Inducing Clusters Deep Kernel Gaussian Process for Longitudinal Data

Junjie Liang, Weijieying Ren, Hanifi Sahar, Vasant Honavar

The Pennsylvania State University
jliang282@outlook.com, {wjr5337, szh6071, vuh14}@psu.edu

Abstract

We consider the problem of predictive modeling from irregularly and sparsely sampled longitudinal data with unknown, complex correlation structures and abrupt discontinuities. To address these challenges, we introduce a novel inducing clusters longitudinal deep kernel Gaussian Process (ICDKGP). ICDKGP approximates the data generating process by a zero-mean GP with a longitudinal deep kernel that models the unknown complex correlation structure in the data and a deterministic non-zero mean function to model the abrupt discontinuities. To improve the scalability and interpretability of ICDKGP, we introduce *inducing clusters* corresponding to centers of clusters in the training data. We formulate the training of ICDKGP as a constrained optimization problem and derive its evidence lower bound. We introduce a novel relaxation of the resulting problem which under rather mild assumptions yields a solution with error bounded relative to the original problem. We describe the results of extensive experiments demonstrating that ICDKGP substantially outperforms the state-of-the-art longitudinal methods on data with both smoothly and non-smoothly varying outcomes.

Introduction

Longitudinal data, consisting of repeated, often irregularly sampled observations, of variables and outcomes for a set of individuals (Liang et al. 2021, 2020), are ubiquitous in many applications, e.g., predictive modeling of health risks from electronic health records data, or educational outcomes from activity logs in online platforms. Such data display complex, often unknown, correlation structures: longitudinal correlation (LC) across time, cluster correlations (CC) across individuals, or multi-level correlations (MC) (Liang et al. 2021). Gaussian processes (GP) (Williams and Rasmussen 2006) offer an attractive framework for predictive modeling from such data (Liang et al. 2021). Existing GP models optimize kernel parameters under the assumption that the longitudinal outcome being modeled is sufficiently smooth. However, in many real-world applications, the longitudinal outcome can show abrupt discontinuities, e.g., due to unobserved transitions between hidden states e.g., healthy versus sick.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Key Contributions. We consider the problem of predictive modeling from irregularly and sparsely sampled longitudinal data with unknown, complex correlation structures and abrupt discontinuities. Specifically, we approximate the data generating process by a zero-mean GP with a longitudinal deep kernel to recover the complex correlation structure in the data and a deterministic non-zero mean function that helps account for the discontinuities in the observed outcomes. We improve the scalability of GP predictions by replacing inducing points (Titsias 2009) with *inducing clusters* (set to centers of clusters in the training data), thereby substantially reducing the number of inducing points needed and performing regression and clustering simultaneously without increasing computational complexity. We show that inducing clusters mimic the mean-field assumption that is often used in variational inference with sparse GP while enhancing both the scalability and interpretability of the learned GP. We formulate the problem of training the resulting inducing clusters deep kernel Gaussian process (ICDKGP) as a constrained optimization problem and derive its evidence lower bound (ELBO). We introduce a novel relaxation of the resulting problem which under rather mild assumptions yields a solution with error bounded relative to that of the original problem. Through extensive experiments with both simulated and real-world data, we show that ICDKGP substantially outperforms the state-of-the-art (SOTA) baselines in terms of both predictive accuracy and correlation structure recovery.

Related Work

GP for Longitudinal Data Analysis. *Gaussian processes* (GP) (Williams and Rasmussen 2006; Cheng et al. 2019), offer an attractive approach for predictive modeling from longitudinal data. GP dispenses with assumptions about the parametric form of the data generating process using parameterized kernels to model complex, a priori unknown correlation structure in the data: If a function $f : \mathcal{X} \to \mathbb{R}$ has a GP prior $f \sim \mathcal{GP}(\mu, k_\theta)$ where μ is the mean function and $k_\theta(\cdot, \cdot)$ is a kernel function parameterized by θ, then any finite collection of components of f (denoted as $\mathbf{f}$) has a multivariate Gaussian distribution $(\mathbf{f}|X) \sim \mathcal{N}(\boldsymbol{\mu}_X, K_{XX})$, where $\boldsymbol{\mu}_X$ is the mean vector, and $(K_{XX})_{ij} = k_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the covariance matrix. A zero mean GP, given a kernel with a *universal approximation property*, *e.g.,* dot product, RBF, polynomial, or Matérn kernel (Williams and Rasmussen

2006), and sufficient training data can approximate any *sufficiently well-behaved* function with arbitrarily high accuracy (Micchelli, Xu, and Zhang 2006). GP can accommodate data sampled at irregularly spaced time points via interpolation (Liang et al. 2021). Previous work on GP models for longitudinal data focuses primarily on the design of suitable kernels to account for the complex correlation structure in the data. Cheng et al. (2019) introduced an additive kernel optimized to yield the desired predictive performance. Timonen et al. (2019) explored the use of a heterogeneous kernel for modeling random effects in non-Gaussian data. Liang et al. (2021) introduced an efficient method for learning a deep kernel that models both time-varying and time-invariant effects. Chen et al. (2020) used a transformer network and kernel warping to fuse information from multiple data sources.

Inducing Points for Speeding up GP. A body of work has been proposed to speed up GP using inducing points (reviewed in (Liu et al. 2020)). Titsias (2009) and Hensman, Matthews, and Ghahramani (2015) showed how to combine inducing points and variational inference to reduce the computational complexity of GP for regression and classification respectively. Wilson et al. (2016) introduced an efficient way to sample the inducing points. Shi, Titsias, and Mnih (2020) showed how to leverage inducing points to speed up a GP that is expressed as a sum of two independent GPs. Much previous work on inducing points was aimed at approximating the inverse of the covariance matrix K_{XX}. Subsequent work was aimed at improving the selection of inducing points to as to enhance the efficiency and accuracy of the resulting GP.

