Bayesian Analysis (2023) **TBA**, Number TBA, pp. 1–25

Warped Gradient-Enhanced Gaussian Process Surrogate Models for Exponential Family Likelihoods with Intractable Normalizing Constants*

Quan Vu[†,‡], Matthew T. Moores[‡], and Andrew Zammit-Mangion[‡]

Abstract. Markov chain Monte Carlo methods for exponential family models with intractable normalizing constant, such as the exchange algorithm, require simulations of the sufficient statistics at every iteration of the Markov chain, which often result in expensive computations. Surrogate models for the likelihood function have been developed to accelerate inference algorithms in this context. However, these surrogate models tend to be relatively inflexible, and often provide a poor approximation to the true likelihood function. In this article, we propose the use of a warped, gradient-enhanced, Gaussian process surrogate model for the likelihood function, which jointly models the sample means and variances of the sufficient statistics, and uses warping functions to capture covariance nonstationarity in the input parameter space. We show that both the consideration of nonstationarity and the inclusion of gradient information can be leveraged to obtain a surrogate model that outperforms the conventional stationary Gaussian process surrogate model when making inference, particularly in regions where the likelihood function exhibits a phase transition. We also show that the proposed surrogate model can be used to improve the effective sample size per unit time when embedded in exact inferential algorithms. The utility of our approach in speeding up inferential algorithms is demonstrated on simulated and real-world data.

Keywords: autologistic model, delayed-acceptance MCMC, exchange algorithm, hidden Potts model, importance sampling, nonstationarity.

1 Introduction

Methods for statistical inference usually require the likelihood function to be evaluated pointwise, up to an unknown normalizing constant. However, many important exponential family models have an intractable likelihood that cannot be evaluated, but that can be easily simulated from. In the case of the Potts model (Potts, 1952) used for image analysis, and the exponential random graph model (ERGM; Frank and Strauss,

*Quan Vu was supported by a University Postgraduate Award from the University of Wollongong, Australia. Andrew Zammit-Mangion's research was supported by an Australian Research Council (ARC) Discovery Early Career Research Award, DE180100203.

†Research School of Finance, Actuarial Studies and Statistics, Australian National University, Australia, quan.vu@anu.edu.au

‡School of Mathematics and Applied Statistics, University of Wollongong, Australia, mmoores@uow.edu.au, azm@uow.edu.au

 https://doi.org/10.1214/23-BA1400

1986) used for social network analysis, the likelihood function features a phase transition where, on a small region of the parameter space, the model behavior changes rapidly from one phase (known as the ordered phase) to another phase (known as the disordered phase). This property makes inference with these models even more challenging.

A growing body of literature is concerned with computational methods for inference with models that have intractable likelihoods. For example, pseudo-marginal methods (Beaumont, 2003; Andrieu and Roberts, 2009) make use of an unbiased estimate of the likelihood function, while in approximate Bayesian computation (ABC; Tavaré et al., 1997; Pritchard et al., 1999), summary statistics are used to compare simulated pseudo-data at given parameter values to observed data. One of the most popular approaches involves the use of Markov chain Monte Carlo (MCMC) with auxiliary variables. Two algorithms in this class include that introduced by Møller et al. (2006), and the exchange algorithm (Murray et al., 2006). These MCMC methods require simulation of pseudo-data from the likelihood at each iteration of the Markov chain. In practice, Gibbs or Swendsen–Wang (SW) algorithms (Swendsen and Wang, 1987) are used for simulating the sufficient statistics. Even these algorithms can be computationally expensive when the data dimension is large, rendering inference infeasible in many applications.

To make inference for these models computationally tractable, surrogate likelihoods are often employed to approximate the true likelihood function. Such methods can speed up inference, as they do not require expensive simulations of the sufficient statistics at every iteration. For example, Boland et al. (2018) used a deterministic function to emulate ratios of normalizing constants, while Moores et al. (2020) used a deterministic function to emulate the sufficient statistics. An attractive way for constructing surrogate models for the likelihood function is through Gaussian processes, which are flexible, probabilistic models. Gaussian process emulators were proposed in the context of computer experiments for modeling computationally expensive functions (e.g., Sacks et al., 1989; Kennedy and O'Hagan, 2000). Gaussian process emulators were subsequently used as surrogate models in approximate Bayesian computation by Meeds and Welling (2014); Wilkinson (2014); Järvenpää et al. (2018) and Järvenpää et al. (2021). Drovandi et al. (2018) and Park and Haran (2020) employed Gaussian process surrogate models for facilitating Bayesian computation in an MCMC context.

Typically, Gaussian process surrogate models are constrained to be stationary. However, if the likelihood function undergoes a phase transition, the sufficient statistics can abruptly change with small changes in the input parameters at the transition; this sudden change in behavior is synonymous with nonstationarity when modeling using stochastic processes. In Section 4.1 we show that using a stationary Gaussian process surrogate model for the sufficient statistics may lead to large errors when emulating said sufficient statistics. The surrogate model is often used directly, instead of the true likelihood function, in an MCMC algorithm, resulting in an inexact-approximate algorithm. In such cases, inferential accuracy heavily depends on the accuracy of the surrogate model.

In this paper we build a surrogate model to emulate the sufficient statistics for making computationally-efficient inference with exponential family models that have an intractable normalizing constant. However, to rectify the aforementioned problems,

we propose using a nonstationary Gaussian process to emulate the sufficient statistics, specifically one based on the warped Gaussian process model introduced by Zammit-Mangion et al. (2022) and Vu et al. (2022). The main contribution here, over these and related works that also consider nonstationary Gaussian process models (e.g. Järvenpää et al., 2018; Aushev et al., 2022) is the incorporation of gradient information in our model (see Laurent et al., 2019, for a review). This modification leads to a multivariate Gaussian process that *jointly* models the means and the variances of the sufficient statistics. We show that gradient-enhanced nonstationary Gaussian process surrogate models offer a large improvement over both univariate stationary, and nonstationary, Gaussian process models when emulating the sufficient statistics, particularly in the vicinity of phase transitions. We illustrate the use of our surrogate model in importance sampling (Everitt et al., 2017; Vihola et al., 2020) and delayed-acceptance MCMC (Christen and Fox, 2005; Sherlock et al., 2017), which target the exact posterior distribution over the parameters. We show that our proposed methodology may be used to good effect in both the complete-data setting and the incomplete-data setting.

The remainder of the article is organized as follows. In Section 2, we present our general approach for modeling the sufficient statistics to approximate the likelihood of exponential family models with intractable normalizing constant. In Section 3, we introduce the gradient-enhanced nonstationary Gaussian process surrogate models for the sufficient statistics, and detail the algorithms that make use of the surrogate models for inference. In Section 4, we demonstrate the use of the surrogate models on three data sets. Section 5 concludes.

2 Background

In this section, we define the models our approach is suitable for, and detail the likelihood function that we approximate with surrogate models in Section 3.

2.1 Intractable Likelihood

In this paper we consider models for which the likelihood function can be written in the following, exponential family, form,

$$p(\mathbf{z} \mid \boldsymbol{\beta}) = \frac{\exp\left\{\boldsymbol{\beta}^T \mathbf{s}(\mathbf{z})\right\}}{\mathcal{C}(\boldsymbol{\beta})}, \tag{1}$$

where the normalizing constant is given by

$$\mathcal{C}(\boldsymbol{\beta}) = \sum_{\mathbf{z} \in \mathcal{Z}} \exp\left\{\boldsymbol{\beta}^T \mathbf{s}(\mathbf{z})\right\}, \tag{2}$$

$\mathbf{z} = (z_1, \ldots, z_N)^T$ are the observed data, $\mathbf{s}(\mathbf{z}) = \left(s^{(1)}(\mathbf{z}), \ldots, s^{(D)}(\mathbf{z})\right)^T$ are the sufficient statistics, and $\boldsymbol{\beta} = (\beta^{(1)}, \ldots, \beta^{(D)})^T$ are the natural parameters. When the set of all possible observed data, $\mathcal{Z}$, is large, the computational cost of evaluating the sum (2)

becomes infeasible. There are models of this form that will benefit from the methodology developed in this work. These include the Potts model and the related autologistic model (Besag, 1974) that we discuss in Sections 4.1 and 4.3, respectively, and the exponential random graph model (ERGM, e.g., Robins et al., 2007).

