Inducing Clusters Deep Kernel Gaussian Process for Longitudinal Data

Junjie Liang, Weijieying Ren, Hanifi Sahar, Vasant Honavar

The Pennsylvania State University

jiang282@outlook.com, {wjr5337, szh6071, vuh14} @psu.edu

Examples of Longitudinal Data and their Applications

Electronic Health Record (EHR)
- Substance Abuse
- Mental Health
- Obesity

Social survey data
- Attitude changes
- Behavioral changes

Urban data
- Event prediction
- Rental management

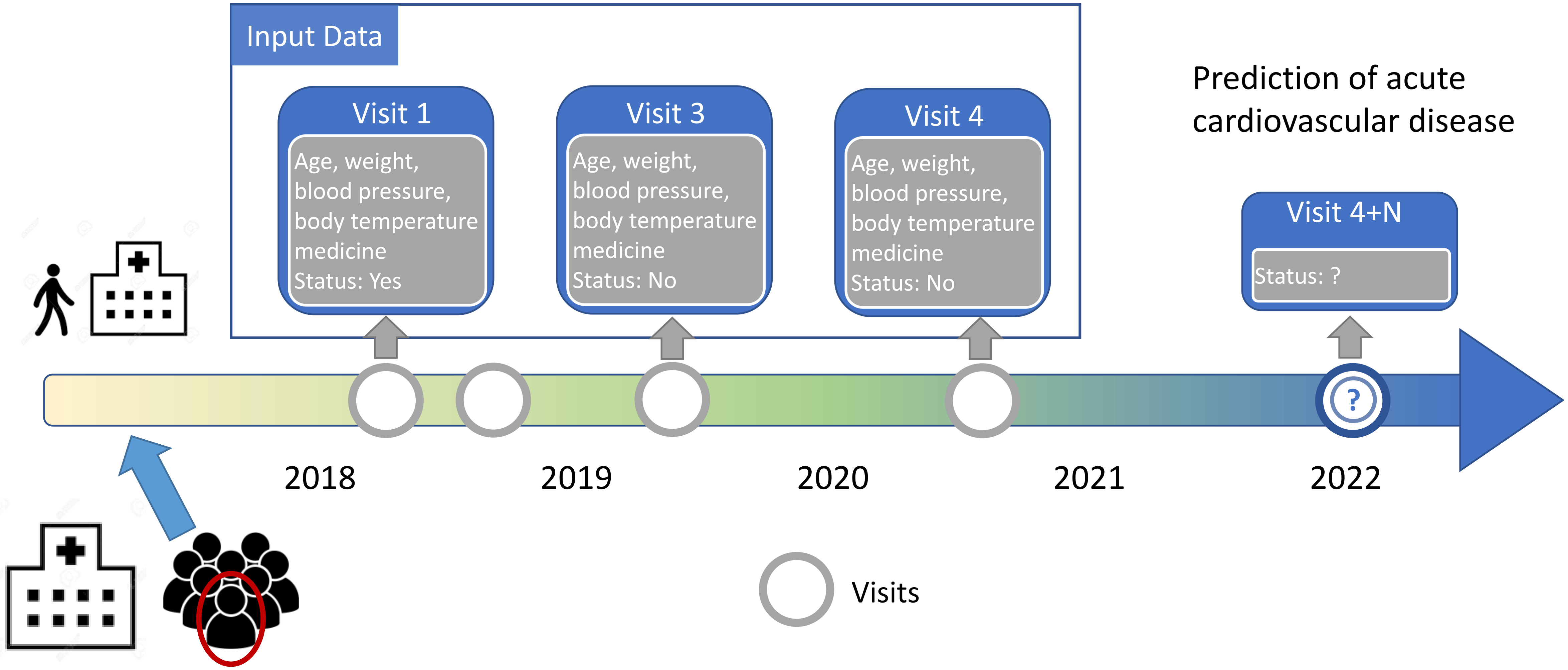
News data
- Entity analysis
- Event classification
- Knowledge extraction