Zero Mean and Non-zero Mean GP. It has been observed that the zero mean assumption is tantamount to asserting that all available prior knowledge can be effectively incorporated into the form of the GP kernel. The kernel parameters are then optimized to obtain the desired predictive performance (Iwata and Ghahramani 2017; Chung et al. 2020; De Ath, Fieldsend, and Everson 2020). However, Iwata and Ghahramani (2017) showed that when the training data are scarce, a zero mean GP produces outcome predictions that approach zero in regions with no training data. Recent work has demonstrated the benefits of GP models with non-zero mean functions. For example, Chung et al. (2020) utilized a recurrent neural network (RNN) to model the mean function of a GP to capture population-average effects from longitudinal electronic health records. De Ath, Fieldsend, and Everson (2020) compared GP models with linear and quadratic mean functions and found that when the outcome to be predicted is discontinuous, nonlinear mean functions tended to outperform constant or linear mean functions.

GP Variants for Non-stationary Data GP models have been extended to cope with non-stationary data using non-stationary kernels (Noack and Sethian 2022; Tompkins, Oliveira, and Ramos 2020), or non-stationary covariance functions (Paciorek and Schervish 2003; Paun, Husmeier, and Torney 2023), or kernel interpolation methods (Graßhoff, Jankowski, and Rostalski 2020) to cope with non-stationary data. Other work has explored sparsity-inducing kernels (Noack et al. 2023), and GP regression models that incorporate domain knowledge to cope with non-stationary longitudinal gene expression data (Cheng et al. 2019; Vantini et al. 2022). The current work shares some similarities with these methods in that it decomposes the underlying kernel into time-invariant and time-variant components, but differs from them in terms of how it achieves scalability, namely, by generalizing inducing points to inducing clusters.

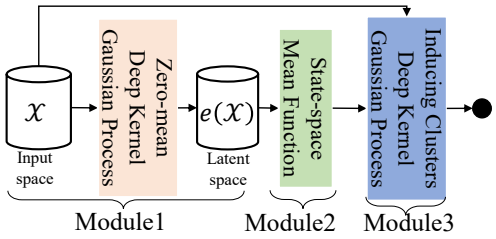

Figure 1: Overview of ICDKGP. Module 1 involves a zero-mean deep kernel Gaussian Process responsible for projecting input data into a latent space; Module 2 fits a mean function with the latent space; Module 3 integrates inducing clusters with the mean function for final prediction.

Inducing Clusters Deep Kernel GP

After (Liang et al. 2021), we use $\mathcal{D} = (X, \boldsymbol{y})$ to denote a longitudinal data set where $X \in \mathbb{R}^{N\times P}$ is the covariate matrix and $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$ is the vector of measured outcomes. A row $\boldsymbol{x}_{it}$ in X denotes the observation for individual i at time index t. Because the observations for each individual are irregularly time-sampled, for each i, we have a sub-matrix $X_i \in \mathbb{R}^{N_i\times P} \subset X$, where N_i denotes the number of observations available for individual i. Denoting the number of individuals in $\mathcal{D}$ by I, we use $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_I^\top)^\top$ to represent the outcomes associated with $X = (X_1^\top, \cdots, X_I^\top)^\top$.

ICDKGP consists of three main modules as shown in Fig. 1: Module 1, a zero mean deep kernel GP that projects the input data to a latent space; Module 2, a deterministic state-space mean function to the latent space. Module 3, integration of the mean function and the data using ICDKGP.

Module 1: Zero Mean Deep Kernel GP. Module 1 employs a deep kernel by combining the expressive power of deep neural networks with the flexibility of a non-parametric kernel. The deep kernel eliminates the need for ad-hoc heuristics or trial-and-error since it learns to fit the correlation structure in the data. Formally, let $e : \mathbb{R}^P \to \mathbb{R}^D$ be a general DNN-based encoder function, and $g : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ be a valid kernel function, then for any pair of data points $\boldsymbol{x}_i, \boldsymbol{x}_j$, the deep kernel is computed as $k_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) = g(e(\boldsymbol{x}_i), e(\boldsymbol{x}_j))$. We adopt the longitudinal deep kernel proposed in (Liang et al. 2021) to model the unknown multilevel correlation structure of longitudinal data. Specifically, the kernel function k_θ is the sum of two components:

$$k_\theta(\boldsymbol{x}_{it}, \boldsymbol{x}_{jq}) = {\alpha^{(v)}}^2 k_{RBF}(e^{(v)}(\boldsymbol{x}_{it}), e^{(v)}(\boldsymbol{x}_{jq})) + {\alpha^{(i)}}^2 k_{RBF}(e^{(i)}(\boldsymbol{x}_{it}), e^{(i)}(\boldsymbol{x}_{jq})) \quad (1)$$

Here, $e^{(v)}$ and $e^{(i)}$ respectively denote the encoder networks that model the time-varying and time-invariant components

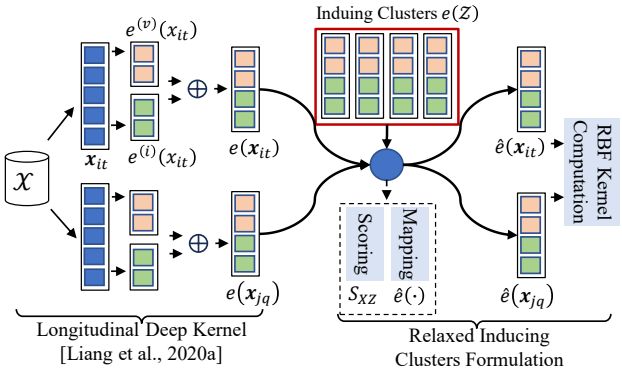

Figure 2: Graphical explanation of kernel computation in ICDKGP. The initial latent representation $e(X)$ is computed the same way as in (Liang et al. 2021). The updated form of latent representation $\hat{e}(\mathcal{X})$ is obtained by computing the proximity between the data points and inducing clusters, followed by the proximity mapping. The kernel value is computed based on the updated latent representation.

of the unknown correlation structure in a latent space (See Fig. 2.). For notation brevity, let $e(x)$ be the concatenation between $e^{(v)}(\boldsymbol{x})$ and $e^{(i)}(\boldsymbol{x})$, The longitudinal kernel k_θ in (1) can be rewritten on the latent space $e(\mathcal{X})$. An explanation of kernel computation is shown on the left panel of Fig. 2.

