Journal of Machine Learning Research 26 (2025) 1-34 Submitted 3/23; Revised 12/24; Published 2/25

Bayesian Multi-Group Gaussian Process Models for Heterogeneous Group-Structured Data

Didong Li DIDONGLI@UNC.EDU
Department of Biostatistics
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA

Andrew Jones AJONES788@GMAIL.COM
Department of Computer Science,
Princeton University
Princeton, NJ 08540, USA

Sudipto Banerjee SUDIPTO@UCLA.EDU
Department of Biostatistics
University of California, Los Angeles
Los Angeles, CA 90095, USA

Barbara Engelhardt BENGELHARDT@STANFORD.EDU
Gladstone Institutes
San Francisco, CA 94158, USA
Department of Biomedical Data Science
Stanford University
Stanford, CA 94305, USA

Editor: Debdeep Pati

Abstract

Gaussian processes are pervasive in functional data analysis, machine learning, and spatial statistics for modeling complex dependencies. Scientific data are often heterogeneous in their inputs and contain multiple known discrete groups of samples; thus, it is desirable to leverage the similarity among groups while accounting for heterogeneity across groups. We propose multi-group Gaussian processes (MGGPs) defined over $\mathbb{R}^p \times \mathscr{C}$, where $\mathscr{C}$ is a finite set representing the group label, by developing general classes of valid (positive definite) covariance functions on such domains. MGGPs are able to accurately recover relationships between the groups and efficiently share strength across samples from all groups during inference, while capturing distinct group-specific behaviors in the conditional posterior distributions. We demonstrate inference in MGGPs through simulation experiments, and we apply our proposed MGGP regression framework to gene expression data to illustrate the behavior and enhanced inferential capabilities of multi-group Gaussian processes by jointly modeling continuous and categorical variables.

Keywords: Mixed data, covariance functions, Gaussian processes, semiparametric regression

©2025 Didong Li, Andrew Jones, Sudipto Banerjee, and Barbara E. Engelhardt.

License: CC-BY 4.0, see `https://creativecommons.org/licenses/by/4.0/`. Attribution requirements are provided at `http://jmlr.org/papers/v26/23-0291.html`.

1. Introduction

Gaussian processes (GPs, Rasmussen and Williams (2005)) are widely used in modeling complex dependent data in diverse inferential settings including nonlinear regression (Ghosal and Van der Vaart, 2017), spatial statistics (Stein, 1999), classification problems (Bernardo et al., 1998) and, increasingly, in deep learning and reinforcement learning applications (Damianou and Lawrence, 2013; Deisenroth et al., 2013). A GP endows an uncountable collection of random variables with a probability law so that any finite subset is multivariate Gaussian. This is achieved through a real-valued positive definite function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which acts as a covariance function or kernel.

We develop GPs for analyzing "multi-group" data, where measurements belong to one of k groups. Examples include biological measurements from distinct tissues or cell types (Consortium et al., 2020; Regev et al., 2017); geospatial data from multiple locations defined by discrete demarcations, such as state or country borders (Pan et al., 2020); and census data from people of different races, ethnicities, and genders (Bureau, 2020). Models built on Euclidean domains do not account for the discrete set of groups. While GPs on non-Euclidean manifolds and graphs have attracted recent attention (Niu et al., 2019; Dunson et al., 2020; Li et al., 2023) and machine learning (Borovitskiy et al., 2020, 2021), the multi-group setting remains largely unaddressed.

For flexibly modeling k-group data, we seek a stochastic process over $\mathcal{X} : \mathbb{R}^p \times \mathscr{C}$ to drive the inference, where $\mathscr{C} = \{c_1, \cdots, c_k\}$ is a finite set representing group labels. We specifically extend three existing approaches over Euclidean domains (Park and Choi, 2010): Separate Gaussian processes (SGPs), Union Gaussian processes (UGPs), and Hierarchical Gaussian processes (HGPs). The SGP assumes independence across groups. Therefore, the across-group correlation is set to zero: $K((x, c_i), (x', c_j)) = 0$ if $i \neq j$. The SGP is equivalent to modeling each group with a separate, independent GP. The UGP assumes the same dependencies within and across groups, so the covariance function does not depend on the members of $\mathscr{C}$, i.e., $K((x, c_i), (x', c_j)) = K_0(x, x')$. It is equivalent to modeling all groups jointly with a single GP. The HGP accommodates both across- and within-group dependencies, where all within-group dependencies are assumed to be identical, and all across-group dependencies are assumed to be identical as well. Here, $K((x, c_i), (x', c_j)) = K_0(x, x') + 1_{\{c_i = c_j\}} K_1(x, x')$, where K_0 and K_1 are real-valued positive definite functions (Park and Choi, 2010; Hensman et al., 2013). Each of the above models build covariance functions based upon a standard GP over a Euclidean domain. These models are often used in practice for their simplicity and extend beyond the context of GPs (Tsherniak et al., 2017). However, each of these three GPs types impose restrictive conditions and fail to model heterogeneous between-group dependencies. If the model is misspecified, the inferential performance of these processes will be unsatisfactory, especially when the overall sample size is small or the groups are imbalanced in terms of sample size.

We introduce multi-group Gaussian process (MGGP) models (Section 2) through the construction of positive definite covariance functions over $\mathcal{X} := \mathbb{R}^p \times \mathscr{C}$ (Section 3). The multi-group process flexibly models heterogeneity across and within groups, leveraging varying levels of similarity between groups and allows us to exploit prior or expert knowledge. Since the multi-group structure is encoded in the covariance function, we can use existing methods and computational algorithms for fitting standard GP models.

The MGGP contributes to the literature on joint modeling of continuous and categorical variables (Dunson et al., 2003; Dunson, 2000; Teimourian et al., 2015; Ru et al., 2020; Schulam et al., 2015; Murray and Reiter, 2016; Leroy et al., 2022, 2023) by avoiding additive, hierarchical, and mixture models entirely. Instead, the MGGP explicitly modeling dependencies within and among the continuous and categorical variables. This flexibility allows straightforward conditioning on multiple categorical partitions, offering a tractable conditional posterior, and enables us to exploit dependencies between groups when some groups have small sample sizes. The inferential benefits of our approach are illustrated using maximum likelihood and full Bayesian inference through simulation experiments and an analysis of gene expression data (Section 4). We conclude with some remarks in Section 6. The Appendix includes proofs of theoretical results, code, data and additional analysis.

2. Multi-Group Gaussian Process Regression Models

We consider a dependent variable $y(x; c_j)$ generated from a latent stochastic process over $\mathbb{R}^p \times \mathscr{C}$ for inputs $x \in \mathbb{R}^p$ and group j through the model

$$y(x; c_j) = \mu(x; c_j) + Z(x; c_j) + \epsilon(x; c_j) \;, \quad \epsilon(x; c_j) \overset{ind}{\sim} N(0, \tau_j^2) \;, \tag{1}$$

where $\mu(x; c_j)$ is a mean function, $Z(x; c_j)$ is a zero-centered latent process, and $\epsilon(x; c_j)$ is a zero-centered white-noise process capturing measurement error or fine-scale variation with group-specific variances. The mean function can be further modeled, if appropriate, as $\mu(x; c_j) = f_j(x)^{\mathrm{T}} \beta_j$, where $f_j(x)$ is a $q_j \times 1$ vector of design variables possibly, but not necessarily, depending on x, and each β_j is a $q_j \times 1$ vector of group-specific regression coefficients. This specification accommodates predictors or other explanatory variables that need neither be continuous nor reside within $\mathbb{R}^p$.

Equation (1) includes a parametric specification through the mean function and a nonparametric specification through the latent process. Our focus in this paper is not so much on modeling $\mu(x; c_j)$, which can be built from standard linear model specifications, as it is on $Z(x; c_j) : \mathcal{X} \longrightarrow \mathbb{R}$, where $\mathcal{X} = \mathbb{R}^p \times \mathscr{C}$. We will specify $Z(\cdot; \cdot)$ to be a GP with zero mean and covariance function $K((x; c_j), (x'; c_{j'})) : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ so that $K((x; c_j), (x'; c_{j'})) = \mathrm{cov}(Z(x; c_j), Z(x'; c_{j'}))$ is a positive-definite covariance function. We consider settings where data arise over a finite, possibly imbalanced, set of points $\{(x_i; c_j)\}$ for $i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, k$. Each group can have a different number, n_j, of inputs. Given the covariance function, the realizations of the process over the finite set of points is the $n \times 1$ vector $Z = (Z_1^{\mathrm{T}}, \ldots, Z_k^{\mathrm{T}})^{\mathrm{T}}$, where $n = \sum_{j=1}^k n_j$ and $Z_j = (Z(x_1; c_j), \ldots, Z(x_{n_j}; c_j))^{\mathrm{T}}$ follows a multivariate Gaussian distribution with an $n \times 1$ zero vector as the mean and an $n \times n$ covariance matrix K, whose (j, j')th block is given by the $n_j \times n_{j'}$ matrix $K_{jj'}$ with (i, i') element $K((x_i; c_j), (x_{i'}; c_{j'}))$ for $i = 1, 2, \ldots, n_j$ and $i' = 1, 2, \ldots, n_{j'}$.

Equation (1) enables likelihood-based inference and can be extended to a Bayesian hierarchical framework (Cressie and Wikle, 2011; Banerjee et al., 2014). Assuming, for elucidation purposes only, that $\mu(x; c_j) = f_j(x)^{\mathrm{T}} \beta_j$, a Bayesian model specifies the joint distribution

$$p(\{\tau_j^2\}, \theta, \{\beta_j\}) \times N(Z \,|\, 0, K_\theta) \times \prod_{j=1}^{k} \prod_{i=1}^{n_j} N(y(x_i; c_j) \,|\, f_j(x_i)^{\mathrm{T}} \beta_j + Z(x_i; c_j), \tau_j^2) \ , \tag{2}$$

where θ denotes parameters in the covariance function, and $p(\{\tau_j^2\}, \theta, \{\beta_j\})$ is the prior on the model parameters. Inference on these parameters and the latent process $Z(\cdot)$ proceeds by drawing samples from the posterior distribution $p(\{\tau_j^2\}, \theta, \{\beta_j\}, Z \,|\, \{y(x_i; c_j)\}, \{f_j(x_i)\})$, which is proportional to Equation (2).

Sampling from the joint posterior distribution including the process realizations Z will be challenging due to the dimension of Z. Exploiting the Gaussian likelihood, we work with the collapsed likelihood after integrating out Z from Equation (2), which yields

$$p(\tau, \theta, \beta \,|\, y, F) \propto p(\tau, \theta, \beta) \times N(y \,|\, F\beta, K_\theta + D_\tau) \ , \tag{3}$$

where y is the $n \times 1$ vector of observations, $y(x_i; c_j)$, constructed analogous to Z, F is an $n \times q$ block-diagonal matrix, $q = \sum_{j=1}^{k} q_j$, with $n_j \times q_j$ blocks $F_j = (f_j(x_1), \ldots, f_j(x_{n_j}))^{\mathrm{T}}$, $\beta = (\beta_1^{\mathrm{T}}, \ldots, \beta_k^{\mathrm{T}})^{\mathrm{T}}$ is the $q \times 1$ vector of stacked regression coefficients, $\tau = \{\tau_j^2\}$ is the collection of error variances, and D_τ is the $n \times n$ diagonal matrix with $\tau_j^2 I_{n_j}$ as $n_j \times n_j$ diagonal blocks. Markov chain Monte Carlo (MCMC) algorithms sample more efficiently from Equation (3) because of the reduced parameter space relative to Equation (2).

We sample from $p(Z \,|\, y, F) = \mathbb{E}[p(Z \,|\, \{\tau, \theta, \beta\}, y, F)]$ to carry out inference on the latent process, where the expectation $\mathbb{E}[\cdot]$ is taken with respect to the posterior distribution in Equation (3); we draw one $Z \sim p(Z \,|\, \{\tau, \theta, \beta\}, y, F)$ for each posterior drawn value of $\{\tau, \theta, \beta\}$. This is straightforward because $p(Z \,|\, \{\tau, \theta, \beta\}, y, F)$ is of the form $N(Mm, M)$, where $M^{-1} = K_\theta^{-1} + D_\tau^{-1}$ and $m = y - F\beta$, and the draws need to be made using only the post-convergence samples of $\{\tau, \theta, \beta\}$.

To estimate the latent process at an unobserved input $x_0 \in \mathbb{R}^p$ for a given group $c_j \in \mathscr{C}$, we evaluate the Bayesian posterior predictive distribution

$$p(Z(x_0; c_j) \,|\, \{y(x_i; c_j)\}, \{f_j(x_i)\}) \propto \int p(Z(x_0; c_j) \,|\, Z, \theta) \times p(Z, \{\tau, \theta, \beta\} \,|\, y, F) dZ d\{\tau, \theta, \beta\} \ , \tag{4}$$

where we use the conditional independence $p(Z(x_0; c_j) \,|\, Z, \{\tau, \theta, \beta\}, y, F) = p(Z(x_0; c_j) \,|\, Z, \theta)$ derived from the hierarchical model in Equation (2). We sample from Equation (4) by drawing one $Z(x_0; c_j) \sim p(Z(x_0; c_j) \,|\, Z, \theta)$ for each drawn posterior sample of Z and θ, where $p(Z(x_0; c_j) \,|\, Z, \theta)$ is Gaussian with mean $K_\theta((x_0; c_j); \cdot)^{\mathrm{T}} K_\theta^{-1} Z$, $K_\theta((x_0; c_j); \cdot)$ is the $n \times 1$ vector with elements $K_\theta((x_0; c_j), (x_i, c_{j'}))$ for $j' = 1, \ldots, k$ and $i = 1, 2, \ldots, n_{j'}$, and variance $K_\theta((x_0; c_j), (x_0; c_j)) - K_\theta((x_0; c_j); \cdot)^{\mathrm{T}} K_\theta^{-1} K_\theta((x_0; c_j); \cdot)$. To predict $Y(x_0; c_j)$, we sample from the predictive distribution $p(Y(x_0; c_j) \,|\, y, F)$ by drawing one $Y(x_0; c_j) \sim N(f_j(x_0)^{\mathrm{T}} \beta_j + Z(x_0; c_j), \tau_j^2)$ for each posterior sample of $\{\beta_j, \tau_j^2\}$ and $Z(x_0; c_j)$.