2.2 Approximate Likelihood

We follow the approach of Price et al. (2018), and approximate the computationally intractable likelihood in (1) using a Bayesian synthetic likelihood. Specifically, we approximate the intractable likelihood as a multivariate normal distribution of the sufficient statistics, with mean $\boldsymbol{\mu}(\boldsymbol{\beta})$ and covariance $\boldsymbol{\Sigma}(\boldsymbol{\beta})$. This yields the synthetic likelihood function $\tilde{p}(\mathbf{z} \mid \boldsymbol{\beta}) = \mathcal{N}(\mathbf{s}(\mathbf{z}); \boldsymbol{\mu}(\boldsymbol{\beta}), \boldsymbol{\Sigma}(\boldsymbol{\beta}))$. Note that if the sufficient statistics are highly non-Gaussian, there are more robust synthetic likelihood approaches that one can use (e.g., An et al., 2020; Frazier and Drovandi, 2021). We further assume that the sufficient statistics in $\mathbf{s}(\mathbf{z})$ are mutually independent, that is, we let $\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \text{diag}(\{\sigma^{2(d)}, d = 1, \ldots, D\})$, so that

$$\tilde{p}(\mathbf{z} \mid \boldsymbol{\beta}) = \prod_{d=1}^{D} \mathcal{N}(s^{(d)}(\mathbf{z}); \mu^{(d)}(\boldsymbol{\beta}), \sigma^{2(d)}(\boldsymbol{\beta})). \tag{3}$$

To evaluate our synthetic likelihood function $\tilde{p}(\mathbf{z} \mid \boldsymbol{\beta}^*)$ for any $\boldsymbol{\beta}^*$, we build surrogate models for the mean functions $\{\mu^{(d)}(\boldsymbol{\beta}^*)\}_d$ and variance functions $\{\sigma^{2(d)}(\boldsymbol{\beta}^*)\}_d$. We first simulate pseudo-data (using the SW algorithm) at a fixed set of parameters for fitting the surrogate model. Specifically, consider a fixed set of p parameter values $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p\}$. For each $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$, we generate q simulations of the sufficient statistics $\{\mathbf{s}_{j,1}, \ldots, \mathbf{s}_{j,q}\}$, where $\mathbf{s}_{j,k} = (s_{j,k}^{(1)}, \ldots, s_{j,k}^{(D)})^T$. Then, we obtain the sample means and the sample variances of the simulations at $\boldsymbol{\beta}_j$, that is, we compute

$$\begin{aligned} m^{(d)}(\boldsymbol{\beta}_j) &= \frac{1}{q} \sum_{k=1}^{q} s_{j,k}^{(d)}, \\ v^{(d)}(\boldsymbol{\beta}_j) &= \frac{1}{q-1} \sum_{k=1}^{q} (s_{j,k}^{(d)} - m^{(d)}(\boldsymbol{\beta}_j))^2, \end{aligned} \tag{4}$$

for $j = 1, \ldots, p$, and $d = 1, \ldots, D$. The sample means $\{m^{(d)}(\boldsymbol{\beta}_j)\}_{j,d}$ and the sample variances $\{v^{(d)}(\boldsymbol{\beta}_j)\}_{j,d}$ are treated as (noisy) observations of the true means $\{\mu^{(d)}(\boldsymbol{\beta}_j)\}_{j,d}$ and variances $\{\sigma^{2(d)}(\boldsymbol{\beta}_j)\}_{j,d}$, respectively.

We note that, from (1), for $d = 1, \ldots, D$,

$$\begin{aligned} \mu^{(d)}(\boldsymbol{\beta}) &= \text{E}_{\boldsymbol{\beta}}(s^{(d)}(\mathbf{z})) = \frac{\partial}{\partial \beta^{(d)}} \log \mathcal{C}(\boldsymbol{\beta}), \\ \sigma^{2(d)}(\boldsymbol{\beta}) &= \text{Var}_{\boldsymbol{\beta}}(s^{(d)}(\mathbf{z})) = \frac{\partial^2}{\partial \beta^{(d)2}} \log \mathcal{C}(\boldsymbol{\beta}), \end{aligned} \tag{5}$$

which implies $\sigma^{2(d)}(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta^{(d)}} \mu^{(d)}(\boldsymbol{\beta})$. This property motivates us to model the means and variances jointly in our surrogate model. Such models are often referred to as gradient-enhanced models (Laurent et al., 2019).

2.3 Prior Distribution

In this paper we employ independent, bounded uniform priors for the model parameter(s) $\boldsymbol{\beta}$, largely for computational convenience and to facilitate comparison with other related techniques that also employ uniform priors (e.g. Møller et al., 2006; Everitt, 2012; Lyne et al., 2015; Järvenpää et al., 2021). Other priors should be considered by practitioners that take into account the context of their application. For example, in the case of the (one-parameter) Potts model described in Section 4.1, it is often reasonable to assume that neighboring variables (pixels) in a lattice are positively correlated. A prior distribution that excludes negative values of the inverse temperature parameter β is therefore reasonable with this model. In practice, there will also be a value of β, β_{crit} say, beyond which realizations of the k-state Potts model will have fewer than k unique labels with high probability. For this reason, it can be beneficial to put an upper bound on β, or at least penalize large values by using an exponential prior for β; this penalized complexity prior would be similar to those discussed by Simpson et al. (2017).

In many applications, there is also ample prior information available. For example, in the case of Landsat data, such as those considered in the example of Section 4.2, satellite imagery is available from 1972 to present (Wulder et al., 2022), and the historical data could be used to construct a prior distribution. In these cases, calibrated log-normal, gamma, truncated normal, or scaled beta distributions are all suitable candidates. Expert elicitation can also be used to construct informative priors (French, 2022). Prior information is particularly useful in our context as it helps us select the appropriate fixed set of parameters $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p\}$ for fitting the Gaussian process surrogates (e.g., by sampling from the prior distribution).

3 The Surrogate Model

In Section 3.1, we introduce the gradient-enhanced nonstationary Gaussian process surrogate model, which uses both deformation functions and gradient information to improve the fit to the surrogate means. In Section 3.2, we describe approaches for evaluating the surrogate synthetic likelihood at arbitrary parameter values using the fitted surrogate model. Section 3.3 presents a few ways with which one could use the surrogate synthetic likelihood for both inexact-approximate and exact-approximate Bayesian inference.

3.1 Gaussian Process Surrogate Models

In this section, we introduce the Gaussian process surrogate models for the sufficient statistics. The Gaussian process surrogates use the observed sample means $\{m^{(d)}(\boldsymbol{\beta}_j)\}_{j,d}$

and sample variances $\{v^{(d)}(\boldsymbol{\beta}_j)\}_{j,d}$ from (4) to emulate the true means and variances at any $\boldsymbol{\beta}^*$.

It is important that simulator "noise" is accounted for in surrogate models (e.g., Gramacy, 2020, Chap. 5). This is relevant for our surrogate models, since we build them using sample (and not exact) means and variances. We therefore model each $m^{(d)}(\boldsymbol{\beta}_j)$ and $v^{(d)}(\boldsymbol{\beta}_j)$ as a noisy observation of the (true) surrogate mean $\mu^{(d)}(\boldsymbol{\beta})$ and surrogate variance $\sigma^{2(d)}(\boldsymbol{\beta})$, respectively. Since the sample mean and the sample variance are asymptotically normally distributed around the mean and variance of the sufficient statistics, respectively, we let

$$\begin{aligned} m^{(d)}(\boldsymbol{\beta}_j) &\equiv \mu^{(d)}(\boldsymbol{\beta}_j) + \epsilon_{\mu j}^{(d)}, \\ v^{(d)}(\boldsymbol{\beta}_j) &\equiv \sigma^{2(d)}(\boldsymbol{\beta}_j) + \epsilon_{\sigma j}^{(d)}, \end{aligned} \tag{6}$$

for $j = 1, \ldots, p$, and $d = 1, \ldots, D$, where $\epsilon_{\mu j}^{(d)} \sim \mathcal{N}(0, \tau_{\mu j}^{2(d)})$ and $\epsilon_{\sigma j}^{(d)} \sim \mathcal{N}(0, \tau_{\sigma j}^{2(d)})$ are Gaussian noise terms with variances $\tau_{\mu j}^{2(d)}$ and $\tau_{\sigma j}^{2(d)}$, respectively. Here, we assume that the noise terms are independent, but these could also be modeled as more general Gaussian processes (e.g., Gramacy, 2020, Chap. 10). For simplicity, here we estimate the variances $\tau_{\mu j}^{2(d)}$ and $\tau_{\sigma j}^{2(d)}$ under the assumption that the underlying sufficient statistics are Gaussian. Hence, we set

$$\tau_{\mu j}^{2(d)} = \frac{v^{(d)}(\boldsymbol{\beta}_j)}{q} \qquad \text{and} \qquad \tau_{\sigma j}^{2(d)} = \frac{2v^{2(d)}(\boldsymbol{\beta}_j)}{q-1}.$$

Equation (6) does not respect nonnegativity of the variance parameter; this was not found to be a problem in practice, and more complicated models could be considered (at some computational cost) if needed.

Note that although one can obtain both the sample means and variances from the observations (see (4)), one may choose to build just one surrogate model: one for the surrogate means or one for the surrogate variances. For example, Moores et al. (2020) built a surrogate model for the variances and used the mean–variance relationship given in (5) to find the surrogate means by integration. Park (2021) built a surrogate model for the means, and fixed the surrogate covariance matrices to be equal to the nearest empirical covariance matrices. One could also employ coupled mean-variance Gaussian process models in this context (e.g., Chapter 10 of Gramacy, 2020); however, these do not take advantage of the mean-variance relationship given in (5).

In what follows we consider different Gaussian process surrogate models: first, a stationary Gaussian process surrogate model for the means, and then a nonstationary Gaussian process surrogate model for the means. Finally, we build a novel joint Gaussian process surrogate for the means and variances that takes advantage of the relationship in (5).

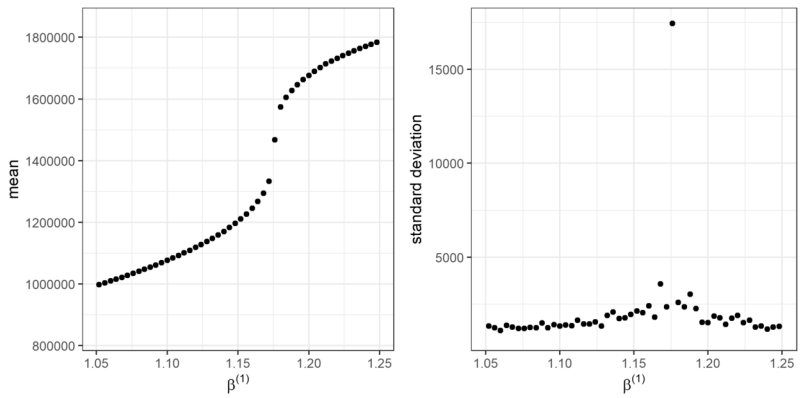

Figure 1: Empirical means (left panel) and standard deviations (right panel) of the simulated sufficient statistics for different values of $\beta^{(1)}$ for the Potts model on a 1000×1000 image with 5 labels.