Module 2: Mean Function for GP. We model GP mean by a deterministic state-space function which allows the resulting GP to be expressed as the sum of a deterministic function and a zero mean GP: $f = f_\perp + f_\parallel$, where $f_\perp(X) = \boldsymbol{\mu}_X$ and $f_\parallel \sim \mathcal{GP}(0, k_\theta)$. The choice of a deterministic mean function allows us to relegate correlation estimation to Module 1 because $var[f] = var[f_\parallel] = k_\theta(\cdot,\cdot)$, with Module 2 (the mean function) $f_\perp$ accounts for the non-smoothly time-varying outcomes. Let $\mathcal{X}' = e(X) \in \mathbb{R}^{N\times D}$ be the output from module 1. The state-space mean model maintains K learnable hidden state encodings $C = \{\boldsymbol{c}_k\}_{k=1}^K$. Given data $\boldsymbol{x}' \in \mathcal{X}'$, we first obtain its state representation by comparing and mapping x' to the state encodings based on a proximity score measured by the dot product, such that

$$v(\boldsymbol{x}') = C^\top \text{softmax}(C\boldsymbol{x}') \tag{2}$$

Then the mean prediction (or logit) is obtained by processing the state encoding $v(\boldsymbol{x}')$ through a multi-layer neural network with the structure: Input $\xrightarrow[\text{GeLU}]{\text{MLP}}$ Hidden $\xrightarrow{\text{MLP}}$ Output. We use the mean squared prediction error as the loss.

Module 3: Inducing Kernel GP. Recall that inducing points are used to increase the efficiency of the GP posterior by reducing the effective number of rows in X, from N to M ($M \ll N$), where M is the number of *inducing points*. Let $\mathbf{u} = \{u_m\}_{m=1}^M$ be the collection of inducing points and Z be their feature vectors, then solving the GP usually involves maximizing the following ELBO (Wilson et al. 2016):

$$\log p(\boldsymbol{y}) \geq \mathbb{E}_{q(\mathbf{f},\mathbf{u})}[\log p(\boldsymbol{y}|\mathbf{f})] - \text{KL}[q(\mathbf{f},\mathbf{u})||p(\mathbf{f},\mathbf{u})] \tag{3}$$

where $p(\boldsymbol{y}|\mathbf{f})$ is the likelihood model, $p(\mathbf{u}) = \mathcal{N}(\boldsymbol{\mu}_Z, K_{ZZ})$ and $q(\mathbf{u}) = \mathcal{N}(m_Z, S)$ are the observational prior and variational prior for inducing points respectively. The joint variational distribution $q(\mathbf{f},\mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u})$.

Unlike the previous use of inducing points primarily for speeding up GP, we use them to enhance the interpretability of ICDKGP. Specifically, we choose inducing points that correspond to cluster centers of the training data such that (i) the cluster centers are approximately mutually independent and (ii) each data point is assigned to its nearest cluster centers with high confidence. To see how these conditions can be enforced, recall the joint signal distributions:[1]

$$p(\mathbf{f},\mathbf{u}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_X\\ \boldsymbol{\mu}_Z\end{bmatrix}, \begin{bmatrix}K_{XX} & K_{XZ}\\ K_{XZ}^\top & K_{ZZ}\end{bmatrix}\right) \tag{4}$$

$$q(\mathbf{f},\mathbf{u}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_X + A(\boldsymbol{m}_z - \boldsymbol{\mu}_Z)\\ \boldsymbol{m}_Z\end{bmatrix}, \begin{bmatrix}V & AS\\ SA^\top & S\end{bmatrix}\right) \tag{5}$$

where $A = K_{XZ}K_{ZZ}^{-1}, V = K_{XX} - AK_{XZ}^\top + ASA^\top$. Since the KL divergence term in (3) acts as a soft constraint to minimize the difference between the observational distribution $p(\mathbf{f},\mathbf{u})$ and variational distribution $q(\mathbf{f},\mathbf{u})$, we enforce the cluster centers constraint on both distributions p and q as follows: (i) K_{ZZ} and S should both be an almost diagonal matrix (Hari 1999) and (ii) all but one of the elements of each row of K_{XZ} and AS should be approximately zero. These two constraints have to be enforced on the joint covariance matrix while ensuring that it remains symmetric positive definite (SPD). Because the covariance matrix is specified by a kernel function, the SPD condition is guaranteed. Hence, we turn our attention to ensuring that the kernel parameters are chosen to enforce the above two constraints. Note that the constraint on S is trivial to enforce because we can simply parameterize S to be a diagonal matrix, which is equivalent to applying the mean-field approximation (*i.e.,* fully factorizing $q(\mathbf{u})$) (Hensman, Matthews, and Ghahramani 2015). With this parameterization, it is easy to show that if all constraints on K_{ZZ} and K_{XZ} hold, so does the constraint on AS.

The preceding observations lead to the following constrained optimization problem:

$$\arg\max_{\Theta} \mathcal{L}_1 = \mathbb{E}_{q(\mathbf{f},\mathbf{u})}[\log p(\boldsymbol{y}|\mathbf{f})] - \text{KL}[q(\mathbf{u})||p(\mathbf{u})] \tag{6}$$

$$\text{s.t.} \quad \max \text{diag}(BB^\top) \leq \epsilon, \quad \max \text{diag}(CC^\top) \leq \epsilon \tag{6a}$$

where $B = K_{ZZ} - \text{diag}(K_{ZZ}), C = K_{XZ} - D \circ K_{XZ}$ with D as a masking matrix defined by $D_{xz} = \begin{cases}1, & D_{xz} = \max_j D_{xj}\\ 0, & \text{otherwise}\end{cases}$. Here, ϵ is a hyperparameter that specifies the threshold for the constraints; and '$\circ$' denotes the Hadamard (element-wise) product. Solving the constrained optimization problem in (6) is hard because the masking matrix D has zero gradients everywhere. Hence, in what follows, we introduce a relaxed version of (6).

[1]Result of (5) is derived by applying the Gaussian Identities (Williams and Rasmussen 2006) on $q(\mathbf{f},\mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u})$.