We need valid positive-definite functions to serve as $K_\theta((x; c_j), (x'; c_{j'}))$. This is crucial for the above inferential framework as it ensures that the matrix K_θ in Equation (2) will be positive definite for any finite set of distinct elements, observed or unobserved, in $\mathbb{R} \times \mathscr{C}$. An advantage of driving the inference through a latent process is the convenience of predictive inference for the underlying process and the response at new inputs. Therefore, we focus upon the construction of valid covariance functions to specify MGGPs.

The computational bottleneck arises from the dimension of K_θ in GP models for large data sets. There is a substantial literature on various approaches to build models that scale up to massive data sets by building low-rank or sparsity-inducing processes (see, e.g. Wikle, 2010; Banerjee, 2017; Heaton et al., 2019, for expository treatments) from any valid covariance function. While our current focus is not specifically on processes that scale inference to massive data sets, we note that constructing MGGPs using valid covariance functions renders the resulting processes as "scalable-ready" since low-rank or sparsity-inducing variants may be derived using existing approaches.

3. Multi-Group Gaussian Processes

Proofs of all theoretical results presented below are in the Appendix.

3.1 Separable multi-group GPs

We start with a simple case where the covariance function over $\mathbb{R}^p \times \mathscr{C}$ is separable.

Definition 1 *K is said to be separable if $K((x, c_i), (x', c_j)) = K_{\mathbb{R}^p}(x, x')K_{\mathscr{C}}(c_i, c_j)$, where $K_{\mathbb{R}^p}$ and $K_{\mathscr{C}}$ are over $\mathbb{R}^p$ and $\mathscr{C}$, respectively.*

Note that K is positive definite if and only if both $K_{\mathbb{R}^p}$ and $K_{\mathscr{C}}$ are positive definite. Constructing valid GPs over $\mathbb{R}^p$ is well known, so we focus on covariance functions $K_{\mathscr{C}}$, i.e., GPs over a categorical set. Also, $\mathscr{C}$ being finite, any function on $\mathscr{C} \times \mathscr{C}$ is completely determined by the $k \times k$ positive definite matrix C with elements $C_{ij} = K_{\mathscr{C}}(c_i, c_j)$.

Proposition 2 *$K_{\mathscr{C}}$ is positive definite if and only if C is a positive definite matrix.*

Thus, a positive definite function on $\mathbb{R}^p$ and a positive definite matrix $C \in \mathbb{R}^{k\times k}$ ensures a separable positive definite function on $\mathcal{X}$. Homogeneous kernels arise as a special case.

Definition 3 *A function $K_{\mathscr{C}} : \mathscr{C} \times \mathscr{C} \to \mathbb{R}$ is said to be homogeneous if $K_{\mathscr{C}}(c_i, c_j) = K_0(1_{\{c_i \neq c_j\}})$ for some function K_0 on $\{0, 1\}$.*

A homogeneous process is completely determined by two scalars $a := K_{\mathscr{C}}(c_i, c_i)$, $b := K_{\mathscr{C}}(c_i, c_j)$ with $c_i \neq c_j$, which represent the within-group and across-group associations, respectively. Without loss of generality, we assume $a = 1$; otherwise, we can rescale $K_{\mathscr{C}}$. In this case, both within-group and between-group correlations are constants. A homogeneous process is appropriate if we only want to distinguish pairs of observations in the same group from those in different groups, while the specific group identities are irrelevant. Hence, K is homogeneous if it is isotropic with respect to the discrete metric $d(c_i, c_j) = 1_{\{c_i \neq c_j\}}$.

Corollary 4 *Let $K_{\mathscr{C}}$ be homogeneous, then $K_{\mathscr{C}}$ is positive definite if and only if $-\frac{1}{k-1} \leq b \leq 1$, where $b = K_{\mathscr{C}}(c_i, c_j)$ with $i \neq j$.*

The inequality $-\frac{1}{k-1} \leq b \leq 1$ implies that across-group correlations should not dominate the within-group correlations, which is intuitively reasonable. Separable models provide computational benefits because the resulting covariance matrix for Z can be expressed as a Kronecker product $K_{\mathbb{R}} \otimes K_{\mathscr{C}}$. However, such covariance functions tend to have "ridges" or discontinuities (Stein, 2005) that lead to worse inference. They also assume that the same covariance structure ($K_{\mathbb{R}^p}$) is retained for all groups, which is restrictive in terms of accommodating associations for different pairs of inputs in $\mathbb{R}^p \times \mathscr{C}$.

3.2 Isotropic multi-group GPs

In order to discuss "isotropic" covariance functions on $\mathbb{R}^p \times \mathscr{C}$, we endow $\mathscr{C}$ with additional structure. To facilitate our development, we introduce a metric d on $\mathscr{C}$ so that $(\mathscr{C}, d)$ is a metric space.

Definition 5 *Given two metric spaces $(\mathcal{Y}, d)$ and $(\mathcal{Y}', d')$, a GP on $\mathcal{Y} \times \mathcal{Y}'$ is said to be semi-isotropic if $K((x_1, x_1'), (x_2, x_2')) = K_0(d(x_1, x_2), d'(x_1', x_2'))$.*

Intuitively, a semi-isotropic process is isotropic in $\mathbb{R}^d$ and $\mathscr{C}$ separately. Isotropy implies semi-isotropy, but the other direction does not hold in general. In practice, d is usually obtained from domain knowledge including prior, exterior, or expert knowledge. Thus, if $\mathscr{C} = \{\text{North Carolina, New Jersey, California}\}$, then the distance can be the geographical distance between the centroids of these states. As another example, if $\mathscr{C}$ represents human tissue types, then d can be constructed using prior biomedical knowledge; two tissue types from the same organ (e.g., brain) might tend to be more similar to each other than from different organs (e.g., brain and liver). If $\mathscr{C}$ is a weighted graph, then the graph distance serves as a valid metric (Bouttier et al., 2003). If domain knowledge is unavailable, a default noninformative distance, $d_{ij} = 1 - \delta_{ij}$, implying that all groups are equidistant, can be adopted. Extensions to unknown d_{ij}, which are instead treated as parameters, are discussed in Appendix I. Our next result creates a large family of semi-isotropic covariance functions on $\mathcal{X} = \mathbb{R}^p \times \mathscr{C}$.

Theorem 6 *Assume the Gram matrix defined as $G_{ij} := \frac{1}{2}\left(d(c_1, c_i) + d(c_1, c_j) - d(c_i, c_j)\right)$ is positive semi-definite, then if $\varphi : \mathbb{R}_+ \to \mathbb{R}$ is a completely monotone function and $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is a positive function with a completely monotone derivative, then:*

$$K((x, c_i), (x', c_j)) = \frac{\sigma^2}{\left(\psi(d_{ij}^2)\right)^{p/2}} \varphi\left(\frac{\|x - x'\|^2}{\psi(d_{ij}^2)}\right). \tag{5}$$

is a valid covariance function, where $\sigma^2 > 0$ is the spatial variance. In particular, if $d(c_i, c_j) = 1 - \delta_{ij}$ is the discrete metric, then

$$K((x, c_i), (x', c_j)) = \frac{\sigma^2}{\alpha^{\frac{p}{2}(1-\delta_{ij})}} \varphi\left(\frac{\|x - x'\|^2}{\alpha^{1-\delta_{ij}}}\right) \tag{6}$$

is a valid covariance function, where $\sigma^2 \in (0, 1]$ is the spatial variance and $\alpha > 0$ controls the interaction between $\mathbb{R}^p$ and $\mathscr{C}$.

A simple form for G emerges when $d(c_i, c_j) = 1 - \delta_{ij}$. The resulting Gram matrix is $G = \begin{pmatrix} 0 & 0_{1\times(k-1)} \\ 0_{(k-1)\times 1} & \widetilde{G} \end{pmatrix}$, where $\widetilde{G} = \frac{1}{2}\mathrm{Id}_{k-1} + \frac{1}{2}1_{(k-1)\times(k-1)}$ and $1_{m\times n}$ denotes the $m \times n$ matrix of ones. Some candidates for completely monotone functions ϕ and positive functions with completely monotone derivatives ψ are in Table 1 (Gneiting, 2002). This class of covariance functions is also known as the "Gneiting class". Selecting ϕ and ψ

$\phi(t)$	$\psi(t)$
$\exp(-ct^{\gamma})$	$(at^{\alpha}+1)^{\beta}$
$(2^{\nu-1}\Gamma(\nu))^{-1}(ct^{1/2})^{\nu}K_{\nu}(ct^{1/2})$	$\log(at^{\alpha}+b)/\log b$
$(1+ct^{\gamma})^{-\nu}$	$(at^{\alpha}+\beta)/(\beta(at^{\alpha}+1))$
$2^{\nu}(\exp(ct^{1/2})+\exp(-ct^{1/2}))^{-\nu}$	

Table 1: **Candidate functions for completely monotone functions ϕ and positive functions with completely monotone derivatives ψ.** Here, $a, c, \nu > 0$, $b > 1$, $0 < \alpha, \beta, \gamma \leq 1$.

from Table 1, we obtain the following semi-isotropic covariance functions on $\mathscr{X}$ (for more covariance functions, see Section H):

$$K((x,c_i),(x',c_j)) = \frac{\sigma^2}{(a^2d_{ij}^2+1)^{p/2}}\exp\left\{-\frac{b^2\|x-x'\|^2}{a^2d_{ij}^2+1}\right\}, \tag{7}$$

$$K((x,c_i),(x',c_j)) = \begin{cases} \frac{\sigma^2 2c^{p/2}}{(a^2d_{ij}^2+1)^{\nu}(a^2d_{ij}^2+c)^{p/2}\Gamma(\nu)}\left\{\frac{b}{2}\left(\frac{a^2d_{ij}^2+1}{a^2d_{ij}^2+c}\right)^{1/2}\|x-x'\|\right\}^{\nu} & \\ \times K_{\nu}\left(b\left(\frac{a^2d_{ij}^2+1}{a^2d_{ij}^2+c}\right)^{1/2}\|x-x'\|\right) & x \neq x' \\ \frac{\sigma^2 c^{p/2}}{(a^2d_{ij}^2+1)^{\nu}(a^2d_{ij}^2+c)^{p/2}} & x = x' \end{cases}, \tag{8}$$

$$K((x,c_i),(x',c_j)) = \frac{\sigma^2 c^{p/2}}{(a^2d_{ij}^2+1)^{1/2}(a^2d_{ij}^2+c)^{p/2}}\exp\left\{-b\left(\frac{a^2d_{ij}^2+1}{a^2d_{ij}^2+c}\right)^{1/2}\|x-x'\|\right\} \tag{9}$$

In the above functions, $\sigma^2 > 0$ is the spatial variance, $a \geq 0$ is the group similarity scale, $b \geq 0$ is the feature scale, $c \geq 0$ is the separability scale, and $\nu > 0$ is a smoothness parameter. The covariance function in Equation (7) is analogous to the squared exponential or radial basis functions (RBFs). The covariance function in Equation (8) is the analogue of the Matérn covariance function. In particular, the covariance function in Equation (9) is a special case of Equation (8) when $\nu = 1/2$, which is the exponential covariance function. The covariance function in Equation (8) becomes separable when $c = 1$. We supply a table summarizing the kernel constructions in Appendix H.

It is important to clarify that ϕ and ψ are legitimate choices within the Gneiting class, and modelers can select one independent of the other from the provided list. The final decision on these parameters should be based on the specific problem at hand. We also remark that our proposed framework is broader than the Gneiting class, and we do not imply that the Gneiting class encompasses all possible choices. Appendix H provides more examples of kernels within and beyond the Gneiting class.

3.3 Stationary multi-group GPs

We now weaken isotropy. Since $(\mathscr{C}, d)$ does not admit a natural algebraic structure, we start with $k = 2$, which appears frequently in practice including in data sets where the two groups are male/female, adults/children, treatment/control, and so on. For the non-isotropic case, $\mathscr{C}$ can be identified with $\mathbb{Z}_2$ when $k = 2$, an Abelian group. In this setting, K is said to be stationary if $K((x, d), (x', l)) = K_0(x - x', d - l)$. We use K instead of K_0 for simplicity, where K is characterized by $K_w = K(\cdot, 0)$, the within-group covariance function, and $K_c = K(\cdot, 1)$, the cross-group covariance function. Any covariance function on $\mathbb{R}^p \times \mathbb{Z}_2$ determines two covariance functions on $\mathbb{R}^p$. On the other hand, not all pairs of covariance functions on $\mathbb{R}^p$ define a valid covariance function on $\mathbb{R}^p \times \mathbb{Z}_2$. In order to construct a valid covariance function on $\mathbb{R}^p \times \mathbb{Z}_2$, we need a sufficient condition for K to be positive definite.

Theorem 7 *Let K_w and K_c be two positive definite functions on $\mathbb{R}^p$ with spectral densities ρ_w and ρ_c such that $K(x, 0) = K_w(x) = \int_{\mathbb{R}^p} e^{-2\pi i\omega x}\rho_w(\omega)d\omega$, $K(x, 1) = K_c(x) = \int_{\mathbb{R}^p} e^{-2\pi i\omega x}\rho_c(\omega)d\omega$. Then, $K(x, l) = \begin{cases} K_w(x) & l = 0 \\ K_c(x) & l = 1 \end{cases}$ is positive definite on $\mathbb{R}^p \times \mathbb{Z}_2$ if and only if $\rho_w \geq \rho_c$.*

Example 1 *Recall the multi-group RBF in Equation (7). The two spectral densities are $\rho_w(\omega) = \sigma^2 \left(\frac{\pi}{b^2}\right)^{\frac{p}{2}} \exp\left\{-\frac{\pi^2\|\omega\|^2}{b^2}\right\}$ and $\rho_c(\omega) = \sigma^2 \left(\frac{\pi}{b^2}\right)^{\frac{p}{2}} \exp\left\{-\frac{\pi^2(a^2+1)\|\omega\|^2}{b^2}\right\}$, where $\rho_w \geq \rho_c$.*

The stationary MGGP assumes homogeneity in the within-group correlation. To account for heterogeneity, we introduce a weaker semi-stationary MGGP. This semi-stationary process is stationary in $\mathbb{R}^p$ but not in $\mathscr{C}$.