Stationary Gaussian Process Surrogate

We first consider a stationary Gaussian process for modeling the means of the sufficient statistics. In this first case, independently, we let

$$\mu^{(d)}(\cdot) = \tilde{g}^{(d)}(\cdot) = b^{(d)}(\cdot) + g^{(d)}(\cdot), \quad d = 1, \ldots, D, \tag{7}$$

where $b^{(d)}(\cdot)$ is the trend, and $g^{(d)}(\cdot)$ is a zero-mean stationary Gaussian process. This model serves as a baseline, against which we will evaluate the more sophisticated models we discuss next.

Nonstationary Gaussian Process Surrogate

Phase transition is a property in many models with intractable likelihoods, such as the Potts model and the ERGM. Figure 1 shows the sample means and sample standard deviations of the sufficient statistic of the Potts model (the sum of all neighbor pairs with the same label) for a 1000×1000 image with 5 labels (see Section 4.1 for details). When the parameter $\beta^{(1)}$ is between 1.15 and 1.2, the mean of the sufficient statistic changes value rapidly with small changes in the parameters; this is in contrast to other regions in the parameter space where the rate of change is relatively slow. This region in the parameter space where the mean of the sufficient statistic changes rapidly is often called a phase transition.

As we show in Section 4.1, a stationary Gaussian process surrogate model cannot adequately capture the nonstationary behavior of the means of the sufficient statistics at the phase transition. We therefore also consider a nonstationary Gaussian process for the surrogate means. There are several approaches that could be adopted to model nonstationarity; here, we use the deformation (or warping) approach (Sampson and Guttorp, 1992) where nonstationarity is obtained via a deformation (or warping) of the

parameter space. The warping function $\mathbf{f}(\cdot)$ maps the parameter space onto a new domain on which a stationary Gaussian process is modeled, which induces a nonstationary Gaussian process on the original domain. We construct a flexible warping function $\mathbf{f}(\cdot)$ by linking several simple warping functions (such as basis functions) through composition, as detailed by Zammit-Mangion et al. (2022). Warping of the parameter space is a natural way forward when modeling the means of the sufficient statistics, as the warping function can stretch the parameter space when these undergo a phase transition and shrink the parameter space in other regions, in such a way that on the warped space it would be reasonable to model the process as stationary.

Our nonstationary model for the means of the sufficient statistics is given by

$$\mu^{(d)}(\cdot) = \tilde{g}^{(d)}(\cdot) = b^{(d)}(\cdot) + g^{(d)}(\mathbf{f}^{(d)}(\cdot)), \tag{8}$$

where $b^{(d)}(\cdot)$ is the trend, $g^{(d)}(\cdot)$ is a zero-mean stationary Gaussian process, and $\mathbf{f}^{(d)}(\cdot)$ is a deformation/warping function given by $\mathbf{f}^{(d)}(\cdot) \equiv \mathbf{f}_n^{(d)} \circ \cdots \circ \mathbf{f}_1^{(d)}(\cdot)$ where each $\mathbf{f}_i^{(d)}(\cdot) = \boldsymbol{\Phi}^{(d)}(\cdot)\boldsymbol{\eta}_i^{(d)}$ is constructed using basis functions $\boldsymbol{\Phi}^{(d)}(\cdot)$ and weights $\boldsymbol{\eta}_i^{(d)}$, that need to be estimated. See Zammit-Mangion et al. (2022) for more details on the basis functions and estimation.

Gradient-Enhanced Nonstationary Gaussian Process Surrogate

As shown in (5), the variance of the sufficient statistic, $\sigma^{2(d)}(\boldsymbol{\beta})$, is the derivative of the mean of the statistic, $\mu^{(d)}(\boldsymbol{\beta})$, with respect to the d^{th} dimension of $\boldsymbol{\beta}$. This motivates us to use a gradient-enhanced Gaussian process to jointly model the means and the variances of the sufficient statistics, to ultimately improve the fit to the surrogate means (Riihimäki and Vehtari, 2010; Laurent et al., 2019). This joint model is expected to improve the quality of the fit to the means of the sufficient statistics as, in this case, the sample variances are also informative on the means of the sufficient statistics.

We model the means of the sufficient statistics using nonstationary Gaussian processes as in (8). We then model the variances as the derivatives of the respective surrogate means. Therefore, our joint model of the mean $\mu^{(d)}(\cdot)$ and the variance $\sigma^{2(d)}(\cdot)$ is a bivariate Gaussian process (Banerjee et al., 2003),

$$\begin{pmatrix} \mu^{(d)}(\cdot) \\ \sigma^{2(d)}(\cdot) \end{pmatrix} = \begin{pmatrix} \tilde{g}^{(d)}(\cdot) \\ \tilde{h}^{(d)}(\cdot) \end{pmatrix} = \begin{pmatrix} b^{(d)}(\cdot) \\ c^{(d)}(\cdot) \end{pmatrix} + \begin{pmatrix} g^{(d)}(\mathbf{f}^{(d)}(\cdot)) \\ h^{(d)}(\mathbf{f}^{(d)}(\cdot)) \end{pmatrix}, \tag{9}$$

where $b^{(d)}(\cdot)$ is the trend, $c^{(d)}(\cdot) = \frac{\partial}{\partial \beta^{(d)}} b^{(d)}(\cdot)$, $g^{(d)}(\cdot)$ is a zero-mean stationary Gaussian process, $\mathbf{f}^{(d)}(\cdot)$ is a deformation function, and $h^{(d)}(\cdot) = \frac{\partial}{\partial \beta^{(d)}} g^{(d)}(\cdot)$. This joint Gaussian process is a special case of the multivariate nonstationary Gaussian process with shared warping function proposed by Vu et al. (2022).

The cross-covariance matrix function of the joint Gaussian process in (9) depends on the covariance function of $g^{(d)}(\cdot)$, which we denote by $K^{(d)}(\cdot,\cdot)$. The cross-covariance matrix function of the joint Gaussian process in (9), which we denote by $\mathbf{C}^{(d)}(\cdot,\cdot)$, is

then given by

$$\mathbf{C}^{(d)}(\boldsymbol{\beta}_j, \boldsymbol{\beta}_l) = \begin{pmatrix} \mathrm{cov}(\tilde{g}^{(d)}(\boldsymbol{\beta}_j), \tilde{g}^{(d)}(\boldsymbol{\beta}_l)) & \mathrm{cov}(\tilde{g}^{(d)}(\boldsymbol{\beta}_j), \tilde{h}^{(d)}(\boldsymbol{\beta}_l)) \\ \mathrm{cov}(\tilde{h}^{(d)}(\boldsymbol{\beta}_j), \tilde{g}^{(d)}(\boldsymbol{\beta}_l)) & \mathrm{cov}(\tilde{h}^{(d)}(\boldsymbol{\beta}_j), \tilde{h}^{(d)}(\boldsymbol{\beta}_l)) \end{pmatrix},$$

where,

$$\begin{aligned}
\mathrm{cov}(\tilde{g}^{(d)}(\boldsymbol{\beta}_j), \tilde{g}^{(d)}(\boldsymbol{\beta}_l)) &= K^{(d)}(\mathbf{f}^{(d)}(\boldsymbol{\beta}_j), \mathbf{f}^{(d)}(\boldsymbol{\beta}_l)), \\
\mathrm{cov}(\tilde{h}^{(d)}(\boldsymbol{\beta}_j), \tilde{g}^{(d)}(\boldsymbol{\beta}_l)) &= \frac{\partial K^{(d)}(\mathbf{f}^{(d)}(\boldsymbol{\beta}_j), \mathbf{f}^{(d)}(\boldsymbol{\beta}_l))}{\partial \beta_j^{(d)}}, \\
\mathrm{cov}(\tilde{g}^{(d)}(\boldsymbol{\beta}_j), \tilde{h}^{(d)}(\boldsymbol{\beta}_l)) &= \frac{\partial K^{(d)}(\mathbf{f}^{(d)}(\boldsymbol{\beta}_j), \mathbf{f}^{(d)}(\boldsymbol{\beta}_l))}{\partial \beta_l^{(d)}}, \\
\mathrm{cov}(\tilde{h}^{(d)}(\boldsymbol{\beta}_j), \tilde{h}^{(d)}(\boldsymbol{\beta}_l)) &= \frac{\partial^2 K^{(d)}(\mathbf{f}^{(d)}(\boldsymbol{\beta}_j), \mathbf{f}^{(d)}(\boldsymbol{\beta}_l))}{\partial \beta_j^{(d)} \partial \beta_l^{(d)}}.
\end{aligned}$$

Our model does not naturally enforce positivity of the predicted variances, however this was not found to be a problem in practice. If this does become an issue, one could use basis functions when constructing the surrogate models, and add constraints to the weights in order to ensure positivity everywhere (Zammit-Mangion et al., 2022). However, basis function surrogate models tend to be less flexible than (full-rank) Gaussian processes.