IoT data
- Stress analysis
- Sleep analysis

Financial Data
- Stock crash
- Loan rate
- Credit risk

Manufacturing data
- Predictive maintenance

Predictive Modeling for Longitudinal Data

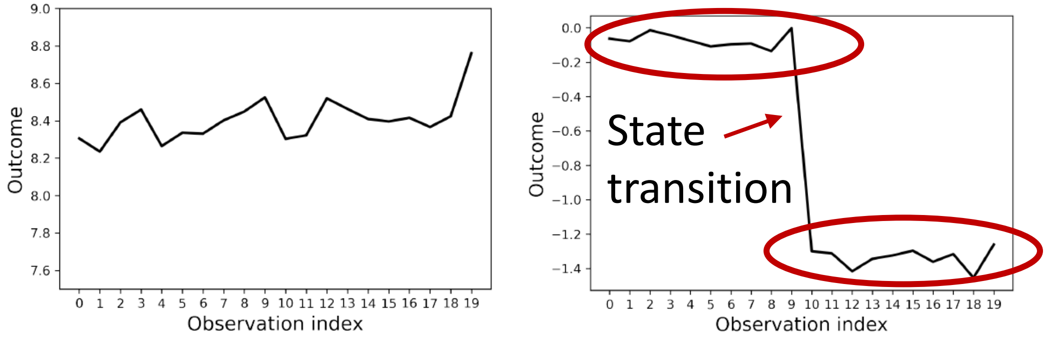


Motivation and Research Goals

In longitudinal study, it is common to see abrupt changes that seemingly indicates a discontinuous curve due to various reasons.

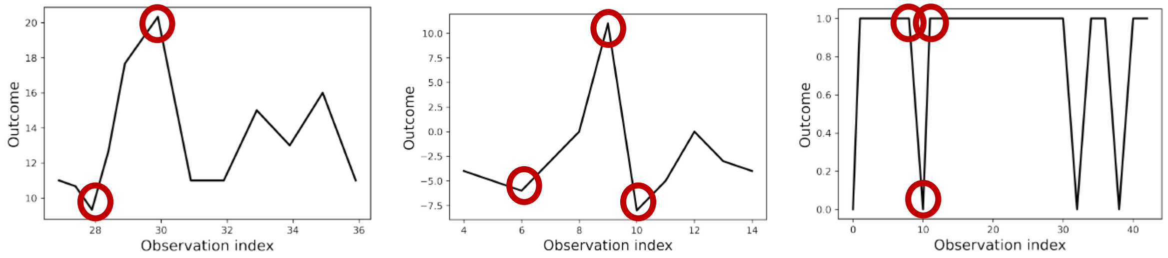
Challenges with H-D longitudinal data

- Kernel learning is difficult when training data is limited.
- Most existing works rely on kernels that embeds a continuous/finite differentiable functional space.



(a) Smooth observations (b) Non-smooth observations

Figure 1: Simulated examples of outcomes from a single individual.



(a) TADPOLE (b) SWAN (c) GSS

Figure 2: Real world examples of non-smooth outcome transitions over time from the observations for a single individual.

Problem Definition

- Goal: Make accurate outcome prediction while accounting for the complex, unknown multilevel data correlation.
  - Learn $p(\boldsymbol{y}|X) \sim N(\boldsymbol{\mu}, \Sigma)$, make prediction using $\boldsymbol{\mu}$, estimate correlation using Σ

$$p(\boldsymbol{y}|X) = \int p(\boldsymbol{y}|\mathbf{f})p(\mathbf{f}|X)d\mathbf{f}$$

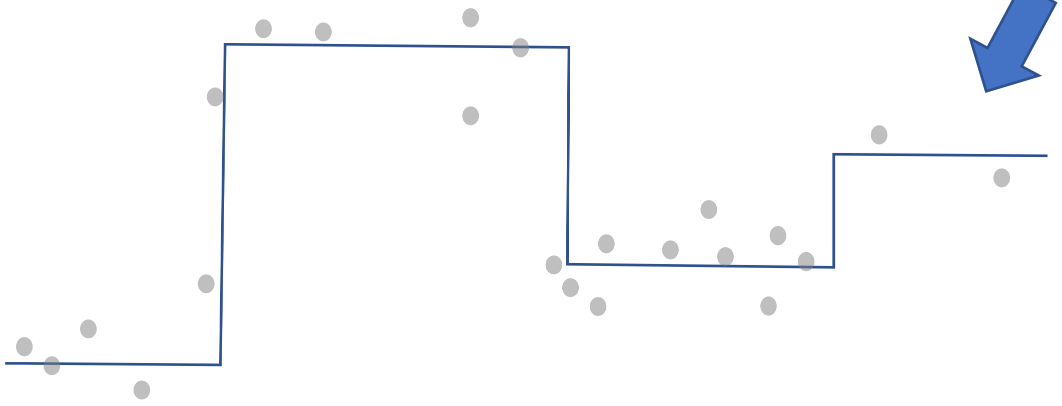
$$(\boldsymbol{y}|\mathbf{f}) \sim N(\mathbf{f}, \sigma^2 I)$$

