Proceedings of the International Conference on Statistics: Theory and Applications (ICSTA'25)
Paris, France – August 17 - 19, 2025
Paper No. 152
DOI: 10.11159/icsta25.152

A Comparative Analysis of Deep Gaussian Processes and Multivariate Bayesian Spline-Based Methods for Simulating Multidimensional Surfaces

Callum Macaulay, Dirk Husmeier, Vinny Davies
University of Glasgow – Glasgow, Scotland
Callum.Macaulay@glasgow.ac.uk; Dirk.Husmeier@glasgow.ac.uk; Vinny.Davies@glasgow.ac.uk

Abstract - Statistical surrogates facilitate inference of a complex system when direct observation would disturb its underlying processes or when modelling the entire system is prohibitively expensive. Gaussian processes (GPs) are a common non-parametric surrogate model, with fast predictions and powerful uncertainty quantification. However, they struggle to capture surfaces that violate its standard assumptions. Deep Gaussian processes (DGPs) relax GP assumptions through a hierarchical structure which introduces model flexibility by warping and rotating the input space between GP layers, improving robustness and predictive accuracy for challenging surface dynamics. Tuning hyperparameters for multiple GP layers requires a higher computational complexity scaling with sample size. This can be reduced through the Vecchia sparse covariance approximation facilitating large training samples, while reducing predictive accuracy. This paper evaluates the computational efficiency and prediction accuracy of DGPs for simulating two smooth multidimensional surfaces against Bayesian multivariate spline-based methods, a common alternative surface estimation approach. In smaller samples, DGPs outperformed the comparison methods in prediction accuracy for both target functions. However, with increasing sample size necessitating the Vecchia approximation, the relative predictive advantage of the DGPs over competitors deteriorated, collapsing for the largest sample. This comparative reduction in performance with the Vecchia approximation appears to be caused by poor mixing of latent hyperparameters. The DGP computation time was approximately 10 times longer than the comparison methods, with a smaller MCMC effective sample size. To conclude, DGP predictions outperformed competing Bayesian spline-based methods in smaller samples, but offered no predictive advantage in larger samples that necessitated the Vecchia approximation.

Keywords: Deep Gaussian Process, Bayesian Splines

1. Introduction

The goal of measuring a complex system to predict its behavior at an unknown position or future state is often hindered in situations where directly observing the system would disrupt its underlying processes or when a full simulation of the system is prohibitively expensive. In such situations, employing a statistical computationally efficient, low-dimensional surrogate to approximate the system's behavior via training input-output mappings offers a cheaper modelling approach to prediction and uncertainty quantification. For example, a simplified cardiac surrogate model inferred from noninvasive imaging via cardiac magnetic resonance facilitates a low-resolution approximation to simulate the biomechanical system associated with myocardial infarction for prognosis and risk quantification [1], [2]. Such surrogates can be pre-trained using representative simulated data, allowing faster inference upon a patient's arrival at a clinic [3]. Accurate pathology identification requires a model with multiple biomarkers [4], thus necessitating flexible multivariate methods such as Deep Gaussian processes (DGPs) and multivariate spline-based methods to estimate the resultant complex multidimensional surfaces.

Gaussian Processes (GPs) are non-parametric multivariate models commonly used as low-cost surrogates for classification, regression and optimization problems in clinical settings when uncertainty quantification is desired [5], [6]. They provide accurate predictions for well-behaved data; however, real-life systems often violate the standard GP assumptions of stationarity and isotropic covariance and smoothness. Separable GPs avoid the isotropy assumption by partitioning the covariance kernel into a product of separate covariance kernels for each input dimension, allowing anisotropic modelling [7], [8].

Deep Gaussian processes (DGPs) further relax GP assumptions through a hierarchical structure, wherein an unobserved latent inner GP layer feeds an output into an observed outer GP layer. The structure introduces flexibility by facilitating non-linear warping and rotation of the original input space while a latent separable GP layer enables anisotropic modelling [9], [10], improving robustness and predictive accuracy for challenging surface dynamics that violate standard GP assumptions [11]. The flexibility from an additional GP layer necessitates tuning more hyperparameters which may be estimated through Markov chain Monte Carlo (MCMC), requiring many matrix inversions whose computational complexity scales with sample size, posing a computational challenge for large datasets. Computational requirements can be reduced through the Vecchia sparse covariance matrix approximation, which restricts the covariance associations to local data points, affording faster computation and facilitating modelling large samples while reducing predictive accuracy [10]. To evaluate whether DGPs are worth their computational cost, and, furthermore, how the Vechhia approximation limits prediction accuracy, their performance must be assessed against competing approaches such as multivariate spline-based methods.