Relaxed Formulation of ICDKGP. To better exploit the cluster structure, we redefine the latent representation of each data point $x \in X$, *i.e.,* $e(x)$, through a soft mapping of the inducing cluster that is closest to it (see Fig. 2):

$$\hat{e}(x) = s_{xz^*} e(z^*) + (1 - s_{xz^*}) e(x) \tag{7}$$

where $s_{xz} \in S_{XZ}$ is the proximity score between x and z. The kernel function in (1) offers a natural measure of pairwise proximity S_{XZ} between data points in X and data points in Z. As such, we define $S_{XZ} = \text{softmax}(K_{XZ}/\tau)$, where τ is a hyperparameter (temperature). Smaller τ forces an approximately row-wise one-hot structure on S_{XZ}. Clearly, $S_{XZ} > 0$ and $\sum_z s_{xz} = 1$. Let and $z^* = \arg\max_z S_{xz}$ and $s_{xz^*} = \max_z s_{xz}$. Based on this new data representation, we can relax the constrained optimization problem (6) as follows:

$$\arg\max_{\Theta} \mathcal{L}_2 = \mathbb{E}_{q(\mathbf{f},\mathbf{u})}[\log p(\boldsymbol{y}|\mathbf{f})] - \text{KL}[q(\mathbf{u})||p(\mathbf{u})] \tag{8}$$

$$\text{s.t.} \quad \max \text{diag}(BB^\top) \leq \epsilon, \quad 1 - \min_{x \in \mathcal{X}} s_{xz^*} \leq \eta \tag{8a}$$

ϵ and η are the hyperparameters. The constraint (6a) in (6) is now replaced by (8a) in (8). We can view the proximity score s_{XZ} as a smooth approximation of the masking matrix D, which renders optimizing $\mathcal{L}_2$ subject to (8a) easier than the original problem of optimizing $\mathcal{L}_1$ subject to (6a). The following lemmas explicate the relationship between $\mathcal{L}_1$ and $\mathcal{L}_2$ and between the solution of (8) and that of (6):

Lemma 1. *A feasible solution that maximizes $\mathcal{L}_2$ yields a feasible solution that maximizes $\mathcal{L}_1$ as $\eta \to 0$.*

Proof. Let $\text{diag}(BB^\top) = \boldsymbol{r}$ and $r^* = \max \boldsymbol{r}$. According to (6a), $r^* \leq \epsilon$. As $\eta \to 0$, we have $\min_x s_{xz^*} \to 1$, meaning $\forall x, s_{xz^*} \to 1$, thus per (7), $\hat{e}(x) \to e(z^*)$. From (1), we have $k_\theta(\hat{e}(\boldsymbol{x}), e(\boldsymbol{z})) \to k_\theta(e(\boldsymbol{z}^*), e(\boldsymbol{z}))$. Hence, the covariance K_{XZ} is reduced to K_{Z^*Z}. Therefore, $\max \text{diag}(CC^\top) = \max \text{diag}(BB^\top) = r^* \leq \epsilon$. Hence, as $\eta \to 0$, a feasible solution of $\mathcal{L}_2$ is also a feasible solution of $\mathcal{L}_1$. □

Lemma 1 states that as $\eta \to 0$, solving $\mathcal{L}_2$ results in a feasible solution for $\mathcal{L}_1$. From (7), we can achieve this by setting $\forall \boldsymbol{x} \in \mathcal{X}, e(\boldsymbol{z}^*) = e(\boldsymbol{x})$, i.e., making the inducing clusters coincide with the data points. This is not feasible in practice. Lemma 2 offers a way out of this difficulty.

Lemma 2. *Maximizing $\mathcal{L}_2$ subject to* (8a) *becomes equivalent to maximizing $\mathcal{L}_1$ subject to* (6a) *when the training data form apparent Mixture of Gaussian (MoG) distributions around the inducing clusters in the latent space $e(\mathcal{X})$.*

Proof. Let the center of the cluster to which a data point x belongs be $z^* = \arg\max_z K_{xz}$. Since we assume that the training data form MoG distributions in the latent space, with the generative model for MoG, the data point $\boldsymbol{x} \in \mathcal{X}$ can be generated with $e(\boldsymbol{x}) = e(\boldsymbol{z}^*) + \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ follows a zero-mean Multivariate Gaussian Distribution. Substituting this into (7), we have $\hat{e}(\boldsymbol{x}) = e(\boldsymbol{z}^*) + (1 - s_{xz^*})\boldsymbol{\xi}$. Clearly, we have $\mathbb{E}[e(x)] = \mathbb{E}[\hat{e}(x)], \det(\mathbb{V}[e(\boldsymbol{x})]) \geq \det(\mathbb{V}[\hat{e}(\boldsymbol{x})])$. As the clusters become apparent in the latent space, the following conditions hold: (i) all data points $e(\boldsymbol{x})$ fall closer and closer to their center, making $\det(\mathbb{V}[\boldsymbol{\xi}]) \to 0$, thus $e(\boldsymbol{x}) \to \hat{e}(\boldsymbol{x}) \to e(\boldsymbol{z}^*)$. Following (1) and the discussion in lemma 1, the second part in (6a) is redundant when the first part holds; (ii) $s_{xz^*} \to 1$, thus the second part in (8a) holds for any $\eta > 0$. Since the second part in both (6a) and (8a) will always hold when the first part holds, we can simply drop the second part in both formulations, thus making $\mathcal{L}_2$ equivalent to $\mathcal{L}_1$. □

Assumptions and Discussions. Lemmas 1 and 2 show that to make the relaxed version of the optimization problem equivalent to the original, one can either increase the number of inducing points or assume the latent space obeys a clustering structure with MoG distribution. In the absence of cluster structure in the latent space, lemma 1 ensures that with small η, it is feasible to optimize $\mathcal{L}_1$ to Θ. In the presence of a strong cluster structure in the latent space, lemma 2 shows that it suffices to optimize $\mathcal{L}_2$. Note that with small ϵ, η, the constraints (8a) would force a cluster structure in $\hat{e}(\mathcal{X})$ regardless of the structure in $e(\mathcal{X})$. Because lemma 2 requires that data points in the latent space follow a Gaussian Distribution around its cluster centers, we enforce a Gaussian prior on the latent space $e(\mathcal{X})$ (See (9) and more details in Appendix).