$$f \sim f_{\perp} + \mathcal{GP}(\mathbf{0}, k_{\theta})$$

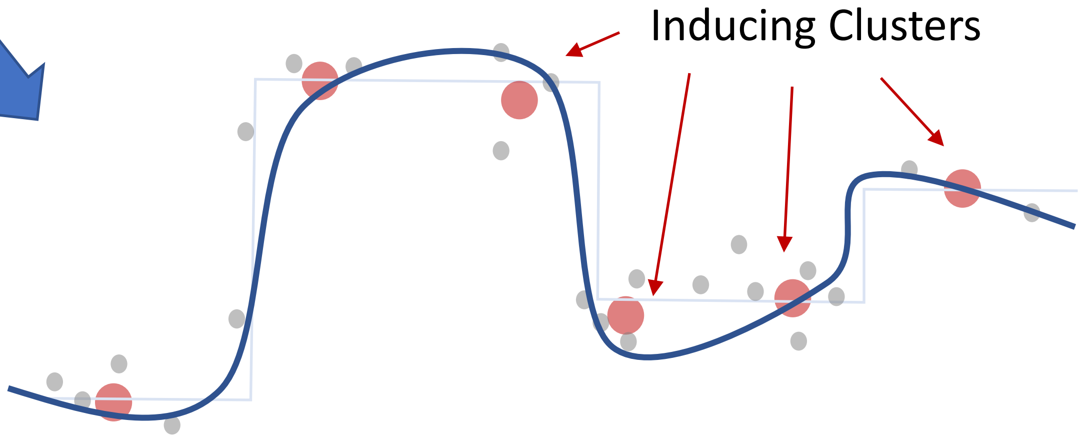
Proposed Method

Decompose the GP into a deterministic mean function and a zero-mean GP with deep kernel

$$f = f_{\perp} + g_{\theta}$$



$f_{\perp}$: State-space mean function



g_{θ}: Inducing Clusters GP

Inducing Clusters Explained

- Inducing clusters = Inducing points + Interpretation
  - Inducing points reduces the computational complexity of GP
  - Interpretation cares about the structure of the input data and tries to force inducing points to locate at the cluster centers of the input data

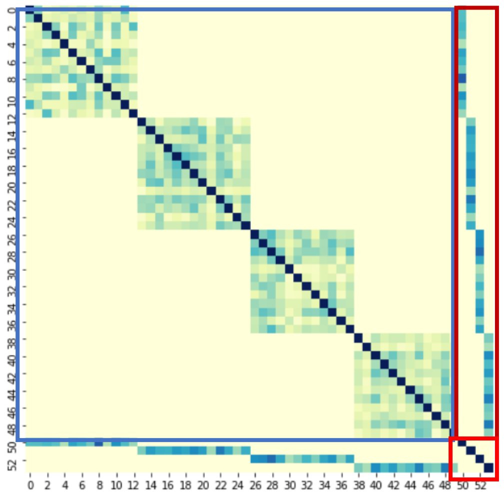
$$\Sigma = \begin{bmatrix} K_{XX}, K_{XZ} \\ K_{XZ}^T, K_{ZZ} \end{bmatrix}$$

Σ is SPD

Input data

Inducing clusters

K_{XX} K_{XZ}



K_{ZZ}

Constraint1. Each input data is correlated to one inducing cluster → K_{XZ} is approximately row-wise single-entry

Constraint2. Inducing clusters are mutually independent → K_{ZZ} is almost diagonal

Constraint ELBO for Proposed Method

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

$$p(\mathbf{f},\mathbf{u}) = \mathcal{N}\left(\begin{bmatrix}\mu_X \\ \mu_Z\end{bmatrix}, \begin{bmatrix}K_{XX} & K_{XZ} \\ K_{XZ}^\top & K_{ZZ}\end{bmatrix}\right) \quad (4)$$

$$q(\mathbf{f},\mathbf{u}) = \mathcal{N}\left(\begin{bmatrix}\mu_X + A(m_z - \mu_Z) \\ m_Z\end{bmatrix}, \begin{bmatrix}V & AS \\ SA^\top & S\end{bmatrix}\right) \quad (5)$$

where $A = K_{XZ}K_{ZZ}^{-1}, V = K_{XX} - AK_{XZ}^\top + ASA^\top$. Since

we have two intuitions:

1. K_{ZZ}, S both almost diagonal
2. K_{XZ}, AS both almost row-wise single-entry

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

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

where $B = K_{ZZ} - \mathrm{diag}(K_{ZZ}), C = K_{XZ} - D \circ K_{XZ}$ 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).