Definition 8 *K is said to be semi-stationary if $K((x, c_i), (x', c_j)) = K_0(x - x', c_i, c_j)$ where K_0 is defined on $\mathbb{R}^p \times \mathscr{C} \times \mathscr{C}$.*

A semi-stationary process is appropriate for applications where groups are expected to have different within-group correlations, but the process is stationary once the group is fixed. For semi-stationary MGGPs, K is determined by $K_0(x) = K(x, 0, 0)$, $K_c(x) = K(x, 0, 1) = K(x, 1, 0)$ and $K_1 = K(x, 1, 1)$, where $K_0 \neq K_1$ in general; otherwise K becomes stationary.

Theorem 9 *Let K_0, K_c and K_1 be positive definite functions on $\mathbb{R}^p$ with spectral densities ρ_0, ρ_c, and ρ_1. Then $K(x, l, l') = \begin{cases} K_0(x) & l = l' = 0 \\ K_c(x) & l + l' = 1 \\ K_1(x) & l = l' = 1 \end{cases}$ is positive definite on $\mathbb{R}^p \times \mathbb{Z}_2$ if and only if $\rho_0\rho_1 \geq \rho_c^2$.*

Data sets with more than two groups are ubiquitous in scientific applications. Hence, we generalize the above theory to $k > 2$ groups. The difficulty here is that $\mathscr{C}$ does not admit a natural group structure for $k > 2$. A straightforward solution would be to identify $\mathscr{C}$ with $\mathbb{Z}_k$, but the modular structure of $\mathbb{Z}_k$, i.e., $1 - 0 = 2 - 1 = \cdots k - 1 - (k - 1 - 1) \neq k - (k - 1)$, is not satisfied in practice. Hence, Bochner's Theorem (Rudin, 2017), which characterizes positive definite functions on locally compact Abelian groups, is not applicable. Theorem 9 draws an equivalence between Bochner's Theorem on $\mathbb{R}^p \times \mathbb{Z}_2$ and Cramér's Theorem on $\mathbb{R}^p$ with $k = 2$. Hence, we can use bivariate GPs to construct two-group GPs. Furthermore, given that Cramér's Theorem (Cramér, 1940) holds for a general k-variate GP, we can develop a general theory for an arbitrary number of groups with $k > 2$, which we do below. We draw similarities between the MGGP with k groups and k-variate GPs, also known as multi-task GPs. Recall that a k-variate random field $\widetilde{Z}$ on $\mathcal{Y}$ is characterized by its cross-covariance function $\widetilde{K} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^{k \times k}$: $\mathrm{Cov}(\widetilde{Z}(x), \widetilde{Z}(x')) = \widetilde{K}(x, x')$.

Theorem 10 *Let $\mathscr{G}$ be the space of all Gaussian random fields on $\mathcal{Y} \times \mathscr{C}$, where $\mathscr{C} = \{c_1, \cdots, c_k\}$ and $\mathscr{V}$ is the space of all Gaussian k-variate random fields on $\mathcal{Y}$. Then $\Phi : \mathscr{G} \to \mathscr{V}$, $(\Phi(Z))_i(x) \coloneqq Z(x, c_i)$, $\forall Z \in \mathscr{G}$ is a bijection, and its inverse Φ^{-1} is given by $\Phi^{-1} : \mathscr{V} \to \mathscr{G}$, $(\Phi^{-1}(\widetilde{Z}))(x, c_i) = \widetilde{Z}_i(x)$, $\forall \widetilde{Z} \in \mathscr{V}$. The correspondence between the covariance function of Z and the cross-covariance function of $\widetilde{Z}$ is given by $K((x, c_i), (x', c_j)) = \widetilde{K}(x, x')_{ij}$.*

Therefore, constructing a k-variate GP will produce a k-group GP, and vice versa. Existing constructions of multivariate GPs can be applied (see, e.g., Cressie and Huang, 1999; Gneiting et al., 2010; Apanasovich and Genton, 2010; Gelfand and Banerjee, 2010; Genton and Kleiber, 2015). While MGGPs and multi-task GPs are mathematically equivalent, they focus on different aspects of statistical learning. The multi-task GP focuses on predicting multiple tasks simultaneously by borrowing information across groups. On the other hand, the MGGP models multiple groups that may or may not have shared underlying structure by learning the kernel parameters explicitly. We prove the following related result.

Theorem 11 *Let $K : \mathbb{R}^p \times \mathscr{C} \times \mathscr{C} \to \mathbb{R}$ be a function with $K_{ij} = K(\cdot, c_i, c_j)$ being stationary on $\mathbb{R}^p$ and spectral densities ρ_{ij}. Then, K is positive definite, hence defines a semi-stationary GP on $\mathbb{R}^p \times \mathscr{C}$, if and only if $\rho(\omega) = \{\rho(\omega)\}_{i,j=1}^k$ is positive semi-definite for any $\omega \in \mathbb{R}^p$.*

Theorem 9 is a special case of Theorem 11 when $k = 2$, but can be proved differently. As a result, the connection between Bochner's Theorem on $\mathbb{R}^p \times \mathscr{C}$ and Cramér's Theorem on $\mathbb{R}^p$ is analogous to the relationship between multi-group and multivariate GPs. The MGGP is a non-trivial generalization of existing processes that allows substantial group heterogeneity by accommodating a variety of flexible covariance functions.

3.4 Multivariate multi-group Gaussian processes

The construction of the MGGP can be extended to multivariate, or multi-output GPs. Let Z be a k'-variate GP on $\mathbb{R}^p \times \mathscr{C}$ with cross-covariance function $K : \mathbb{R}^p \times \mathscr{C} \times \mathbb{R}^p \times \mathscr{C} \to \mathbb{R}^{k' \times k'}$, that is, $\mathrm{Cov}(Z(x, c_i), Z(x', c_j)) = K((x, c_i), (x, c_j))$. Similar to the construction of the MGGP, first we assume that there exists a metric d' between output variables with a positive semi-definite Gram matrix.

Theorem 12 *If $\varphi : \mathbb{R}_+ \to \mathbb{R}$ is a completely monotone function and $\psi_1, \psi_2 : \mathbb{R}_+ \to \mathbb{R}_+$ are positive functions with completely monotone derivatives, then*

$$K((x, c_i), (x', c_j))_{kl} = \frac{\sigma^2}{\left(\psi_1\left(\frac{d_{ij}^2}{\psi_2(d'^2_{kl})}\right)\right)^{p/2} \left(\psi_2(d'^2_{kl})\right)^{1/2}} \varphi\left(\frac{\|x - x'\|^2}{\psi_1\left(\frac{d_{ij}^2}{\psi_2(d'^2_{jl})}\right)}\right)$$

is a valid cross-covariance function, where $\sigma^2 > 0$ is the spatial variance.

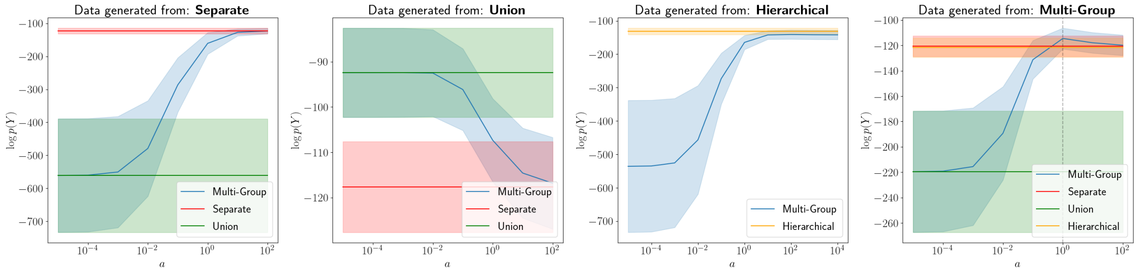

Figure 1: **Comparison between the *Multi-Group*, *Separate*, *Union*, and *Hierarchical* process models**. Using two-group data generated from each of the four models, we computed the log marginal likelihood of the data under each model. For the Multi-Group model, we used the covariance function in Equation (7) and used a range of different values the parameter a. In the rightmost plot, the dashed vertical line indicates the true value of a used for data generation. We used an RBF kernel, which does not have an a parameter, for the *Separate* and *Union* models. We repeated this experiment 20 times, and the bands in each plot represent 95% confidence intervals.

4. Simulations

4.1 Comparing the MGGP with related models on simulations

We conducted an experiment to assess the MGGP's ability to recover the Separate, Union, and Hierarchical Gaussian processes as special cases. We generated data from each of these models using Equation (1) with $k = 2$ groups. We specified a zero mean, i.e., $\mu(x; c_j) = 0$ for both groups, and specified the latent process using covariance functions for the three models. We set $b = \sigma^2 = a = 1$ in Equation (7). (Note that a is only used in the generation of data from the MGGP.) We also assumed $\tau_1^2 = \tau_2^2 = \tau^2$ in Equation (1) and used $\tau^2 = 0.1$ to generate our data. Using these settings, we generated $n_1 = n_2 = 100$ measurements for each group.

We computed the log marginal likelihood of the data, i.e., $N(y \,|\, 0, K_\theta + D_\tau)$, under each model for each data set. For the SGP, UGP and HGP, we used the RBF, $K(x, x') = \sigma^2 \exp\{-b^2 \|x - x'\|^2\}$. For the Multi-Group model we used the "multi-group" RBF in Equation (7). When computing the likelihood under each model, we fix b, σ^2, and τ^2 to their true values; for a we use a grid of values, $a = 10^{-5}, 10^{-4}, \ldots, 10^{2}$, and we specify $D_\tau = \tau^2 I_n$.

Our MGGP performs on par with the SGP, UGP and HGP in the expected regimes (Figure 1). The MGGP matches the performance (as measured by the log marginal likelihood) of the SGP when a is large, and the MGGP matches the performance of the UGP as $a \to 0$. For data generated from the MGGP, we find that the likelihood peaks at the true value of a and is higher than all other models at this value. These results i) serve as a demonstration of the role of a; ii) confirm numerically that the MGGP recovers these alternative models in certain regimes; and iii) suggest that the MGGP is a viable generalization of the other three models.

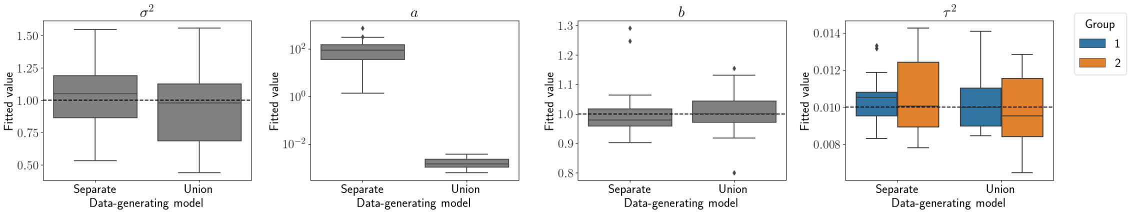

Figure 2: **Covariance function parameter estimation**. Using data generated from the Separate and Union processes, we fit the Multi-Group process by finding the MLEs for the true parameters of the kernel function in Equation (7). The boxes cover the interquartile range; the lower border of the box is the 25th percentile, the upper border of the box is the 75th percentile, and the middle line is the median. The whiskers extend 1.5 times the interquartile range in each direction.

4.2 Estimation and inference for the MGGP

In our previous experiment, we used the multi-group covariance function in Equation (7) with a fixed value of a. In practice, we will need to estimate a and all other covariance parameters from the data. Next, we assess the parameter estimates of the MGGP using both maximum likelihood and fully Bayesian posterior inference.

We first conducted an experiment where we generated data from the SGP and UGP as in the previous section. We maximize the collapsed or marginalized likelihood corresponding to Equation (1), i.e., $\mathcal{N}(y \,|\, 0, K_\theta + \tau^2 I_n)$, with respect to $\theta = \{a, b, \sigma^2\}$, and a common measurement error variance τ^2, where θ corresponds to the three parameters in the multi-group covariance function in Equation (7). We used a conjugate gradient ascent algorithm (Nocedal and Wright, 2006) to obtain the joint estimates of $\{\theta, \tau^2\}$ and executed the algorithm in Python using the JAX software framework (Bradbury et al., 2018) designed for fast computation, compilation, and automatic differentiation. Experiments were run on an internal computing cluster using a 320 NVIDIA P100 Graphical Processing Unit. The maximum likelihood estimates (MLEs) for a were consistently high for data generated from the SGP and low for the UGP data, as expected (Figure 2, middle panel). Additionally, our estimation was able to capture the true values for σ^2 and b (Figure 2, left and right panels).

Next, we generated four data sets from the MGGP for $a \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$ with sample size fixed at 100. We optimized all parameters jointly by maximizing the marginal multi-group likelihood and examined the estimated value of a for each. We repeated this experiment ten times and found that we could consistently estimate a reasonable value of a (Figure 3). While the estimated values did not exactly coincide with the true values, they showed a desirable monotone relationship. These results reveal that likelihood-based parameter estimation is feasible in the multi-group model and that existing estimation and computational algorithms, such as gradient ascent, can be successfully applied to multi-group models. A formal proof of the consistency of the MLE for a and other kernel parameters, as well as the consistency of the posterior distribution, remains challenging. Consistently estimating GP kernel parameters, even in Euclidean domains, is well-recognized in the lit-

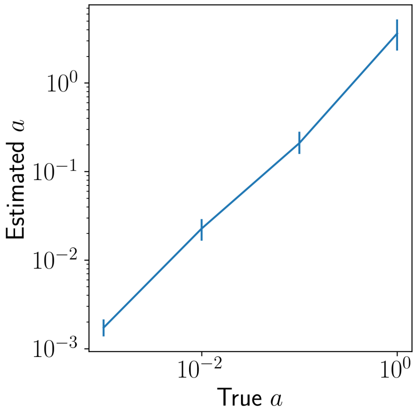

Figure 3: **Recovering a with the MGGP MLE**. We generated synthetic data from the MGGP at different values for a and subsequently fit the MGGP to these data. We fix all other parameters to their true values. We are able to recover a close approximation of the true value of a.

erature as a difficult problem (Zhang, 2004; Tang et al., 2021; Li, 2022; Loh and Sun, 2023). This issue remains an important area for future research, as discussed in Section 6.