Optimization. The straightforward way to optimize $\mathcal{L}_2$ subject to (8a) involves working with its dual using the technique of Lagrangian multipliers. For inducing clusters to work as intended, ϵ and η to must be sufficiently small. Alternatively, we can solve an unconstrained optimization of:

$$\arg\min_{\Theta} \mathcal{L}_3 = -\mathbb{E}_{q(\mathbf{f},\mathbf{u})}[\log p(\boldsymbol{y}|\mathbf{f})] + \text{KL}[q(\mathbf{u})||p(\mathbf{u})]$$
$$+ \lambda_1 \max \text{diag}(BB^\top) - \lambda_2 \min_{x \in \mathcal{X}} \max_{z \in \mathcal{Z}} s_{xz} - \lambda_3 \mathcal{L}_{\text{Gau}}(e(\mathcal{X})) \tag{9}$$

where $\lambda_i, i = 1, 2, 3$ are regularization coefficients. $\mathcal{L}_{\text{Gau}}$ is the Gaussian prior applied to latent representation on the training data $e(\mathcal{X})$ (See Appendix for details). Karush-Kuhn-Tucker (KKT) conditions imply that there is a one-to-one mapping between λ_1 and ϵ, and between λ_2 and η, provided the constraints (8a) in $\mathcal{L}_2$ are active.[2]. For non-Gaussian likelihood, the expectation term in $\mathcal{L}_3$ is intractable. However, we can use Monte Carlo sampling and reparameterization to obtain an efficient approximation of the likelihood. That is, we can draw T samples from $(\mathbf{f}, \mathbf{u}) \sim q(\mathbf{f}, \mathbf{u})$ then approximate the expectation as $\frac{1}{T}\sum_{i=1}^{T} \log p(\boldsymbol{y}|\mathbf{f}^{(i)})$.

Experiments and Results

Experimental Setup. We compare ICDKGP to several state-of-the-art methods for predictive modeling from longitudinal data on both simulated and real-world benchmark data sets. The experiments are designed to answer the following research questions: (RQ1) How does the performance of ICDKGP compare with state-of-the-art baselines on standard longitudinal regression tasks? (RQ2) Can ICDKGP better

[2]The same trick is often used to convert constraint ℓ_1-regularized regression to unconstrained LASSO regression (Hastie, Tibshirani, and Wainwright 2019).

recover complex a priori unknown correlation structure in the longitudinal data? (RQ3) To what extent does the performance of ICDKGP depend on the mean function and inducing clusters?

To answer RQ1, we conducted experiments with simulated data with smooth and non-smooth target functions and several real-world data sets. We evaluated the performance of each model on each regression task using 10 independent runs, using 50%, 20%, and 30% of data for training, validation, and testing respectively. Following (Liang et al. 2020), we report the mean and standard deviation of R^2 between the predicted outcomes and actual outcomes. $\mathrm{R}^2 = 1 - \frac{MSE \text{ of model}}{MSE \text{ of mean}}$ (where MSE denotes the mean squared error). R^2 measures the relative improvement of the model's regression accuracy over a baseline that uses the mean of outcomes over the training set as its prediction. The sign of R^2 indicates whether the model performs better than the baseline.

To answer RQ2, because the true underlying correlation structure of the real-world data is not known, we used simulated data with a known correlation structure. Following the procedure described in (Liang et al. 2021), we compare the learned correlation matrix with the known ground truth correlation matrix for the simulated data.

To answer RQ3, we compared ICDKGP with two of its variants: (i) DKGP, a deep kernel GP regression model using standard inducing points and the state-space mean function (Module 1 and Module 2); and (ii) DKGP-ZM, DKGP with zero means (Module 1). We also compared the visualizations of the latent representations learned by ICDKGP ($\hat{e}(X)$) and DKGP ($e(X)$) using a 2-dimensional T-SNE plot.

Simulated Data. We simulated longitudinal data with the desired correlation structure. Specifically, we set the outcome $\boldsymbol{y} = f(X) + \boldsymbol{\epsilon}$ where $f(X)$ is a non-linear transformation of the observed covariate matrix X and the residual $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma)$. Correlation type and smoothness are varied by manipulating Σ (Liang et al. 2021). To simulate a smooth target function with longitudinal correlation, we set Σ to a block diagonal matrix with non-zero entries for within-individual observations. To simulate a smooth target function with multi-level correlation, we first split the individuals into C clusters and assign non-zero entries for the data points in the same cluster. To simulate a non-smooth target function across the observations per individual, we split the observations for each individual into 2 clusters and assign non-zero entries for the data points in the same cluster. Following (Liang et al. 2021), we simulated longitudinal data consisting of 30 covariates for 40 individuals to obtain 20 observations per individual. We vary the number of clusters C from 2 to 5. Details of simulated data generation are given in the Appendix.

Real-world Data. We used three real-world longitudinal data sets, each with some degree of discontinuities. (i) The **SWAN** (Sutton-Tyrrell et al. 2005) data is taken from a longitudinal study of women's health in midlife. We trained models to predict the adjusted CESD score, which is often used to screen for depression. (ii) **GSS** (Smith et al. 2015) data is taken from a 30-year longitudinal study designed to monitor, understand, and explain changes in the attitudes and behaviors of Americans. Specifically, we trained models to predict the self-reported happiness of individuals; (iii) **TADPOLE** (Marinescu et al. 2018) data is taken from the Alzheimer's Disease Neuroimaging Initiative (ADNI) longitudinal study of individuals at high for Alzheimer's Disease. We focused on predicting the ADAS-Cog13 score from the TADPOLE data using demographic features and MRI measures. When all the methods ran to completion on a data set, their regression performance was compared on the entire data set. When baseline methods fail to run to completion on these real-world data sets, following (Liang et al. 2021), we compared the regression methods on a subset of each real-world data set consisting of 50 individuals with the largest number of observations. We report execution failure if a method fails to converge within 48 hours or generates an execution error. Additional details, e.g., on how the data were pre-processed, etc., are provided in the Appendix.

State-of-the-Art Baseline Methods. We compared ICDKGP with the following SOTA baselines: (i) Conventional longitudinal regression models, *i.e.,* **GLMM** (Bates et al. 2015), **GEE** (Inan and Wang 2017); (ii) State-of-the-art (SOTA) longitudinal regression models, *i.e.,* **LMLFM** (Liang et al. 2020) and **L-DKGPR** (Liang et al. 2021); (iii) SOTA GP models for general regression, *i.e.,* **SKIPGP**, exact GP with scalable kernel interpolation for product kernels (Gardner et al. 2018), **SVGP**, stochastic variational GP regression (Hensman, Matthews, and Ghahramani 2015) and its variant **DSVGP** that incorporates a deep kernel.