Relaxed Constraint ELBO

- Consider redefining the representation of X by soft mapping from its closest inducing points

Original: $e(x) = e_\gamma(x)$

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

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

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

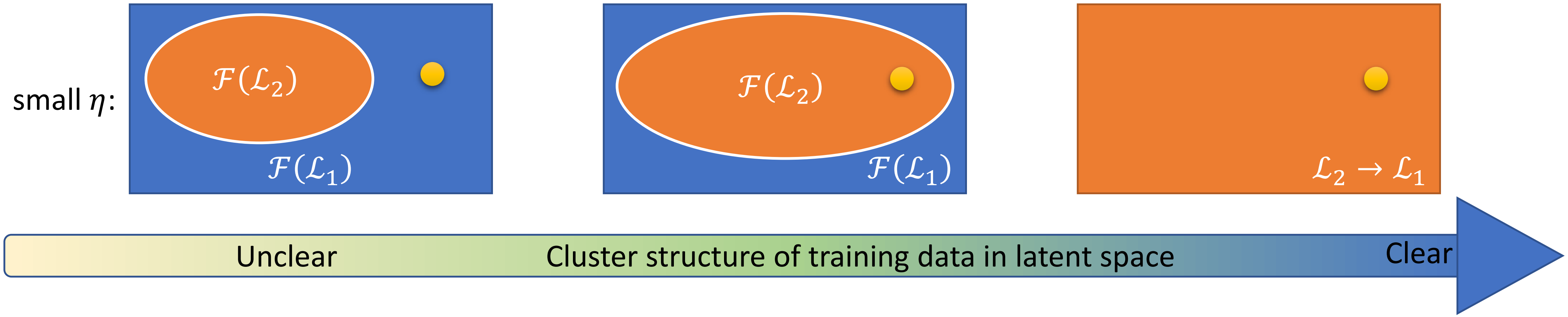
→

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

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

Theoretical Analysis on Relaxed Constraint ELBO

- **Lemma 1. Solution of $\mathcal{L}_2$ is feasible for $\mathcal{L}_1$ when $\boldsymbol{\eta} \to \mathbf{0}$.**
- **Lemma 2. $\mathcal{L}_2$ converges to $\mathcal{L}_1$ when training data form apparent Mixture of Gaussian (MoG) distributions around the latent space $\boldsymbol{e}(\mathcal{X})$.**
- Lemma 1 defines the worst-case scenario while Lemma 2 defines the best-case scenario



Experiment Questions

- RC1. How does performance of ICDKGP compare with SOTA LDA baselines?
- RC2. Can ICDKGP better recover complex correlation structure in longitudinal data?
- RC3. To what extent does the performance of ICDKGP depend on the mean function and inducing clusters?

Data sets and Baselines

Datasets	N	I	P
Simulated	1600	40	30
SWAN	28405	3300	137
GSS	59599	4510	1553
TADPOLE	8771	1681	24

- Data:
  - Simulated data.
  - Three real-world data sets.
- Baselines:
  - Conventional longitudinal models: GLMM; GEE
  - State-of-the-art longitudinal models: LMLFM; L-DKGPR
  - Gaussian Process models: SKIPGP, SVGP, DSVGP

Answering RC1.

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

Answering RC2.

