Doubly Mixed-Effects Gaussian Process Regression

Jun Ho Yoon **Daniel P. Jeong** **Seyoung Kim**
Computational Biology Department
School of Computer Science
Carnegie Mellon University
{junhoy, danielje, sssykim}@cs.cmu.edu

Abstract

We address the multi-task Gaussian process (GP) regression problem with the goal of decomposing input effects on outputs into components shared across or specific to tasks and samples. We propose a family of mixed-effects GPs, including doubly and translated mixed-effects GPs, that performs such a decomposition, while also modeling the complex task relationships. Instead of the tensor product widely used in multi-task GPs, we use the direct sum and Kronecker sum for Cartesian product to combine task and sample covariance functions. With this kernel, the overall input effects on outputs decompose into four components: fixed effects shared across tasks and across samples and random effects specific to each task and to each sample. We describe an efficient stochastic variational inference method for our proposed models that also significantly reduces the cost of inference for the existing mixed-effects GPs. On simulated and real-world data, we demonstrate that our approach provides higher test accuracy and interpretable decomposition.

1 INTRODUCTION

Gaussian processes (GPs) have been widely used for multi-task regression due to their strength as flexible nonparametric Bayesian models that also model uncertainty in prediction. Many of the previous multi-task GP regression methods used the tensor product to combine a covariance function for a single-task GP with a task correlation matrix (Liu et al., 2018, 2020; Bonilla et al., 2008). This approach, used in intrinsic models of coregionalization (IMC; Bonilla et al., 2008) and linear models of coregionalization (LMC; Goulard and Voltz, 1992), has the advantage that the estimations of the covariance function and the task correlation matrix decouple, which makes inference efficient, especially with stochastic variational inference (Titsias, 2009; Hoffman et al., 2013; Hensman et al., 2013). However, these and other multi-task GPs, such as collaborative multi-output GPs (COGPs; Nguyen and Bonilla, 2014) or convolved GPs (CVGPs; Álvarez and Lawrence, 2011), did not provide an explicit decomposition of input effects on outputs into meaningful and interpretable components.

As an alternative, mixed-effects GPs (Pillonetto et al., 2010; Wang and Khardon, 2012; Chung et al., 2020; Tonner et al., 2020) have been proposed to decompose input effects on outputs into fixed effects shared across all tasks and random effects specific to each task. However, they were too restrictive to model the complex dependencies across tasks, as the input effects on outputs were either the same across tasks as in fixed effects or independent across tasks as in random effects.

In this paper, we introduce a family of mixed-effects GPs for multi-task regression, including doubly mixed-effects GP and translated mixed-effects GP, which combines the advantages of both multi-task GPs with tensor-product kernels and mixed-effects GPs. Our doubly mixed-effects GP models the complex inter-task and inter-sample relationships and decomposes the input effects on the output functions into four components (Fig. 1): fixed effects shared by samples and by tasks and random effects specific to each sample and to each task. Excluding the sample-specific random effects leads to a translated mixed-effects GP that can model the task-specific translation of the functions for all samples.

Our approach overcomes the limitations of the existing multi-task GPs with Kronecker product or tensor prod-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

uct by combining the task and sample covariance functions with direct sum (Duvenaud et al., 2011; Yukawa, 2015; Pravesh and Roi, 2020) and Kronecker sum (Greenewald et al., 2019; Yoon and Kim, 2020; Zhang, 2020). Both the direct sum and Kronecker sum are the Cartesian product rather than the tensor product. We show that the direct sum leads to fixed effects shared across tasks and samples and that the Kronecker sum leads to random effects specific to each task and sample. Unlike IMC, our approach does not suffer from the cancellation of inter-task transfer for noiseless observations with block design, known as autokrigeability (Bonilla et al., 2008; Wackernagel, 2003).

We develop a stochastic variational inference method for doubly and translated mixed-effects GPs. Our approach breaks down the Cartesian-product kernel into fixed- and random-effects components in the variational distribution and introduces inducing points in each component for efficient inference. For mixed-effects GPs, in comparison to the previous method (Wang and Khardon, 2012), we reduce the expensive cost of inversion of the covariance matrix $\mathcal{O}(n^3p^3)$ for n samples and p tasks to $\mathcal{O}(n^3)$ for exact inference, and reduce $\mathcal{O}(n^3p)$ to $\mathcal{O}(m_X^3)$ with m_X inducing points for variational inference. On simulated and real-world data, we demonstrate that our approach provides accurate predictive models and an interpretable decomposition of input effects on outputs into four fixed and random effects components across tasks and samples.

2 DOUBLY AND TRANSLATED MIXED-EFFECTS GPs

We introduce a doubly mixed-effects GP for learning fixed and random effects across samples and tasks, and from this model, derive a translated mixed-effects GP. We consider a functional mapping from d inputs $\boldsymbol{x} \in \mathbb{R}^d$ to p outputs $f : \mathbb{R}^d \to \mathbb{R}^p$ with a GP prior. Let $\boldsymbol{g}_k$ be an r-dimensional task descriptor for the kth task. We model $f(\boldsymbol{x}_i, \boldsymbol{g}_k)$, the kth output for the ith sample with the task descriptor $\boldsymbol{g}_k$ and input $\boldsymbol{x}_i$, with two sets of mixed effects:

$$f(\boldsymbol{x}_i, \boldsymbol{g}_k) = \bar{f}_X(\boldsymbol{x}_i) + \bar{f}_G(\boldsymbol{g}_k) + \tilde{f}_X^k(\boldsymbol{x}_i) + \tilde{f}_G^i(\boldsymbol{g}_k). \quad (1)$$

In Eq. (1), $\bar{f}_X$ and $\tilde{f}_X^k$ are fixed effects shared across tasks and random effects specific to the kth task, respectively, whereas $\bar{f}_G$ and $\tilde{f}_G^i$ are fixed effects shared across samples and random effects for the ith sample, respectively (Fig. 1). Each component in Eq. (1) has its own zero-mean GP prior,

$$\bar{f}_X \sim \mathcal{GP}(0,\, \bar{k}_X(\boldsymbol{x}, \boldsymbol{x}')), \quad \tilde{f}_X^k \sim \mathcal{GP}(0,\, \tilde{k}_X(\boldsymbol{x}, \boldsymbol{x}')),$$
$$\bar{f}_G \sim \mathcal{GP}(0,\, \bar{k}_G(\boldsymbol{g}, \boldsymbol{g}')), \quad \tilde{f}_G^i \sim \mathcal{GP}(0,\, \tilde{k}_G(\boldsymbol{g}, \boldsymbol{g}')),$$

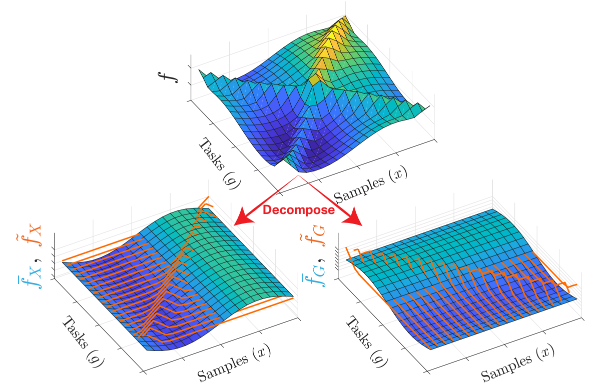


Figure 1: Illustration of the decomposition by doubly mixed-effects GPs. The overall input effects on multiple outputs (top) are decomposed into four components: task-wise fixed effects $\bar{f}_X$ (surface plot, left), task-wise random effects $\tilde{f}_X^k$'s (orange lines, left), sample-wise fixed effects $\bar{f}_G$ (surface plot, right), and sample-wise random effects $\tilde{f}_G^i$'s in Eq. (1) (orange lines, right).

where $\bar{k}_X(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}(\bar{f}_X(\boldsymbol{x}), \bar{f}_X(\boldsymbol{x}'))$, $\tilde{k}_X(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}(\tilde{f}_X^k(\boldsymbol{x}), \tilde{f}_X^k(\boldsymbol{x}'))$, $\bar{k}_G(\boldsymbol{g}, \boldsymbol{g}') = \mathrm{Cov}(\bar{f}_G(\boldsymbol{g}), \bar{f}_G(\boldsymbol{g}'))$, and $\tilde{k}_G(\boldsymbol{g}, \boldsymbol{g}') = \mathrm{Cov}(\tilde{f}_G^i(\boldsymbol{g}), \tilde{f}_G^i(\boldsymbol{g}'))$ are covariance functions. The random effects share the same kernels $\tilde{k}_X$ and $\tilde{k}_G$, and are mutually independent across tasks and across samples, i.e., $\mathrm{Cov}(\tilde{f}_X^i(\boldsymbol{x}), \tilde{f}_X^j(\boldsymbol{x}')) = 0$ and $\mathrm{Cov}(\tilde{f}_G^i(\boldsymbol{g}), \tilde{f}_G^j(\boldsymbol{g}')) = 0$ if $i \neq j$.

The doubly mixed-effects GP prior in Eq. (1) is equivalent to a GP prior that combines sample and task covariance functions through the direct-sum and Kronecker-sum operators, both of which are Cartesian products. Assume input data $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ for n samples, output data $\boldsymbol{Y} \in \mathbb{R}^{p\times n}$, task descriptors $\boldsymbol{G} \in \mathbb{R}^{r\times p}$, and a $p \times n$ matrix $\boldsymbol{F}$ with the (k, i)th element $[\boldsymbol{F}]_{ki} = f(\boldsymbol{x}_i, \boldsymbol{g}_k)$. Then, vec$(\boldsymbol{F})$ after stacking the columns of $\boldsymbol{F}$ into a vector has the following multivariate Gaussian distribution,

$$\mathrm{vec}(\boldsymbol{F}) \sim \mathcal{N}(\boldsymbol{0},\, \boldsymbol{K}), \quad (2)$$

where the $np \times np$ kernel matrix $\boldsymbol{K}$ is

$$\boldsymbol{K} = \bar{\boldsymbol{K}}_X \boxplus \bar{\boldsymbol{K}}_G + \tilde{\boldsymbol{K}}_X \oplus \tilde{\boldsymbol{K}}_G. \quad (3)$$

The direct-sum operator $\boxplus$ and the Kronecker-sum operator $\oplus$ above are defined as

$$\bar{\boldsymbol{K}}_X \boxplus \bar{\boldsymbol{K}}_G = \bar{\boldsymbol{K}}_X \otimes \mathbb{1}_{p,p} + \mathbb{1}_{n,n} \otimes \bar{\boldsymbol{K}}_G, \quad (4a)$$
$$\tilde{\boldsymbol{K}}_X \oplus \tilde{\boldsymbol{K}}_G = \tilde{\boldsymbol{K}}_X \otimes \boldsymbol{I}_p + \boldsymbol{I}_n \otimes \tilde{\boldsymbol{K}}_G, \quad (4b)$$

using the Kronecker-product operator $\otimes$, an $a \times a$ all-one matrix $\mathbb{1}_{a,a}$, and an $a \times a$ identity matrix $\boldsymbol{I}_a$. $\bar{\boldsymbol{K}}_X$, $\tilde{\boldsymbol{K}}_X$, $\bar{\boldsymbol{K}}_G$, and $\tilde{\boldsymbol{K}}_G$ are kernel matrices with the (i, j)th elements $[\bar{\boldsymbol{K}}_X]_{ij} = \bar{k}_X(\boldsymbol{x}_i,\, \boldsymbol{x}_j)$, $[\tilde{\boldsymbol{K}}_X]_{ij} = \tilde{k}_X(\boldsymbol{x}_i,\, \boldsymbol{x}_j)$,

$[\bar{\boldsymbol{K}}_G]_{ij} = \bar{k}_G(\boldsymbol{g}_i, \boldsymbol{g}_j)$, and $[\tilde{\boldsymbol{K}}_G]_{ij} = \tilde{k}_G(\boldsymbol{g}_i, \boldsymbol{g}_j)$. Then, the output data $\boldsymbol{Y} \in \mathbb{R}^{p\times n}$ are modeled as the noisy observations of $\boldsymbol{F}$, i.e., $\mathrm{vec}(\boldsymbol{Y}) \sim \mathcal{N}(\mathrm{vec}(\boldsymbol{F}), \sigma^2\boldsymbol{I}_{np})$.