Turning to the full Bayesian analysis, we generated a simulated data set in the same manner as above for the MGGP described in Section 2. We include group-specific intercepts for the $k = 2$ groups, denoted as $\beta = (\beta_1, \beta_2)^{\mathrm{T}}$, and use $\beta_1 = 1, \beta_2 = 2$ along with group-specific noise variances $\tau_1^2 = 0.1, \tau_2^2 = 0.3$ to generate the data. We form the $(n_1 + n_2) \times 2$ binary design matrix F in order to apply the group-specific intercept in the model. For computational efficiency, we fit the collapsed posterior distribution in Equation (3). With $\theta = \{a, b, \sigma^2\}$, the prior distribution in Equation (3) is specified as

$$\begin{aligned} p(\theta, \{\tau_1^2, \tau_2^2\}, \beta) = {} & IG(a \mid \alpha_a, \alpha'_a) \times IG(b \mid \alpha_b, \alpha'_b) \times IG(\sigma^2 \mid \alpha_\sigma, \alpha'_\sigma) \\ & \times \prod_{j=1}^{2} IG(\tau_j^2 \mid \alpha_{\tau_j}, \alpha'_{\tau_j}) \times N(\beta \mid \mu_\beta, V_\beta), \end{aligned} \tag{10}$$

where we set $\alpha_a = \alpha'_a = \alpha_b = \alpha'_b = \alpha_{\tau_1} = \alpha'_{\tau_1} = \alpha_{\tau_2} = \alpha'_{\tau_2} = 5$, $\alpha_\sigma = \alpha'_\sigma = 1$, $\mu_\beta = 0$ and $V_\beta^{-1} = \mathrm{I}$. We set the values of these parameters for simplicity, but note that, in practice, one may use external information, if available, to elicit prior information. For example, a shrinkage prior on a could be used if there is evidence that the groups are similar.

For inference, we sample from the posterior distribution in Equation (3) using a Hamiltonian No U-Turn Sampling (Hoffman et al., 2014) algorithm as implemented in the `Stan` programming environment (Stan Development Team, 2020; Riddell et al., 2021). We ran four chains with dispersed initial values for $1,200$ iterations each. Convergence was diagnosed after 200 iterations using visual inspection of autocorrelation plots (Figure 10) and computation of Gelman-Rubin R-hat and Monte Carlo standard errors. The subsequent $4,000$ samples were retained for posterior inference.

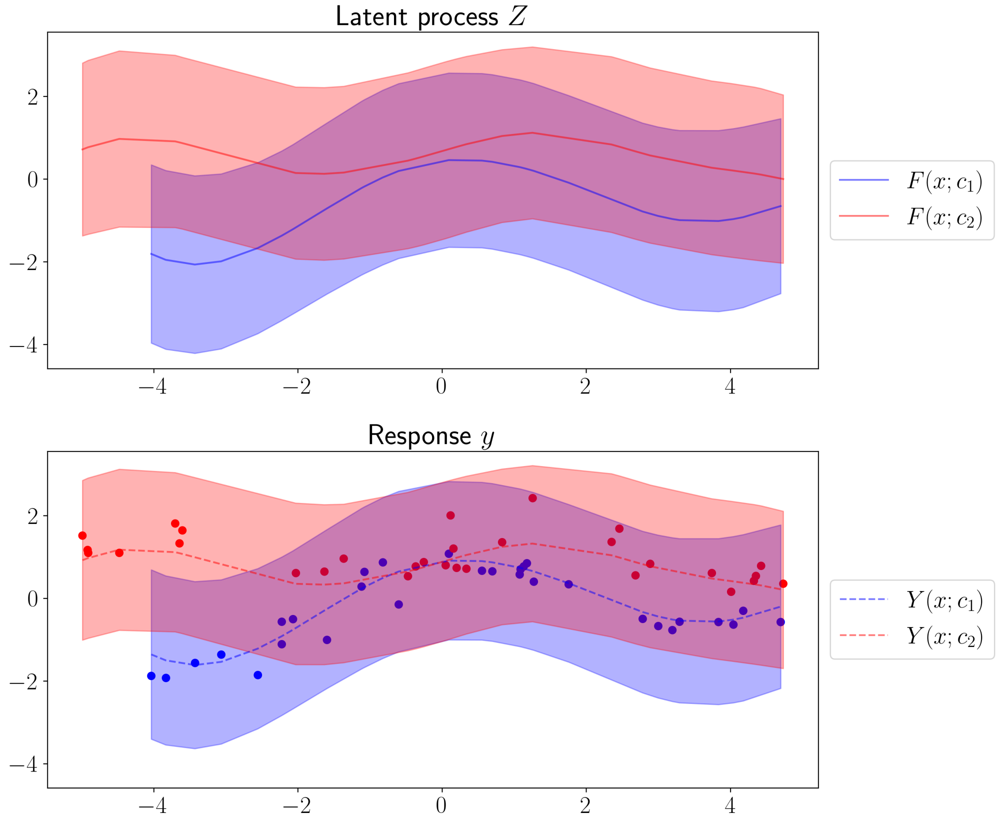

Figure 4: **Posterior predictive distribution from the multi-group Gaussian process**. The points represent training data; the solid lines show the means of the latent processes $F(x; c)$; the dashed lines represent the predictive means of $Y(x; c)$; and the shaded areas around the lines are twice the standard deviation of the posterior predictive distribution at the corresponding input points.

The posterior median and 95% credible intervals show that the covariance function parameters capture their true values (Figure 11, Table 2). We also sample from the posterior predictive distribution, $p(Y(x_0; c_j) \,|\, y, F)$ (see Section 2), for a collection of new inputs or test cases (Figure 4). Because all of the MGGP assumptions are encoded in the covariance function, any appropriate method for estimation and inference in standard GPs can be applied.

We next evaluated the MGGP in terms of predicting held-out values in a GP regression task. We generated data from a GP regression model, as in Equation (1), using the SGP, UGP, HGP, and MGGP. We fit these models to each of the data sets using 50% of the data for training, and we test our predictions over the remaining data. We use the predictive mean $\mu^\star = K_{X^\star X} K_{XX}^{-1} y$ as a point prediction for each of the $n^\star$ held-out samples, where $K_{X^\star X}$ is the $n^\star \times n$ matrix of covariance function evaluations for each pair of test and training samples, and K_{XX} is the $n \times n$ matrix of covariance function evaluations for each pair of training samples. We center the data for each group around their mean. We compute

Parameter	True	Posterior percentiles
a	1.0	1.66 (0.82; 4.29)
b	1.0	0.47 (0.29; 0.88)
σ^2	1.0	1.41 (0.76; 2.98)
τ_1	0.1	0.14 (0.09; 0.21)
τ_2	0.3	0.29 (0.2; 0.47)
β_1	1.0	0.73 (-0.38; 1.66)
β_2	2.0	1.86 (0.77; 2.8)

Table 2: **Parameter posterior summaries for simulated Bayesian analysis.** Posterior summaries are presented as $50(2.5; 97.5)$ percentiles in the third column.

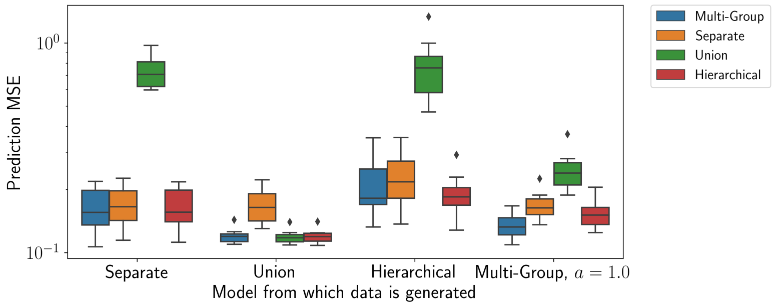

Figure 5: **GP predictions with simulated data**. We generate data from each of the four models—SGP, UGP, HGP and MGGP—and fit each of these models to all data sets. Prediction error (MSE) was computed on a held-out data set. Figure 9 shows an analogous experiment using the Matérn covariance function.

the mean squared error (MSE) of the predictions, $E = \frac{1}{n^\star}\sum_{i=1}^{n^\star}(y_i - \mu_i^\star)^2$ to evaluate the quality of predictive inference.

We find that the MGGP emulates the performance of the SGP, UGP and HGP on their respective simulated data sets (Figure 5). With data generated from the MGGP itself, the MGGP substantially outperforms the other models. While the SGP can be expected to perform well when each group has a large sample size, a primary benefit of MGGPs is their ability to share information across similar groups when the (group) sample size is limited. Thus, we expect MGGPs to excel over SGP when some groups have a small number of samples, but are closely related to other groups.

To test this claim, we conducted another multi-group regression experiment in which we sought to predict the held-out values for one group that contained few samples. Specifically, we generated synthetic data consisting of three groups, where group 1 and group 2 are similar to one another, and group 3 is dissimilar from the other two. We then generate a series of data sets, with sample size of group 2 and 3 being 50 and varying the number of samples in group 1 to take values in $\{5, 10, 30, 50\}$. Then, we fit the SGP, UGP, HGP and MGGP to each of the data sets, using 50% of the data for training, and testing predictions on the other 50%. We find that the MGGP model outperforms the other methods, especially when

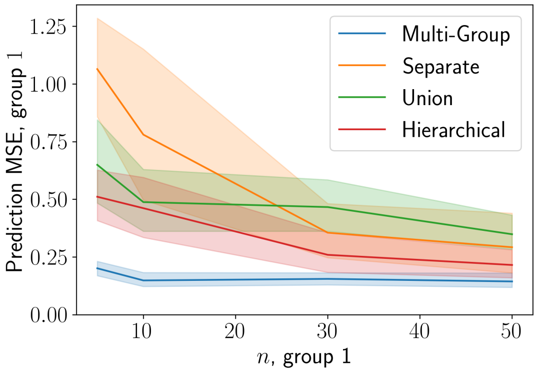

Figure 6: **Prediction using simulated data with imbalanced groups**. We perform a prediction experiment with $k = 3$ groups. To generate a series of data sets, we fix the sample size of groups c_2 and c_3 to be 50, and we vary the sample size of group c_1.

the sample size for group 1 is small (Figure 6). This result shows that the MGGP appears to thrive when the sample size for some groups is limited, as it most effectively leverages information from similar groups while acknowledging the group structure.

4.3 Partially observed coordinates

A key benefit of our process-based modeling framework is its versatility in dealing with situations where we have partially observed coordinates. This setting arises in situations where not all group labels $c_j \in \mathcal{C}$ have yielded measurements on an identical set of $x_j \in \mathcal{X}$, which leads to an imbalance. Despite this, inference still proceeds seamlessly using the framework described in Section 2. In fact, the simulated data in our experiments reflect this exact setup: the x_js within each group rarely overlapped across groups, meaning that almost all x_js were exclusive to a single group. For example, the observed measurements for two groups in our previous example have minimal overlap (Figure 4), and the MSE values (Figure 5) were computed under this disjoint groups of x_js setup.

5. Application to GTEx tissue samples

We applied the MGGP to a large gene expression data set collected by the Genotype-Tissue Expression (GTEx) project (Consortium et al., 2020). The GTEx v8 data contain measurements from 17,382 samples that span 52 tissue types collected from 838 human donors; see Appendix K for a full list of tissue types and the sample size for each tissue. Along with gene expression profiling, a variety of additional metadata characteristics are collected, including demographic variables and tissue health measurements.

In these experiments, we use GP regression models to analyze the relationship between a sample's gene expression profile and its *ischemic time*–the duration of time between death and tissue collection. Previous work has shown a robust relationship between gene expression and ischemic time (Musella et al., 2013; Ferreira et al., 2018); however, whether this relationship exhibits tissue-specific patterns remains largely untested. In these experiments, the groups correspond to tissue of origin for each sample.

As an initial test with the GTEx data, we applied the MGGP to samples from just two tissue types at a time. These experiments aim to validate that MGGP regression can appropriately model known associations across similar groups and estimate pairwise similarity between groups.

In a preliminary experiment, we examined three tissue types: anterior cingulate cortex ($n = 172$), frontal cortex ($n = 200$), and coronary artery ($n = 238$). First, for each of the three pairs of tissues, we fit the MGGP model with MLEs, as described in Section 4.2, using the multi-group RBF covariance function (Equation (7)). In this experiment, we fixed a to one value in a preset range, and found the MLEs of the remaining parameters. This experiment aims to justify our interpretation of a using a real data set where we know the similarity between certain groups (i.e., tissues). In practice, a is estimated using an MLE in all other simulations and applications. Using these MLEs and the fixed a, we then computed the log marginal likelihood of the data, $\log p(y \mid X, a, \widehat{b}, \widehat{\sigma^2}, \widehat{\tau^2}) = -\frac{k}{2} 2\pi - \frac{1}{2}\det(K_{XX} + \widehat{\tau^2} I) - \frac{1}{2} y^{\mathrm{T}} (K_{XX} + \widehat{\tau^2} I)^{-1} y$, where K_{XX} is the $(n_1 + n_2) \times (n_1 + n_2)$ matrix of covariance function evaluations for each pair of samples. We also fit the SGP and UGP for each pair of tissues using the standard RBF, and computed the log marginal likelihood of the data under these models.

Examining the log marginal likelihood across varying values of a (Figure 7), we found that two brain tissue types that are expected to be similar to one another—anterior cingulate and frontal cortex—showed a higher marginal likelihood under small values of a ($a \lessapprox 0.01$), while tissues that have unique expression patterns—anterior cingulate cortex and coronary artery—showed a higher marginal likelihood under large values of a ($a \gtrapprox 10$). In both cases, the MGGP gracefully recovered the Separate and Union marginal likelihoods for $a \to \infty$ and $a \to 0$, respectively. This result implies that MGGP is a viable strategy not only for sharing information across groups, but also for quantifying the group relationships themselves.