Bayesian multivariate spline-based methods are common semi-parametric approaches to model complex surfaces [12]. They approximate a target surface through tensor products of piecewise polynomials defined over local intervals. The multivariate spline coefficients are estimated from observations given a prior distribution, and posterior distributions allow for greater quantification. Bayesian multivariate spline-based methods can model challenging surfaces through flexible implementations such as allowing smoothing parameters to vary across input dimensions to facilitate modelling anisotropic surface estimates. The flexibility and powerful predictions of multivariate spline-based methods make them a good comparative approach to evaluate the effectiveness of DGPs.

Previously, Bach and Klein (2022) [12] evaluated modern Bayesian multivariate spline-based methods. This paper extends their work by evaluating the predictive accuracy and computational efficiency of DGPs compared to multivariate spline-based methods. The accuracy and efficiency for estimating four-dimensional isotropic and anisotropic smooth surfaces of DGPs is compared to standard GPs, separable GPs and Bayesian multivariate spline-based methods across various sample sizes.

2. Methods

The methods considered are trained on an observation dataset $\mathcal{D} = \{\mathbf{X},\mathbf{y}\}$ of inputs $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}, \mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})$, and noisy outputs $\mathbf{y} = (y_1, ..., y_N)$ for sample size $i = 1, ..., N$. The target function $\mathbf{y} = f(\mathbf{X})$ is estimated as the predictive mean function $\hat{\mathbf{y}} = f(\mathbf{X})$.

2.1. Gaussian Processes

A Gaussian process (GP) defines an infinite set of jointly multivariate Gaussian random variables. GPs are non-parametric probabilistic machine learning models trained on $\mathcal{D}$ with multivariate Gaussian $\mathbf{y}$ generating a predictive mean function $\hat{\mathbf{y}} = f_{GP}(\mathbf{X}) \sim \mathcal{N}(m(\mathbf{X}), \mathbf{K} + \sigma_n^2\mathbf{I})$ for prediction and uncertainty quantification in regression and classification [8], [13]. Here, the mean is assumed to be $m(\mathbf{X}) = \mathbf{0}$ for simplicity. In the Bayesian framework, the GP prior over functions is multivariate Gaussian and prior belief about covariance structure and smoothness is encoded through a kernel function, $\mathbf{K} = k(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})$, given hyperparameters $\boldsymbol{\theta}$, which generates a positive semi-definite covariance matrix. When $f(\mathbf{X})$ is expected to be smooth, the squared exponential kernel,

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\ell^2}\right),$$

is a common choice. This has three hyperparameters, $\boldsymbol{\theta} = (\ell, \sigma_f^2, \sigma_n^2)$, lengthscale, ℓ, models the decay in correlation between points as distance increases, controlling the smoothness versus ‘wiggliness’ of GP random variables. A constant lengthscale imposes the assumptions of isotropic correlation and stationarity. Signal variance, σ_f^2, models variance within $f_{GP}(\mathbf{X})$, controlling GP random variables amplitude and noise variance, σ_n^2, models iid Gaussian additive noise within the data. $\boldsymbol{\theta}$ is typically learned from the data by minimizing the negative marginal likelihood function

$$-\log p(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X}) = \frac{1}{2}\mathbf{y}^\top(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K} + \sigma_n^2\mathbf{I}| + \frac{N}{2}\log(2\pi).$$

This closed-form expression requires $(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}$, with computational complexity required for inversions growing in $\mathcal{O}(N^3)$, which presents challenges for large samples. This computational demand can be reduced by the Vecchia approximation for the covariance matrix [14]. The Vecchia approximation requires, possibly randomly, ordered data. It conditions $\mathbf{x}_i$ only on the $m << n$ nearest neighbours earlier in the ordering. This induces a sparse covariance structure of conditionally independent non-neighbour pairs, reducing the complexity of inversions to $\mathcal{O}(Nm^3)$ [10]. The upper-lower Cholesky factorisation of $\mathbf{K}$ is used to estimate $\hat{y}_i = f_{GP}(\mathbf{x}_i)$ for more sparsity [15]. Sparse covariance approximations facilitating larger samples [14], [16] often provide reasonable mean estimates, but poorer uncertainty quantification from lost covariance information [17].

2.2. Separable Gaussian Processes

Separable GPs avoid the isotropy assumption by partitioning the input space into up to $p = dim(\mathbf{X}) \geq 2$, with separate hyperparameters $\boldsymbol{\theta}_d$ in each dimension [8]. The covariance kernel is a product of kernels for the p segments:

$$k(\mathbf{x}, \mathbf{x}') = \prod_{d=1}^{p} k_d(\mathbf{x}, \mathbf{x}')$$

where the d-th kernel, $k_d(\mathbf{x}, \mathbf{x}')$, depends only on the inputs $(\mathbf{x}, \mathbf{x}')$ of the subset d. Estimating a larger number of hyperparameters requires greater computation, but is not prohibitive compared to similar methods [8]. Separable GPs circumvent the isotropy assumption; however, their implementation is inflexible for target surfaces that exhibit complex dynamics.