The direct sum and Kronecker sum in Eq. (3) model doubly fixed effects and doubly random effects, respectively, combining the task effects and sample effects through a Cartesian product. In the direct sum in Eq. (4a), $\bar{\boldsymbol{K}}_X \otimes \mathbb{1}_{p,p}$ forces the rows of $\boldsymbol{F}$ to have the same fixed effects for all tasks, while $\mathbb{1}_{n,n} \otimes \bar{\boldsymbol{K}}_G$ forces the columns of $\boldsymbol{F}$ to have the same fixed effects for all samples. In the Kronecker sum in Eq. (4b), $\tilde{\boldsymbol{K}}_X \otimes \boldsymbol{I}_p$ encodes the random effects specific to each row of $\boldsymbol{F}$ for each task, whereas $\boldsymbol{I}_n \otimes \tilde{\boldsymbol{K}}_G$ encodes the independent random effects for each column or sample.

The kernel matrix of doubly mixed-effects GP in Eqs. (4a) and (4b) can be viewed as a special case of that of LMC. IMC used Eq. (2) with the tensor-product kernel matrix

$$\boldsymbol{K} = \boldsymbol{K}_X \otimes \boldsymbol{\Sigma}_G, \tag{5}$$

where $\boldsymbol{K}_X$ is an $n \times n$ kernel matrix for samples, and $\boldsymbol{\Sigma}_G$ is a $p \times p$ free-form positive semi-definite matrix modeling correlation among tasks. IMC was generalized to LMC with $\boldsymbol{K} = \sum_{s=1}^{S} \boldsymbol{K}_X^s \otimes \boldsymbol{\Sigma}_G^s$ with a set of kernel matrices $\{\boldsymbol{K}_X^s\}_{s=1}^S$ and task covariance matrices $\{\boldsymbol{\Sigma}_G^s\}_{s=1}^S$. Eqs. (4a) and (4b) can be obtained by fixing either the kernel or coregionalization matrix of LMC to an identity or all-one matrix.

However, there are several advantages to using the Cartesian product over the tensor product. First, the Cartesian product is known as a sparser and more interpretable counterpart to the tensor product (Kalaitzis et al., 2013; Imrich et al., 2008). Second, our model allows a decomposition of input effects on outputs into fixed and random effects across samples and tasks as we show in prediction below, providing insights into the input-output relationships that IMC and LMC cannot. Finally, as we detail later in this section, unlike IMC, our model does not suffer from the undesirable property known as autokrigeability, where inter-task transfer does not occur in prediction when the data is noiseless with block design (Wackernagel, 2003; Bonilla et al., 2008).

Translated Mixed-Effects GPs We obtain the translated mixed-effects GP by modifying the doubly mixed-effects GP in Eq. (1) to exclude the sample-specific random effects $\tilde{f}_G^i$:

$$f(\boldsymbol{x}_i, \boldsymbol{g}_k) = \bar{f}_X(\boldsymbol{x}_i) + \bar{f}_G(\boldsymbol{g}_k) + \tilde{f}_X^k(\boldsymbol{x}_i).$$

We use the term "translated," because the mixed-effects GP $f = \bar{f}_X + \tilde{f}_X^k$, which we describe below, is combined with the fixed effects $\bar{f}_G$ that plays the role of task-specific translation.

Mixed-Effects GPs Our doubly mixed-effects GP reduces to the mixed-effects GP (Pillonetto et al., 2010; Wang and Khardon, 2012; Chung et al., 2020), when we modify Eq. (1) to $f(\boldsymbol{x}_i, \boldsymbol{g}_k) = \bar{f}_X(\boldsymbol{x}_i) + \tilde{f}_X^k(\boldsymbol{x}_i)$ to include only task-wise mixed effects. This is equivalent to Eq. (2) with $\boldsymbol{K} = \bar{\boldsymbol{K}}_X \otimes \mathbb{1}_{p,p} + \tilde{\boldsymbol{K}}_X \otimes \boldsymbol{I}_p$.

If observations are available for all tasks for each sample, the covariance matrix in Eq. (2) can be conveniently written in terms of Cartesian products as in Eqs. (4a) and (4b) for all variations of mixed-effects GPs. Such a block design is commonly used in multi-task GPs, including LMC and IMC (van der Wilk et al., 2020). For our doubly and translated mixed-effects GPs and the existing mixed-effects GPs, it is not necessary to have a block design, where observations are available for all tasks per sample. However, in doubly mixed-effects GPs, observations for only few tasks per sample may not provide enough statistical power to model $\tilde{f}_G^i$ in Eq. (1) in a meaningful way. In an extreme case, with an observation for only one task per sample, doubly mixed-effects GPs cannot model $\tilde{f}_G^i$ but mixed-effects GPs and translated mixed-effects GPs can still model all of their random and fixed effects.

Prediction Given the doubly mixed-effects GP prior with the decomposition in Eq. (1), the posterior predictive distribution also decomposes into four components. Given n' new samples $\boldsymbol{X}^* = [\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_{n'}^*] \in \mathbb{R}^{d\times n'}$, the predicted outputs for existing tasks $\boldsymbol{F}^* \in \mathbb{R}^{p\times n'}$ have a similar decomposition

$$\boldsymbol{F}^* = \bar{\boldsymbol{F}}_{X^*} + \bar{\boldsymbol{F}}_G + \tilde{\boldsymbol{F}}_{X^*} + \tilde{\boldsymbol{F}}_G, \tag{6}$$

where $[\bar{\boldsymbol{F}}_{X^*}]_{ki} = \bar{f}_X(\boldsymbol{x}_i^*)$, $[\bar{\boldsymbol{F}}_G]_{ki} = \bar{f}_G(\boldsymbol{g}_k)$, $[\tilde{\boldsymbol{F}}_{X^*}]_{ki} = \tilde{f}_X^k(\boldsymbol{x}_i^*)$, and $[\tilde{\boldsymbol{F}}_G]_{ki} = \tilde{f}_G^i(\boldsymbol{g}_k)$. The distribution of $\boldsymbol{F}^*$ is given as Gaussian with the mean and covariance matrix formed by the sum of the means and covariance matrices of the four component-wise posterior predictive distributions,

$$\begin{aligned}
&p(\mathrm{vec}(\bar{\boldsymbol{F}}_{X^*}) \mid \boldsymbol{y}) = \mathcal{N}\big((\bar{\boldsymbol{K}}_{X^*X} \otimes \mathbb{1}_{p,p})\boldsymbol{\Gamma}\boldsymbol{y},\\
&\quad \bar{\boldsymbol{K}}_{X^*} \otimes \mathbb{1}_{p,p} - (\bar{\boldsymbol{K}}_{X^*X} \otimes \mathbb{1}_{p,p})\boldsymbol{\Gamma}(\bar{\boldsymbol{K}}_{XX^*} \otimes \mathbb{1}_{p,p})\big),\\
&p(\mathrm{vec}(\bar{\boldsymbol{F}}_G) \mid \boldsymbol{y}) = \mathcal{N}\big((\mathbb{1}_{n',n} \otimes \bar{\boldsymbol{K}}_G)\boldsymbol{\Gamma}\boldsymbol{y},\\
&\quad \mathbb{1}_{n',n'} \otimes \bar{\boldsymbol{K}}_G - (\mathbb{1}_{n',n} \otimes \bar{\boldsymbol{K}}_G)\boldsymbol{\Gamma}(\mathbb{1}_{n,n'} \otimes \bar{\boldsymbol{K}}_G)\big),\\
&p(\mathrm{vec}(\tilde{\boldsymbol{F}}_{X^*}) \mid \boldsymbol{y}) = \mathcal{N}\big((\tilde{\boldsymbol{K}}_{X^*X} \otimes \boldsymbol{I}_p)\boldsymbol{\Gamma}\boldsymbol{y},\\
&\quad \tilde{\boldsymbol{K}}_{X^*} \otimes \boldsymbol{I}_p - (\tilde{\boldsymbol{K}}_{X^*X} \otimes \boldsymbol{I}_p)\boldsymbol{\Gamma}(\tilde{\boldsymbol{K}}_{XX^*} \otimes \boldsymbol{I}_p)\big),\\
&p(\mathrm{vec}(\tilde{\boldsymbol{F}}_G) \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{n'} \otimes \tilde{\boldsymbol{K}}_G),
\end{aligned}$$

where $\boldsymbol{\Gamma} = (\boldsymbol{K} + \sigma^2\boldsymbol{I}_{np})^{-1}$, $\boldsymbol{y} = \mathrm{vec}(\boldsymbol{Y})$, $[\bar{\boldsymbol{K}}_{X^*X}]_{ij} = \bar{k}_X(\boldsymbol{x}_i^*, \boldsymbol{x}_j)$, and $[\tilde{\boldsymbol{K}}_{X^*X}]_{ij} = \tilde{k}_X(\boldsymbol{x}_i^*, \boldsymbol{x}_j)$. The posterior predictive distribution for the sample-specific random effects $\tilde{\boldsymbol{F}}_G$ is simply their prior $p(\tilde{\boldsymbol{F}}_G) = \prod_i p(\tilde{f}_G^i)$.

In the noiseless case with $\sigma^2 = 0$, the posterior predictive means under Eq. (6) depend on the entire observation vector $\boldsymbol{y}$. This implies that there exists inter-task transfer and that the model does not suffer from autokrigeability. To see this, notice that when $\sigma^2 = 0$, we have $\boldsymbol{\Gamma} = \boldsymbol{K}^{-1}$ and this inversion cannot be distributed to each term in $\boldsymbol{K}$ in Eq. (3), whereas with the Kronecker-product kernel in Eq. (5) the inversion can be distributed as $\boldsymbol{K}^{-1} = \boldsymbol{K}_X^{-1} \otimes \boldsymbol{\Sigma}_G^{-1}$, causing the inter-task transfer to cancel (Bonilla et al., 2008).