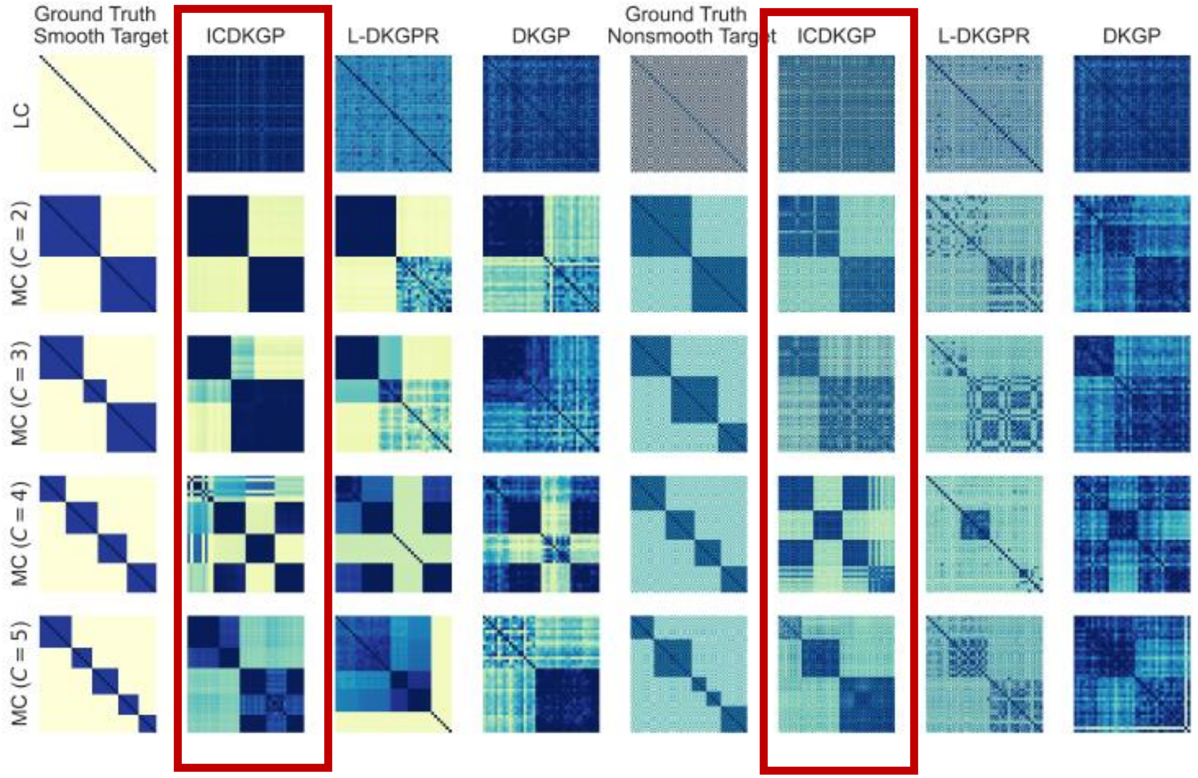
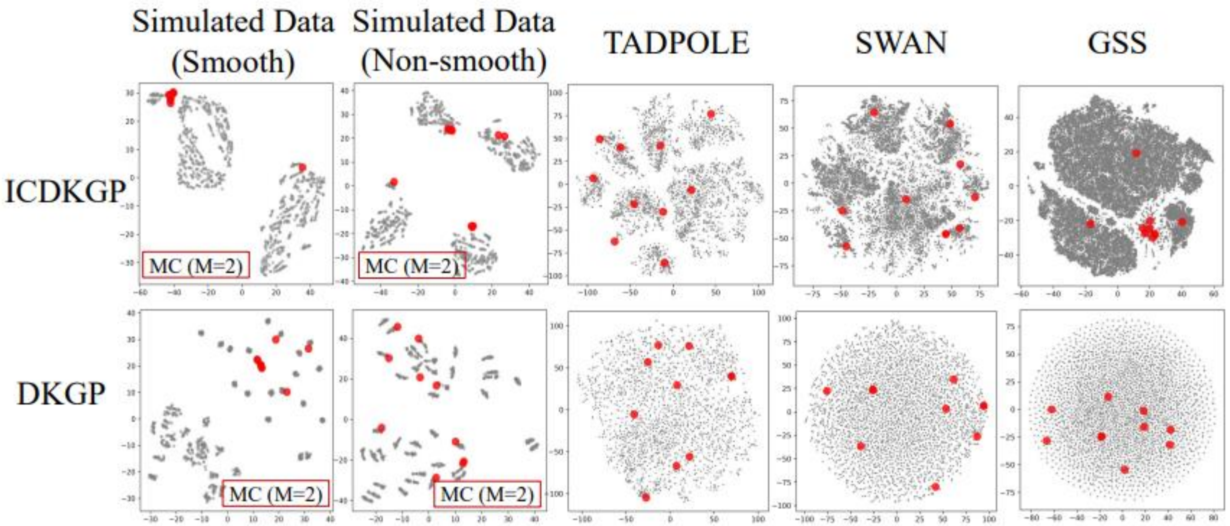


Figure 3: Recovering correlation structure: Comparison of IDDKGP with SOTA baselines on simulated data.

Answering RQ3.

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



Conclusions & Future Works

- We proposed ICDKGP to handle longitudinal data with smooth/non-smooth outcomes
- We introduce and formulate inducing clusters, featuring interpretability of inducing points related technique.
- Future works
  - Show the broad applicability of inducing clusters by applying it to general ML problems.

Thanks for your attention!