We also conduct a similar experiment where we obtain MLEs of a (along with all other model and covariance parameters) from the data. Here, we apply the model to all 52 tissue types. We fit the MGGP for every pair of tissues, and extract $\widehat{a}_{MLE}$ for each pair. This experiment yields $\frac{1}{2}(52 \times 51) = 1326$ estimated values of a (one for each pair of tissue types). The estimated values of a reflect many of the expected relationships between the tissue types (Figure 8). Notably, we report 11 regions of the brain yielding low values for a, which suggests that gene expression in these tissue types changes in a similar manner as ischemic time changes.

Finally, we conduct a fully Bayesian analysis of the GTEx data using the MGGP model. Here, we analyze the 11 brain tissue types, which comprise a total of 2218 samples, using Equation 3. For ease of visualization and demonstration, we use the expression of just one gene, *TXNIP*, as our explanatory variable, and use each sample's ischemic time as the response as before. We use the same modeling approach as in our simulation study (Section 4.2), adapting the model for 11 groups. Again, we fit the collapsed likelihood in Equation 3 incorporating group-specific intercepts in F and group-specific variances in D_τ. We run four chains with dispersed initial values for 300 iterations each. Convergence is diagnosed

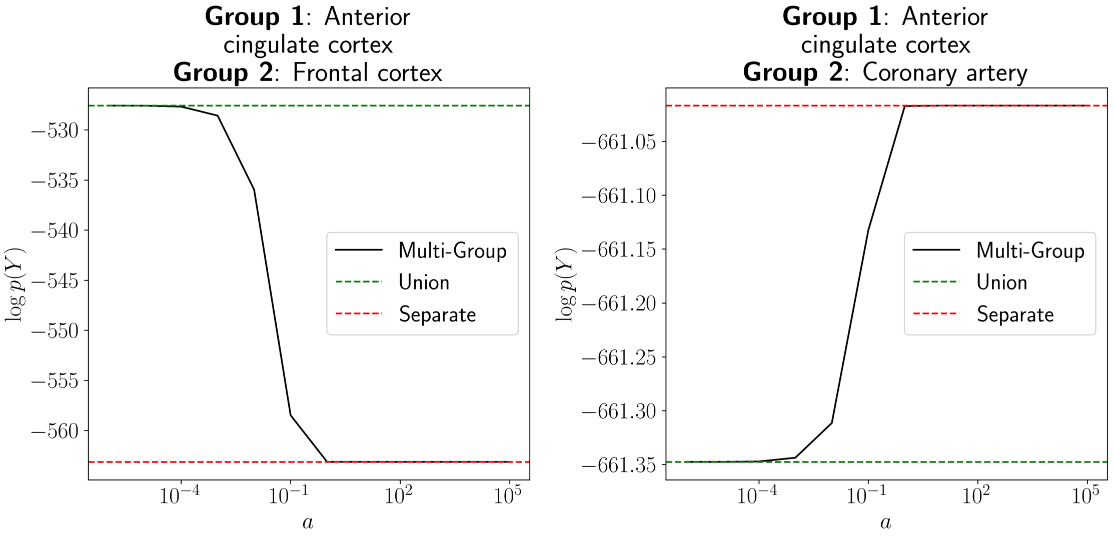

Figure 7: **Likelihood of GTEx gene expression data under the MGGP model**. For each pair of tissues, we computed the log marginal likelihood of the data under the MGGP model with a set to be a range of values. Similar tissue types (e.g., anterior cingulate cortex and frontal cortex) prefer low values of a, while more dissimilar tissues (e.g., anterior cingulate cortex and coronary artery) prefer high values of a.

after 100 iterations using visual inspection of autocorrelation plots. The subsequent 800 samples are retained for posterior inference.

Using posterior summaries of the covariance function parameters, intercepts, and noise variances (Table 3), we find that the MGGP successfully models these relationships across groups. Moreover, the predictive processes demonstrate the similarities and differences between the groups (Supplementary Figures 13, 14). For example, we see that while all brain regions tend to exhibit a similar relationship between gene expression levels and ischemic time, this relationship shows a distinct trend in the cerebellum and putamen. We find that the MGGP predictive processes capture these subtle group relationship differences.

6. Discussion

We develop multi-group Gaussian process (MGGP) models as a flexible approach for modeling complex dependencies in data sets with subgroup structure. We present several options for constructing valid covariance functions on $\mathbb{R}^p \times \mathscr{C}$, and we show that this structure generalizes existing GP models. We emphasize that the MGGP novelty is in the construction of the covariance functions, enabling all GP inference strategies applicable to MGGPs. We demonstrate the behavior of the MGGP through several simulation experiments and an application to gene expression data with ischemic time measurements for 52 distinct tissues.

Several future directions remain to be explored. First, this paper lays the groundwork for developing new positive definite covariance functions on $\mathbb{R}^p \times \mathscr{C}$. An interesting direction is to construct covariance functions whose within-group and between-group correlations

Parameter	**Posterior percentiles**	**Parameter**	**Posterior percentiles**
σ^2	0.49 (0.28; 0.82)	β_{11}	-0.02 (-0.48; 0.67)
b	0.63 (0.32; 1.16)	τ_1^2	0.79 (0.63; 1.0)
a	0.45 (0.28; 0.93)	τ_2^2	0.97 (0.81; 1.14)
β_1	-0.04 (-0.51; 0.56)	τ_3^2	0.89 (0.75; 1.12)
β_2	-0.08 (-0.58; 0.47)	τ_4^2	0.86 (0.74; 1.07)
β_3	0.01 (-0.46; 0.61)	τ_5^2	0.87 (0.75; 1.01)
β_4	0.64 (0.07; 1.17)	τ_6^2	0.89 (0.74; 1.09)
β_5	0.7 (0.12; 1.28)	τ_7^2	0.85 (0.68; 1.05)
β_6	-0.0 (-0.46; 0.59)	τ_8^2	0.86 (0.74; 1.01)
β_7	-0.04 (-0.46; 0.52)	τ_9^2	0.8 (0.67; 1.0)
β_8	-0.1 (-0.48; 0.46)	τ_{10}^2	0.89 (0.67; 1.09)
β_9	0.18 (-0.3; 0.68)	τ_{11}^2	0.88 (0.69; 1.14)
β_{10}	-0.03 (-0.47; 0.51)		

Table 3: **Parameter posterior summaries for Bayesian analysis of GTEx data.** Posterior summaries are presented as $50(2.5; 97.5)$ percentiles in the second column. Subscripts on parameter names indicate group labels.

exhibit fundamentally different structure (e.g., the within-group correlation may be Matérn families, while the between group correlation may be RBF families). Second, as briefly mentioned in Section 2, recent advances in classes of GPs that scale learning to massive data sets can be applied to the MGGP. For example, sparsity-inducing GPs have received much attention recently (see, e.g., Datta et al., 2016; Katzfuss and Guinness, 2021; Peruzzi et al., 2022), and such methods can be applied to the class of multi-group models presented here. Third, a linear, yet nonseparable, MGGP kernel remains to be explored: A naive approach is to assign different linear coefficients to different groups; however, this leads to a separable kernel.Fourth, there are opportunities to explore alternate GP representations by adopting Mercer's theorem, i.e., an eigenfunction-based decomposition, to help build new covariance functions. The main challenge here is to identify the suitable eigenfunctions with both continuous and categorical components. Fifth, the consistency of MLEs and the posterior consistency of kernel parameters remain open and challenging problems. Sixth, the multi-group kernels can serve as valid cross-covariance functions for multivariate spatial processes by treating the categories as indices for the elements of a vector process. They present sparser parametric forms than cross-covariance functions resulting from the widely employed linear models of coregionalization in spatial statistics (see, e.g., Gelfand et al., 2004; Banerjee and Johnson, 2006; Guhaniyogi et al., 2013, with applications in agronomy, ecology and environmental sciences) and can also serve as alternatives in process-based factor models (Zhang and Banerjee, 2022; Davies et al., 2022) and as candidates to build highly multivariate graphical GPs (Dey et al., 2022). Finally, given that the MGGPs are well-defined stochastic processes, they can be introduced in any process-based model, perhaps replacing more customary choices that do not allow both continuous and categorical variables simultaneously. Thus, there could be benefits from using a multi-group process in

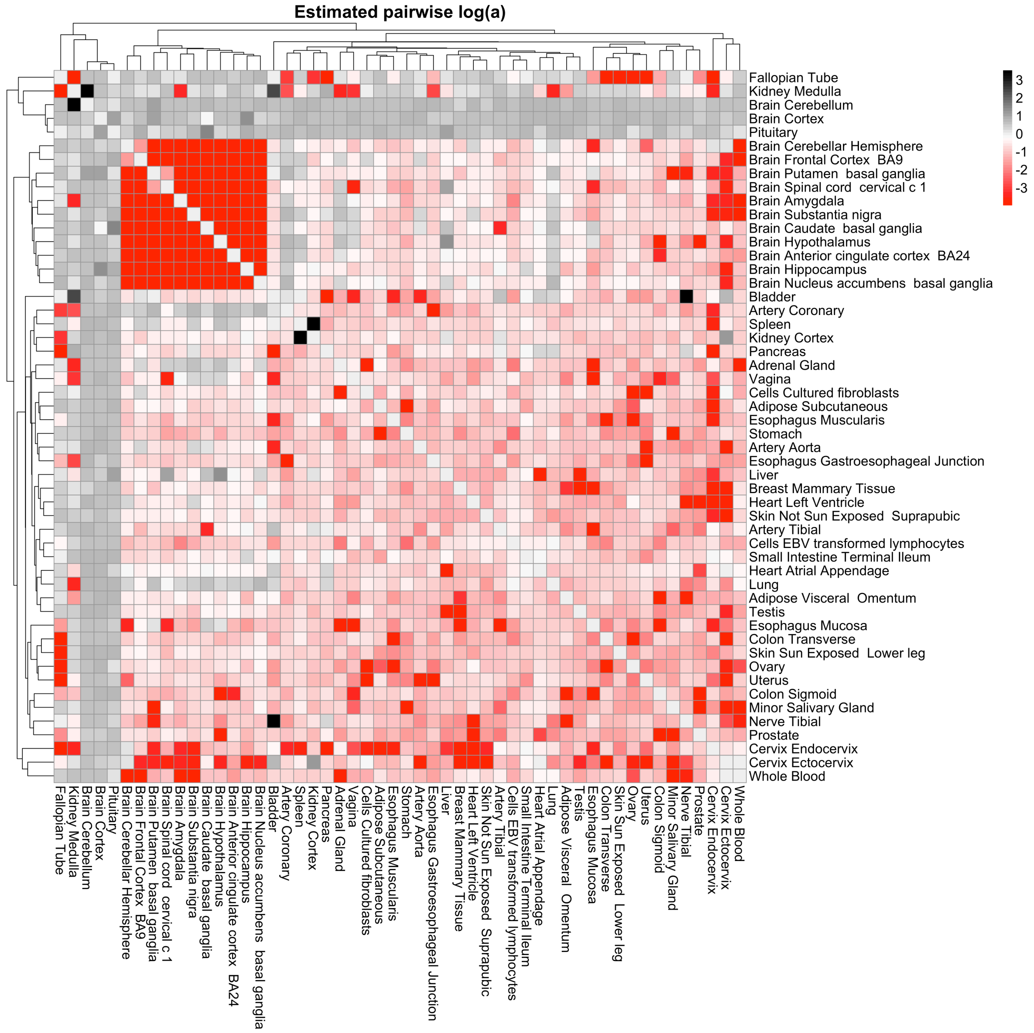

Figure 8: **Estimation of a for each pair of GTEx tissue types**. Cell ij in the heatmap represents $\log_{10}(a_{ij})$, where a_{ij} is the MLE of a when fitting the MGGP model using tissues i and j. Lower values of a (red) indicate higher similarity, while higher values of a (black) indicate lower similarity.

classification, latent variable models, and other model types. We envision the MGGP being a flexible tool in diverse contexts.

Acknowledgements

BEE was funded in part by grants from the Parker Institute for Cancer Immunology (PICI), the Chan-Zuckerberg Institute (CZI), NIH NHGRI R01 HG012967, and NIH NHGRI R01 HG013736. BEE is a CIFAR Fellow in the Multiscale Human Program. DL was supported by R01 AG079291, R56 LM013784, R01 HL149683, UM1 TR004406, R01 LM014407, and P30 ES010126.

Conflicts of Interest

BEE is on the Scientific Advisory Board for ArrePath Inc; she consults for Neumora.

References

Tatiyana V Apanasovich and Marc G Genton. Cross-covariance functions for multivariate random fields based on latent dimensions. *Biometrika*, 97(1):15–30, 2010.

Sudipto Banerjee. High-Dimensional Bayesian Geostatistics. *Bayesian Analysis*, 12(2):583 – 614, 2017. doi: 10.1214/17-BA1056R.

Sudipto Banerjee and Gregg A. Johnson. Coregionalized single- and multiresolution spatially varying growth curve modeling with application to weed growth. *Biometrics*, 62(3):864–876, 2006. doi: https://doi.org/10.1111/j.1541-0420.2006.00535.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2006.00535.x`.

Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC Press, Boca Raton, FL, 2nd edition, 2014.

J Bernardo, J Berger, APAFMS Dawid, A Smith, et al. Regression and classification using Gaussian process priors. *Bayesian Statistics*, 6:475, 1998.

Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Deisenroth. Matérn Gaussian processes on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 33, pages 12426–12437, 2020.

Viacheslav Borovitskiy, Iskander Azangulov, Alexander Terenin, Peter Mostowsky, Marc Deisenroth, and Nicolas Durrande. Matérn Gaussian processes on graphs. In *International Conference on Artificial Intelligence and Statistics*, pages 2593–2601. PMLR, 2021.

Jérémie Bouttier, Philippe Di Francesco, and Emmanuel Guitter. Geodesic distance in planar graphs. *Nuclear Physics B*, 663(3):535–567, 2003.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

U.S. Census Bureau. 2020 census. U.S. Department of Commerce, February 2020.

GTEx Consortium et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

Harald Cramér. On the theory of stationary random processes. *Annals of Mathematics*, pages 215–230, 1940.

Noel Cressie and Hsin-Cheng Huang. Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448):1330–1339, 1999.

Noel Cressie and Christopher K Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, Hoboken, NJ, 2011.

Andreas Damianou and Neil D Lawrence. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215. PMLR, 2013.

Abhirup Datta, Sudipto Banerjee, Andrew O Finley, and Alan E Gelfand. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812, 2016.