Implementation and Supplementary Material. Full implementation details and the Appendix can be accessed through https://github.com/junjieliang672/ICDKGP/blob/main/ICDKGP-AAAI24_appendix.pdf.

Results

We proceed to describe the results of our experiments that answer our research questions RQ1-RQ3.

Predictive Performance of ICDKGP vs SOTA Baselines on Simulated Data. Table 1 summarizes the results of our experiments on simulated data with both smooth and non-smooth target functions. We see that ICDKGP substantially outperforms all of the GP models for general regression (*i.e.,* SVGP, DSVGP and SKIPGP) and most longitudinal models (GLMM, GEE, and LMLFM) when the data exhibit multi-level correlations (MC). This result is in part explained by the fact that GEE and GLMM are designed for settings where the correlation structure in the data is known; LMLFM handles only a special case of MC where cluster correlation exists only among individuals observed at the same time points. Although SVGP, DSVGP, and SKIPGP can handle data with arbitrary stationary correlation structure, they lack to ability to handle time-invariant effects. Though L-DKGPR delivers the best performance among the SOTA baselines, because of its zero mean assumption, it fails to learn a kernel that is expressive enough to model the target function. In contrast, ICDKGP can better model the target function thereby outperforming the SOTA baselines.

Predictive Performance of ICDKGP vs SOTA Baselines on Real-World Data. Table 2 summarizes how ICDKGP

Target Type	Method	LC	MC($C=2$)	MC($C=3$)	MC($C=4$)	MC($C=5$)
Smooth	ICDKGP	84.2±2.9	**99.5±0.5**	**99.5±0.3**	**99.5±0.3**	**99.6±0.3**
	L-DKGPR	**86.0±0.2**	91.3±0.2	**99.6±0.2**	**99.8±0.2**	**99.8±0.2**
	LMLFM	54.7±15.1	-138.3±121.9	-48.3±123.6	22.6±49.0	36.2±41.1
	SVGP	78.5±3.1	-102.7±83.1	-102.7±83.1	-51.6±41.5	-36.4±35.2
	DSVGP	51.1±10.9	-138.3±126.4	-30.6±21.3	-27.4±27.8	-5.8±3.3
	SKIPGP	17.4±40.6	-104.9±86.5	-67.2±36.1	-85.0±40.2	-77.3±36.9
	GLMM	5.3±27.9	-656.3±719.8	-801.4±507.4	-684.1±491.3	-528.7±313.5
	GEE	59.0±24.5	-636.1±606.0	-703.6±465.8	-665.6±554.3	-516.5±457.5
Non-smooth	ICDKGP	**84.2±5.2**	**89.1±0.5**	**89.6±2.9**	**92.0±5.4**	**93.1±3.2**
	L-DKGPR	76.8±17.8	62.7±41.9	75.0±12.0	89.6±5.5	83.4±7.8
	LMLFM	76.4±8.8	70.8±1.9	69.4±3.6	73.1±4.6	69.2±7.3
	SVGP	69.2±13.6	31.2±20.4	26.0±28.7	19.3±26.3	10.2±19.9
	DSVGP	78.5±16.9	35.0±28.0	31.5±29.6	20.7±30.3	9.7±24.5
	SKIPGP	68.4±13.5	31.2±20.1	28.1±25.0	19.8±25.5	12.1±18.2
	GLMM	66.8±15.9	18.7±26.0	11.4±38.4	1.9±30.6	-10.9±27.1
	GEE	71.6±14.9	29.3±24.6	25.8±28.8	17.7±30.1	5.0±23.6

Table 1: Regression accuracy R^2 (%) comparison on simulated data over different correlation structures.

Data sets	N	I	P	ICDKGP	L-DKGPR	LMLFM	SVGP	DSVGP	SKIPGP	GLMM	GEE
TADPOLE^S	595	50	24	**53.8±5**	44.0±6	8.7±5	-0.5±4	-1.7±5	-6.7±26	50.8±6	-11.4±5
SWAN^S	550	50	137	**47.9±4**	46.8±5	38.6±4	-24.3±8	19.9±3	-36.8±10	40.1±8	46.4±8
GSS^S	1.5K	50	1.6K	**25.3±3**	19.1±4	15.3±1	8.9±6	6.0±13	NI	NC	-4.6±4
TADPOLE^L	8.7K	1.7K	24	63.1±2	**64.9±1**	10.4±1	21.3±1	14.1±4	OOM	61.9±2	17.6±1
SWAN^L	28.4K	3.3K	137	**54.2±0**	52.5±0	48.6±2	46.4±0	46.1±1	OOM	NC	NC
GSS^L	59.6K	4.5K	1.6K	**56.4±1**	**56.9±0**	54.8±2	55.6±0	45.8±4	OOM	NC	NC

Table 2: Regression accuracy R^2 (%) on real-world data. We use N to denote the number of data points in the data set, I the number of individuals, and P the number of features. For models that fail during the inference we use 'NI' to denote numerical issues, 'NC' for failure to converge within 48 hours, and 'OOM' for out-of-memory issues.

compares with SOTA baselines on real-world data sets. Comparison of SVGP and DSVGP suggests no clear advantage in replacing a standard RBF kernel with a deep kernel without also adding elements that accommodate, as in the case of ICDKGP, the complex correlation structure of longitudinal data. We further observe that when N (number of individuals represented in the data set) is small, ICDKGP outperforms L-DKGPR by a wide margin, whereas when N is large, L-DKGPR catches up with ICDKGP. We conclude that the mean function used in ICDKGP offers an advantage over zero mean models like L-DKGPR, especially when N is small.