When data are available only for a subset of the tasks, given a single observation for the kth task and ith new sample y^*_{ki}, the posterior predictive distribution for the sample-specific random effects becomes

$$p([\tilde{\boldsymbol{F}}_G]_{:i} \mid y^*_{ki}) = \mathcal{N}\left(\tilde{\boldsymbol{K}}_{Gg_k}\gamma^*_{ki}y^*_{ki},\, \tilde{\boldsymbol{K}}_G - \gamma^*_{ki}\tilde{\boldsymbol{K}}_{Gg_k}\tilde{\boldsymbol{K}}^T_{Gg_k}\right),$$

where $[\tilde{\boldsymbol{F}}_G]_{:i}$ is the ith column of $\tilde{\boldsymbol{F}}_G$, $\tilde{\boldsymbol{K}}_{Gg_k}$ is the kernel matrix between $\boldsymbol{G}$ and $\boldsymbol{g}_k$, and $\gamma^*_{ki} = ([\bar{\boldsymbol{K}}_{X^*}]_{ii} + [\tilde{\boldsymbol{K}}_{X^*}]_{ii} + [\bar{\boldsymbol{K}}_G]_{kk} + [\tilde{\boldsymbol{K}}_G]_{kk} + \sigma^2)^{-1}$.

For mixed-effects GPs, compared to the previous inference methods (Wang and Khardon, 2012), we describe an approach that significantly reduces the computational cost for exact inference. The bottleneck operation in posterior inference in Eq. (6) is the inversion $\boldsymbol{\Gamma} = (\boldsymbol{K} + \sigma^2\boldsymbol{I}_{np})^{-1}$ with the time cost $\mathcal{O}(n^3p^3)$ as described in Wang and Khardon (2012). The following theorem shows that this cost can be significantly reduced to $\mathcal{O}(n^3)$ involving only an inversion of $n \times n$ matrices (proof in Supplementary Material A).

Theorem 1. *In mixed-effects GPs, $\boldsymbol{\Gamma}$ can be obtained as follows:*

$$(\boldsymbol{K} + \sigma^2\boldsymbol{I}_{np})^{-1} = \frac{1}{p}\Big((p\bar{\boldsymbol{K}}_X + \tilde{\boldsymbol{K}}_X + \sigma^2\boldsymbol{I}_n)^{-1} - (\tilde{\boldsymbol{K}}_X + \sigma^2\boldsymbol{I}_n)^{-1}\Big) \otimes \mathbb{1}_{p,p} + (\tilde{\boldsymbol{K}}_X + \sigma^2\boldsymbol{I}_n)^{-1} \otimes \boldsymbol{I}_p.$$

In other previous works, to reduce the cost $\mathcal{O}(n^3p^3)$ to $\mathcal{O}(np^3)$, fixed effects were modeled with linear models (Shi et al., 2012) or with deep neural networks (Chung et al., 2020) and only random effects were modeled with GPs. With our simple strategy described above, we are able to reduce the cost, while keeping the GP prior for the fixed effects.

3 VARIATIONAL INFERENCE

The computational bottleneck for learning and inference in the doubly mixed-effects GPs is the inversion of the large $np \times np$ matrix $(\boldsymbol{K} + \sigma^2\boldsymbol{I}_{np})$ in $\mathcal{O}(n^3p^3)$ time. To improve the computational efficiency, we adopt the sparse variational GP framework with mini-batch training (Hensman et al., 2013). We set up the variational distribution such that the posterior predictive distribution for each of the four random and fixed effects components in Eq. (6) involves only task or sample covariance matrices, not the large Cartesian product of the two matrices. Our approach collapses the large covariance matrix in exact inference such that matrix inversion is performed only on the smaller individual task and sample covariance matrices.

We define inducing points $\mathcal{U} = \{\bar{\boldsymbol{u}}_X,\, \tilde{\boldsymbol{u}}_X^{1:p},\, \bar{\boldsymbol{u}}_G,\, \tilde{\boldsymbol{u}}_G^{1:n}\}$ for fixed and random effects and a set of latent functions $\mathcal{Z} = \{\bar{\boldsymbol{Z}}_X,\, \tilde{\boldsymbol{Z}}_X,\, \bar{\boldsymbol{Z}}_G,\, \tilde{\boldsymbol{Z}}_G\}$ evaluated at m_X and m_G inducing inputs for samples and for tasks. We assume that the fixed and random effects have the same m_X for samples and m_G for tasks, but this can be relaxed. Let $\mathcal{F} = \{\bar{\boldsymbol{f}}_X,\, \tilde{\boldsymbol{f}}_X^{1:p},\, \bar{\boldsymbol{f}}_G,\, \tilde{\boldsymbol{f}}_G^{1:n}\}$ denote the collection of fixed and random effects given the input data. Then, we approximate the posterior $p(\mathcal{F}, \mathcal{U} \mid \boldsymbol{y})$ with the following variational distribution,

$$q(\mathcal{F},\mathcal{U}) = p(\bar{\boldsymbol{f}}_X \mid \bar{\boldsymbol{u}}_X)q(\bar{\boldsymbol{u}}_X)\prod_{k=1}^{p} p(\tilde{\boldsymbol{f}}_X^k \mid \tilde{\boldsymbol{u}}_X^k)q(\tilde{\boldsymbol{u}}_X^k) \cdot p(\bar{\boldsymbol{f}}_G \mid \bar{\boldsymbol{u}}_G)q(\bar{\boldsymbol{u}}_G)\prod_{i=1}^{n} p(\tilde{\boldsymbol{f}}_G^i \mid \tilde{\boldsymbol{u}}_G^i)q(\tilde{\boldsymbol{u}}_G^i),$$

with the following independent variational priors on the fixed and random effects

$$\begin{aligned} q(\bar{\boldsymbol{u}}_X) &= \mathcal{N}(\bar{\boldsymbol{m}}_X,\, \bar{\boldsymbol{S}}_X), \quad & q(\tilde{\boldsymbol{u}}_X^k) &= \mathcal{N}(\tilde{\boldsymbol{m}}_X^k,\, \tilde{\boldsymbol{S}}_X^k), \\ q(\bar{\boldsymbol{u}}_G) &= \mathcal{N}(\bar{\boldsymbol{m}}_G,\, \bar{\boldsymbol{S}}_G), \quad & q(\tilde{\boldsymbol{u}}_G^i) &= \mathcal{N}(\tilde{\boldsymbol{m}}_G^i,\, \tilde{\boldsymbol{S}}_G^i). \end{aligned} \tag{7}$$

We construct the variational distribution such that the four doubly mixed-effects components are independent of each other by assigning separate inducing inputs, mean vector, and free-form covariance matrix to each component. The limitation of this variational approximation is that it may be more prone to overfitting than the exact posterior because each component has its own variational parameters. However, as we show in our experiments in Section 4, our methods outperformed other methods on test accuracy.

Then, we optimize the evidence lower bound (ELBO) for doubly mixed-effects GP

$$\mathcal{L} = \mathbb{E}_{q(\mathcal{F})}\big[\log p(\boldsymbol{y} \mid \mathcal{F})\big] - \mathrm{KL}\left[q(\mathcal{U}) \,||\, p(\mathcal{U})\right]. \tag{8}$$

If task descriptors $\boldsymbol{g}$ in Eq. (1) are not available, the inducing points over tasks $\bar{\boldsymbol{u}}_G$ and $\tilde{\boldsymbol{u}}_G^{1:n}$ cannot be defined. However, by allowing the task covariances $\bar{k}_G(\boldsymbol{g}, \boldsymbol{g}') = \mathrm{Cov}(\bar{f}_G(\boldsymbol{g}), \bar{f}_G(\boldsymbol{g}'))$ and $\tilde{k}_G(\boldsymbol{g}, \boldsymbol{g}') = \mathrm{Cov}(\tilde{f}_G^i(\boldsymbol{g}), \tilde{f}_G^i(\boldsymbol{g}'))$ to be free-form covariances $\bar{\boldsymbol{\Sigma}}_G$ and $\tilde{\boldsymbol{\Sigma}}_G$, the variational parameters with inducing points over samples and the free-form covariances can be optimized. Alternatively, our model can be combined with methods for learning features to jointly extract latent task features.

Prediction Unlike the exact posterior in Eq. (6), the variational posterior predictive distribution $q(\mathcal{F})$

involves inversions of smaller $m_X \times m_X$ and $m_G \times m_G$ matrices. Given new samples $\boldsymbol{X}^*$, the variational posterior predictive distributions for $\bar{\boldsymbol{f}}_{X^*}$, $\bar{\boldsymbol{f}}_G$, and $\tilde{\boldsymbol{f}}^k_{X^*}$ can be obtained from $q\big(\bar{\boldsymbol{f}}_{X^*}\big) = \int p(\bar{\boldsymbol{f}}_{X^*} \mid \bar{\boldsymbol{u}}_X)q(\bar{\boldsymbol{u}}_X)d\bar{\boldsymbol{u}}_X$ and similarly for the other random and fixed effects,

$$\begin{aligned} q\big(\bar{\boldsymbol{f}}_{X^*}\big) &= \mathcal{N}\big(\bar{\boldsymbol{K}}_{X^*\bar{Z}_X}\bar{\boldsymbol{K}}^{-1}_{\bar{Z}_X}\bar{\boldsymbol{m}}_X, \\ &\qquad \bar{\boldsymbol{K}}_{X^*} + \bar{\boldsymbol{K}}_{X^*\bar{Z}_X}\bar{\boldsymbol{K}}^{-1}_{\bar{Z}_X}(\bar{\boldsymbol{S}}_X - \bar{\boldsymbol{K}}_{\bar{Z}_X})\bar{\boldsymbol{K}}^{-1}_{\bar{Z}_X}\bar{\boldsymbol{K}}_{\bar{Z}_X X^*}\big), \\ q\big(\bar{\boldsymbol{f}}_G\big) &= \mathcal{N}\big(\bar{\boldsymbol{K}}_{G\bar{Z}_G}\bar{\boldsymbol{K}}^{-1}_{\bar{Z}_G}\bar{\boldsymbol{m}}_G, \\ &\qquad \bar{\boldsymbol{K}}_G + \bar{\boldsymbol{K}}_{G\bar{Z}_G}\bar{\boldsymbol{K}}^{-1}_{\bar{Z}_G}(\bar{\boldsymbol{S}}_G - \bar{\boldsymbol{K}}_{\bar{Z}_G})\bar{\boldsymbol{K}}^{-1}_{\bar{Z}_G}\bar{\boldsymbol{K}}_{\bar{Z}_G G}\big), \\ q\big(\tilde{\boldsymbol{f}}^k_{X^*}\big) &= \mathcal{N}\big(\tilde{\boldsymbol{K}}_{X^*\tilde{Z}_X}\tilde{\boldsymbol{K}}^{-1}_{\tilde{Z}_X}\tilde{\boldsymbol{m}}^k_X, \\ &\qquad \tilde{\boldsymbol{K}}_{X^*} + \tilde{\boldsymbol{K}}_{X^*\tilde{Z}_X}\tilde{\boldsymbol{K}}^{-1}_{\tilde{Z}_X}(\tilde{\boldsymbol{S}}^k_X - \tilde{\boldsymbol{K}}_{\tilde{Z}_X})\tilde{\boldsymbol{K}}^{-1}_{\tilde{Z}_X}\tilde{\boldsymbol{K}}_{\tilde{Z}_X X^*}\big). \end{aligned}$$

Notice that the variational posterior predictive distribution for the sample-specific random effects $\tilde{\boldsymbol{f}}^i_G$ is just the prior distribution because $p(\tilde{\boldsymbol{f}}^i_G \mid \tilde{\boldsymbol{u}}^j_G) = p(\tilde{\boldsymbol{f}}^i_G)$ for $i \neq j$. This prediction has complexity $\mathcal{O}(pm_X^2 + m_X^3 + m_G^3)$.