Tilman M. Davies, Sudipto Banerjee, Adam P. Martin, and Rose E. Turnbull. A nearest-neighbour gaussian process spatial factor model for censored, multi-depth geochemical data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(4):1014–1043, 05 2022. ISSN 0035-9254. doi: 10.1111/rssc.12565. URL `https://doi.org/10.1111/rssc.12565`.

Marc Peter Deisenroth, Dieter Fox, and Carl Edward Rasmussen. Gaussian processes for data-efficient learning in robotics and control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):408–423, 2013.

Debangan Dey, Abhirup Datta, and Sudipto Banerjee. Graphical Gaussian process models for highly multivariate spatial data. *Biometrika*, 109(4):993–1014, 12 2022. ISSN 1464-3510. doi: 10.1093/biomet/asab061. URL `https://doi.org/10.1093/biomet/asab061`.

David B Dunson. Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):355–366, 2000.

David B Dunson, Zhen Chen, and Jean Harry. A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics*, 59(3):521–530, 2003.

David B Dunson, Hau-Tieng Wu, and Nan Wu. Diffusion based Gaussian processes on restricted domains. *arXiv preprint arXiv:2010.07242*, 2020.

Pedro G Ferreira, Manuel Muñoz-Aguirre, Ferran Reverter, Caio P Sá Godinho, Abel Sousa, Alicia Amadoz, Reza Sodaei, Marta R Hidalgo, Dmitri Pervouchine, Jose Carbonell-Caballero, et al. The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nature Communications*, 9(1):1–15, 2018.

A. E. Gelfand and S. Banerjee. Multivariate spatial process models. In A.E. Gelfand, P.J. Diggle, M. Fuentes, and P Guttorp, editors, *Handbook of Spatial Statistics*, pages 495–516. Boca Raton, FL: CRC Press, 2010.

Alan E. Gelfand, Alexandra M. Schmidt, Sudipto Banerjee, and C. F. Sirmans. Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST*, 13 (2):263–312, 2004.

Marc G Genton and William Kleiber. Cross-covariance functions for multivariate geostatistics. *Statistical Science*, 30(2):147–163, 2015.

Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

Tilmann Gneiting. Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97(458):590–600, 2002.

Tilmann Gneiting, William Kleiber, and Martin Schlather. Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105 (491):1167–1177, 2010.

Rajarshi Guhaniyogi, Andrew O. Finley, Sudipto Banerjee, and Richard K. Kobe. Modeling complex spatial dependencies: Low-rank spatially varying cross-covariances with application to soil nutrient data. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(3):274–298, Sep 2013. ISSN 1537-2693. doi: 10.1007/s13253-013-0140-3. URL https://doi.org/10.1007/s13253-013-0140-3.

M.J. Heaton, A. Datta, A.O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, F. Gerber, R. Gramacy, D. Hammerling, M. Katzfuss, F. Lindgren, D. Nychka, F. Sun, and A. Zammit-Mangion. Methods for analyzing large spatial data: A review and comparison. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425, 2019. doi: 10.1007/s13253-018-00348-w.

James Hensman, Neil D Lawrence, and Magnus Rattray. Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC bioinformatics*, 14(1):1–12, 2013.

Matthew D Hoffman, Andrew Gelman, et al. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

Matthias Katzfuss and Joseph Guinness. A general framework for Vecchia approximations of Gaussian processes. *Statistical Science*, 36(1):124 – 141, 2021. doi: 10.1214/19-STS755.

Arthur Leroy, Pierre Latouche, Benjamin Guedj, and Servane Gey. Magma: inference and prediction using multi-task Gaussian processes with common mean. *Machine Learning*, 111(5):1821–1849, 2022.

Arthur Leroy, Pierre Latouche, Benjamin Guedj, and Servane Gey. Cluster-specific predictions with multi-task Gaussian processes. *Journal of Machine Learning Research*, 24(5): 1–49, 2023.

Cheng Li. Bayesian fixed-domain asymptotics for covariance parameters in a Gaussian process model. *The Annals of Statistics*, 50(6):3334–3363, 2022.

Didong Li, Wenpin Tang, and Sudipto Banerjee. Inference for gaussian processes with matérn covariogram on compact riemannian manifolds. *Journal of Machine Learning Research*, 24(101):1–26, 2023.

Wei-Liem Loh and Saifei Sun. Estimating the parameters of some common Gaussian random fields with nugget under fixed-domain asymptotics. *Bernoulli*, 29(3):2519–2543, 2023.

Hiroshi Maehara. Euclidean embeddings of finite metric spaces. *Discrete Mathematics*, 313 (23):2848–2856, 2013.

Jared S Murray and Jerome P Reiter. Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516):1466–1479, 2016.

Valeria Musella, Paolo Verderio, James Francis Reid, Sara Pizzamiglio, Manuela Gariboldi, Maurizio Callari, Milione Massimo, Loris De Cecco, Silvia Veneroni, Marco Alessandro Pierotti, et al. Effects of warm ischemic time on gene expression profiling in colorectal cancer tissues and normal mucosa. *PLOS One*, 8(1):e53406, 2013.

Mu Niu, Pokman Cheung, Lizhen Lin, Zhenwen Dai, Neil Lawrence, and David Dunson. Intrinsic Gaussian processes on complex constrained domains. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):603–627, 2019.

Jorge Nocedal and Stephen Wright. *Numerical optimization.* Springer Science & Business Media, 2006.

Tianyu Pan, Guanyu Hu, and Weining Shen. Identifying latent groups in spatial panel data using a Markov random field constrained product partition model. *arXiv preprint arXiv:2012.10541*, 2020.

Sunho Park and Seungjin Choi. Hierarchical Gaussian process regression. In *Proceedings of 2nd Asian Conference on Machine Learning*, pages 95–110. JMLR Workshop and Conference Proceedings, 2010.

Michele Peruzzi, Sudipto Banerjee, and Andrew O. Finley. Highly scalable Bayesian geostatistical modeling via meshed Gaussian processes on partitioned domains. *Journal of the American Statistical Association*, 117(538):969–982, 2022. doi: 10.1080/01621459.2020.1833889.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* The MIT Press, Cambridge, MA, first edition, 2005.

Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. Science forum: the human cell atlas. *eLife*, 6:e27041, 2017.

Allen Riddell, Ari Hartikainen, and Matthew Carter. PySTAN (3.0.0). PyPI, March 2021.

Binxin Ru, Ahsan Alvi, Vu Nguyen, Michael A Osborne, and Stephen Roberts. Bayesian optimisation over multiple continuous and categorical inputs. In *International Conference on Machine Learning*, pages 8276–8285. PMLR, 2020.

Walter Rudin. *Fourier analysis on groups.* Courier Dover Publications, 2017.

Peter Schulam, Fredrick Wigley, and Suchi Saria. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Stan Development Team. Stan modeling language users guide and reference manual, 2020.

Michael L Stein. *Interpolation of spatial data: Some theory for kriging*. Springer Science & Business Media, 1999.

Michael L Stein. Space–time covariance functions. *Journal of the American Statistical Association*, 100(469):310–321, 2005.

Wenpin Tang, Lu Zhang, and Sudipto Banerjee. On identifiability and consistency of the nugget in Gaussian spatial process models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):1044–1070, 2021.

M Teimourian, T Baghfalaki, M Ganjali, and D Berridge. Joint modeling of mixed skewed continuous and ordinal longitudinal responses: a Bayesian approach. *Journal of Applied Statistics*, 42(10):2233–2256, 2015.

Aviad Tsherniak, Francisca Vazquez, Phil G Montgomery, Barbara A Weir, Gregory Kryukov, Glenn S Cowley, Stanley Gill, William F Harrington, Sasha Pantel, John M Krill-Burger, et al. Defining a cancer dependency map. *Cell*, 170(3):564–576, 2017.

Christopher K. Wikle. Low-rank representations for spatial processes. *Handbook of Spatial Statistics*, pages 107–118, 2010. Gelfand, A. E., Diggle, P., Fuentes, M. and Guttorp, P., editors, Chapman and Hall/CRC, pp. 107-118.

Hao Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004.

Lu Zhang and Sudipto Banerjee. Spatial factor modeling: A bayesian matrix-normal approach for misaligned data. *Biometrics*, 78(2):560–573, 2022. doi: https://doi.org/10.1111/biom.13452. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13452`.

Appendix A. Proof of Proposition 2

Proof

First, assume K_2 is positive definite, and let the observation set be $\{c_1, \cdots, c_k\}$. Then the covariance matrix is exactly C, which is positive definite as well.

Then assume C is positive definite, and define X be a random variable that follows a k-dimension Gaussian distribution: $X \sim N(0, C)$. Given observations $a_1, \cdots, a_n \subset \{c_1, \cdots, c_k\}$, let Y be a n-dimensional random variable such that $Y_i = X_{a_i}$, then $K(a_i, a_j) = C_{a_i,a_j} = \mathrm{Cov}(X_{a_i}, X_{a_j}) = \mathrm{Cov}(Y_i, Y_j)$. That is, K is positive definite.

■

Appendix B. Proof of Corollary 4

Proof By Proposition 2, K_2 is positive definite if and only if C is positive definite. Under the assumption of homogeneity, $C = \begin{bmatrix} 1 & b & \cdots & b \\ b & 1 & \cdots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \cdots & 1 \end{bmatrix}$. We can rewrite $C = (1-b)\mathrm{I}_k + b1_{k\times k}$. Recall that $1_{k\times k}$ has eigenvalue p with multiplicity 1, and the corresponding eigenvector is 1_k. Moreover, 0 is another eigenvalue with multiplicity $k-1$. As a result, eigenvalues of Σ are $1-b+bk$ with multiplicity 1 and $1-b$ with multiplicity -1. We can conclude that C is positive definite if and only if $-\frac{1}{k-1} \le b \le 1$. ■

Appendix C. Proof of Theorem 6

Restricting the GP to a subset of the original domain is again a GP (Rasmussen and Williams, 2005). Hence, if we can isometrically embed $\mathscr{C}$ to an Euclidean space $\mathbb{R}^{p'}$ by some mapping $\iota : \mathscr{C} \to \mathbb{R}^{p'}$ such that $d_{ij} := d(c_i, c_j) = \|\iota(c_i) - \iota(c_j)\|$, then an isotropic GP on $\mathbb{R}^{p'}$ induces a GP on the image $\iota(\mathscr{C})$. To construct valid covariance functions, the explicit form of ι is not necessary. As long as such embedding exists, we can construct semi-isotropic covariance functions.

Lemma 13 (Maehara (2013)) *Let $G_{ij} = \frac{1}{2}\left(d(c_1, c_i) + d(c_1, c_j) - d(c_i, c_j)\right)$ be the $k \times k$ Gram matrix of d. Then there exists an isotropic embedding $\iota : \mathscr{C} \to \mathbb{R}^{p'}$ if and only if G is positive semi-definite with rank at most p'.*

Proof By the isometric embedding, finding K is equivalent to finding a semi-isotropic covariance function K_0 in $\mathbb{R}^p \times \mathbb{R}^{p'}$. For any completely monotone function $\varphi : \mathbb{R}_+ \to \mathbb{R}$ and positive function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ with a completely monotone derivative, Gneiting (2002) proved that $K_0((x_1, x_1'), (x_2, x_2')) = \frac{\sigma^2}{\left(\psi(\|x_1'-x_2'\|^2)\right)^{p/2}}\varphi\left(\frac{\|x_1-x_2\|^2}{\psi(\|x_1'-x_2'\|^2)}\right)$ is positive definite for any $\sigma^2 > 0$. As a result, $K((x, c_i), (x', c_j)) := K_0((x, \iota(c_i)), (x', \iota(c_j)))$ is positive definite. ■

Appendix D. Proof of Theorem 7

Recall the general form of Bochner's Theorem for a locally compact Abelian group:

Lemma 14 (Bochner's Theorem) *Let G be a locally compact Abelian group and $\widehat{G}$ be its dual group, then for any continuous positive-definite function K on G, there exists a unique positive measure μ on $\widehat{G}$ such that*

$$K(g) = \int_{\widehat{G}} \xi(g) d\mu(\xi).$$

Note that $G = \mathbb{R}^p \times \mathbb{Z}_2$ is a locally compact Abelian group and the dual group is $\widehat{G} = \mathbb{R}^p \times U_2$, where U_2 is the group of second roots of unity, that is, $U_2 = \{1, -1\}$. $\widehat{G}$ acts on G as

$$(\omega, z)((x, l)) = e^{-2\pi i \omega x} z^l.$$

The spectral measure of K, denoted by μ on $\mathbb{R}^p \times U_2$ splits as $\mu_1 \times \mu_2$ where μ_1 is a measure on $\mathbb{R}^p$ and μ_2 is a measure on U_2. Then we claim that

$$K(x, l) = \sum_{z \in U_2} \int_{\mathbb{R}^p} e^{-2\pi i \omega x} z^l \rho(\omega, z) d\omega,$$

where $\rho(\omega, 1) = \frac{1}{2}\rho_w(\omega) + \frac{1}{2}\rho_c(\omega)$ and $\rho(\omega, -1) = \frac{1}{2}\rho_w(\omega) - \frac{1}{2}\rho_c(\omega)$. We derive the left hand side from the right hand side. First consider the case when $l = 0$:

$$\begin{aligned}
&\sum_{z \in U_2} \int_{\mathbb{R}^p} e^{-2\pi i \omega x} z^0 \rho(\omega, z) d\omega \\
&= \frac{1}{2} \int_{\mathbb{R}^p} e^{-2\pi i \omega x} (\rho_w(\omega) + \rho_c(\omega)) d\omega + \frac{1}{2} \int_{\mathbb{R}^p} e^{-2\pi i \omega x} (\rho_w(\omega) - \rho_c(\omega)) d\omega \\
&= \int_{\mathbb{R}^p} e^{-2\pi i \omega x} \rho_w(\omega) d\omega \\
&= K_w(x) = K(x, 0).
\end{aligned}$$

Similarly, when $l = 1$,

$$\begin{aligned}
&\sum_{z \in U_2} \int_{\mathbb{R}^p} e^{-2\pi i \omega x} z^1 \rho(\omega, z) d\omega \\
&= \frac{1}{2} \int_{\mathbb{R}^p} e^{-2\pi i \omega x} (\rho_w(\omega) + \rho_c(\omega)) d\omega - \frac{1}{2} \int_{\mathbb{R}^p} e^{-2\pi i \omega x} (\rho_w(\omega) - \rho_c(\omega)) d\omega \\
&= \int_{\mathbb{R}^p} e^{-2\pi i \omega x} \rho_c(\omega) d\omega \\
&= K_c(x) = K(x, 1).
\end{aligned}$$

As a result, K is positive definite $\iff$ ρ is a positive measure $\iff$ $\rho_w \geq \rho_c$, and the Theorem follows.