Correlation Structure Recovery by ICDKGP vs. SOTA Baselines. Results of experiments with estimation of the underlying correlation structure of longitudinal data are shown in Fig. 3. We see that for both smooth and non-smooth target functions, ICDKGP outperforms SOTA baselines in accurately recovering the underlying correlation structure. Each of the models shows a degradation in their performance in the non-smooth setting as compared to the smooth setting (e.g., L-DKGPR col. 3 compared to col. 7 in Fig. 3). This suggests that accurate recovery of the correlation structure is harder in the case of non-smoothly time-varying outcomes. The comparison of methods e.g., DKGP, that use a non-zero

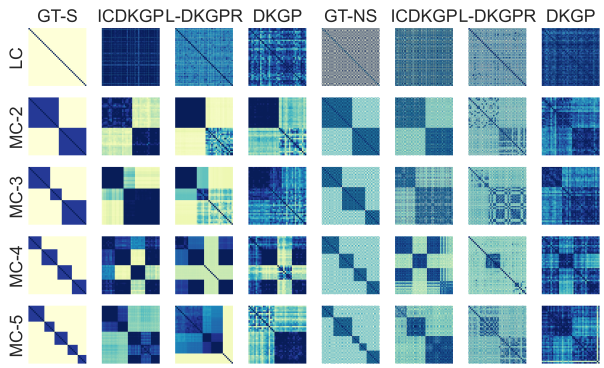

Figure 3: Recovering correlation structure: Comparison of IDDKGP with SOTA baselines on simulated data. GT-S stands for ground truth on smooth function; GT-NS stands for ground truth on nonsmooth function.

mean function with methods that do not, e.g., L-DKGPR, suggest the benefits of non-zero mean in modeling non-smoothly time-varying outcomes. (col. 7 vs. col. 8 in Fig. 3). As observed in (Liang et al. 2021), accurate recovery of the unknown correlation structure is extremely challenging in the absence of a strong prior on the kernel structure. Deep kernels

Data sets	ICDKGP	DKGP		DKGP-ZM		Mean Function
	M=10	M=10	M=100	M=10	M=100	
TADPOLE^L	**63.1±1.8**	60.7±3.9	60.2±4.1	58.9±3.1	61.7±3.4	57.1±3.9
SWAN^L	**54.2±0.2**	**54.4±0.5**	**54.8±0.8**	43.3±3.8	52.8±0.9	**53.9±0.6**
GSS^L	**56.4±0.9**	53.9±1.3	53.7±1.2	48.2±10.2	51.9±5.4	54.8±0.7

Table 3: Effects on the regression accuracy R^2 (%) of different components of ICDKGP on full real-world data.

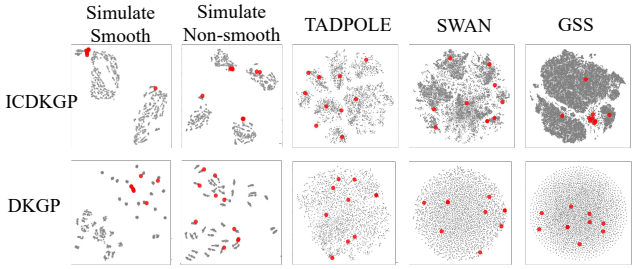

Figure 4: T-SNE visualization of the learned latent representation of ICDKGP and DKGP. The grey and the red dots represent the data points and the inducing cluster centers respectively.

offer a versatile data-driven approach to modeling complex correlation structures. Comparison of ICDKGP with DKGP shows the benefits of inducing clusters in recovering the underlying complex correlation structure from data. Since ICDKGP, DKGP, and L-DKGPR all work within the MLE framework, they search for a kernel that maximizes the likelihood. When the optimal solution is surrounded by a large number of local maxima, it is easy to get stuck in one of such local maxima. Based on the results in Fig. 3, we conjecture that inducing clusters offers a useful prior to constraining the search space, thus delivering better results. We further observe that when inducing clusters are replaced by inducing points as in DKGP, we see a significant drop in performance. We conclude that inducing clusters help to recover complex multi-level correlation structures in the data.

Role of Non-zero Mean Function. Table 3 shows results of comparison of DKGP (deep kernel GP with non-zero mean function) with DKGP-ZM (the with zero mean counterpart of DKGP) shows that DKZP-ZM needs an order of magnitude more inducing points to match the predictive performance of DKGP. Not surprisingly, the mean function serves as an empirical prior that improves the model's predictions.

Role of Inducing Clusters. The results summarized in Table 3 show that ICDKGP consistently outperforms DKGP, albeit by a small margin. The 2-dimensional T-SNE plots of the learned latent representations of the data are shown in Fig. 4. Here, we find that in the case of simulated data, ICDKGP finds a much clearer cluster structure (red points) that matches the ground truth (cluster of grey points) whereas DKGP fails to do so. In the case of real-world data, ICDKGP finds substantially more convincing clusters compared to DKGP. Specifically, the significant performance improvement of ICDKGP over SOTA baselines on TADPOLE data is explained by its ability to learn complex correlation structure (Fig. 4, col. 3).

Summary and Discussion

Summary. We proposed ICDKGP, a novel inducing clusters based longitudinal deep kernel Gaussian process for predictive modeling of irregularly time-sampled, non-smooth, longitudinal data with complex, multi-level correlation structures. ICDKGP decomposes the underlying GP as a sum of a deterministic mean function that reflects the discontinuities in the observed outcomes and a zero mean GP equipped with a longitudinal deep kernel to recover the complex correlation structure in the data. To limit the degrees of freedom (or complexity) of the model and enhance its interpretability, we constrain the inducing points to be the learned centers of clusters in the training data. We formulated the problem of training ICDKGP as a constrained optimization problem and derived its evidence lower bound. We provided a practical solution to the problem based on a novel relaxation whose solutions provably approximate the solution of the original problem under mild assumptions. The results of extensive experiments demonstrate that the predictive models produced by ICDKGP outperform those obtained by SOTA baselines in terms of their ability to accurately predict outcomes and to recover the underlying correlation structure from the data. Our experiments also show the contributions of inducing clusters and of non-zero mean function to ICDKGP's performance advantages over SOTA baselines.

Discussion. Although this paper focuses on predictive modeling and correlation structure recovery from longitudinal data, the inducing clusters based formulation of deep kernel GP should be more broadly applicable to machine learning problems that are amenable to GP-based solutions. Work in progress is investigating such GP models across a broad range of machine learning problems. Also of interest are introducing constraints to confine the learned latent space to Mixture of Gaussians (e.g., (Jiang et al. 2017)) and applications of ICDKGP as well as its variants to challenging real-world applications, e.g., electronic health records.

Acknowledgments

This work was funded in part by grants from the National Science Foundation (2226025, 2041759), the National Center for Advancing Translational Sciences, and the National Institutes of Health (UL1 TR002014).