3.2 Predicting the Surrogate Likelihood Elsewhere in the Parameter Space

After fitting the surrogate models, we can predict the means and variances of the sufficient statistics for an arbitrary parameter vector $\boldsymbol{\beta}^*$ through Gaussian conditioning. Using the joint model in (9), we can obtain predictions of both the means and the variances of the sufficient statistics. However, while we found that the variance information was very important for improving the estimates of the parameters when *fitting* the Gaussian process, we found little benefit in using them to predict the means and the variances, and predicting the variances as the derivative of the mean predictions sufficed. Focusing on the prediction of the means also facilitates comparison between the different surrogate models in Section 3.1, two of which do not consider the variance of the sufficient statistics.

We consider two approaches for predicting the means and variances of the sufficient statistics with our surrogate models. In the first approach, we let the surrogate mean be equal to the prediction mean of the Gaussian process. That is, for some possibly new parameter value $\boldsymbol{\beta}^*$, we let

$$\tilde{\mu}^{(d)}(\boldsymbol{\beta}^*) = \mathrm{E}[\tilde{g}^{(d)}(\boldsymbol{\beta}^*) | \mathbf{M}^{(d)}], \quad d = 1, \ldots, D,$$

where $\mathbf{M}^{(d)} = \{m^{(d)}(\boldsymbol{\beta}_1), \ldots, m^{(d)}(\boldsymbol{\beta}_p)\}$. Then, we set the surrogate variance to be

equal to the gradient of the surrogate mean, that is, we set

$$\tilde{\sigma}^{2(d)}(\boldsymbol{\beta}^*) = \frac{\partial}{\partial \beta^{(d)*}} \mathrm{E}[\tilde{g}^{(d)}(\boldsymbol{\beta}^*) | \mathbf{M}^{(d)}], \quad d = 1, \ldots, D.$$

Then, the surrogate synthetic likelihood is given by,

$$\tilde{p}(\mathbf{z} \mid \boldsymbol{\beta}^*) = \prod_{d=1}^{D} \mathcal{N}(s^{(d)}(\mathbf{z}); \tilde{\mu}^{(d)}(\boldsymbol{\beta}^*), \tilde{\sigma}^{2(d)}(\boldsymbol{\beta}^*)).$$

The second approach we propose accounts for the uncertainty in the surrogate model. In this approach, we sample r realizations of the surrogate means, $\{\hat{\mu}_1^{(d)}(\boldsymbol{\beta}^*), \ldots, \hat{\mu}_r^{(d)}(\boldsymbol{\beta}^*)\}$, $d = 1, \ldots, D$, from the fitted Gaussian process. Then, for each realization $\hat{\mu}_i^{(d)}(\boldsymbol{\beta}^*)$, we evaluate the surrogate variance

$$\hat{\sigma}_i^{2(d)}(\boldsymbol{\beta}^*) = \frac{\partial}{\partial \beta^{(d)*}} \hat{\mu}_i^{(d)}(\boldsymbol{\beta}^*), \quad i = 1, \ldots, r.$$

The surrogate synthetic likelihood is then given by,

$$\tilde{p}_i(\mathbf{z} \mid \boldsymbol{\beta}^*) = \prod_{d=1}^{D} \mathcal{N}(s^{(d)}(\mathbf{z}); \hat{\mu}_i^{(d)}(\boldsymbol{\beta}^*), \hat{\sigma}_i^{2(d)}(\boldsymbol{\beta}^*)), \quad i = 1, \ldots, r.$$

Because we sample independently from the fitted Gaussian process, the surrogate likelihood for all the realizations is then just the average of the surrogate likelihoods for each realization. Specifically,

$$\tilde{p}(\mathbf{z} \mid \boldsymbol{\beta}^*) \approx \frac{1}{r} \sum_{i=1}^{r} \tilde{p}_i(\mathbf{z} \mid \boldsymbol{\beta}^*).$$

3.3 Inference Using the Surrogate Model

Once we obtain the surrogate synthetic likelihood, there are different ways we can use it when making Bayesian inference. The simplest way is to just substitute the approximately-true likelihood (typically obtained from simulation via the SW algorithm), with the surrogate likelihood in the Metropolis-Hastings ratio in the exchange algorithm, as in Moores et al. (2020) and Park and Haran (2020). This is computationally efficient, as it precludes the need for further simulations of the sufficient statistics, but it is an inexact-approximate method. One exact-approximate MCMC method, which uses the surrogate likelihood to reduce the computational cost, is delayed-acceptance MCMC (Christen and Fox, 2005). The delayed-acceptance algorithm is a two-stage algorithm, wherein the first step involves using the surrogate likelihood to quickly reject any poor proposals: Only good proposals accepted in the first stage are passed through to the second stage, where similarly to the exchange algorithm, auxiliary variables are used to accept or reject the proposals. This algorithm prevents one from wasting computational resources on simulations of sufficient statistics at a poor proposal. The delayed-acceptance algorithm is detailed in Algorithm 1.

Algorithm 1: Delayed-acceptance MCMC.

Denote samples from the target posterior distribution as $\tilde{\boldsymbol{\beta}}_i : i = 0, \ldots, T$, where T is the number of iterations. Initialize $\tilde{\boldsymbol{\beta}}_0$. For $i = 0, \ldots, (T-1)$ do

1 Propose a new value $\boldsymbol{\beta}^{\text{new}}$ from a proposal distribution $q(\boldsymbol{\beta}^{\text{new}} \mid \tilde{\boldsymbol{\beta}}_i)$.

2 Stage 1: Pass $\boldsymbol{\beta}^{\text{new}}$ to Stage 2 with probability

$$\min\left(1, \frac{q(\tilde{\boldsymbol{\beta}}_i \mid \boldsymbol{\beta}^{\text{new}}) p(\boldsymbol{\beta}^{\text{new}}) \tilde{p}(\mathbf{z} \mid \boldsymbol{\beta}^{\text{new}})}{q(\boldsymbol{\beta}^{\text{new}} \mid \tilde{\boldsymbol{\beta}}_i) p(\tilde{\boldsymbol{\beta}}_i) \tilde{p}(\mathbf{z} \mid \tilde{\boldsymbol{\beta}}_i)}\right).$$

3 Stage 2: Simulate pseudo-data $\mathbf{x}^{\text{new}}$ from the likelihood $p(\mathbf{x}^{\text{new}} \mid \boldsymbol{\beta}^{\text{new}})$. Accept $\boldsymbol{\beta}^{\text{new}}$ with probability

$$\min\left(1, \frac{\tilde{p}(\mathbf{z} \mid \tilde{\boldsymbol{\beta}}_i) \exp\{\boldsymbol{\beta}^{\text{new}^T} \mathbf{s}(\mathbf{z})\} \exp\{\tilde{\boldsymbol{\beta}}_i^T \mathbf{s}(\mathbf{x}^{\text{new}})\}}{\tilde{p}(\mathbf{z} \mid \boldsymbol{\beta}^{\text{new}}) \exp\{\tilde{\boldsymbol{\beta}}_i^T \mathbf{s}(\mathbf{z})\} \exp\{\boldsymbol{\beta}^{\text{new}^T} \mathbf{s}(\mathbf{x}^{\text{new}})\}}\right).$$

4 If $\boldsymbol{\beta}^{\text{new}}$ is accepted at Stage 2, then set $\tilde{\boldsymbol{\beta}}_{i+1} = \boldsymbol{\beta}^{\text{new}}$, otherwise, if $\boldsymbol{\beta}^{\text{new}}$ is not accepted at either Stage 1 or Stage 2, then set $\tilde{\boldsymbol{\beta}}_{i+1} = \tilde{\boldsymbol{\beta}}_i$.

An alternative to MCMC methods is importance sampling (e.g., Everitt et al., 2017). Importance sampling is an exact-approximate method, yet it affords a significant improvement in computational time, as the simulations of sufficient statistics (required to determine the importance weights), can be performed in parallel. The importance sampling algorithm shown in Algorithm 2 uses auxiliary variables in a similar fashion to Møller et al. (2006). In our approach, we choose the proposal distribution for importance sampling to be the *surrogate posterior distribution*, which we define to be that posterior distribution obtained when using the surrogate likelihood function directly in place of the true likelihood function. A fast approach to obtain the surrogate posterior is by using grid approximation, wherein one evaluates the surrogate posterior density at regular points on a grid (e.g., Gelman et al., 2013); this is feasible in the low-dimensional settings we consider. Once the grid approximation is found, it is straightforward to draw samples from the surrogate posterior distribution prior to reweighing using the simulations of the sufficient statistics.

4 Data Examples

In this section, we show examples that demonstrate the utility of our proposed Gaussian process surrogate models when making inference on parameters that appear in the Potts, the hidden Potts, the autologistic models, and the Kent distribution, using both simulated and real-world data sets. All MCMC computations were done on a computer with a 6-core Intel Core i7-8700 @3.2 GHz, 32 GB RAM, and an NVIDIA GeForce GTX 1600 GPU. All computations with the importance sampling algorithm were done on a server with 64 cores in Intel Xeon E5-2683 @2.1 GHz processors, 256 GB RAM,

Algorithm 2: Importance sampling.

Denote samples from the target posterior distribution as $\tilde{\boldsymbol{\beta}}_i : i = 1, \ldots, T$, where T is the number of samples.

1 Sample $\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_T$ from the surrogate posterior distribution $\tilde{p}(\mathbf{z} \mid \boldsymbol{\beta})p(\boldsymbol{\beta})$.

2 For each $\tilde{\boldsymbol{\beta}}_i$, simulate pseudo-data $\mathbf{x}_i$ from the likelihood $p(\mathbf{x}_i \mid \tilde{\boldsymbol{\beta}}_i)$.