Appendix E. Proof of Theorem 9

The generalized spectral measure of K, denoted by μ (with density ρ) on $\mathbb{R}^p \times U_2 \times U_2$, splits as $\mu_1 \times \mu_2$ where μ_1 is a measure on $\mathbb{R}^p$ and μ_2 is a positive semi-definite measure on $U_2 \times U_2$. Then

$$K(x, l, l') = \sum_{z \in U_2} \sum_{z' \in U_2} \int_{\mathbb{R}^p} e^{-2\pi i \omega x} z^l \overline{z'^{l'}} \rho(\omega, z, z') d\omega.$$

We claim that $\rho(\omega, z, z') = \begin{cases} \frac{1}{4}\left(\rho_0(\omega) + 2\rho_c(\omega) + \rho_0(\omega)\right) & z = z' = 1 \\ \frac{1}{4}\left(\rho_0(\omega) - \rho_1(\omega)\right) & zz' = -1 \\ \frac{1}{4}\left(\rho_0(\omega) - 2\rho_c(\omega) + \rho_1(\omega)\right) & z = z' = -1. \end{cases}$.

We work with the right hand side. First consider the case when $l = l' = 0$:

$$\begin{aligned}
&\sum_{z\in U_2}\sum_{z'\in U_2}\int_{\mathbb{R}^p} e^{-2\pi i\omega x} z^0\overline{z'^0}\rho(\omega,z,z')d\omega \\
&\quad = \frac{1}{4}\int_{\mathbb{R}^p} e^{-2\pi i\omega x}\left\{\rho_0(\omega)+2\rho_c(\omega)+\rho_1(\omega)\right. \\
&\qquad\qquad \left.+\rho_0(\omega)-\rho_1(\omega)+\rho_0(\omega)-\rho_1(\omega)+\rho_0(\omega)-2\rho_c(\omega)+\rho_1(\omega)\right\}d\omega \\
&\quad = \int_{\mathbb{R}^p} e^{-2\pi i\omega x}\rho_0(\omega)d\omega = K_0(x) = K(x,0,0).
\end{aligned}$$

Similarly, when $l = 0, l' = 1$,

$$\begin{aligned}
&\sum_{z\in U_2}\sum_{z'\in U_2}\int_{\mathbb{R}^p} e^{-2\pi i\omega x} z^0\overline{z'^1}\rho(\omega,z,z')d\omega \\
&\quad = \frac{1}{4}\int_{\mathbb{R}^p} e^{-2\pi i\omega x}\left(\rho_0(\omega)+2\rho_c(\omega)+\rho_1(\omega)\right. \\
&\qquad\qquad \left.-\rho_0(\omega)+\rho_1(\omega)+\rho_0(\omega)-\rho_1(\omega)-\rho_0(\omega)+2\rho_c(\omega)-\rho_1(\omega)\right)d\omega \\
&\quad = \int_{\mathbb{R}^p} e^{-2\pi i\omega x}\rho_c(\omega)d\omega = K_c(x) = K(x,0,1).
\end{aligned}$$

When $l = l' = 1$,

$$\begin{aligned}
&\sum_{z\in U_2}\sum_{z'\in U_2}\int_{\mathbb{R}^p} e^{-2\pi i\omega x} z^1\overline{z'^1}\rho(\omega,z,z')d\omega \\
&\quad = \frac{1}{4}\int_{\mathbb{R}^p} e^{-2\pi i\omega x}\left(\rho_0(\omega)+2\rho_c(\omega)+\rho_1(\omega)-\rho_0(\omega)+\rho_1(\omega)-\rho_0(\omega)\right. \\
&\qquad\qquad \left.+\rho_1(\omega)+\rho_0(\omega)-2\rho_c(\omega)+\rho_1(\omega)\right)d\omega \\
&\quad = \int_{\mathbb{R}^p} e^{-2\pi i\omega x}\rho_1(\omega)d\omega = K_1(x) = K(x,1,1).
\end{aligned}$$

As a result, K is positive definite $\iff$ ρ is a positive semi-definite measure $\iff$ $(\rho_0+2\rho_c+\rho_1)(\rho_0-2\rho_c+\rho_1)-(\rho_0-\rho_1)^2 = 4(\rho_0\rho_1-\rho_c^2)\geq 0$, and the Theorem follows.

Appendix F. Proof of Theorem 10

Proof Given a Gaussian random field Z on $\mathcal{Y}\times\mathscr{C}$ with covariance function K, we prove that $\widetilde{Z} := \Phi(Z)$ is a Gaussian k-variate random field on $\mathcal{Y}$ with cross-covariance function

$\widetilde{K}$. It suffices to check that $\mathrm{Cov}(\widetilde{Z}(x), \widetilde{Z}(x')) = \widetilde{K}(x, x')$, for any $x, x' \in \mathcal{Y}$.

$$\begin{aligned}
\mathrm{Cov}(\widetilde{Z}(x), \widetilde{Z}(x')) &= \mathrm{Cov}([Z(x, c_1), \cdots, Z(x, c_k)]^{\mathrm{T}}, [Z(x', c_1), \cdots, Z(x', c_k)]^{\mathrm{T}}) \\
&= \begin{bmatrix} K((x,c_1),(x',c_1)) & K((x,c_1),(x',c_2)) & \cdots & K((x,c_1),(x',c_k)) \\ K((x,c_2),(x',c_1)) & K((x,c_2),(x',c_2)) & \cdots & K((x,c_2),(x',c_k)) \\ \vdots & \vdots & \ddots & \vdots \\ K((x,c_k),(x',c_1)) & K((x,c_k),(x',c_2)) & \cdots & K((x,c_k),(x',c_k)) \end{bmatrix} \\
&= \begin{bmatrix} \widetilde{K}_{11} & \widetilde{K}_{12} & \cdots & \widetilde{K}_{1k} \\ \widetilde{K}_{21} & \widetilde{K}_{22} & \cdots & \widetilde{K}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{K}_{k1} & \widetilde{K}_{k2} & \cdots & \widetilde{K}_{kk} \end{bmatrix} \\
&= \widetilde{K}(x, x').
\end{aligned}$$

Then assume $\widetilde{Z}$ is a Gaussian k-variate random field on $\mathcal{Y}$. We prove that $Z \colon \Phi^{-1}(\widetilde{Z})$ is a Gaussian random field on $\mathcal{Y} \times \mathscr{C}$ with covariance function K. It suffices to check that $\mathrm{Cov}(Z(x, c_i), Z(x', c_j)) = K((x, c_i), (x', c_j))$ for any $x, y \in \mathcal{Y}$, $i, j = 1, \cdots, k$. Then,

$$\begin{aligned}
\mathrm{Cov}(Z(x, c_i), Z(x', c_j)) &= \mathrm{Cov}(\widetilde{Z}_i(x)), \widetilde{Z}_i(x'))) \\
&= \widetilde{K}(x, x')_{ij} \\
&= K((x, c_i), (x', c_j)).
\end{aligned}$$

■

Appendix G. Proof of Theorem 12

Similar to the proof of Theorem 6, it suffices to construct a covariance function on $\mathbb{R}^p \times \mathbb{R}^{p'} \times \mathbb{R}^{p''}$. Then the Theorem follows Proposition 1 in Apanasovich and Genton (2010) naturally.

Appendix H. More covariance functions

We provide more semi-isotropic covariance functions below from Cressie and Huang (1999) derived from spectral densities.

$$K((x, c_i), (x', c_j)) = \frac{\sigma^2}{(ad_{ij}+1)^{p/2}} \exp\left\{-\frac{b^2\|x-x'\|^2}{ad_{ij}+1}\right\}, \tag{11}$$

$$K((x, c_i), (x', c_j)) = \frac{\sigma^2(a^2 d_{ij}^2+1)}{[(a^2 d_{ij}^2+1)^2 + b^2\|x-x'\|^2]^{\frac{p+1}{2}}}, \tag{12}$$

$$K((x, c_i), (x', c_j)) = \frac{\sigma^2(ad_{ij}+1)}{[(ad_{ij}+1)^2 + b^2\|x-x'\|^2]^{\frac{p+1}{2}}}, \tag{13}$$

$$K((x, c_i), (x', c_j)) = \sigma^2 \exp\left\{-a^2 d_{ij}^2 - b^2\|x-x'\|^2 - c d_{ij}^2\|x-x'\|^2\right\}, \tag{14}$$

$$K((x, c_i), (x', c_j)) = \sigma^2 \exp\left\{-ad_{ij} - b^2\|x-x'\|^2 - c d_{ij}\|x-x'\|^2\right\}. \tag{15}$$

Kernel	Class	Details
(7)	Gneiting	$\phi(t) = e^{bt^{1/2}}$, $\psi(t) = a^2 t + 1$
(8)	Gneiting	$\phi(t) = \frac{(bt^{1/2}/\sqrt{c})^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(bt^{1/2}\sqrt{c})$, $\psi(t) = \frac{a^2 t + c}{c(a^2 t + 1)}$
(9)	Gneiting	Set $\nu = 1/2$ in (8)
(11)	Cressie & Huang	Example 2
(12)	Cressie & Huang	Example 3
(13)	Cressie & Huang	Example 4
(14)	Cressie & Huang	Example 5
(15)	Cressie & Huang	Example 6

Table 4: **Summary of constructions of kernel (7-9), (11-15).**

Table 4 summaries the kernel constructions:

Appendix I. Extension to unknown distance d_{ij}

In the situation where d_{ij} is unknown, we instead parameters the kernels discussed in main sections (Equations (7-9), (11-15)) as $a_{ij} := ad_{ij}$, so that the parameters are (σ^2, b, A) where $A = (a_{ij})$ is a k by k matrix. The constraints on A are from the metric axioms:

$$A_{ii} = 0,\ A_{ij} = A_{ji},\ A_{ij} + A_{jl} \geq A_{il},\ \forall i, j, l = 1, \cdots, k.$$

The MLE is then over the new parameter space, under the above constraints. Fortunately, these constraints are all linear, so optimization methods such as L-BFGS can still be applied. In this case, the parameter a_{ij} is purely data-driven, and can be interpreted as the dissimilarity between group i and group j, as desired.

Appendix J. Code

Our GitHub repository is `https://github.com/andrewcharlesjones/multi-group-GP`. This repository contains downloadable code for the models and experiments to reproduce the analysis in the paper. We provide a Python package for model fitting, computing covariance functions and carrying out estimation and prediction.

Appendix K. GTEx Data

The GTEx data can be downloaded from the GTEx portal: `https://gtexportal.org/home/datasets`. We use 52 tissue types, listed below, although some experiments use a subset of these tissue types. The sample size for each tissue type is shown in square brackets. Adipose Subcutaneous [644], Adipose Visceral (Omentum) [539], Adrenal Gland [254], Artery Aorta [424], Artery Coronary [238], Artery Tibial [657], Bladder [21], Brain Amygdala [147], Brain Anterior cingulate cortex (BA24) [172], Brain Caudate (basal ganglia) [230], Brain Cerebellar Hemisphere [208], Brain Cerebellum [241], Brain Cortex [255], Brain Frontal Cortex (BA9) [200], Brain Hippocampus [188], Brain Hypothalamus [193], Brain Nucleus accumbens (basal ganglia) [232], Brain Putamen (basal ganglia) [194], Brain Spinal cord (cervical c-1) [155], Brain Substantia nigra [133], Breast Mammary Tissue [456], Cells Cultured fibroblasts [444], Cells EBV-transformed lymphocytes [174], Cervix Ectocervix [9], Cervix Endocervix [10], Colon Sigmoid [373], Colon Transverse [406], Esophagus Gastroesophageal Junction [375], Esophagus Mucosa [551], Esophagus Muscularis [515], Fallopian Tube [9], Heart Atrial Appendage [422], Heart Left Ventricle [428], Kidney Cortex [85], Kidney Medulla [4], Liver [224], Lung [573], Minor Salivary Gland [162], Nerve Tibial [615], Ovary [180], Pancreas [37], Pituitary [283], Prostate [245], Skin Not Sun Exposed (Suprapubic) [595], Skin Sun Exposed (Lower leg) [665], Small Intestine Terminal Ileum [187], Spleen [233], Stomach [359], Testis [359], Uterus [142], Vagina [156], Whole Blood [139].

Appendix L. Supplementary figures

copy.png

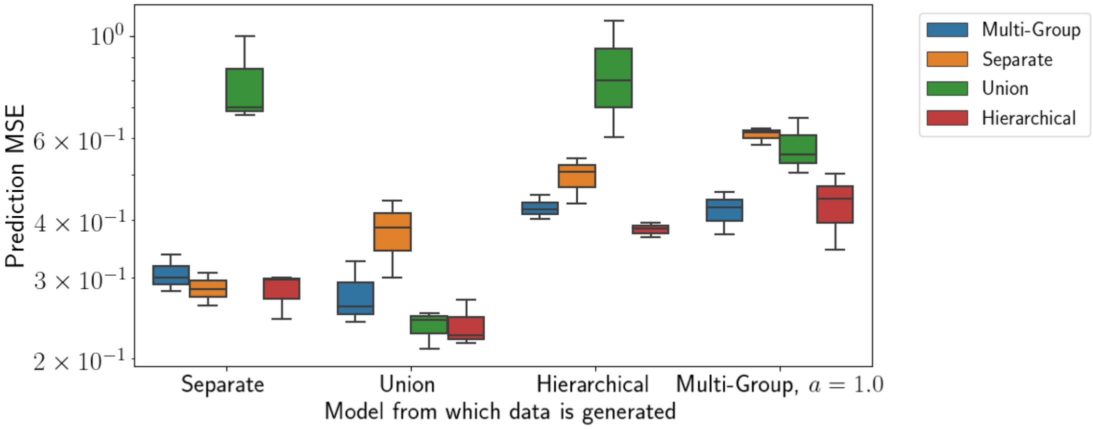

Figure 9: **Prediction experiment with synthetic data using a Matérn covariance function, as described in Section 4.2.** The covariance function parameters were set as $\sigma^2 = b = 1$, $\nu = 1/2$.