2.3. Deep Gaussian Processes

Deep Gaussian Processes (DGPs) extend 1-layer GPs through a hierarchical structure of GP input-output mapping [11]. A two-layer DGP feeds an input through an unobserved latent inner separable GP layer $f_{GP_W}(\mathbf{X})$, whose outputs then feed into the observed outer layer $f_{GP_Y}(\mathbf{W})$ with an output equivalent to a standard GP:

$$\mathbf{W} = f_{GP_W}(\mathbf{X}) \sim \mathcal{N}(\mathbf{0}, K_{\boldsymbol{\phi}}(\mathbf{X}, \mathbf{X})) \qquad f_{GP_Y}(\mathbf{W}) \sim \mathcal{N}(\mathbf{0}, K_{\boldsymbol{\theta}}(\mathbf{W}, \mathbf{W}) + \sigma_n^2 \mathbf{I})$$

where $\mathbf{W}$ is an $n \times p$ matrix whose columns correspond to the p input dimensions of the latent space while $K_{\boldsymbol{\phi}}$ and $K_{\boldsymbol{\theta}}$ are the inner and outer kernels with hyperparameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ respectively. The separable latent layer facilitates anisotropic modeling while performing non-linear transformations warping and rotating $\mathbf{X}$, allowing DGPs to flexibly model challenging non-stationary dynamics [18].

The posterior predictive distribution can be taken by marginalising over the inner layer. However, this tends to be analytically intractable; therefore, MCMC is usually used to approximate the posterior distribution. For example, the approach we consider [10], [19] leverages a routine of elliptical slice sampling [20] of the latent layers and the Metropolis-Hastings algorithm within a Gibbs sampling routine to approximate the posterior hyperparameters. This requires many matrix inversions, posing computational challenges for even moderately sized training samples; thus, the Vecchia approximation may be implemented similarly to 1-layer GPs [9]. Due to their flexibility and efficiency through the Vecchia approximation, DGPs have demonstrated good predictive accuracy when modelling multidimensional surfaces that violate the standard GP assumptions [11]. To evaluate DGP performance in practice, we must consider their predictive accuracy and computational efficiency compared to competing methods such as commonly used Bayesian multivariate spline-based methods.

2.4. Bayesian Multivariate Spline-Based Methods

Bayesian multivariate spline-based methods estimate $f(\mathbf{X})$ through tensor products of piecewise polynomials defined over local intervals which can model the shape and smoothness. In the Bayesian perspective, given a multivariate spline coefficient prior and observations, the resultant posterior distribution over potential multivariate splines provides uncertainty quantification. The predictions are taken as an average of the potential multivariate splines, weighted by their posterior probability. We consider a series of multivariate spline-based methods, given in bold, initially compared by Bach and Klein (2022), who developed a novel adaptive Metropolis-Hastings (MH) algorithm to update the smoothing coefficient sampler. The methodology is tailored to modelling anisotropic smooth multidimensional surfaces through one-dimensional multivariate splines, allowing smoothness to vary across dimensions. The smoothing coefficient prior used are inverse-Gamma (**BK-IG**), Weibull (**BK-WB**) and Weibull with prior scaling (**BK-WB-PS**) then estimates its posterior distribution through adaptive MH.

These models were evaluated against the following competing Bayesian multivariate spline-based methods, which may be limited in modelling anisotropic, high-dimensional, or large datasets. **Jagam** [21] uses the R package `mgcv` [22] for mixed generalised additive models with automatic smoothness estimation with the MCMC sampler `JAGS` [23], which is effective for two-dimensional smoothing, but has prohibitive computation costs in higher dimensions. The R package **Rstanarm** [24] provides a similar implementation with an alternative roughness penalty that also lacks predictive accuracy and is computationally inefficient when $p > 2$. **Bamlss** [25] uses slice sampling with a stepping out procedure [26] to estimate smoothing coefficients for each dimension; however, it is similarly impractically slow when $p > 2$. An implementation in the R package **BayesX** [27], [28] models anisotropy by dividing the training data into two isotropic groups. Although it is faster than the other methods when $p > 2$, it cannot capture higher-dimensional anisotropic surface dynamics. The **BK-WB-PS** approach tended to outperform the prediction accuracy and computational efficiency of the comparative methods and therefore sets the performance benchmark for comparison with DGPs.