3 Calculate the importance weight for each of the samples,

$$w_i = \frac{\exp\{\tilde{\boldsymbol{\beta}}_i^T \mathbf{s}(\mathbf{z})\} \exp\{\hat{\boldsymbol{\beta}}^T \mathbf{s}(\mathbf{x}_i)\}}{\exp\{\tilde{\boldsymbol{\beta}}_i^T \mathbf{s}(\mathbf{x}_i)\} \tilde{p}(\mathbf{z} \mid \tilde{\boldsymbol{\beta}}_i)},$$

where $\hat{\boldsymbol{\beta}}$ is fixed at an arbitrary point. Similar to Møller et al. (2006), we choose the maximum likelihood estimate, which we approximate using the surrogate likelihood.

4 Normalize the weights

$$\tilde{w}_i = \frac{w_i}{\sum_{i=1}^T w_i}.$$

5 Obtain the weighted posterior samples from which we can then make inference on $\boldsymbol{\beta}$.

and an NVIDIA GeForce GTX TITAN GPU, in order to take advantage of algorithm parallelization. Code to reproduce the results in this section can be found at https://github.com/quanvu17/warped_gradient_enhanced_GP_surrogate.

4.1 Potts Model

The Potts model is often used for analyzing spatial dependence between neighboring labels in images. In the Potts model, the probability of observing a specific combination of labels is defined as

$$p(\mathbf{z} \mid \beta^{(1)}) = \frac{\exp\big(\beta^{(1)} \sum_{u \sim v} \delta(z_u, z_v)\big)}{\mathcal{C}(\beta^{(1)})}, \tag{10}$$

where $z_u, u = 1, \ldots, N$, is the label of pixel u, $u \sim v$ denotes the neighboring pixels of pixel u, and $\delta(\cdot)$ is the Kronecker delta function. The sufficient statistic of the Potts model, $s(\mathbf{z}) = \sum_{u \sim v} \delta(z_u, z_v)$, is the count of pairs of neighboring pixels that have the same label. The normalizing constant $\mathcal{C}(\beta^{(1)}) = \sum_{\mathbf{z} \in \mathcal{Z}} \exp\big(\beta^{(1)} \sum_{u \sim v} \delta(z_u, z_v)\big)$ involves a summation over all possible combinations of the labels over all the pixels, and therefore is computationally infeasible to evaluate. The Potts model undergoes a phase transition from a disordered state (where most neighboring pixels do not share the same label) to an ordered state (where most neighboring pixels share the same label) near a critical value of the parameter $\beta^{(1)}$, which is often referred to as the inverse temperature parameter. As $p(\mathbf{z} \mid \beta^{(1)})$ involves a computationally intractable sum and

also contains a phase transition, we use the proposed gradient-enhanced nonstationary Gaussian process surrogate model to approximate this probability.

In this experiment we consider the Potts model for a 1000×1000 size image, where the number of labels is $k = 5$. The likelihood function exhibits a phase transition around $\beta^{(1)} = 1.175$; see Figure 1.

Comparison of Gaussian Process Surrogate Models

We chose a bounded uniform prior distribution on the interval $[0.9, 1.3]$ for the inverse temperature parameter $\beta^{(1)}$ (see Section 2.3 for a discussion on the choice of prior distributions). We chose $p = 51$ equally-spaced points on this interval for training, and 50 points (in between the training data) for testing. For each of the 101 values of $\beta^{(1)}$ we simulated sufficient statistics using the SW algorithm.

We fit the surrogate models introduced in Section 3.1 to the training data set:

1. A stationary Gaussian process model (S-GP),
2. A nonstationary Gaussian process model (NS-GP),
3. A gradient-enhanced nonstationary Gaussian process model (GE-NS-GP),

where the trend $b^{(1)}(\cdot)$ was fixed to be the mean of the observed sample means, and the stationary Gaussian process $g^{(1)}(\cdot)$ was set to have the Matérn 3/2 covariance function

$$K^{(1)}(\boldsymbol{\beta}_j, \boldsymbol{\beta}_l) = \text{cov}(g^{(1)}(\boldsymbol{\beta}_j), g^{(1)}(\boldsymbol{\beta}_l)) = \xi^{2(1)}(1 + a^{(1)}\|\mathbf{h}\|)\exp\big(-a^{(1)}\|\mathbf{h}\|\big),$$

where $\xi^{2(1)}$ is the process variance parameter, $a^{(1)}$ is the process scale parameter, and $\mathbf{h} \equiv \boldsymbol{\beta}_l - \boldsymbol{\beta}_j$. For Models 2 and 3 we used an axial warping unit for $f^{(1)}(\cdot)$ (Zammit-Mangion et al., 2022) comprised of 100 sigmoid basis functions.

For comparison purposes we also ran the parametric functional approximate Bayesian (PFAB) algorithm (Moores et al., 2020), which uses a surrogate parametric model that takes into account specific properties of the Potts likelihood (e.g., the critical value, which can be calculated exactly for a 2D lattice). PFAB is specifically designed for the Potts model and is therefore a good candidate for comparison. Note that our Gaussian surrogate models are more general as they can be easily used with other exponential-family models (e.g., with the autologistic model and the Kent distribution, as we show in Section 4.3 and Section 4.4, respectively).

Fitting of the Gaussian process surrogate models 1–3 only took a few seconds using maximum likelihood, while running the PFAB algorithm took nearly 30 minutes. The predicted means and variances of the sufficient statistics at the testing locations were then compared to those of the simulated sufficient statistics at these testing locations. The mean absolute prediction error (MAPE) and root mean square prediction error (RMSPE) are shown in Table 1. The PFAB algorithm produced substantially worse predictions than all of the Gaussian process models for both the mean and the variance,

Model	PFAB	S-GP	NS-GP	GE-NS-GP
MAE (mean, $\times 10^3$)	4.224	1.258	1.007	0.817
RMSPE (mean, $\times 10^3$)	12.581	3.554	2.978	1.528
MAE (variance, $\times 10^6$)	6.387	5.810	5.783	5.699
RMSPE (variance, $\times 10^7$)	4.185	3.819	3.792	3.748

Table 1: Comparison of the different surrogate models.

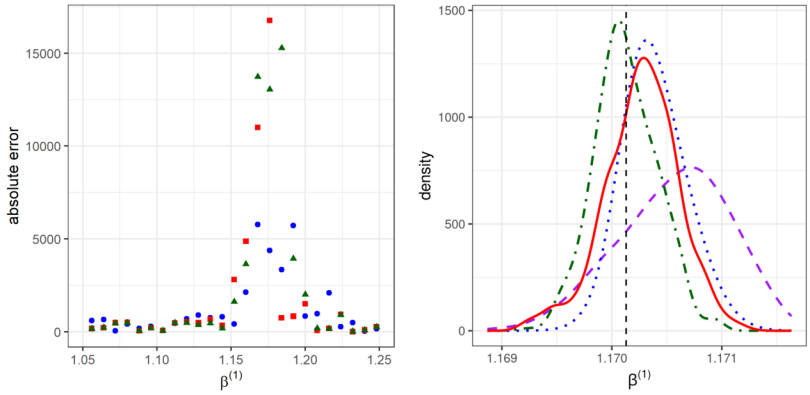

Figure 2: Left panel: Absolute errors of the means of the sufficient statistics for the surrogate models. Green triangle: S-GP. Red square: NS-GP. Blue dot: GE-NS-GP. Right panel: Posterior distribution of the parameter $\beta^{(1)}$ for the Potts model. Purple dashed line: GE-NS-GP surrogate posterior. Blue dotted line: Importance sampling. Green dot-dashed line: Delayed-acceptance MCMC. Red solid line: Exchange algorithm. Vertical dashed line: True parameter value.

despite it being specifically designed for the Potts model. Focusing just on the Gaussian process models, we see that in terms of predicting the mean, the stationary GP performed the worst, while the gradient-enhanced nonstationary GP model performed the best. The left panel of Figure 2 shows the absolute errors for Models 1–3. All three models perform similarly in regions far away from the phase transition. However, near the phase transition, we see that the stationary model results in large errors, while the gradient-enhanced nonstationary model results in smaller errors. Table 1 shows that the GE-NS-GP model also generates slightly better predictions of the variances, when compared to those of the S-GP and NS-GP models, although the improvement is less substantial.

Inference

In this sub-section, we conduct experiments where we infer the inverse temperature parameter from images using the Gaussian process surrogate models. We set the true inverse temperature parameter $\beta^{(1)}$ by simulating from a normal distribution centered at the true critical value $\beta^{(1)} = 1.1744$ (see Potts, 1952), with standard deviation 0.05:

Method	Surrogate	Importance sampling	Delayed-acceptance	Exchange
Posterior mean	1.17056	1.17039	1.17013	1.17027
Posterior SD	0.000497	0.000252	0.000285	0.000333
Time (hours)	3×10^{-6}	0.7	25.0	31.0
ESS/hour	3.6×10^{9}	79.8	7.2	5.7

Table 2: Posterior distribution for the Potts example of Section 4.1.

The simulated value was 1.1701. We then generated five different images from this parameter value to show the repeatability of our experiment.