For mixed-effects GP, our approach performs prediction for a new sample with the cost of matrix inversion $\mathcal{O}(m_X^3)$, significantly less than $\mathcal{O}(pn^3)$ previously discussed in Wang and Khardon (2012). While Wang and Khardon (2012) also used sparse variational inference, they introduced inducing points only for fixed effects, and thus, an inversion of $n \times n$ matrix was required to infer random effects for each of the p tasks.

Mini-Batch Training Since the elements of $\boldsymbol{Y}$ are conditionally independent given the mixed effects, and the expected log-likelihood term in Eq. (8) decomposes across samples and tasks, we can optimize the ELBO by using noisy estimates of the gradient (Hoffman et al., 2013; Hensman et al., 2013). We subsample b_X samples and b_G task descriptors into mini-batches of shape (b_X, b_G) and estimate the expected log-likelihood term in Eq. (8) as the expected log-likelihood only for the corresponding observations in $\boldsymbol{Y}$ scaled by $\frac{np}{b_X b_G}$.

Identifiable Parameters The diagonal elements of two matrices combined with Kronecker sum are known to be unidentifiable as $\boldsymbol{A} \oplus \boldsymbol{B} = (\boldsymbol{A} + c\boldsymbol{I}) \oplus (\boldsymbol{B} - c\boldsymbol{I})$ for any $c \in \mathbb{R}$ (Greenewald et al., 2019), and similarly for the elements in direct sum since $\boldsymbol{A} \boxplus \boldsymbol{B} = (\boldsymbol{A} + c\mathbb{1}) \boxplus (\boldsymbol{B} - c\mathbb{1})$. However, the following theorem states that doubly mixed-effects GPs do not suffer from this unidentifiability (proof in Supplementary Material A).

Theorem 2. *The ELBO in Eq.* (8) *is minimized by a unique* $\{\bar{\boldsymbol{S}}_X, \bar{\boldsymbol{S}}_G, \tilde{\boldsymbol{S}}^{1:p}_X, \tilde{\boldsymbol{S}}^{1:n}_G\}$. *In other words, the ELBO changes if we substitute* $\{\bar{\boldsymbol{S}}_X, \bar{\boldsymbol{S}}_G, \tilde{\boldsymbol{S}}^{1:p}_X, \tilde{\boldsymbol{S}}^{1:n}_G\}$ *with* $\{\bar{\boldsymbol{S}}_X + c\mathbb{1}, \bar{\boldsymbol{S}}_G - c\mathbb{1}, \tilde{\boldsymbol{S}}^1_X + c\boldsymbol{I}, \ldots, \tilde{\boldsymbol{S}}^p_X + c\boldsymbol{I}, \tilde{\boldsymbol{S}}^1_G - c\boldsymbol{I}, \ldots, \tilde{\boldsymbol{S}}^n_G - c\boldsymbol{I}\}$ *for any* $c \in \mathbb{R}$.

Extensions Our doubly and translated mixed-effects GPs can be extended in a straightforward manner. The SOLVE-GP framework (Shi et al., 2020) can be adopted by decomposing the fixed and random effects into orthogonal components. Our approach can be extended to deep GPs (Damianou and Lawrence, 2013; Bui et al., 2016) by recursively putting our GP prior on the function output of each layer and adopting the doubly stochastic variational inference framework (Salimbeni and Deisenroth, 2017).

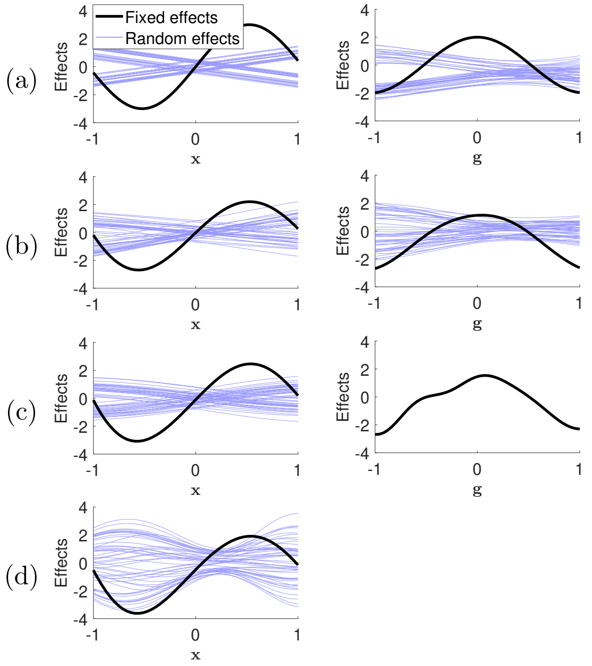


Figure 2: Fixed and random effects estimated by different methods on simulated data. (a) Ground truth, (b) doubly mixed-effects GP, (c) translated mixed-effects GP, and (d) mixed-effects GP. Task-wise mixed effects (left) and sample-wise mixed effects (right).

4 EXPERIMENTS

We compare the performance of doubly and translated mixed-effects GPs with those of mixed-effects GP, IMC, LMC, COGP, and CVGP on simulated and four real-world datasets.

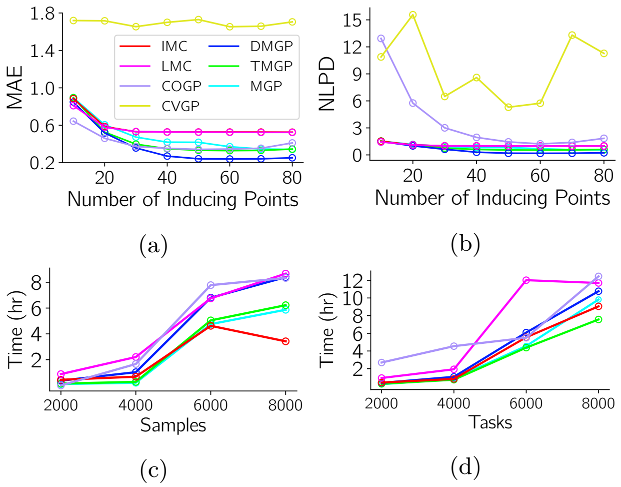
We implemented all GPs with mixed effects to be compatible with the GPflow framework (Matthews et al., 2017) in TensorFlow (Abadi et al., 2015). For LMC, we used the implementation of stochastic variational inference available in GPflow (van der Wilk et al., 2020), where in $\boldsymbol{K} = \sum_{s=1}^S \boldsymbol{K}^s_X \otimes \Sigma^s_G$, each Σ^s_G has rank 1. For IMC, we modified the implementation of LMC such that with $S = 1$, Σ^1_G can have an arbitrary rank L. For all models, we used squared exponential kernels $SE(\sigma^2, \ell) = \sigma^2 \cdot \exp(-\frac{1}{2\ell^2}\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2)$. We initialized the kernel hyperparameters to $\sigma^2 = 1$, $\ell = 1$, the inducing inputs to the cluster centers identified by k-means, and the variational parameters to zero-mean and prior covariance matrix of the given inducing points. With the Adam optimizer (Kingma and Ba, 2015), we used the default learning rate $\eta = 10^{-3}$ and momentum hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We computed the absolute percent change in the average ELBO evaluated for 10 successive iterations and trained until we observed five drops below tolerance $\epsilon = 10^{-4}$. For COGP and CVGP, we modified the codes provided by the authors and used the same convergence criterion on their objectives, ELBO for COGP and log-likelihood for CVGP. For COGP and CVGP, we chose the number of latent GPs between 1 and 5 with the best accuracy.

For the models with mixed effects, we compared the decomposition of input effects on outputs, as the other models do not provide such a decomposition. We evaluated all methods on prediction accuracy using mean absolute error (MAE) and prediction uncertainty using negative log predictive density (NLPD) on heldout test data. For prediction in doubly mixed-effects GPs, we added 10% of the observations in each task in the test data to the training data to allow the model to recover the corresponding sample-specific random effects.

4.1 Simulated Data

We illustrate the decomposition of fixed and random effects by a doubly mixed-effects GP on a single simulated dataset, comparing with those from the other methods with mixed effects. We generated a dataset with 50 tasks and 50 samples and with a single dimensional input x and task descriptor g, both linearly spaced in $[-1, 1]$, from sinusoidal and linear functions, $\bar{f}_X = 3\sin(3x)$, $\bar{f}_G = 2\cos(3g)$, $\tilde{f}_X^{1:25} = -x - 0.5 + \alpha$, $\tilde{f}_X^{26:50} = x + 0.5 - \alpha$, $\tilde{f}_G^{1:25} = -g\cos(g + \phi) - 0.5 + \alpha$, and $\tilde{f}_G^{26:50} = g\cos(g + \phi) - 0.5 - \alpha$, where each $\alpha, \phi \sim \text{Uniform}(0, 1)$. For all methods, we used mini-batches of shape $(10, 10)$ and 10 inducing points for both samples and tasks. Given the true model (Fig. 2(a)), the doubly mixed-effects GP recovered all underlying effects most accurately (Fig. 2(b)). In the translated mixed-effects GP, the missing sample-specific random effects were absorbed into the other three components, most notably into the fixed effect shared across samples, which was made less accurate (Fig. 2(c)). The mixed-effects GP recovered task-specific random effects poorly, as it does not learn the sample-wise mixed effects (Fig. 2(d)).

We compared all methods on test accuracy while varying the number of inducing points. We simulated a dataset with 30 tasks and 100 samples. With x and g linearly spaced in $[-10, 10]$, we sampled each fixed and random effect from a GP prior with the kernels $\bar{k}_X = SE(1, 1)$, $\tilde{k}_X = SE(2, 0.5)$, $\bar{k}_G = SE(3, 1)$, and $\tilde{k}_G = SE(0.1, 2)$, and add noise from $\mathcal{N}(0, 0.1^2)$. We randomly selected 20% of the samples as a test set. We varied $m_X = 10, 20, \ldots, 80$ and $m_G = 10, 20, 30$. We used $L = 15$ for IMC, $S = 15$ for LMC, and one latent GP for both COGP and CVGP. We used mini-batches of shape $(b_X, b_G) = (15, 20)$ for all models with mixed effects, and $(b_X, b_G) = (30, 20)$ for the other methods. Experiments for each setting were repeated 10 times with different initializations. Doubly mixed-effects GP almost always outperformed all other methods in MAE (Fig. 3(a)) and in NLPD (Fig. 3(b)). All mixed-effects models outperformed IMC and LMC: while IMC and LMC required around 40 inducing points to achieve their optimal test accuracy, translated and doubly mixed-effects GPs needed only about half as many inducing points for each component to achieve the same accuracy and performed significantly better with more inducing points. All of the methods showed little variance in MAE and NLPD over 10 repetitions.



(a)

(b)

(c)

(d)

Figure 3: Comparison of methods on simulated data. (a) MAE and (b) NLPD on test data. (c) Runtime, as we vary the number of samples with a fixed number of tasks $p = 200$, and (d) runtime, as we vary the number of tasks with a fixed number of samples $n = 200$.

Next, we compared all methods on computation time. We sampled data from the same GP priors above, either fixing the number of tasks to 200 and varying the number of samples as $[2000, 4000, 6000, 8000]$ or vice versa, with 1.6 million data points for the largest dataset. For each dataset, we fixed the number of inducing points to 30 and averaged the runtime over 5 different initializations. When fixing the number of tasks, we set $L = 100$ for IMC and $S = 100$ for LMC. When fixing the number of samples, we set $L = 500$ for IMC and $S = 500$ for LMC. For COGP, we used five latent GPs, as it resulted in the best performance. We report the runtime on AMD EPYC 7742 CPUs each with 64 cores, 256GB of RAM, 2.25-3.40 GHz clock speed, 256MB of L3 cache, and 8 memory channels.