3. Simulation

3.1. Isotropic and Anisotropic Target Functions

We follow Bach and Klein's (2022) procedure, by simulating $\mathbf{y} \sim f_g(\mathbf{x}_i), g = \{1, 2\}$ from training inputs, sampled uniformly in $\mathbf{x}_i \in [0, 1]^3$ and $\mathbf{y}$ with variance $\sigma^2 = 0.5^2$ for 10 replicates in $N = \{100, 250, 500, 1000\}$ as the sample size alters the interpretation of signal-to-noise [29]. An isotropic target function $f_1(\mathbf{x}_i) = \sin(2\pi||\mathbf{x}_i||_2)$, where $||\mathbf{x}_i||_2$ is the L2 norm euclidean distance between the i-th observation, $i = \{1, ..., N\}$, and an anisotropic target function $f_2(\mathbf{x}_i) = \sin(2\pi\sqrt{3x_1 + x_2^2 + x_3^3/3})$ were estimated using standard/separable GPs, DGP and multivariate spline-based methods. The prediction accuracy was measured as the log-mean squared error, $\text{logMSE} = \log\left(\frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2\right)$, $k = \{1, .., 1000\}$, between the prediction $\hat{y}_k$ and the true value $y_k = f_g(\mathbf{x}_k)$.

3.2. Model Specifications

Using the R package deepGP [19], the standard/separable GPs and DGP's were fitted with a squared exponential kernel, as $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are smooth. Models implemented without the Vecchia approximation are referred to as "full GPs/DGPs", and those implemented with the Vecchia approximation are "approx GPs/DGPs". Full standard/separable GPs were implemented for $N = \{100, 250\}$; approx standard/separable GPs were implemented for $N = \{500, 1000\}$. Full DGPs were implemented for $N = \{100, 250\}$ and approx DGPs were implemented for $N = \{100, 250, 500, 1000\}$ to assess the impact of the Vecchia approximation on DGPs in low samples. The approx GPs and DGPs were implemented with $m = 10$ the default in deepGP. However, this resulted in logMSE close to 0 for f_2 at $N = 1000$, so these replicates were rerun using $m = 20$ but with little to no improvement and presented.

The posterior distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ were approximated three times for each replicate to assess convergence, with initialisations $\sigma_n^2 = 0.001$, $\ell = 0.1$ for standard kernels, $\boldsymbol{\ell} = (0.1, 0.1, 0.1)$ for separable kernels. MCMC chains ran between 40,000 and 200,000 iterations depending on convergence diagnostics. The between-chain convergence was assessed using Gelman-Rubin $\hat{R}$ [30] and the within-chain stationarity using Geweke [31] and Heidelberg diagnostics [32]. MCMC runs were stopped when diagnostics indicated satisfactory convergence and stationarity. However, approx DGPs tended to indicate non-convergence of inner lengthscales with convergence for outer lengthscale and noise. In these cases, MCMC runs were continued for 10,000 iteration chunks and convergence was assessed. This was repeated until either the total computation time reached approximately 10 hours or 200,000 iterations. For stationary chains, thinning was determined by the lag at which the autocorrelation function (ACF) was non-significant. The MCMC chain was then burned-in to leave an inner lengthscale effective sample size (ESS) of 100. However, for poorly mixed chains, a subjective decision, informed by the ACF and convergence diagnostics, about burn-in and thinning was made to balance stationarity and ESS of the inner lengthscale. Since individual models did not converge satisfactorily, predictions for each replicate were taken from all three instantiations and the average logMSE per replicate was calculated. The deepGP package takes a point estimate, $f_{GP}(\mathbf{x}^*)$, at an unknown location, $\mathbf{x}^*$, as the mean of the predictions for each draw of the posterior distribution.

4. Results

4.1. Prediction Accuracy

Figure 1 shows that full DGPs tended to outperform other methods. In contrast, approx DGPs tended not to offer a predictive advantage over the best multivariate spline-based methods, with accuracy collapsing for f_2 at $N = 1000$. Full standard and separable GPs tended to perform similarly to the best multivariate spline-based methods, while approx standard and separable GPs were less accurate. Assessing the Vecchia approximation, Welch's t-tests [33] for the logMSE of full and approx DGPs in Table 1 shows full DGPs have significantly lower mean logMSE for $N = 250$ after Bonferroni correction.

