). However, for the Poisson model EP was only slightly better than the Laplace approximation (Vanhatalo et al. 2010). Because the Laplace approximation for the logistic Gaussian process performed almost as well as the MCMC approximation, we believe that EP could only slightly improve the performance of LA. The implementation of expectation propagation for non-diagonal W is non-trivial. A quadrature-free moment matching for EP could also be considered for density estimation, in a similar way as was done for multiclass GP classification (Riihimäki et al. 2013), but in our preliminary testing the moment matching of the distributions turned out to be quite slow.

The Gaussian approximations can be improved by considering corrections for the marginal posterior distributions (Rue et al. 2009; Cseke and Heskes 2011). These corrections can be challenging for the LGP model because the likelihood function cannot be factorized into terms depending only on a single latent value. In the future, it would be interesting to see whether similar corrections as considered by Rue et al. (2009), could be extended for LA-LGP where each likelihood term depends on multiple latent values.

The code for LA-LGP, MCMC-LGP and violin plots are available as part of the free

GPstuff toolbox for Matlab and Octave (http://becs.aalto.fi/en/research/bayes/gpstuff/ or http://mloss.org/software/view/451/).

References

Adams, R. P. (2009). "Kernel Methods for Nonparametric Bayesian Inference of Probabilities and Point Processes." Ph.D. thesis, University of Cambridge. 436, 437

Adams, R. P., Murray, I., and MacKay, D. J. (2009). "The Gaussian process density sampler." In *Advances in Neural Information Processing Systems 21*, 9–16. NIPS foundation. 426

Brooks, S. P. and Gelman, A. (1998). "General Methods for Monitoring Convergence of Iterative Simulations." *Journal of Computational and Graphical Statistics*, 7(4): 434–455. 436

Cseke, B. and Heskes, T. (2011). "Approximate marginals in latent Gaussian models." *Journal of Machine Learning Research*, 12: 417–454. 444

Cunningham, J. P., Shenoy, K. V., and Sahani, M. (2008). "Fast Gaussian process methods for point process intensity estimation." In McCallum, A. and Roweis, S. (eds.), *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, 192–199. Omnipress. 430, 431

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1(3): 515–533. 432

Geweke, J. (1989). "Bayesian Inference in Econometric Models Using Monte Carlo Integration." *Econometrica*, 57(6): 1317–1339. 432

Geyer, C. J. (1992). "Practical Markov chain Monte Carlo." *Statistical Science*, 7(4): 473–483. 436

Griebel, M. and Hegland, M. (2010). "A Finite Element Method for Density Estimation with Gaussian Process Priors." *SIAM Journal of Numerical Analysis*, 47(6): 4759–4792. 425

Griffin, J. (2010). "Default priors for density estimation with mixture models." *Bayesian Analysis*, 5(1): 45–64. 427, 436, 438

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag. 431, 433, 434

Hoffmann, M. and Gelman, A. (2013). "The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research*. In press. 435

Jylänki, P., Vanhatalo, J., and Vehtari, A. (2011). "Robust Gaussian Process Regression with a Student-t Likelihood." *Journal of Machine Learning Research*, 12: 3227–3257. 444

Kong, A., Liu, J. S., and Wong, W. H. (1994). "Sequential Imputations and Bayesian Missing Data Problems." *Journal of the American Statistical Association*, 89(425): 278–288. 432

Kundu, S. and Dunson, D. B. (2011). "Latent Factor Models for Density Estimation." ArXiv:1108.2720v2. 443

Lenk, P. J. (1991). "Towards a Practicable Bayesian Nonparametric Density Estimator." *Biometrika*, 78(3): 531–543. 426

— (2003). "Bayesian Semiparametric Density Estimation and Model Verification Using a Logistic-Gaussian Process." *Journal of Computational and Graphical Statistics*, 12(3): 548–565. 426

Leonard, T. (1978). "Density Estimation, Stochastic Processes, and Prior Information." *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(2): 113–146. 425, 426, 427

Neal, R. (1998). "Regression and Classification Using Gaussian Process Priors." In Bernardo, J. M., Berger, J. O., David, A. P., and Smith, A. P. M. (eds.), *Bayesian Statistics 6*, 475–501. Oxford University Press. 435

Nickisch, H. and Rasmussen, C. E. (2008). "Approximations for Binary Gaussian Process Classification." *Journal of Machine Learning Research*, 9: 2035–2078. 444

O'Hagan, A. (1978). "Curve fitting and optimal design for prediction." *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1): 1–42. 428

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). "A Unifying View of Sparse Approximate Gaussian Process Regression." *Journal of Machine Learning Research*, 6(3): 1939–1959. 434

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge: MIT Press. 428, 430, 431, 432, 433

Riihimäki, J., Jylänki, P., and Vehtari, A. (2013). "Nested Expectation Propagation for Gaussian Process Classification with a Multinomial Probit Likelihood." *Journal of Machine Learning Research*, 14: 75–109. 444

Rubin, D. B. (1981). "The Bayesian Bootstrap." *Annals of Statistics*, 9(1): 130–134. 439

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(2): 319–392. 432, 444

Snelson, E. and Ghahramani, Z. (2006). "Sparse Gaussian Processes Using Pseudo-Inputs." In Weiss, Y., Schlkopf, B., and Platt, J. (eds.), *Advances in Neural Information Processing Systems 18*. The MIT Press. 433, 434

Thorburn, D. (1986). "A Bayesian Approach to Density Estimation." *Biometrika*, 73(1): 65–75. 426

Tokdar, S. T. (2007). "Towards a Faster Implementation of Density Estimation With Logistic Gaussian Process Priors." *Journal of Computational and Graphical Statistics*, 16(3): 633–655. 425, 436, 437

Tokdar, S. T. and Ghosh, J. K. (2007). "Posterior consistency of logistic Gaussian process priors in density estimation." *Journal of Statistical Planning and Inference*, 137(1): 34–42. 425

Tokdar, S. T., Zhu, Y. M., and Ghosh, J. K. (2010). "Bayesian Density Regression with Logistic Gaussian Process and Subspace Projection." *Bayesian Analysis*, 5(2): 319–344. 435

van der Vaart, A. W. and van Zanten, J. H. (2009). "Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth." *The Annals of Statistics*, 37(5B): 2655–2675. 425

Vanhatalo, J., Pietiläinen, V., and Vehtari, A. (2010). "Approximate inference for disease mapping with sparse Gaussian processes." *Statistics in Medicine*, 29(15): 1580–1607. 434, 444

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. (2013). "GPstuff: Bayesian Modeling with Gaussian Processes." *Journal of Machine Learning Research*, 14: 1175–1179. 436

Vehtari, A. and Lampinen, J. (2002). "Bayesian Model Assessment and Comparison Using Cross-Validation Predictive Densities." *Neural Computation*, 14(10): 2439–2468. 439

Villani, M. and Larsson, R. (2006). "The Multivariate Split Normal Distribution and Asymmetric Principal Components Analysis." *Communications in Statistics: Theory & Methods*, 35(6): 1123–1140. 432

Williams, C. K. I. and Barber, D. (1998). "Bayesian classification with Gaussian processes." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12): 1342–1351. 430

Acknowledgments