We used the GE-NS-GP surrogate model to do Bayesian inference using importance sampling and delayed-acceptance MCMC, and compared both these methods to the exchange algorithm. We first generated 1000 samples from the surrogate posterior distribution using grid approximation, which took less than one second. Then, for importance sampling, we simulated pseudo-data to reweigh the samples. For both the delayed-acceptance MCMC algorithm and the exchange algorithm, we used 2200 MCMC iterations (including 200 burn-in iterations). The posterior distributions obtained from these methods from one of the five images are shown in the right panel of Figure 2 and summarized in Table 2. Posterior distributions for the other four images are shown in Figure S1 in Section S1 of the Supplementary Material (Vu et al., 2023). We see that running the importance sampler only took 0.7 hours, since it is parallelizable. The sampler also resulted in a much higher effective sample size (ESS) per hour (79.8 samples per hour) than the other exact-approximate methods. Delayed-acceptance took 25.0 hours, a bit less, relatively, than the 31.0 hours of the exchange algorithm, and resulted in a small gain in ESS per hour of 7.2, when compared to 5.7 for the exchange algorithm. In Figure 2, we also show the inexact-approximate posterior distribution (i.e., the surrogate posterior distribution), where we simply use our surrogate model as the likelihood and do grid-based inference. Note that this inexact posterior distribution is slightly different from the posterior distributions obtained from the exact methods. However, the speed at which this approximate posterior distribution was obtained (< 1 second in this case) can make it attractive for situations where computing time is a serious concern.

4.2 Hidden Potts Model

The hidden Potts model is an extension of the Potts model, and is used for analyzing spatial dependence when the labels are not directly observed. The model links the observed pixel intensity, y_u, with the latent label, z_u, through the following relationship

$$p(y_u \mid z_u = \lambda, \mu_\lambda, \sigma^2_\lambda) = \text{Gau}(\mu_\lambda, \sigma^2_\lambda), \quad \lambda = 1, \ldots, k,$$

where the $\{\mu_\lambda\}_\lambda$ and the $\{\sigma^2_\lambda\}_\lambda$ are unknown and equipped with informative prior distributions; see Moores et al. (2020) for details. In this example, we consider a normalized difference vegetation index (NDVI) 1000×1000 image of Brisbane derived from Landsat-8 satellite data on 03 May 2015,[1] shown in the left panel of Figure 3. We ana-

[1]Data available from https://hpc.niasra.uow.edu.au/ckan/dataset/ndvip089r079_20150503

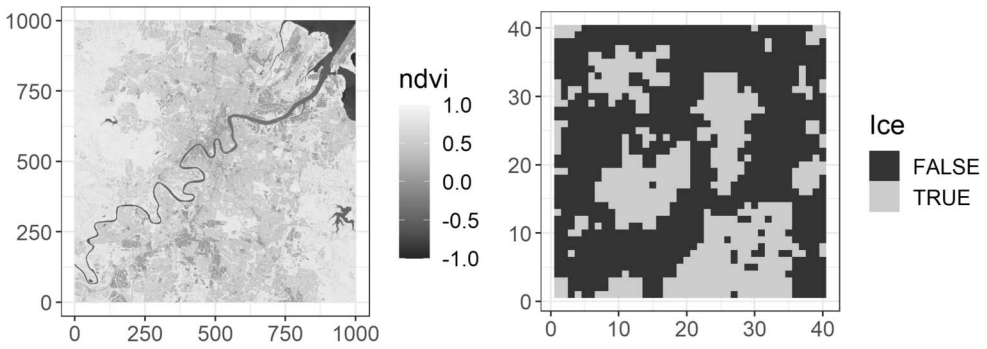

Figure 3: Images used in the analysis in Section 4.2 and Section 4.3, respectively. Left panel: NDVI image of Brisbane derived from Landsat-8 data on 03 May 2015. Right panel: Cropped satellite image of ice floe, originally published in Banfield and Raftery (1992), available in the `R` package `PAWL`.

Method	Surrogate	Delayed-acceptance	Exchange
Posterior mean	1.23593	1.23592	1.23576
Posterior SD	0.000842	0.000777	0.000845
Time (hours)	3.2	14.6	28.3
ESS/hour	112.8	7.8	6.5

Table 3: Posterior distribution for the hidden Potts example of Section 4.2.

lyzed spatial dependence in the data by classifying pixels using five labels: forest, light vegetation, urban area, suburban area, and water. We use the hidden Potts model and make inference on the inverse temperature parameter $\beta^{(1)}$, which determines the spatial dependence between these labels. As we do not directly observe the labels we cannot use importance sampling for this model. We therefore used MCMC to update the latent labels, as well as the parameters $\{\mu_\lambda\}_\lambda$ and $\{\sigma^2_\lambda\}_\lambda$, at each iteration of the chain.

We chose the prior distribution over $\beta^{(1)}$ to be a uniform distribution on the interval $[0.9, 1.3]$. The satellite image in this section has the same dimension and number of labels as the simulated image in Section 4.1; hence we used the same GE-NS-GP model we fitted in Section 4.1 when making inference. We compared the inexact-approximate method to the two MCMC exact-approximate methods we consider in this work: delayed-acceptance MCMC and the exchange algorithm. We ran 2200 MCMC iterations (including 200 burn-in iterations) for each method. The posterior distributions obtained from these methods are shown in Figure 4 and summarized in Table 3. We see that all the posterior distributions are very similar. However, with the same number of iterations, evaluating the surrogate inexact-approximate posterior distribution took 3.2 hours. On the other hand, delayed-acceptance MCMC took 14.6 hours, while the exchange algorithm took 28.3 hours. Unsurprisingly, the highest ESS per hour of 112.8 was obtained with the inexact-approximate method. Delayed acceptance and the exchange algorithm yielded an ESS per hour of 7.8 and 6.5, respectively.

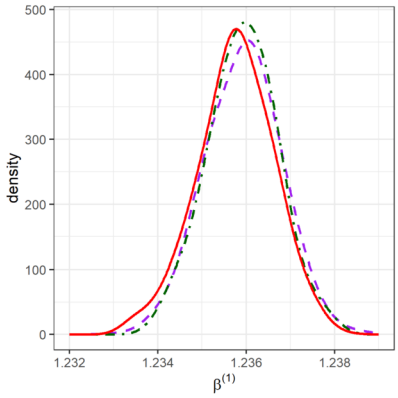

Figure 4: Posterior distribution of the parameter $\beta^{(1)}$ for the hidden Potts model. Purple dashed line: Surrogate posterior. Green dot-dashed line: Delayed-acceptance MCMC. Red solid line: Exchange algorithm.

4.3 Autologistic Model

The autologistic model is another extension of the Potts model, where the sufficient statistics also include the number of pixels associated with each label. The two-label autologistic model, proposed by Besag (1974), has as likelihood function

$$p(\mathbf{z} \mid \boldsymbol{\beta}) = \frac{\exp\left(\beta^{(1)} \sum_u z_u + \beta^{(2)} \sum_{u \sim v} \delta(z_u, z_v)\right)}{\mathcal{C}(\boldsymbol{\beta})},$$

where the label z_u of pixel u takes the value 1 or -1.

We use the autologistic model to analyze spatial dependence in a satellite image of ice floe originally published in Banfield and Raftery (1992); here we consider the cropped 40×40 image (Bornn et al., 2013), available in the `R` package `PAWL`, and shown in the right panel of Figure 3.

We chose the prior distribution to be a uniform distribution on $[-0.2, 0.1] \times [0.7, 1.2]$, and we chose training data from a 7×11 equally-spaced grid on this rectangular domain to fit the GE-NS-GP surrogate model. We fixed the trend $b^{(1)}(\cdot)$, $b^{(2)}(\cdot)$ to the means of the observed sample means for each sufficient statistic, and chose the covariance functions for $g^{(1)}(\cdot)$ and $g^{(2)}(\cdot)$ to be Matérn 3/2 covariance functions, and used axial warping units in each dimension for the deformation functions $\mathbf{f}^{(1)}(\cdot)$ and $\mathbf{f}^{(2)}(\cdot)$.

We first took 2000 samples from the surrogate posterior distribution using grid approximation. For importance sampling, we simulated sufficient statistics at these 2000 values to reweigh the samples. We compared the results to those using the exchange algorithm, which was run for 4200 iterations (200 were discarded as burn-in). The posterior distributions obtained from the two methods are shown in Figure 5 and summarized in Table 4. As we saw in Section 4.1, all posterior distributions are largely similar, and

Method	Surrogate	Importance sampling	Exchange
Posterior mean $\beta^{(1)}$	-0.00167	-0.00154	-0.00208
Posterior SD $\beta^{(1)}$	0.00176	0.00176	0.00272
Posterior mean $\beta^{(2)}$	0.88361	0.88312	0.88315
Posterior SD $\beta^{(2)}$	0.01647	0.01748	0.02028
Time (mins)	0.001	1.8	29.2
ESS/minute $\beta^{(1)}$	3×10^6	201	7.1
ESS/minute $\beta^{(2)}$	3×10^6	201	5.3

Table 4: Posterior distribution for the autologistic example of Section 4.3.

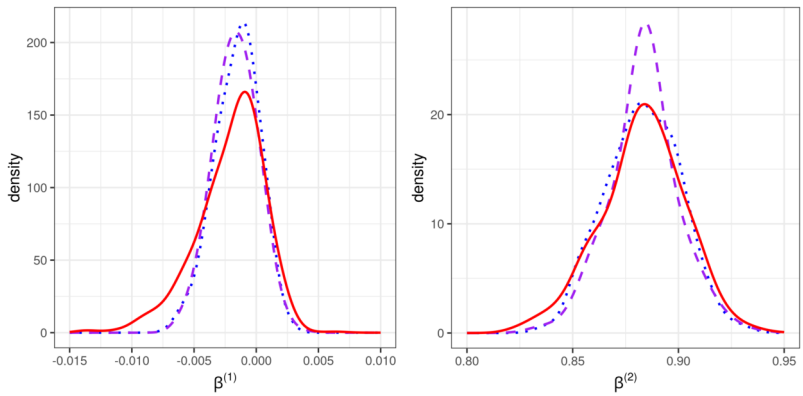

Figure 5: Posterior distribution of the parameter $\beta^{(1)}$ (left panel) and $\beta^{(2)}$ (right panel) for the autologistic model. Purple dashed line: Surrogate posterior. Blue dotted line: Importance sampling. Red solid line: Exchange algorithm.

we again see the huge computational gain of importance sampling over the exchange algorithm in terms of time taken and ESS per unit time.