The mixed-effects GP and translated mixed-effects GP were almost always the fastest (Figs. 3(c) and 3(d)). Doubly mixed-effects GP generally required more time than IMC and the other mixed-effects GPs since it has additional variational parameters for the sample-specific random effects. CVGP ran out of memory on the smallest dataset.

4.2 Real-World Data

On four real-world datasets, we examine the decomposition of fixed and random effects learned by our methods and compare all methods on prediction accuracy for test data. We set the number of inducing points m_X such that they have real-world interpretation: one inducing point per week for the COVID data, per 3 weeks for the NASA temperature data, per 2 weeks for the UK house price data, and per 4 streets for NYC taxi data. For the models with mixed effects, we used half as many inducing points as the models without mixed effects. For IMC and LMC, we set the number of latent GPs to about 10% of the number of tasks, and for translated and doubly mixed-effects GPs, we set the number of inducing points m_G to half of the number of latent GPs in IMC and LMC. For all methods, increasing the number of inducing points did not significantly increase the performance.

United States COVID-19 We obtained the daily confirmed cases of COVID-19 from the COVID-19 Data Repository of the Center for Systems Science and Engineering at Johns Hopkins University (Dong et al., 2020). We used the case counts for 3,091 counties in the United States over 273 days from July 2020 to March 2021. We fit all models to this data, treating days as samples and counties as tasks, and predicted the confirmed cases in all counties for the last 63 days given the first 210 days. We performed the same analysis on the transposed data, treating days as tasks and counties as samples.

In the decomposition, the doubly mixed-effects GP captured how the pandemic evolved in the most meaningful way (Fig. 4). For doubly mixed-effects GP, the fixed effects shared across time captured the regions in California and Florida known for elevated case counts throughout the pandemic (Fig. 4(a) top left), the time-specific random effects recovered the short-term surge in Wisconsin (Oct 2020) and Louisiana and Massachusetts (Jan 2021) (Fig. 4(a) bottom), the fixed effects shared across all counties showed the nationwide fluctuation in cases over time (Fig. 4(a) top right, black curve), and the county-specific random effects recovered the changes in cases over time for each county (Fig. 4(a) top right, red/blue curves). The elevated cases in California and Florida were not captured by

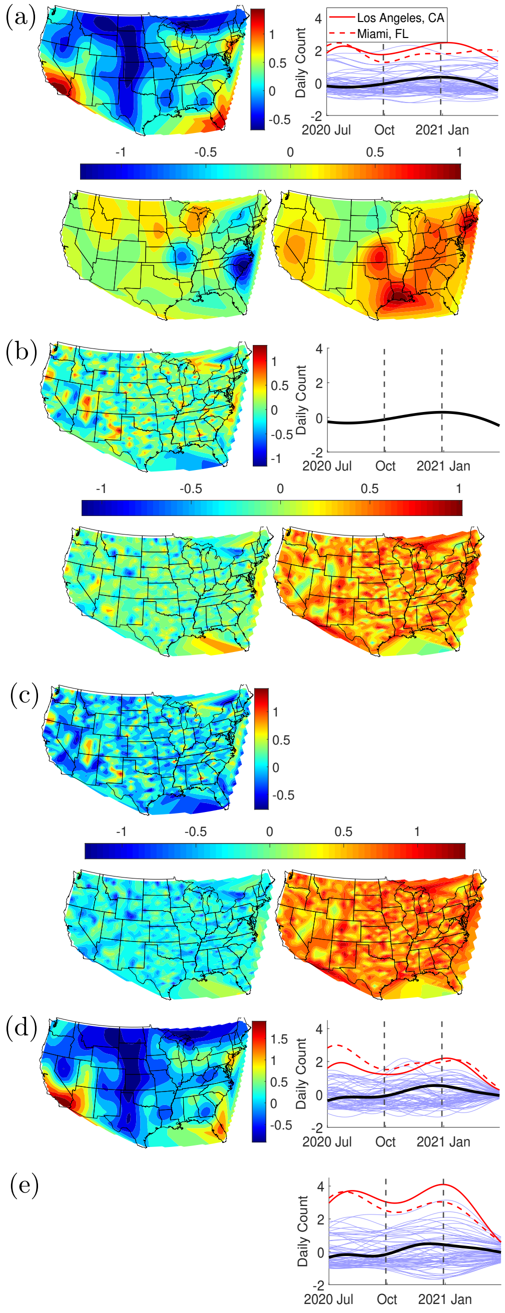


Figure 4: Decomposition of fixed and random effects by different methods for COVID-19 data. (a) Doubly mixed-effects GP, (b) translated mixed-effects GP, and (c) mixed-effects GP, with days as samples and counties as tasks. (d) Translated mixed-effects GP and (e) mixed-effects GP, with the transposed data with counties as samples and days as tasks. In each panel, fixed effects shared across time (top left), time-specific random effects on two days, one in October 2020 and the other in January 2021 (bottom), and fixed effects shared across counties and county-specific random effects for 2% of the counties (black and blue, top right) are shown, if available from the given method.

Table 1: Prediction accuracy on real-world data.

Dataset	(n, p)	MAE							NLPD						
		DMGP	TMGP	MGP	IMC	LMC	COGP	CVGP	DMGP	TMGP	MGP	IMC	LMC	COGP	CVGP
COVID-19	(273, 3091)	**0.3624**	0.5086	0.5028	0.6302	0.6578	0.5164		**0.7127**	1.0026	1.0045	1.3261	1.3691	56.0848	
COVID-19	(3091, 273)	**0.3686**	0.5872	0.6188	0.5872	0.5815	0.5964		**0.7209**	1.1427	1.1844	1.1404	1.1321	9.4386	
NASA Air	(72, 576)	**0.2235**	0.5688	0.5487	0.5849	0.6133	0.5375	0.6281	**0.3682**	1.1721	1.1281	1.2903	1.2643	549.5703	2.7537
UK House	(36, 2290)	**0.1934**	0.2144	0.2332	0.2474	0.2446	0.3075	0.5944	**0.1758**	0.2318	0.2807	1.1459	1.1473	6.8704	6.6488
Taxi Time	(102, 9)	0.1731	**0.1723**	0.1730	0.1788	0.1793	0.6866	0.3723	0.1324	0.1310	**0.1232**	0.1492	0.1424	700.5208	51.4264
Taxi Fare	(102, 9)	**0.0955**	0.0959	0.0963	0.0999	0.0991	0.7916	0.2187	-0.4934	-0.5046	**-0.5053**	-0.4860	-0.4906	434.8118	48.0812

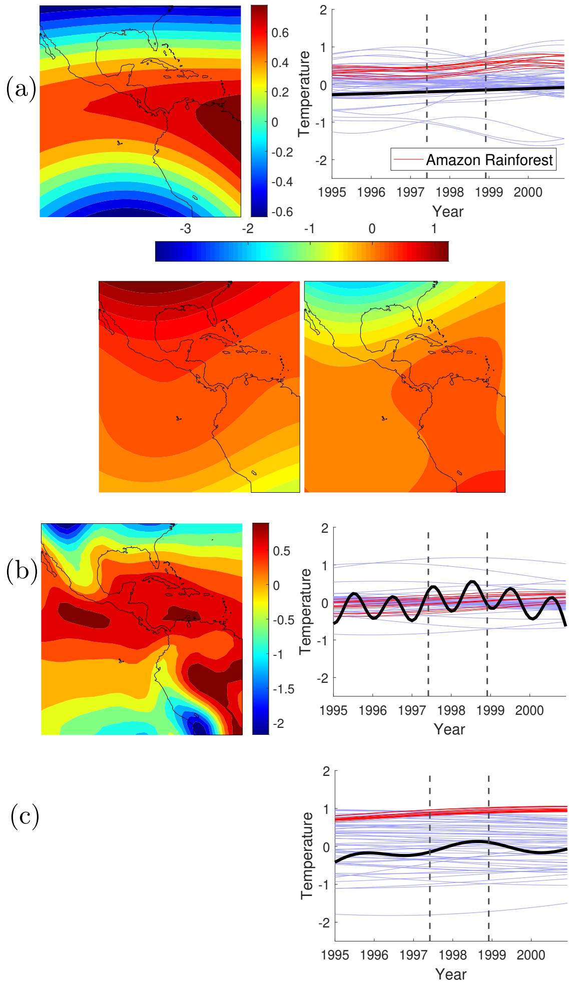


Figure 5: Decomposition of fixed and random effects on the NASA temperature data. (a) Doubly mixed-effects GP, (b) translated mixed-effects GP, and (c) mixed-effects GP. In each panel, the fixed effects shared across time (top left), time-specific random effects for the summer and winter months in the northern hemisphere (bottom), and fixed effects across grids and 2% of grid-specific random effects (black and red/blue, top right) are shown, if available from the given method.

the translated mixed-effects GP (Fig. 4(b) top left) and mixed-effects GP (Fig. 4(c) top left). Only for the transposed data, the translated mixed-effects GP was able to identify these regions (Fig. 4(d) left). Only the doubly mixed-effects GP was able to identify the short-term surge in Wisconsin, Louisiana, and Massachusetts. This suggests that mixed effects across samples and tasks should be considered jointly as in doubly mixed-effects GPs to accurately identify all meaningful spatial/temporal effects on the case counts.

NASA Central America Air Temperature We analyzed air temperature data from the NASA Langley Research Center Atmospheric Sciences Data Center (Murrell, 2010). We obtained air temperature data on a 24-by-24 grid covering Central America, collected for 72 months from 1995 to 2000. With $n = 72$ samples for time points and $p = 24 \times 24$ tasks for grid points, our goal is to predict the temperature of the last 12 months at all grid locations given the data for the first 60 months.

In the decomposition, again, the doubly mixed-effects GP extracted interpretable patterns in temperature change that the other methods failed to model (Fig. 5). For the doubly mixed-effects GP, fixed effects shared across all time points (Fig. 5(a), top left) captured the general trend in the temperature: warmer near the equator and cooler away from the equator. The time-specific random effects (Fig. 5(a), bottom) captured the seasonal effects: the opposite seasons in the southern and northern hemispheres. The fixed effects shared across locations (Fig. 5(a), top right, black curve) did not show notable effects; however, the random effects specific to the Amazon rainforest in Brazil (Fig. 5(a), top right, red curves) showed the location-specific rise in temperature over time due to deforestation. The translated mixed-effects GP (Fig. 5(b)) and mixed-effects GP (Fig. 5(c)) were unable to model the effects from the opposite seasons in the different hemispheres; in fact, the temperature fluctuations in the fixed effects shared across locations (Figs. 5(b) and (c), right, black curves) were not meaningful, as they either corresponded to seasons only in the northern hemisphere or did not correspond to seasons at all, suggesting all four components should be modeled jointly to adequately uncover each component.

New York City Taxi Trip and United Kingdom House Price We provide the experimental details and decompositions in Supplementary Material B.