Table 1: 95% confidence intervals for the difference in mean logMSE of full and approximate DGPs for f_1 and f_2 at $N = \{100, 250\}$ and corresponding p-values.

f_1			f_2		
N	95% CI	p-value	N	95% CI	p-value
100	(-0.85, -0.02)	0.04	100	(-0.37, 0.10)	0.24
250	(-0.68, -0.20)	< 0.01	250	(-0.88, -0.28)	< 0.01

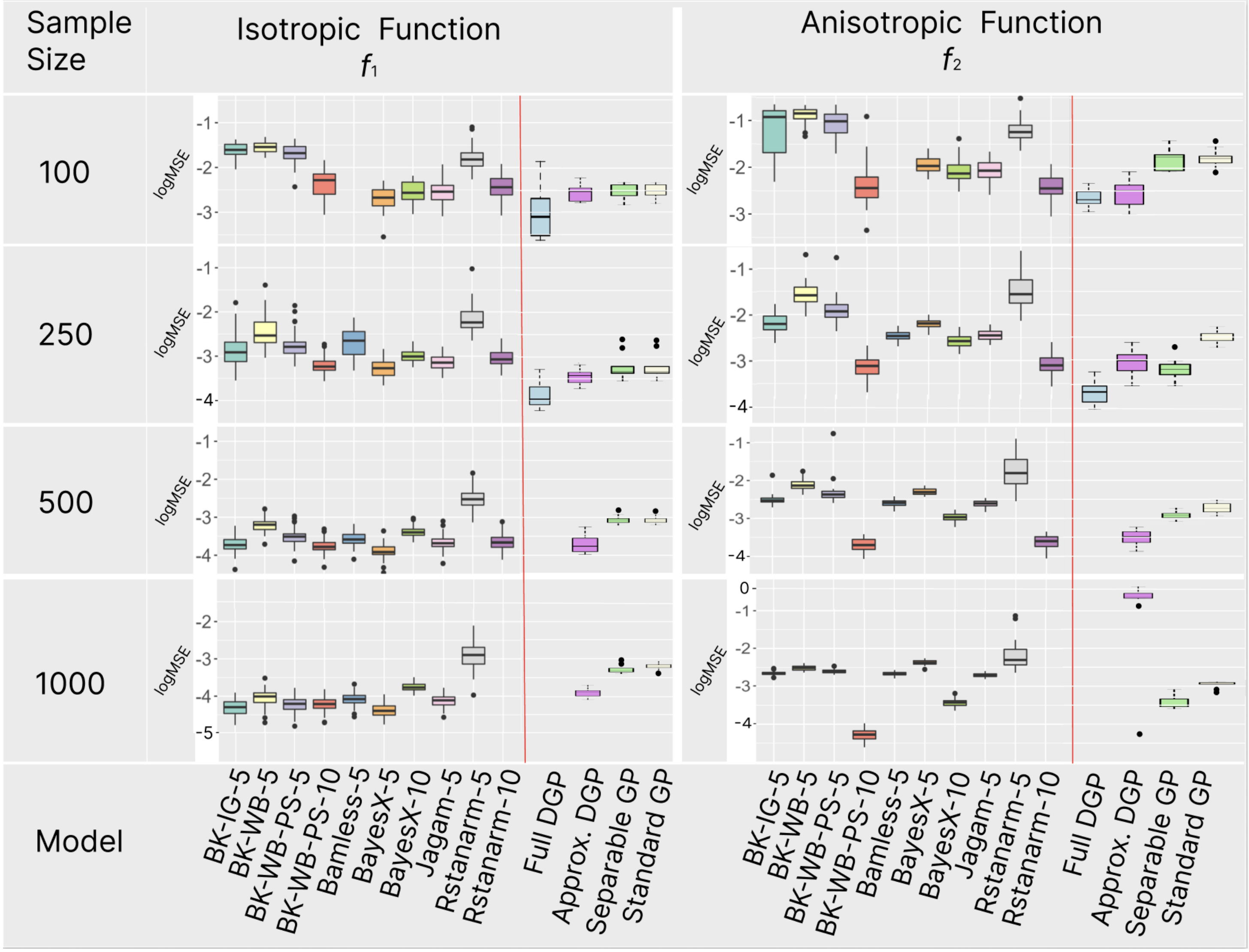

Fig. 1: The logMSE plot for f_1 and f_2 taken directly from Bach and Klein (2022) is altered to include the full and approx standard/separable GP, and DGP results, separated by a red line. The multivariate spline-based methods were implemented using $\{5, 10\}$ dimension marginal basis multivariate splines, with the suffixes **-5** and **-10** respectively, except for **Bamlss** and **Jagam** due to excessive runtime for 10-dimension marginal basis multivariate splines. Full and approx standard/separable GP results are plotted in the same column for parsimony, $N = \{100, 250\}$ are full and $N = \{500, 1000\}$ are approx. The boxplots of GPs/DGPs denote a distribution of mean logMSE across 10 replicates calculated as described in Section 3.2.