4.4 Kent Distribution

In this section, we illustrate the use of our proposed methodology with the Kent distribution (Kent, 1982). This distribution is used to model data on a sphere, and is analogous to a bivariate normal distribution on a Euclidean plane. The likelihood is given by,

$$p(\mathbf{z} \mid \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \kappa, \beta) = \frac{1}{\mathcal{C}(\kappa, \beta)} \exp\{\kappa(\boldsymbol{\gamma}_1'\mathbf{z}) + \beta[(\boldsymbol{\gamma}_2'\mathbf{z})^2 - (\boldsymbol{\gamma}_3'\mathbf{z})^2]\}, \tag{11}$$

where $\mathbf{z} \in \mathbb{S}^2$, $\boldsymbol{\gamma}_1$ determines the mean direction, $\boldsymbol{\gamma}_2$ and $\boldsymbol{\gamma}_3$ are the major and minor axes, κ determines the concentration, and β determines the ellipticity ($0 < \beta < \kappa/2$). To simplify the illustration, we have assumed that the parameters $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2$ and $\boldsymbol{\gamma}_3$ are fixed and known (in this case to $(1,0,0)', (0,1,0)'$, and $(0,0,1)'$, respectively). We can

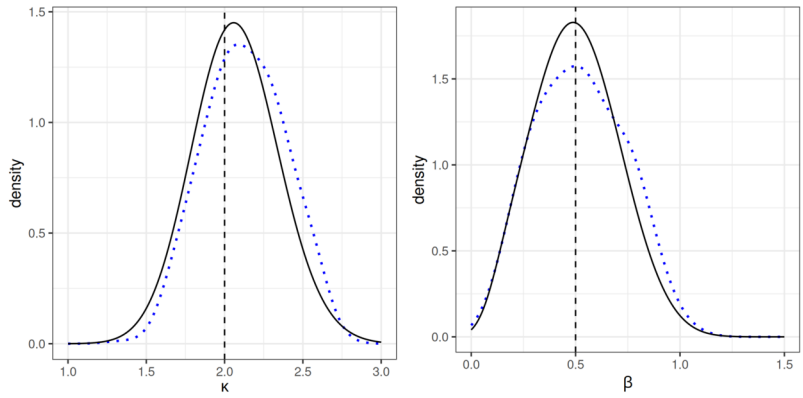

Figure 6: Posterior distribution of the parameter κ (left panel) and β (right panel) for the Kent distribution. Blue dotted line: Importance sampling using the warped gradient-enhanced Gaussian process surrogate model as proposal distribution. Black solid line: Grid-based approximation to the true posterior distribution. Vertical dashed line: True parameter value.

see from (11) that the Kent distribution is an exponential-family model. In this case the normalizing constant $\mathcal{C}(\kappa, \beta)$ is known; hence, with this model we can easily compare the posterior distribution obtained using our warped gradient-enhanced Gaussian process surrogate model with the true posterior distribution evaluated on a fine discretization of the parameter space.

We performed a simulation study using a sample of 100 data points generated from the Kent distribution with parameters $\kappa = 2$ and $\beta = 0.5$ (chosen close to the "phase transition" of the distribution). We used the GE-NS-GP surrogate model to do Bayesian inference using importance sampling with the prior distribution a uniform distribution on the triangular part of the domain $[0.2, 10] \times [0.1, 5]$ that lies below the line $\beta = \kappa/2$. To train the GE-NS-GP, we simulated data from the Kent distribution with parameters κ and β on an equally-spaced grid on this triangular domain. The setup of the surrogate model was otherwise identical to that used in the autologistic model example. For importance sampling we first took 2000 samples from the surrogate posterior distribution (this was done via a grid approximation), and then simulated sufficient statistics at these 2000 values to obtain a weighted sample. The resulting posterior distribution is shown in Figure 6, where we also show the (true) posterior distribution obtained using grid-based methods. As expected, the posterior distributions are very similar.

5 Discussion

In this article, we have introduced a warped gradient-enhanced Gaussian process surrogate model for modeling the means and variances of the sufficient statistics of models with intractable likelihoods. In particular, we showed that the inclusion of nonstation-

arity and gradient information in this surrogate model resulted in smaller errors at the phase transition than the stationary Gaussian process surrogate model, and that the surrogate model can easily be used to speed up parameter inference in exact-approximate methods such as importance sampling and delayed-acceptance MCMC. Note that it is also possible to employ the surrogate model directly without a correction (such as by sampling directly from the surrogate posterior distribution, or in a non-Bayesian approach by maximizing the surrogate likelihood), but this might lead to slightly inaccurate results (as shown in Section 4.1).

In Section 4, we showed examples for the Potts, the hidden Potts, the autologistic, and the Kent models. Inference for other models with exponential-family intractable likelihoods can also be done using our approach. One such model is the exponential random graph model (ERGM). ERGMs have, however, some distinct properties. In particular, there are some ERGMs for which the sufficient statistics do not result in any phase transitions; inference with these models can be done effectively using stationary Gaussian process surrogate models (Park and Haran, 2020). Sufficient statistics in other ERGMs experience steep phase transitions, and simulations at the phase transitions often result in bimodal distributions. Inference with these types of models is a future direction of research.

In this work, we have assumed independence between the sufficient statistics (see (3)). Although this assumption does not seem to have affected our inferences, one can envisage the use of surrogate models where the sufficient statistics are modeled as dependent. One would need a Gaussian process which jointly models all the surrogate means and covariances of the sufficient statistics; in this case, the cross-covariances between the means of the different sufficient statistics are non-zero. The cross-covariances between the variances of the different sufficient statistics are also non-zero. This leads to the question of what is the appropriate form of the cross-covariance function. The answer to this question is non-trivial, as the covariance between two different sufficient statistics can be written as two different gradients of the means of these sufficient statistics with respect to a different dimension of the parameter, that is,

$$\begin{aligned}\mathrm{cov}_{\boldsymbol{\beta}}(s^{(d_1)}(\mathbf{z}), s^{(d_2)}(\mathbf{z})) &= \frac{\partial^2}{\partial\beta^{(d_1)}\partial\beta^{(d_2)}} \log \mathcal{C}(\boldsymbol{\beta}) \\ &= \frac{\partial}{\partial\beta^{(d_1)}}\mu^{(d_2)}(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta^{(d_2)}}\mu^{(d_1)}(\boldsymbol{\beta}).\end{aligned}$$

In future work we will also address the scalability of our approach to high-dimensional parameter space; in this article, we only showed examples with one-dimensional and two-dimensional parameter spaces. In a high-dimensional parameter space, we would need to simulate sufficient statistics at a much larger number of parameter values in order to get a good fit to the means and variances of the sufficient statistics. This will entail investigating the use of approximation methods to Gaussian processes (e.g., Quiñonero-Candela and Rasmussen, 2005) in the context of surrogate likelihood models.

The relationship between the mean and the variance shown in (5) holds only for exponential-family models, and therefore, the gradient-enhanced Gaussian process sur-

rogates is only applicable to models in that class. However, there still are elements in our work that can be used to construct representative synthetic likelihood functions for more general (non-exponential) intractable models. In these cases, we would approximate the likelihood as a multivariate normal distribution of summary statistics. For the means of the summary statistics, we could then use the (warped) nonstationary Gaussian process surrogate model (8) introduced in Section 3.1. For the (log) variances, we could also use a nonstationary Gaussian process surrogate model; that is, we could employ the model

$$\log \sigma^{2(d)}(\cdot) = c^{(d)}(\cdot) + h^{(d)}(\mathbf{f}^{(d)}(\cdot)), \tag{12}$$

where $c^{(d)}(\cdot)$ is a trend, $h^{(d)}(\cdot)$ is a zero-mean stationary Gaussian process, and $\mathbf{f}^{(d)}(\cdot)$ is a warping function. One would need to use a different warping function for the variance than from the mean, since unlike in the exponential-family case there is no pre-defined relationship between the mean and the variance. The resulting model can be seen as a warped version of the coupled mean and variance Gaussian processes presented by Gramacy (2020, Chap. 10). Developing a flexible class of models on these lines for non-exponential family models is the subject of future research.

Supplementary Material

Supplementary Material to "Warped Gradient-Enhanced Gaussian Process Surrogate Models for Exponential Family Likelihoods with Intractable Normalizing Constants" (DOI: 10.1214/23-BA1400SUPP; .pdf).