Prediction Accuracy Doubly mixed-effects GPs achieved higher accuracy than the other methods on most datasets (Table 1). Translated mixed-effects GPs and mixed-effects GPs yielded lower NLPD on NYC taxi datasets, because the taxi trips from Midtown to Upper Manhattan have strong task-wise mixed effects over the streets with increasing fare and time but only weak sample-wise mixed effects over avenues. CVGP ran out of memory even with 5 inducing points on COVID-19 data, so we were unable to obtain the prediction accuracy.

5 CONCLUSION

We introduced doubly and translated mixed-effects GPs as multi-task nonparametric Bayesian regression methods that model mixed effects for samples and tasks, using direct-sum kernels for fixed effects and Kronecker-sum kernels for random effects. We demonstrated that our approach can obtain an interpretable decomposition of input effects on outputs.

There are several limitations of our proposed methods that remain as future work. To handle datasets without task descriptors, doubly mixed-effects GPs could be extended such that they learn the model and latent task descriptors jointly or work with free-form task correlation matrices as in IMC and LMC. In order to allow the model to flexibly learn more complex fixed effects over samples and tasks, the doubly and translated mixed-effects GPs could be generalized to work with mixtures of GPs instead of a single GP. Additionally, our implementation could be extended to handle a non-block design. Finally, to improve the accuracy and efficiency of our stochastic variational inference strategy, our implementation could be extended to use natural gradients for optimization of the variational parameters.

Acknowledgements

This work was supported by NIH 1R21HG011116 and 1R21HG010948. DPJ was supported by the ARCS Foundation. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF ACI-1548562 and ACI-1445606.

References

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`.

M. A. Álvarez and N. D. Lawrence. Computationally efficient convolved multiple output Gaussian processes. *Journal of Machine Learning Research*, 12 (41):1459–1500, 2011.

E. V. Bonilla, K. Chai, and C. K. I. Williams. Multitask Gaussian process prediction. In *Advances in Neural Information Processing Systems*, 2008.

T. Bui, D. Hernandez-Lobato, J. Hernandez-Lobato, Y. Li, and R. Turner. Deep Gaussian processes for regression using approximate expectation propagation. In *International Conference on Machine Learning*, 2016.

I. Chung, S. Kim, J. Lee, K. J. Kim, S. J. Hwang, and E. Yang. Deep mixed effect model using Gaussian processes: A personalized and reliable prediction for healthcare. In *AAAI Conference on Artificial Intelligence*, 2020.

A. Damianou and N. D. Lawrence. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, 2013.

E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet. Infectious diseases*, 20(5):533–534, 2020.

D. K. Duvenaud, H. Nickisch, and C. Rasmussen. Additive Gaussian processes. In *Advances in Neural Information Processing Systems*, 2011.

M. Goulard and M. Voltz. Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, 24(3):269–286, 1992.

K. Greenewald, S. Zhou, and A. Hero III. Tensor graphical lasso (TeraLasso). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(5):901–931, 2019.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence*, 2013.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(4):1303–1347, 2013.

W. Imrich, S. Klavzar, and D. F. Rall. *Topics in graph theory: Graphs and their Cartesian product.* CRC Press, 2008.

A. Kalaitzis, J. Lafferty, N. D. Lawrence, and S. Zhou. The bigraphical lasso. In *International Conference on Machine Learning*, 2013.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

H. Liu, J. Cai, and Y.-S. Ong. Remarks on multi-output Gaussian process regression. *Knowledge-Based Systems*, 144:102–121, 2018.

H. Liu, Y.-S. Ong, X. Shen, and J. Cai. When Gaussian process meets big data: A review of scalable gps. *IEEE transactions on neural networks and learning systems*, 31(11):4405–4423, 2020.

A. G. d. G. Matthews, M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, and J. Hensman. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, pages 1–6, 2017.

P. Murrell. The 2006 data expo of the American Statistical Association. *Computational Statistics*, 25(4): 551–554, 2010.

T. V. Nguyen and E. V. Bonilla. Collaborative multi-output Gaussian processes. In *Uncertainty in Artificial Intelligence*, 2014.

G. Pillonetto, F. Dinuzzo, and G. De Nicolao. Bayesian online multitask learning of Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):193–205, 2010.

K. K. Pravesh and L. Roi. On the expressive power of kernel methods and the efficiency of kernel learning by association schemes. In *International Conference on Algorithmic Learning Theory*, 2020.

H. Salimbeni and M. Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems*, 2017.

J. Shi, B. Wang, E. Will, and R. West. Mixed-effects Gaussian process functional regression models with application to dose–response curve prediction. *Statistics in medicine*, 31(26):3165–3177, 2012.

J. Shi, M. Titsias, and A. Mnih. Sparse orthogonal variational inference for Gaussian processes. In *Artificial Intelligence and Statistics*, 2020.

M. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, 2009.

P. D. Tonner, C. L. Darnell, F. M. L. Bushell, P. A. Lund, A. K. Schmid, and S. C. Schmidler. A Bayesian non-parametric mixed-effects model of microbial growth curves. *PLOS Computational Biology*, 16 (10):1–21, 2020.

M. van der Wilk, V. Dutordoir, S. John, A. Artemev, V. Adam, and J. Hensman. A framework for interdomain and multioutput Gaussian processes. *arXiv:2003.01115*, 2020.

H. Wackernagel. *Multivariate Geostatistics: An Introduction with Applications*. Springer Berlin Heidelberg, 3rd edition, 2003.

Y. Wang and R. Khardon. Sparse Gaussian processes for multi-task learning. *Lecture Notes in Computer Science*, page 711–727, 2012.

J. H. Yoon and S. Kim. EiGLasso: Scalable estimation of Cartesian product of sparse inverse covariance matrices. In *Uncertainty in Artificial Intelligence*, 2020.

M. Yukawa. Adaptive learning in Cartesian product of reproducing kernel Hilbert spaces. *IEEE Transactions on Signal Processing*, 63(22):6037–6048, 2015.

X. Zhang. *Statistical Analysis for Network Data using Matrix Variate Models and Latent Space Models*. Ph.D. disseration, University of Michigan, 2020.

Supplementary Material: Doubly Mixed-Effects Gaussian Process Regression

A PROOF OF THEOREMS

In this section, we present detailed proofs of Theorems 1 and 2.

A.1 Proof of Theorem 1

Proof. We re-write $\mathbf{\Gamma}$ using the Woodbury matrix identity as follows:

$$
\begin{aligned}
\mathbf{\Gamma} &= (\boldsymbol{K} + \sigma^2 \boldsymbol{I}_{np})^{-1} \\
&= \big(\bar{\boldsymbol{K}}_X \otimes \mathbb{1}_{p,p} + \tilde{\boldsymbol{K}}_X \otimes \boldsymbol{I}_p + \sigma^2 \boldsymbol{I}_{np}\big)^{-1} \\
&= \big((\boldsymbol{I}_n \otimes \mathbb{1}_{p,1}) \boldsymbol{I}_n (\bar{\boldsymbol{K}}_X \otimes \mathbb{1}_{1,p}) + (\tilde{\boldsymbol{K}}_X + \sigma^2 \boldsymbol{I}_n) \otimes \boldsymbol{I}_p\big)^{-1} \\
&= (\tilde{\boldsymbol{K}}_X + \sigma^2 \boldsymbol{I}_n)^{-1} \otimes \boldsymbol{I}_p - \big((\tilde{\boldsymbol{K}}_X + \sigma^2 \boldsymbol{I}_n)^{-1} \otimes \mathbb{1}_{p,1}\big)\big(\boldsymbol{I}_n + p\bar{\boldsymbol{K}}_X (\tilde{\boldsymbol{K}}_X + \sigma^2 \boldsymbol{I}_n)^{-1}\big)^{-1}\big(\bar{\boldsymbol{K}}_X \big(\tilde{\boldsymbol{K}}_X + \sigma^2 \boldsymbol{I}_n)^{-1} \otimes \mathbb{1}_{1,p}\big) \\
&= (\tilde{\boldsymbol{K}}_X + \sigma^2 \boldsymbol{I}_n)^{-1} \otimes \boldsymbol{I}_p - (\tilde{\boldsymbol{K}}_X + \sigma^2 \boldsymbol{I}_n)^{-1}\big(\boldsymbol{I}_n + p\bar{\boldsymbol{K}}_X (\tilde{\boldsymbol{K}}_X + \sigma^2 \boldsymbol{I}_n)^{-1}\big)^{-1} \bar{\boldsymbol{K}}_X (\tilde{\boldsymbol{K}}_X + \sigma^2 \boldsymbol{I}_n)^{-1} \otimes \mathbb{1}_{p,p} \\
&= (\tilde{\boldsymbol{K}}_X + \sigma^2 \boldsymbol{I}_n)^{-1} \otimes \boldsymbol{I}_p + \frac{1}{p}\Big((p\bar{\boldsymbol{K}}_X + \tilde{\boldsymbol{K}}_X + \sigma^2 \boldsymbol{I}_n)^{-1} - (\tilde{\boldsymbol{K}}_X + \sigma^2 \boldsymbol{I}_n)^{-1}\Big) \otimes \mathbb{1}_{p,p}.
\end{aligned}
$$

The Woodbury matrix identity was used in the third and the last equalities above. □

A.2 Proof of Theorem 2

Proof. We show that the objective of variational inference, ELBO, in Eq. (8) changes if $\{\bar{\boldsymbol{S}}_X, \bar{\boldsymbol{S}}_G, \tilde{\boldsymbol{S}}_X^{1:p}, \tilde{\boldsymbol{S}}_G^{1:n}\}$ are substituted with $\{\bar{\boldsymbol{S}}_X + c\mathbb{1}, \bar{\boldsymbol{S}}_G - c\mathbb{1}, \tilde{\boldsymbol{S}}_X^1 + c\boldsymbol{I}, \ldots, \tilde{\boldsymbol{S}}_X^p + c\boldsymbol{I}, \tilde{\boldsymbol{S}}_G^1 - c\boldsymbol{I}, \ldots, \tilde{\boldsymbol{S}}_G^n - c\boldsymbol{I}\}$ for any $c \in \mathbb{R}$. We show this for each of the two terms in the ELBO in Eq. (8). For the first term in ELBO that corresponds to the expected log-likelihood term, before this substitution, we have