4.2. Computational Efficiency

The BK-WB-PS-10 approach tends to outperform the computational efficiency of other multivariate spline-based methods, with better predictions. Therefore, it is the performance benchmark. The posterior ESS and system computation time for 1,000 MCMC iterations of full and approx standard/separable GPs and DGPs, and BK-WB-PS-10 have been evaluated. The models were trained on inputs sampled uniformly in $\mathbf{x}_i \in [0, 1]^3$ and outputs $\mathbf{y} \sim f_g(\mathbf{x}_i), g = \{1, 2\}$, with $\sigma^2 = 0.5^2$ and $N = 250$, using R version 4.3.2 [34] on an Apple Mac laptop running on macOS Sequoia, Version 15.1, Build 24B83 with an Apple M1 Pro chip and 8 CPU cores, 14 GPU cores and 16GB RAM [35]. The difference in system time immediately before and after the computation, and the mean ESS are presented in Table 2.

The full standard/separable GPs, approx DGPs and BK-WB-PS-10 computation times have the same order of magnitude,

Table 2: The system computation time (seconds) for 1000 MCMC iterations and resulting posterior distribution effective sample size, averaged over the input dimension, of the primary interest hyperparamers: lengthscale for standard/separable GPs, inner lengthscale for DGPs and smoothing coefficient for BK-WB-PS-10.

Model	**Mean Effective Sample Size**		**Computation Time (s)**	
	f_1	f_2	f_1	f_2
Full GP	69	61	25	25
Approx GP	41	63	3	3
Full Separable GP	111	92	51	51
Approx Separable GP	87	61	5	5
Full DGP	39	22	418	419
Approx DGP	13	17	40	42
BK-WB-PS-10	1000	918	47	50

with approx standard/separable GPs approximately 1/10*th* smaller and full DGPs approximately 10 times larger. Inspecting ESS, BK-WB-PS-10 has near perfect sampling efficiency, with ESS an order of magnitude greater than the next most efficient method, the full separable GP, and two orders of magnitude greater than the least efficient method, the approx DGP. Trace plots and posterior histograms were seen to indicate that the BK-WB-PS-10 sampler reached stationarity within 1000 iterations, while all GP/DGP methods indicated non-stationarity. These metrics show that BK-WB-PS-10 is the most efficient approach, requiring fewer MCMC iterations and less computation time than the GP/DGP methods to map out the posterior distribution. Additionally, full GPs/DGPs have more efficient sampling, but require substantially longer computation.

5. Discussion

We compared predictive accuracy and computational efficiency of DGPs and Bayesian spline-based methods [12]. Evaluating accuracy finds that full DGPs can surpass the predictive performance of standard and separable GPs and spline-based methods, while requiring substantially higher computational investment. That DGP performance was greater than standard and separable GPs for both f_1 and f_2, indicates that DGP's warping and rotating inputs between layers enhanced flexibility beyond simply an anisotropic covariance structure. Greater variability in DGP prediction accuracy compared to standard/separable GP may be linked to the hierarchical structure, as the prediction error in the inner layer is amplified in the outer layer. Additionally, the increased flexibility may impact how random noise influences the number of iterations required for MCMC converge [36]. Full DGPs were powerful in small samples, but the high computational cost prohibits their applicability for large samples, which may limit their use as a low-cost surrogate in clinical settings requiring fast predictions.

In larger samples, approx DGPs showed little to no advantage over the best multivariate spline-based methods while slightly outperforming standard/separable GPs. The relative accuracy of approx DGPs degraded with increasing sample size and collapsed for the largest anisotropic sample, where BK-WB-PS-10 offered a more accurate and cheaper approach. This work aligns with previous conclusions that DGPs offer a powerful approach to simulating smooth surfaces [37], [38] while the Vecchia approximation facilitates larger sample simulations but restricts predictive accuracy [9].

Compared to full DGPs, in smaller samples, approx GP/DGPs' exhibited poorer MCMC mixing with inner lengthscales rarely converging, whereas noise and outer lengthscale indicated no non-convergence within low 10,000's of iterations, suggesting that the sparce covariance contributes towards non-convergence. This was exacerbated in larger samples with increasingly multimodal posterior distributions, resulting in the sampler getting stuck in suboptimal modes or transient behavior between local optima. Poor inner hyperparameter convergence is a persistent problem throughout other hierarchical machine learning methods, such as neural networks [39], [40]. A deep analysis to assess long-term convergence issues would require longer MCMC runs (e.g., 1,000,000 iterations), though this may push beyond practicality.

Future work to adapt the MCMC sampler may provide improvements. For example, introducing a simulated annealing mechanism into the sampler may promote more efficient convergence to a posterior global optimum without exploring low likelihood regions for larger datasets [41]. Potential improvements may also be found by using a hierarchical Vecchia approximation for covariance [42]. This approach is tailored for larger samples and extends the Vecchia approximation by retaining some global covariance structure of a few far away data points in the sparse matrix. This was successfully applied to standard GPs [42] and may improve the predictions and convergence of DGPs. Additional avenues to improve the DGPs worth consideration may be inspired by nearest-neighbour GPs [43], [44] and the Vecchia-Laplace approximation for GPs

[45], which have shown promise in 1-layer GPs but they are yet to be applied to DGPs.