References

An, Z., Nott, D. J., and Drovandi, C. (2020). "Robust Bayesian synthetic likelihood via a semi-parametric approach." *Statistics and Computing*, 30: 543–557. MR4065218. doi: https://doi.org/10.1007/s11222-019-09904-x. 4

Andrieu, C. and Roberts, G. O. (2009). "The pseudo-marginal approach for efficient Monte Carlo computations." *The Annals of Statistics*, 37: 697–725. MR2502648. doi: https://doi.org/10.1214/07-AOS574. 2

Aushev, A., Pesonen, H., Heinonen, M., Corander, J., and Kaski, S. (2022). "Likelihood-free inference with deep Gaussian processes." *Computational Statistics & Data Analysis*, 174: 107529. MR4427015. doi: https://doi.org/10.1016/j.csda.2022.107529. 3

Banerjee, S., Gelfand, A. E., and Sirmans, C. (2003). "Directional rates of change under spatial process models." *Journal of the American Statistical Association*, 98: 946–954. MR2041483. doi: https://doi.org/10.1198/C16214503000000909. 8

Banfield, J. D. and Raftery, A. E. (1992). "Ice floe identification in satellite images using mathematical morphology and clustering about principal curves." *Journal of the American Statistical Association*, 87: 7–16. MR1951635. doi: https://doi.org/10.1198/016214502760047131. 16, 17

Beaumont, M. A. (2003). "Estimation of population growth or decline in genetically monitored populations." *Genetics*, 164: 1139–1160. 2

Besag, J. (1974). "Spatial interaction and the statistical analysis of lattice systems." *Journal of the Royal Statistical Society: Series B*, 36: 192–225. MR0373208. 4, 17

Boland, A., Friel, N., and Maire, F. (2018). "Efficient MCMC for Gibbs Random Fields using pre-computation." *Electronic Journal of Statistics*, 12: 4138–4179. MR3890764. doi: https://doi.org/10.1214/18-EJS1504. 2

Bornn, L., Jacob, P. E., Del Moral, P., and Doucet, A. (2013). "An adaptive interacting Wang–Landau algorithm for automatic density exploration." *Journal of Computational and Graphical Statistics*, 22: 749–773. MR3173740. doi: https://doi.org/10.1080/10618600.2012.723569. 17

Christen, J. A. and Fox, C. (2005). "Markov chain Monte Carlo using an approximation." *Journal of Computational and Graphical Statistics*, 14: 795–810. MR2211367. doi: https://doi.org/10.1198/106186005X76983. 3, 10

Drovandi, C. C., Moores, M. T., and Boys, R. J. (2018). "Accelerating pseudo-marginal MCMC using Gaussian processes." *Computational Statistics & Data Analysis*, 118: 1–17. MR3715260. doi: https://doi.org/10.1016/j.csda.2017.09.002. 2

Everitt, R. G. (2012). "Bayesian parameter estimation for latent Markov random fields and social networks." *Journal of Computational and Graphical Statistics*, 21: 940–960. MR3005805. doi: https://doi.org/10.1080/10618600.2012.687493. 5

Everitt, R. G., Johansen, A. M., Rowing, E., and Evdemon-Hogan, M. (2017). "Bayesian model comparison with un-normalised likelihoods." *Statistics and Computing*, 27: 403–422. MR3599680. doi: https://doi.org/10.1007/s11222-016-9629-2. 3, 11

Frank, O. and Strauss, D. (1986). "Markov graphs." *Journal of the American Statistical Association*, 81: 832–842. MR0860518. 1

Frazier, D. T. and Drovandi, C. (2021). "Robust approximate Bayesian inference with synthetic likelihood." *Journal of Computational and Graphical Statistics*, 30: 958–976. MR4356598. doi: https://doi.org/10.1080/10618600.2021.1875839. 4

French, S. (2022). "From soft to hard elicitation." *Journal of the Operational Research Society*, 73: 1181–1197. 5

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, Boca Raton, FL, third edition. MR3235677. 11

Gramacy, R. B. (2020). *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. CRC Press, Boca Raton, FL. MR4283556. doi: https://doi.org/10.1201/9780367815493. 6, 21

Järvenpää, M., Gutmann, M. U., Vehtari, A., and Marttinen, P. (2018). "Gaussian process modelling in approximate Bayesian computation to estimate horizontal gene transfer in bacteria." *The Annals of Applied Statistics*, 12: 2228–2251. MR3875699. doi: https://doi.org/10.1214/18-AOAS1150. 2, 3

— (2021). "Parallel Gaussian process surrogate Bayesian inference with noisy likelihood evaluations." *Bayesian Analysis*, 16: 147–178. MR4194277. doi: https://doi.org/10.1214/20-BA1200. 2, 5

Kennedy, M. C. and O'Hagan, A. (2000). "Predicting the output from a complex computer code when fast approximations are available." *Biometrika*, 87: 1–13. MR1766824. doi: https://doi.org/10.1093/biomet/87.1.1. 2

Kent, J. T. (1982). "The Fisher-Bingham distribution on the sphere." *Journal of the Royal Statistical Society: Series B*, 44: 71–80. MR0655376. 18

Laurent, L., Le Riche, R., Soulier, B., and Boucard, P.-A. (2019). "An overview of gradient-enhanced metamodels with applications." *Archives of Computational Methods in Engineering*, 26: 61–106. MR3895170. doi: https://doi.org/10.1007/s11831-017-9226-3. 3, 5, 8

Lyne, A.-M., Girolami, M., Atchadé, Y., Strathmann, H., and Simpson, D. (2015). "On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods." *Statistical Science*, 30: 443–467. MR3432836. doi: https://doi.org/10.1214/15-STS523. 5

Meeds, E. and Welling, M. (2014). "GPS-ABC: Gaussian process surrogate approximate Bayesian computation." In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*, 593–602. 2

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). "An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants." *Biometrika*, 93: 451–458. MR2278096. doi: https://doi.org/10.1093/biomet/93.2.451. 2, 5, 11, 12

Moores, M. T., Nicholls, G. K., Pettitt, A. N., and Mengersen, K. (2020). "Scalable Bayesian inference for the inverse temperature of a hidden Potts model." *Bayesian Analysis*, 15: 1–27. MR4050875. doi: https://doi.org/10.1214/18-BA1130. 2, 6, 10, 13, 15

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006). "MCMC for doubly-intractable distributions." In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*, 359–366. 2

Park, J. (2021). "Bayesian indirect inference for models with intractable normalizing functions." *Journal of Statistical Computation and Simulation*, 91: 300–315. MR4198589. doi: https://doi.org/10.1080/00949655.2020.1814286. 6

Park, J. and Haran, M. (2020). "A function emulation approach for doubly intractable distributions." *Journal of Computational and Graphical Statistics*, 29: 66–77. MR4085864. doi: https://doi.org/10.1080/10618600.2019.1629941. 2, 10, 20

Potts, R. B. (1952). "Some generalized order-disorder transformations." *Mathematical Proceedings of the Cambridge Philosophical Society*, 48: 106–109. MR0047571. 1, 14

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2018). "Bayesian synthetic

likelihood." *Journal of Computational and Graphical Statistics*, 27: 1–11. MR3788296. doi: https://doi.org/10.1080/10618600.2017.1302882. 4

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). "Population growth of human Y chromosomes: a study of Y chromosome microsatellites." *Molecular Biology and Evolution*, 16: 1791–1798. 2

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). "A unifying view of sparse approximate Gaussian process regression." *Journal of Machine Learning Research*, 6: 1939–1959. MR2249877. 20

Riihimäki, J. and Vehtari, A. (2010). "Gaussian processes with monotonicity information." In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 645–652. 8

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). "An introduction to exponential random graph (p*) models for social networks." *Social Networks*, 29: 173–191. 4

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). "Design and analysis of computer experiments." *Statistical Science*, 4: 409–423. MR1041765. 2

Sampson, P. D. and Guttorp, P. (1992). "Nonparametric estimation of nonstationary spatial covariance structure." *Journal of the American Statistical Association*, 87: 108–119. 7

Sherlock, C., Golightly, A., and Henderson, D. A. (2017). "Adaptive, delayed-acceptance MCMC for targets with expensive likelihoods." *Journal of Computational and Graphical Statistics*, 26: 434–444. MR3640199. doi: https://doi.org/10.1080/10618600.2016.1231064. 3

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). "Penalising model component complexity: A principled, practical approach to constructing priors." *Statistical Science*, 32: 1–28. MR3634300. doi: https://doi.org/10.1214/16-STS576. 5

Swendsen, R. H. and Wang, J.-S. (1987). "Nonuniversal critical dynamics in Monte Carlo simulations." *Physical Review Letters*, 58: 86–88. 2

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). "Inferring coalescence times from DNA sequence data." *Genetics*, 145: 505–518. 2

Vihola, M., Helske, J., and Franks, J. (2020). "Importance sampling type estimators based on approximate marginal Markov chain Monte Carlo." *Scandinavian Journal of Statistics*, 47: 1339–1376. MR4178196. doi: https://doi.org/10.1111/sjos.12492. 3

Vu, Q., Zammit-Mangion, A., and Cressie, N. (2022). "Modeling nonstationary and asymmetric multivariate spatial covariances via deformations." *Statistica Sinica*, 32: 2071–2093. MR4478190. 3, 8

Vu, Q., Moores, M. T., Zammit-Mangion, A. (2023). "Supplementary Material for "Warped gradient-enhanced Gaussian process surrogate models for exponential family

likelihoods with intractable normalizing constants"." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1400SUPP. 15

Wilkinson, R. (2014). "Accelerating ABC methods using Gaussian processes." In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, 1015–1023. 2

Wulder, M. A., Roy, D. P., Radeloff, V. C., Loveland, T. R., Anderson, M. C., Johnson, D. M., Healey, S., Zhu, Z., Scambos, T. A., Pahlevan, N., et al. (2022). "Fifty years of Landsat science and impacts." *Remote Sensing of Environment*, 280: 113195. 5

Zammit-Mangion, A., Ng, T. L. J., Vu, Q., and Filippone, M. (2022). "Deep compositional spatial models." *Journal of the American Statistical Association*, 117: 1787–1808. MR4528471. doi: https://doi.org/10.1080/01621459.2021.1887741. 3, 8, 9, 13