$$
\begin{aligned}
E &= \mathbb{E}_{q(\mathcal{F})}\big[\log p(\boldsymbol{y} \mid \mathcal{F})\big] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{p}\Bigg[\log \mathcal{N}\Big(\bar{\boldsymbol{k}}_{x_i Z_X} \bar{\boldsymbol{K}}_{Z_X}^{-1} \bar{\boldsymbol{m}}_X + \bar{\boldsymbol{k}}_{g_k Z_G} \bar{\boldsymbol{K}}_{Z_G}^{-1} \bar{\boldsymbol{m}}_G + \tilde{\boldsymbol{k}}_{x_i Z_X} \tilde{\boldsymbol{K}}_{Z_X}^{-1} \tilde{\boldsymbol{m}}_X^k + \tilde{\boldsymbol{k}}_{g_k Z_G} \tilde{\boldsymbol{K}}_{Z_G}^{-1} \tilde{\boldsymbol{m}}_G^i,\, \sigma^2\Big) \\
&\qquad - \frac{1}{2\sigma^2}\Big(\bar{k}_X(\boldsymbol{x}_i, \boldsymbol{x}_i) + \bar{\boldsymbol{k}}_{x_i Z_X} \bar{\boldsymbol{K}}_{Z_X}^{-1} (\bar{\boldsymbol{S}}_X - \bar{\boldsymbol{K}}_{Z_X}) \bar{\boldsymbol{K}}_{Z_X}^{-1} \bar{\boldsymbol{k}}_{Z_X x_i}\Big) \\
&\qquad - \frac{1}{2\sigma^2}\Big(\tilde{k}_X(\boldsymbol{x}_i, \boldsymbol{x}_i) + \tilde{\boldsymbol{k}}_{x_i Z_X} \tilde{\boldsymbol{K}}_{Z_X}^{-1} (\tilde{\boldsymbol{S}}_X^k - \tilde{\boldsymbol{K}}_{Z_X}) \tilde{\boldsymbol{K}}_{Z_X}^{-1} \tilde{\boldsymbol{k}}_{Z_X x_i}\Big) \\
&\qquad - \frac{1}{2\sigma^2}\Big(\bar{k}_G(\boldsymbol{g}_k, \boldsymbol{g}_k) + \bar{\boldsymbol{k}}_{g_k Z_G} \bar{\boldsymbol{K}}_{Z_G}^{-1} (\bar{\boldsymbol{S}}_G - \bar{\boldsymbol{K}}_{Z_G}) \bar{\boldsymbol{K}}_{Z_G}^{-1} \bar{\boldsymbol{k}}_{Z_G g_k}\Big) \\
&\qquad - \frac{1}{2\sigma^2}\Big(\tilde{k}_G(\boldsymbol{g}_k, \boldsymbol{g}_k) + \tilde{\boldsymbol{k}}_{g_k Z_G} \tilde{\boldsymbol{K}}_{Z_G}^{-1} (\tilde{\boldsymbol{S}}_G^i - \tilde{\boldsymbol{K}}_{Z_G}) \tilde{\boldsymbol{K}}_{Z_G}^{-1} \tilde{\boldsymbol{k}}_{Z_G g_k}\Big)\Bigg],
\end{aligned}
$$

whereas after the substitution, we have

$$
\begin{aligned}
E' = E &- \frac{c}{2\sigma^2}\Big(\bar{\boldsymbol{k}}_{x_i Z_X} \bar{\boldsymbol{K}}_{Z_X}^{-1} \mathbb{1} \bar{\boldsymbol{K}}_{Z_X}^{-1} \bar{\boldsymbol{k}}_{Z_X x_i}\Big) - \frac{c}{2\sigma^2}\Big(\tilde{\boldsymbol{k}}_{x_i Z_X} \tilde{\boldsymbol{K}}_{Z_X}^{-2} \tilde{\boldsymbol{k}}_{Z_X x_i}\Big) \\
&+ \frac{c}{2\sigma^2}\Big(\bar{\boldsymbol{k}}_{g_k Z_G} \bar{\boldsymbol{K}}_{Z_G}^{-1} \mathbb{1} \bar{\boldsymbol{K}}_{Z_G}^{-1} \bar{\boldsymbol{k}}_{Z_G g_k}\Big) + \frac{c}{2\sigma^2}\Big(\tilde{\boldsymbol{k}}_{g_k Z_G} \tilde{\boldsymbol{K}}_{Z_G}^{-2} \tilde{\boldsymbol{k}}_{Z_G g_k}\Big).
\end{aligned}
$$

The additional terms in E' compared to E do not cancel out with each other. For the second term in the ELBO in Eq. (8) that corresponds to the KL divergence term, before the substitution, we have

$$
\begin{aligned}
D &= \mathrm{KL}\left[q(\mathcal{U}) \mid\mid p(\mathcal{U})\right] \\
&= \mathrm{KL}\left[q(\bar{\boldsymbol{u}}_X) \mid\mid p(\bar{\boldsymbol{u}}_X)\right] + \sum_{k=1}^{p} \mathrm{KL}\left[q(\tilde{\boldsymbol{u}}_X^k) \mid\mid p(\tilde{\boldsymbol{u}}_X^k)\right] + \mathrm{KL}\left[q(\bar{\boldsymbol{u}}_G) \mid\mid p(\bar{\boldsymbol{u}}_G)\right] + \sum_{i=1}^{n} \mathrm{KL}\left[q(\tilde{\boldsymbol{u}}_G^i) \mid\mid p(\tilde{\boldsymbol{u}}_G^i)\right] \\
&= \frac{1}{2}\Bigg[\left(\mathrm{tr}(\bar{\boldsymbol{K}}_X^{-1}\bar{\boldsymbol{S}}_X) + \bar{\boldsymbol{m}}_X^T \bar{\boldsymbol{K}}_X^{-1} \bar{\boldsymbol{m}}_X + \log \frac{|\bar{\boldsymbol{K}}_X|}{|\bar{\boldsymbol{S}}_X|} \right) + \sum_{k=1}^{p} \left(\mathrm{tr}(\tilde{\boldsymbol{K}}_X^{-1}\tilde{\boldsymbol{S}}_X^k) + \tilde{\boldsymbol{m}}_X^{k\,T} \tilde{\boldsymbol{K}}_X^{-1} \tilde{\boldsymbol{m}}_X^k + \log \frac{|\tilde{\boldsymbol{K}}_X|}{|\tilde{\boldsymbol{S}}_X^k|} \right) \\
&\quad + \left(\mathrm{tr}(\bar{\boldsymbol{K}}_G^{-1}\bar{\boldsymbol{S}}_G) + \bar{\boldsymbol{m}}_G^T \bar{\boldsymbol{K}}_G^{-1} \bar{\boldsymbol{m}}_G + \log \frac{|\bar{\boldsymbol{K}}_G|}{|\bar{\boldsymbol{S}}_G|} \right) + \sum_{i=1}^{n} \left(\mathrm{tr}(\tilde{\boldsymbol{K}}_G^{-1}\tilde{\boldsymbol{S}}_G^i) + \tilde{\boldsymbol{m}}_G^{i\,T} \tilde{\boldsymbol{K}}_G^{-1} \tilde{\boldsymbol{m}}_G^k + \log \frac{|\tilde{\boldsymbol{K}}_G|}{|\tilde{\boldsymbol{S}}_G^i|} \right) \\
&\quad + (p+1)m_X - (n+1)m_G \Bigg],
\end{aligned}
$$

and after the substitution, we have

$$
\begin{aligned}
D' = D + \frac{1}{2}\Bigg(c\Big(&\mathrm{tr}(\bar{\boldsymbol{K}}_X^{-1}\mathbb{1} + p\tilde{\boldsymbol{K}}_X^{-1}) - \mathrm{tr}(\bar{\boldsymbol{K}}_G^{-1}\mathbb{1} + n\tilde{\boldsymbol{K}}_G^{-1}) \Big) \\
&+ \log \frac{|\bar{\boldsymbol{S}}_G - c\mathbb{1}|}{|\bar{\boldsymbol{S}}_X + c\mathbb{1}|} + \sum_{k=1}^{p} \log \frac{1}{|\tilde{\boldsymbol{S}}_X^k + c\boldsymbol{I}|} + \sum_{i=1}^{n} \log \frac{1}{|\tilde{\boldsymbol{S}}_G^i - c\boldsymbol{I}|} \Bigg).
\end{aligned}
$$

Again the additional terms in D' compared to D do not cancel out with each other and change the value of ELBO. Thus, the ELBO is minimized with a unique $\{\bar{\boldsymbol{S}}_X, \bar{\boldsymbol{S}}_G, \tilde{\boldsymbol{S}}_X^{1:p}, \tilde{\boldsymbol{S}}_G^{1:n}\}$. □

B ADDITIONAL RESULTS ON REAL-WORLD DATA

We describe the experimental details for the New York City taxi trip and United Kingdom house price datasets, and present the mixed-effects decomposition in these datasets by different mixed-effects GPs.

New York City Taxi Trip We obtained data on taxi trips in 2009 in New York City from the City website. Given the travel time and fare of the 170,896,052 trips from Lower to Midtown Manhattan, we discretized the latitude and longitude of the destination into 102 bins for latitude spaced between 23rd and 125th St. and 9 bins for longitude between 1st and 11th Ave. Our goal is to predict the travel time and fare of taxi trips ending on randomly chosen 20 streets given the data from trips that end in the other streets.

We show the decomposition of the mixed effects for the New York City taxi travel time data found by different methods with mixed effects in Figure 6. For doubly mixed-effects GP (Fig. 6(a)), the fixed effects shared across avenues that run from the lower left corner to the upper right corner on the map show the increase in travel time as destinations get farther away from Lower Manhattan that is located in the lower left corner on the map (Fig. 6(a), Column 1). The avenue-specific random effects have their highest peaks in the middle because there are limited-access highways along the east and west sides of Manhattan with exits near Central Park to reach areas in Midtown Manhattan (Fig. 6(a), Column 2). The fixed and random effects over streets in doubly mixed-effects GP have relatively small effect sizes (Fig. 6(a), Column 3 and 4). As avenue-wise mixed effects are stronger than street-wise mixed effects, the mixed-effects GP and translated mixed-effects GP are able to capture the fixed and random effects for avenues. For the same reason, the translated mixed-effects GP had higher prediction accuracy than the doubly mixed-effects GPs in Table 1. We found similar results for the New York City travel fare data (Fig. 7).

United Kingdom House Price We analyzed data for house prices over 36 months from 2018 to 2020 in 2,085 locations in the United Kingdom, provided by the UK government website[1]. Our goal is to predict the house prices at all locations in the last 6 months given those in the first 30 months.

[1] It contains HM Land Registry data © Crown copyright and database right 2021. This data is licensed under the Open Government Licence v3.0. Available at https://www.gov.uk/government/collections/price-paid-data

The mixed-effects decomposition of the United Kingdom house price dataset is shown in Figure 8. The doubly mixed-effects GP recovers the overall distribution of house prices in the United Kingdom. London has the highest peak in the fixed effects shared across time (Fig. 8(a), left), but has relatively small seasonal effects on prices in time-specific random effects (Fig. 8(a), middle). The fixed effects shared across all regions are small (Fig. 8(a), right, black curve), but the region-specific random effects show consistently high housing prices in the London area (Fig. 8(a), right, red curves). In the translated mixed-effects GP (Fig. 8(b)), because the time-specific random effects were not modeled, the pattern of consistently high housing prices in London is weaker (Fig. 8(b), right). The mixed-effects GP can capture only the fixed effects shared by all regions and region-specific random effects (Fig. 8(c)).

C SOFTWARE

The software is available at `https://github.com/SeyoungKimLab/DMGP`.