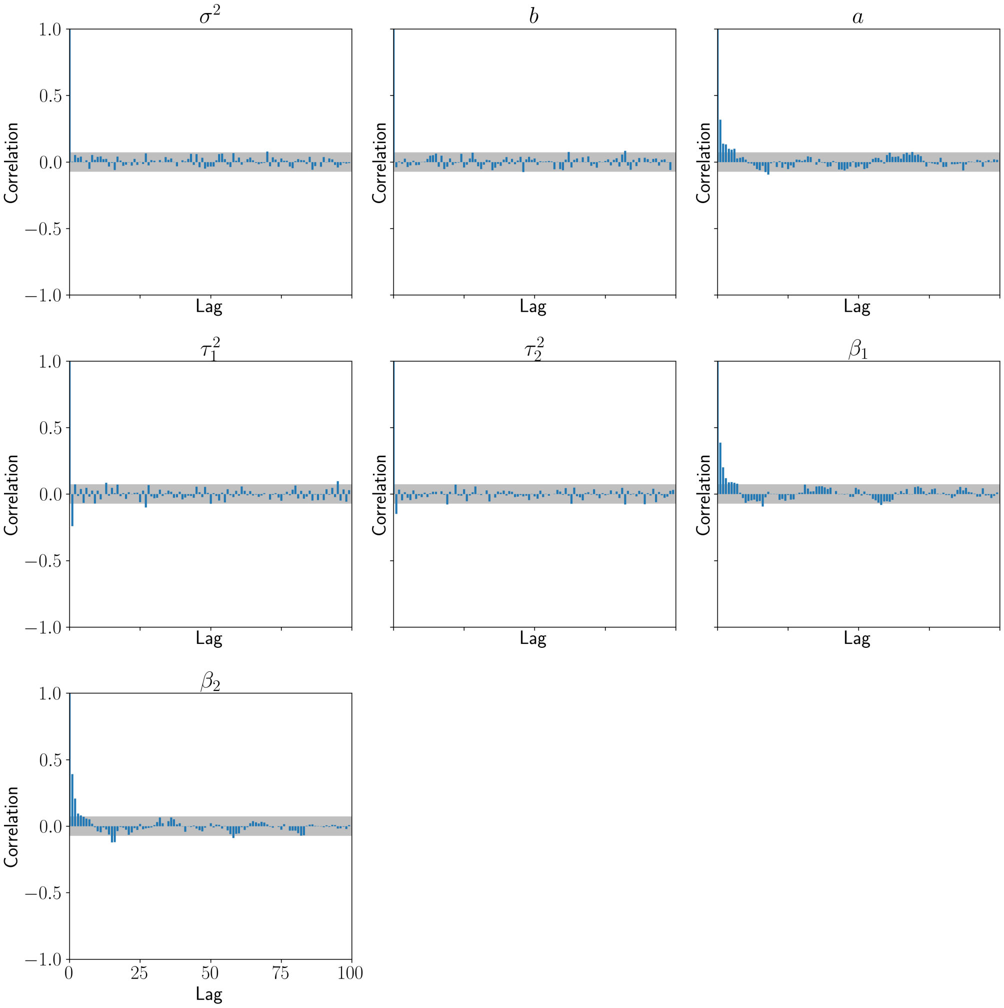

Figure 10: **Autocorrelation plots for MCMC samples from Multi-Group process model's posterior distribution, as described in Section 4.2. Gray bands cover** ±0.05.

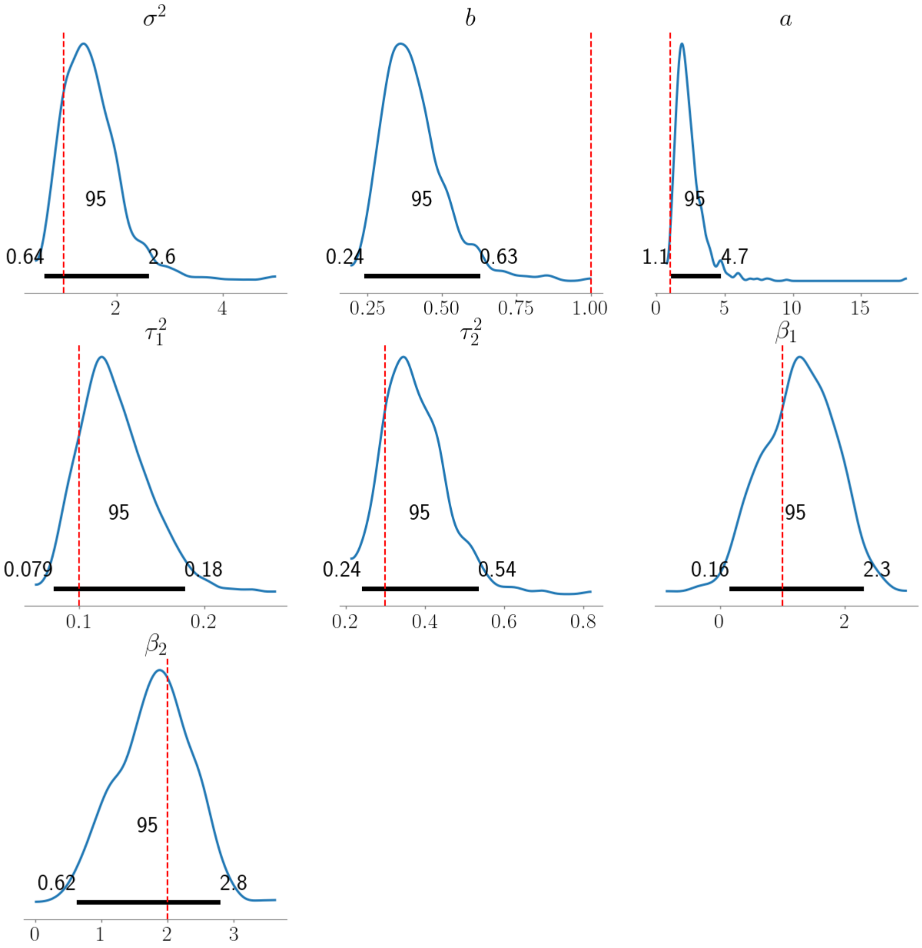

Figure 11: **Posterior samples of covariance function and model parameters from the Bayesian analysis described in Section 4.2**. Curves show the density of posterior samples for each parameter and black horizontal bars show the highest 95% density intervals for each set of samples. Red vertical lines indicate the parameter values used to generate the data.

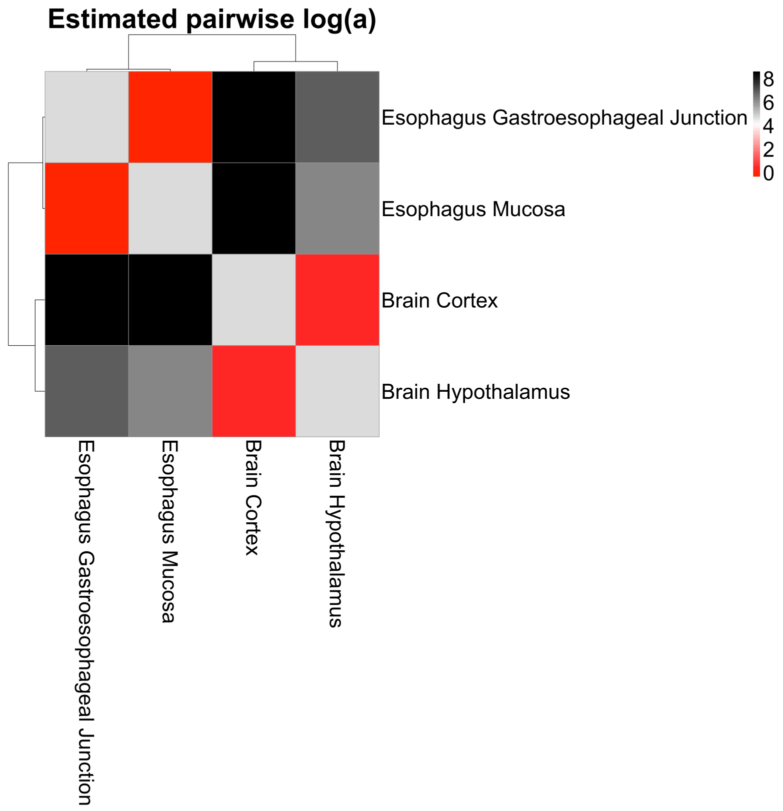

Figure 12: **Estimation of a for each pair of GTEx tissue types**. Cell ij in the heatmap represents $\log_{10}(a_{ij})$, where a_{ij} is the MLE of a when fitting the Multi-Group process using tissues i and j. Lower values of a (red) indicate higher similarity, while higher values of a (black) indicate lower similarity. Here, we allow each group its own noise variance τ_j^2.

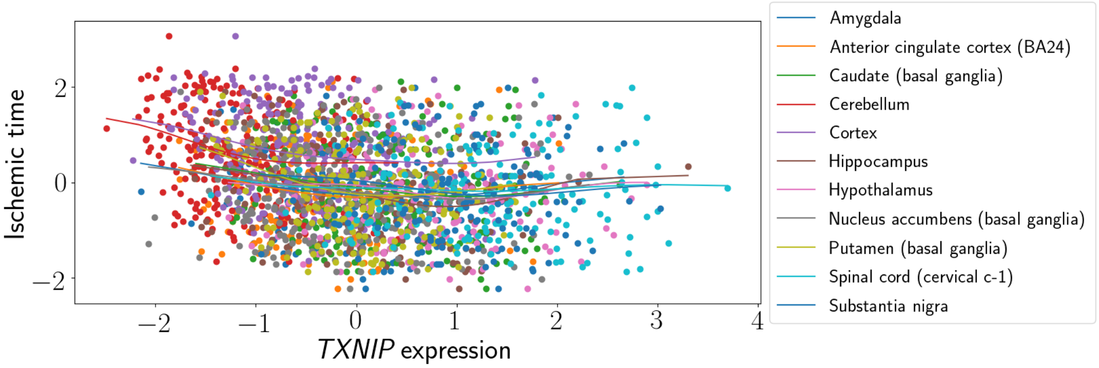

Figure 13: **Bayesian analysis of GTEx data across all brain tissues.** Points show the *TXNIP* expression and ischemic time for brain tissues. Each line shows the mean of the group-specific predictive process estimated using the MGGP.

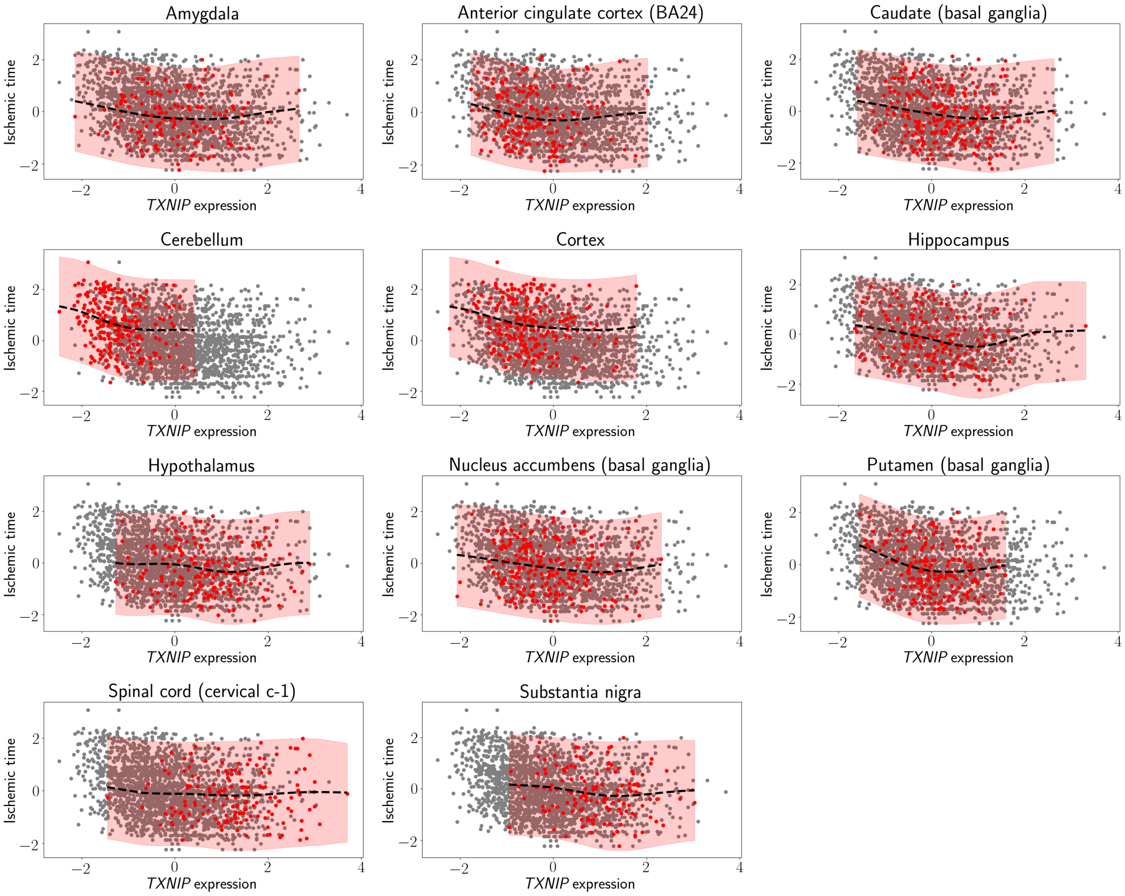

Figure 14: **Bayesian analysis of GTEx data for each brain tissue.** Points belonging to each group are colored in red. Points show the *TXNIP* expression and ischemic time for brain tissues. The dashed line in each panel shows the mean of the group-specific predictive process estimated using the MGGP, and the bands cover twice the standard deviation above and below the mean.