The flexibility to warp and rotate the training data between hierarchical GP layers afforded DGPs a predictive advantage over standard and separable GPs and Bayesian multivariate spline-based methods in small samples. However, the considerably greater computational requirements necessitating the Vecchia approximation for moderate to large sample sizes limited DGPs usefulness as Bayesian multivariate spline-based methods provided cheaper, more accurate predictions. Future work to adapt the posterior sampling mechanism or alter the sparse covariance Vecchia approximation may offer progress.

References

[1] H. Gao, W. G. Li, L. Cai, C. Berry, and X. Y. Luo, "Parameter estimation in a Holzapfel–Ogden law for healthy myocardium," *Journal of engineering mathematics*, vol. 95, pp. 231–248, 2015.

[2] M. Genet, L. C. Lee, and R. Nguyen, "Distribution of normal human left ventricular myofiber stress at end diastole and end systole: a target for in silico design of heart failure treatments," *Journal of applied physiology*, vol. 117, no. 2, pp. 142–152, 2014.

[3] V. Davies, U. Noè, and A. Lazarus, "Fast parameter inference in a biomechanical model of the left ventricle by using statistical emulation," *Journal of the Royal Statistical Society: Applied Statistics*, vol. 68, no. 5, pp. 1555–1576, 2019.

[4] H. Gao, A. Aderhold, K. Mangion, X. Luo, D. Husmeier, and C. Berry, "Changes and classification in myocardial contractile function in the left ventricle following acute myocardial infarction," *Journal of The Royal Society Interface*, vol. 14, p. 20 170 203, Jul. 2017. DOI: 10.1098/rsif.2017.0203.

[5] H. Gao, K. Mangion, D. Carrick, D. Husmeier, X. Luo, and C. Berry, "Estimating prognosis in patients with acute myocardial infarction using personalized computational heart models," *Scientific Reports*, vol. 7, Oct. 2017.

[6] H. Gao, D. Carrick, and C. Berry, "Dynamic finite-strain modelling of the human left ventricle in health and disease using an immersed boundary-finite element method," *IMA journal of applied mathematics*, vol. 79, pp. 978–1010, 2014.

[7] R. B. Gramacy, *laGP: Local Approximate Gaussian Process Regression (R package version 1.5)*, 2016.

[8] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. Cambridge, MA: MIT Press, 2006.

[9] A. E. Sauer, *Deep Gaussian Process Surrogates for Computer Experiments*, 2023.

[10] A. Sauer, A. Cooper, and R. B. Gramacy, "Vecchia-approximated deep Gaussian processes for computer experiments," *Journal of Computational and Graphical Statistics*, vol. 32, no. 3, pp. 824–837, 2023.

[11] A. Damianou and N. D. Lawrence, "Deep Gaussian Processes," in *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR, 2013, pp. 207–215.

[12] P. Bach and N. Klein, "Anisotropic multidimensional smoothing using Bayesian tensor product P-splines," *arXiv preprint arXiv:2211.16218*, 2022.

[13] D. J. MacKay, "Introduction to Gaussian processes," in *NATO ASI series F computer and systems sciences*, 1998.

[14] A. V. Vecchia, "Estimation and model identification for continuous spatial processes," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 50, no. 2, pp. 297–312, 1988.

[15] M. Katzfuss, J. Guinness, and E. Lawrence, "Scaled Vecchia approximation for fast computer-model emulation," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 10, no. 2, pp. 537–554, 2022.

[16] J. P. Bharadiya, "A review of Bayesian machine learning principles, methods, and applications," *International Journal of Innovative Science and Research Technology*, vol. 8, no. 5, pp. 2033–2038, 2023.

[17] A. Marrel and B. Iooss, "Probabilistic surrogate modeling by Gaussian process: A review on recent insights in estimation and validation," *Reliability Engineering & System Safety*, p. 110 094, 2024.

[18] S. D. Barnett, L. J. Beesley, A. S. Booth, R. B. Gramacy, and D. Osthus, "Monotonic warpings for additive and deep Gaussian processes," *arXiv preprint arXiv:2408.01540*, 2024.

[19] A. S. Booth, *deepgp: Bayesian Deep Gaussian Processes using MCMC (R package version 1.1.3)*, https://CRAN.R-project.org/package=deepgp, Accessed 2024-xx-xx, 2024.

[20] I. Murray, R. P. Adams, and D. J. C. MacKay, "Elliptical slice sampling," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR, vol. 9, 2010, pp. 541–548.

[21] S. N. Wood, “Just another Gibbs additive modeler: Interfacing JAGS and mgcv,” *Journal of Statistical Software*, vol. 75, no. 7, pp. 1–15, 2016.