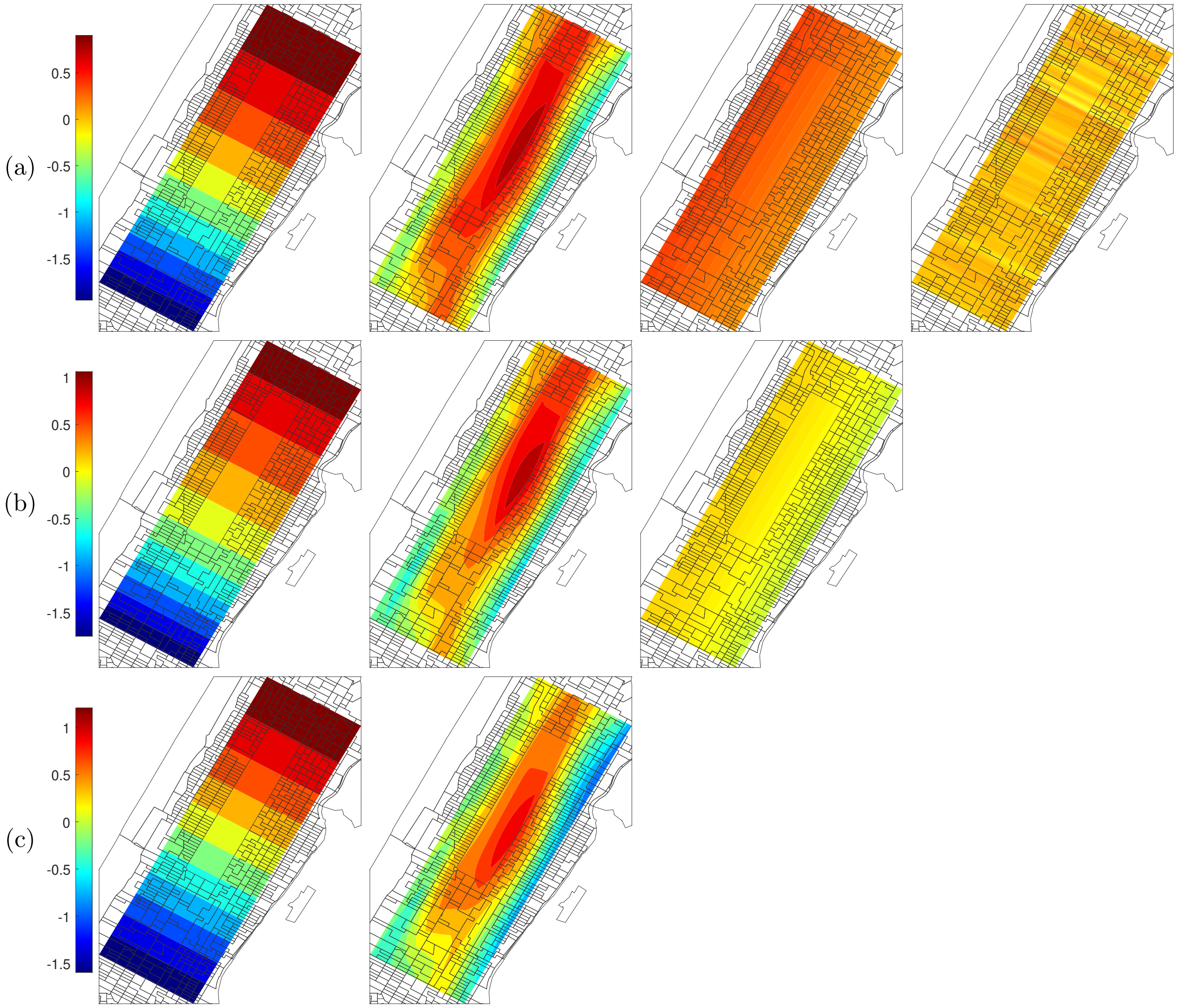


Figure 6: Decomposition of fixed and random effects from different mixed-effects models on the NYC taxi travel time data. (a) Doubly mixed-effects GP, (b) translated mixed-effects GP, and (c) mixed-effects GP. The decomposed fixed and random effects are shown in each of the four columns, if available from the given method. Column 1: fixed effects shared across avenues. Column 2: avenue-specific random effects. Column 3: fixed effects shared across streets. Column 4: street-specific random effects.

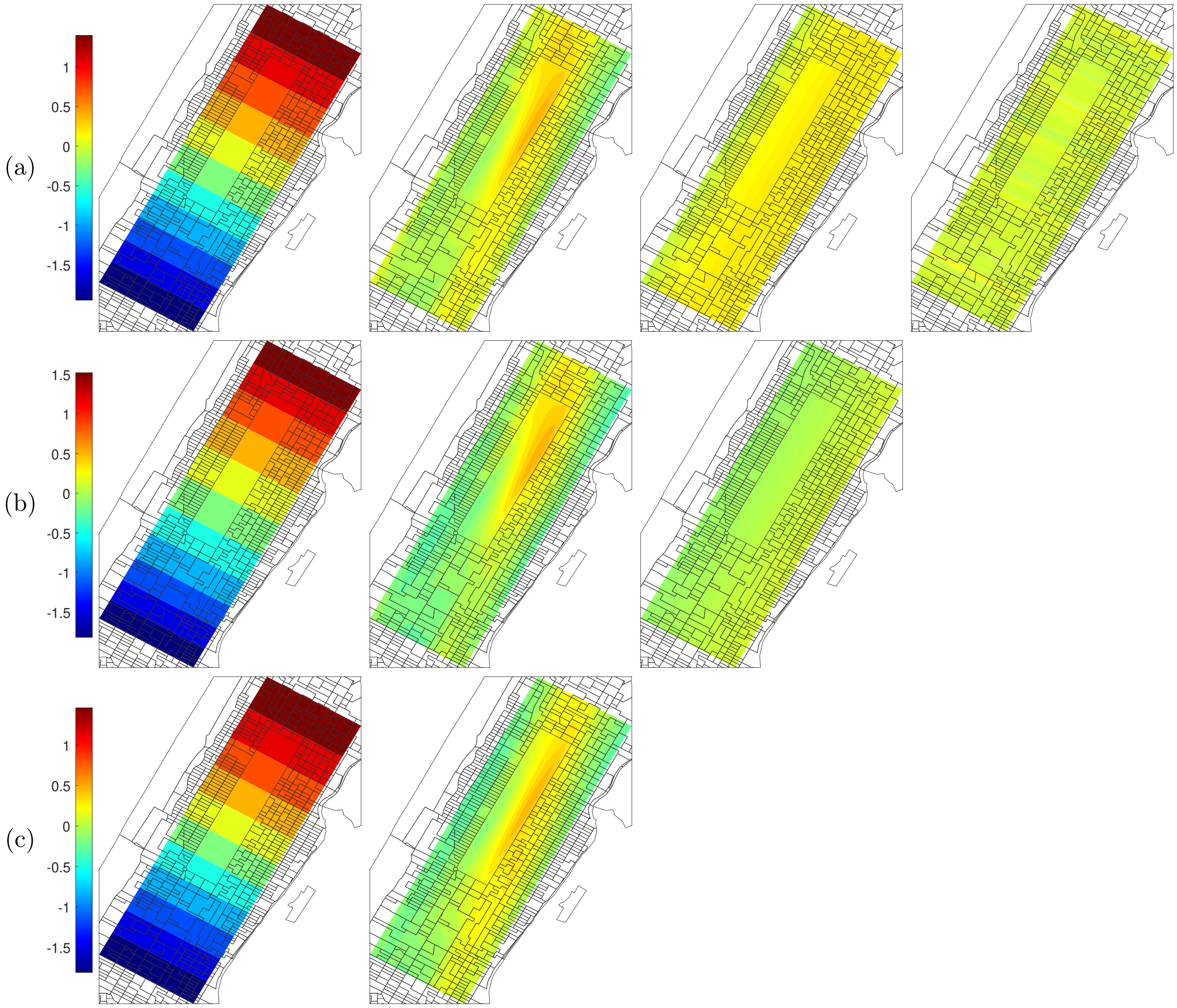


Figure 7: Decomposition of fixed and random effects from different mixed-effects models on the NYC taxi travel fare data. (a) Doubly mixed-effects GP, (b) translated mixed-effects GP, and (c) mixed-effects GP. The decomposed fixed and random effects are shown in each of the four columns, if available from the given method. Column 1: fixed effects shared across avenues. Column 2: avenue-specific random effects. Column 3: fixed effects shared across streets. Column 4: street-specific random effects.

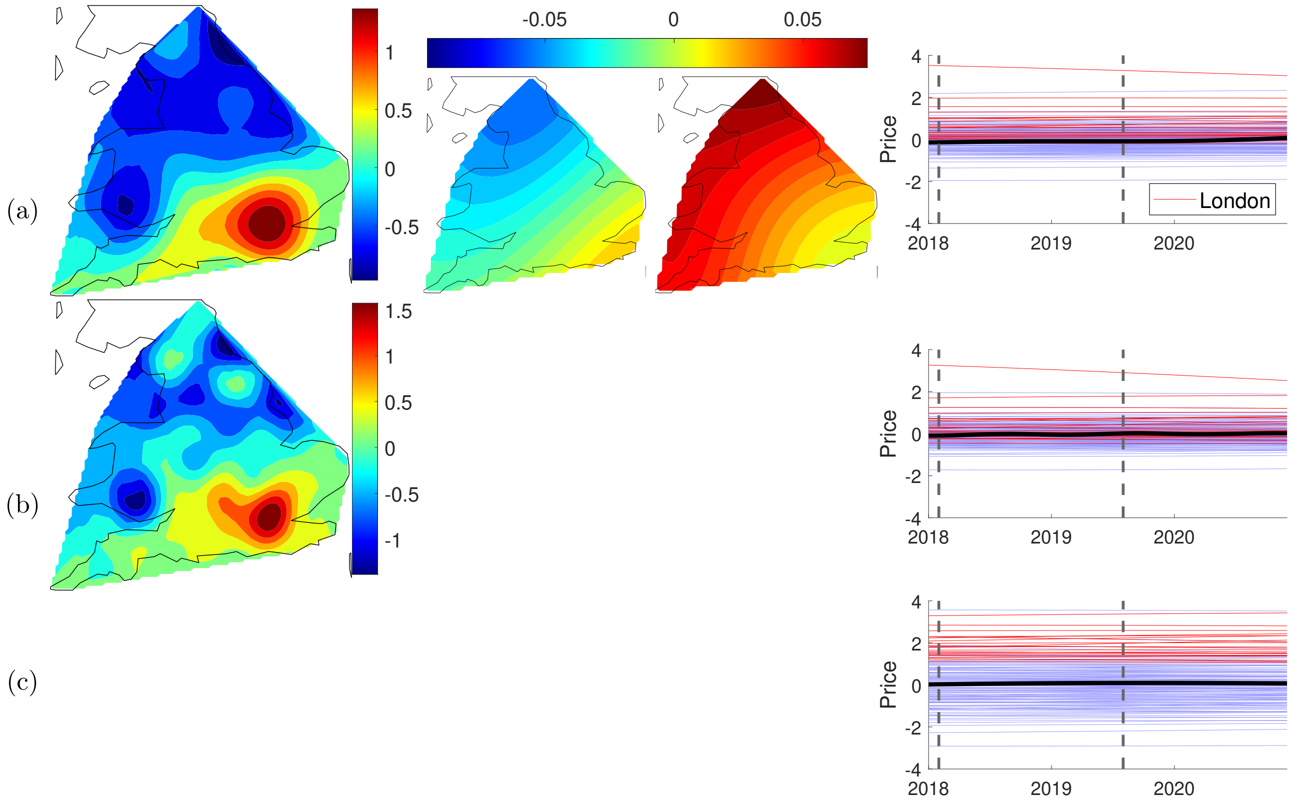


Figure 8: Decomposition of fixed and random effects on the UK house price data. (a) Doubly mixed-effects GP, (b) translated mixed-effects GP, and (c) mixed-effects GP. The decomposed fixed and random effects are shown in each of the three columns, if available from the given method. Left: fixed effects shared across time. Middle: time-specific random effects for two months marked by the black dashed lines in the figure on the right. Right: fixed effects across regions (black) and 2% of region-specific random effects.