References

Bates, D.; Mächler, M.; Bolker, B.; and Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, Articles*, 67(1): 1–48.

Chen, C.; Liang, J.; Ma, F.; Glass, L. M.; Sun, J.; and Xiao, C. 2020. UNITE: Uncertainty-based Health Risk Prediction Leveraging Multi-sourced Data. *arXiv preprint arXiv:2010.11389*.

Cheng, L.; Ramchandran, S.; Vatanen, T.; Lietzén, N.; Lahesmaa, R.; Vehtari, A.; and Lähdesmäki, H. 2019. An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data. *Nature communications*, 10(1): 1798.

Chung, I.; Kim, S.; Lee, J.; Kim, K. J.; Hwang, S. J.; and Yang, E. 2020. Deep Mixed Effect Model Using Gaussian Processes: A Personalized and Reliable Prediction for Healthcare. In *Proc. AAAI*, volume 34, 3649–3657.

De Ath, G.; Fieldsend, J. E.; and Everson, R. M. 2020. What do you mean? the role of the mean function in bayesian optimisation. In *Proc. GECCO*, 1623–1631.

Gardner, J.; Pleiss, G.; Wu, R.; Weinberger, K.; and Wilson, A. 2018. Product kernel interpolation for scalable Gaussian processes. In *Proc. AISTATS*, 1407–1416. PMLR.

Graßhoff, J.; Jankowski, A.; and Rostalski, P. 2020. Scalable Gaussian process separation for kernels with a non-stationary phase. In *Proc. ICML*, 3722–3731. PMLR.

Hari, V. 1999. Structure of almost diagonal matrices. *Mathematical Communications*, 4(2): 135–158.

Hastie, T.; Tibshirani, R.; and Wainwright, M. 2019. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Hensman, J.; Matthews, A.; and Ghahramani, Z. 2015. Scalable variational Gaussian process classification. In *Proc. AISTATS*, 351–360. PMLR.

Inan, G.; and Wang, L. 2017. PGEE: An R Package for Analysis of Longitudinal Data with High-Dimensional Covariates. *R Journal*, 9(1): 393–402.

Iwata, T.; and Ghahramani, Z. 2017. Improving output uncertainty estimation and generalization in deep learning via neural network Gaussian processes. *arXiv preprint arXiv:1707.05922*.

Jiang, Z.; Zheng, Y.; Tan, H.; Tang, B.; and Zhou, H. 2017. Variational deep embedding: an unsupervised and generative approach to clustering. In *Proc IJCAI*, 1965–1972.

Liang, J.; Wu, Y.; Xu, D.; and Honavar, V. G. 2021. Longitudinal deep kernel Gaussian process regression. In *Proc. AAAI*, volume 35, 8556–8564.

Liang, J.; Xu, D.; Sun, Y.; and Honavar, V. 2020. LMLFM: Longitudinal Multi-Level Factorization Machine. In *Proc. AAAI*, volume 34, 4811–4818.

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

Marinescu, R. V.; Oxtoby, N. P.; Young, A. L.; Bron, E. E.; Toga, A. W.; Weiner, M. W.; Barkhof, F.; Fox, N. C.; Klein, S.; Alexander, D. C.; et al. 2018. TADPOLE Challenge: Prediction of Longitudinal Evolution in Alzheimer's Disease. *arXiv preprint arXiv:1805.03909*.

Micchelli, C. A.; Xu, Y.; and Zhang, H. 2006. Universal Kernels. *JMLR*, 7(12).

Noack, M. M.; Krishnan, H.; Risser, M. D.; and Reyes, K. G. 2023. Exact Gaussian processes for massive datasets via non-stationary sparsity-discovering kernels. *Scientific reports*, 13(1): 3155.

Noack, M. M.; and Sethian, J. A. 2022. Advanced stationary and nonstationary kernel designs for domain-aware Gaussian processes. *Communications in Applied Mathematics and Computational Science*, 17(1).

Paciorek, C.; and Schervish, M. 2003. Nonstationary covariance functions for Gaussian process regression. *NeurIPS*, 16.

Paun, I.; Husmeier, D.; and Torney, C. J. 2023. Stochastic variational inference for scalable non-stationary Gaussian process regression. *Statistics and Computing*, 33(2): 44.

Shi, J.; Titsias, M.; and Mnih, A. 2020. Sparse orthogonal variational inference for gaussian processes. In *Proc. AISTATS*, 1932–1942. PMLR.

Smith, T. W.; Marsden, P.; Hout, M.; and Kim, J. 2015. General Social Surveys, 1972–2014. *National Opinion Research Center at the University of Chicago*.

Sutton-Tyrrell, K.; Wildman, R. P.; Matthews, K. A.; Chae, C.; Lasley, B. L.; Brockwell, S.; Pasternak, R. C.; Lloyd-Jones, D.; Sowers, M. F.; Torréns, J. I.; et al. 2005. Sex hormone–binding globulin and the free androgen index are related to cardiovascular risk factors in multiethnic premenopausal and perimenopausal women enrolled in the Study of Women Across the Nation (SWAN). *Circulation*, 111(10): 1242–1249.

Timonen, J.; Mannerström, H.; Vehtari, A.; and Lähdesmäki, H. 2019. An interpretable probabilistic method for heterogeneous longitudinal studies. *arXiv preprint arXiv:1912.03549*.

Titsias, M. 2009. Variational learning of inducing variables in sparse Gaussian processes. In *Proc. AISTATS*, 567–574. PMLR.

Tompkins, A.; Oliveira, R.; and Ramos, F. T. 2020. Sparse spectrum warped input measures for nonstationary kernel learning. *NeurIPS*, 33: 16153–16164.

Vantini, M.; Mannerström, H.; Rautio, S.; Ahlfors, H.; Stockinger, B.; and Lähdesmäki, H. 2022. PairGP: Gaussian process modeling of longitudinal data from paired multi-condition studies. *Computers in Biology and Medicine*, 143: 105268.

Williams, C. K.; and Rasmussen, C. E. 2006. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Wilson, A. G.; Hu, Z.; Salakhutdinov, R. R.; and Xing, E. P. 2016. Stochastic variational deep kernel learning. In *NeurIPS*, 2586–2594.