[22] S. N. Wood, *mgcv: Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation*, 2012.

[23] M. Plummer, K. Hornik, F. Leisch, and A. Zeileis, Eds., *Proceedings of the 3rd international workshop on distributed statistical computing*, 2003.

[24] B. Goodrich, J. Gabry, I. Ali, and S. Brilleman, *rstanarm: Bayesian applied regression modeling via Stan (R package version 2.21.3)*, `https://mc-stan.org/rstanarm/`, 2022.

[25] N. Umlauf, N. Klein, and A. Zeileis, “bamlss: Bayesian additive models for location, scale, and shape (and beyond),” *Journal of Computational and Graphical Statistics*, vol. 27, no. 3, pp. 612–627, 2018.

[26] R. M. Neal, “Slice sampling,” *The Annals of Statistics*, vol. 31, no. 3, pp. 705–767, 2003.

[27] A. Brezger and S. Lang, “Generalized structured additive regression based on Bayesian P-splines,” *Computational Statistics & Data Analysis*, vol. 50, no. 4, pp. 967–991, 2006.

[28] T. Kneib, N. Klein, and S. Lang, “Modular regression—a lego system for building structured additive distributional regression models with tensor product interactions,” *TEST*, vol. 28, no. 1, pp. 1–39, 2019.

[29] T. G. Rudner, O. Key, Y. Gal, and T. Rainforth, “On signal-to-noise ratio issues in variational inference for deep Gaussian processes,” in *International Conference on Machine Learning (ICML)*, PMLR, 2021, pp. 9148–9156.

[30] A. Gelman and D. B. Rubin, “Inference from iterative simulation using multiple sequences,” *Statistical Science*, vol. 7, no. 4, pp. 457–472, 1992.

[31] J. Geweke, “Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments,” in *Bayesian Statistics 4*, J. Bernardo and J. Berger and A. Dawid and A. Smith, Ed., Oxford University Press, 1992, pp. 169–193.

[32] P. Heidelberger and P. D. Welch, “Simulation run length control in the presence of an initial transient,” *Operations Research*, vol. 31, no. 6, pp. 1109–1144, 1983.

[33] Z. L. Lu and K.-H. Yuan, “Welch’s t-test,” in *Encyclopedia of Research Design*, N. J. Salkind, Ed., Sage, 2010, pp. 1551–1556. DOI: `{10.13140/RG.2.1.3057.9607}`.

[34] R Core Team, *R: A Language and Environment for Statistical Computing*, `https://www.R-project.org/`, R Foundation for Statistical Computing, Vienna, Austria, 2023.

[35] Apple Inc., *macOS Sequoia version 15.1 [Computer software]*, `https://www.apple.com/macos/sequoia`, Retrieved from `https://www.apple.com/macos/sequoia`, Cupertino, CA, 2024.

[36] B. Franzke and B. Kosko, “Noise can speed convergence in Markov chains,” *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 84, no. 4, p. 041 112, 2011.

[37] A. S. Booth, S. A. Renganathan, and R. B. Gramacy, “Contour Location for Reliability in Airfoil Simulation Experiments using Deep Gaussian Processes,” *arXiv preprint arXiv:2308.04420*, 2023.

[38] F. Yazdi, D. Bingham, and D. Williamson, “Deep Gaussian Process Emulation and Uncertainty Quantification for Large Computer Experiments,” *arXiv preprint arXiv:2411.14690*, 2024.

[39] K. P. Murphy, *Machine learning: a probabilistic perspective*. Cambridge, MA: MIT Press, 2012.

[40] R. M. Neal, *Bayesian learning for neural networks*. Springer Science & Business Media, 2012, vol. 118.

[41] B. Wang, L. Yan, X. Duan, T. Yu, and H. Zhang, “An integrated surrogate model constructing method: Annealing combinable gaussian process,” *Information Sciences*, vol. 591, pp. 176–194, 2022.

[42] M. Jurek and M. Katzfuss, “Hierarchical sparse Cholesky decomposition with applications to high-dimensional spatio-temporal filtering,” *Statistics and Computing*, vol. 32, no. 1, p. 15, 2022.

[43] A. Datta, “Sparse nearest neighbor Cholesky matrices in spatial statistics,” *arXiv preprint arXiv:2102.13299*, 2021.

[44] I. Grenier and B. Sansó, “Distributed nearest-neighbor Gaussian processes,” *Communications in Statistics-Simulation and Computation*, vol. 52, no. 7, pp. 2886–2898, 2023.

[45] D. Zilber and M. Katzfuss, “Vecchia–Laplace approximations of generalized Gaussian processes for big non-Gaussian spatial data,” *Computational Statistics & Data Analysis*, vol. 153, p. 107 081, 2021.