Longitudinal Deep Kernel Gaussian Process Regression

Presented by: Junjie Liang

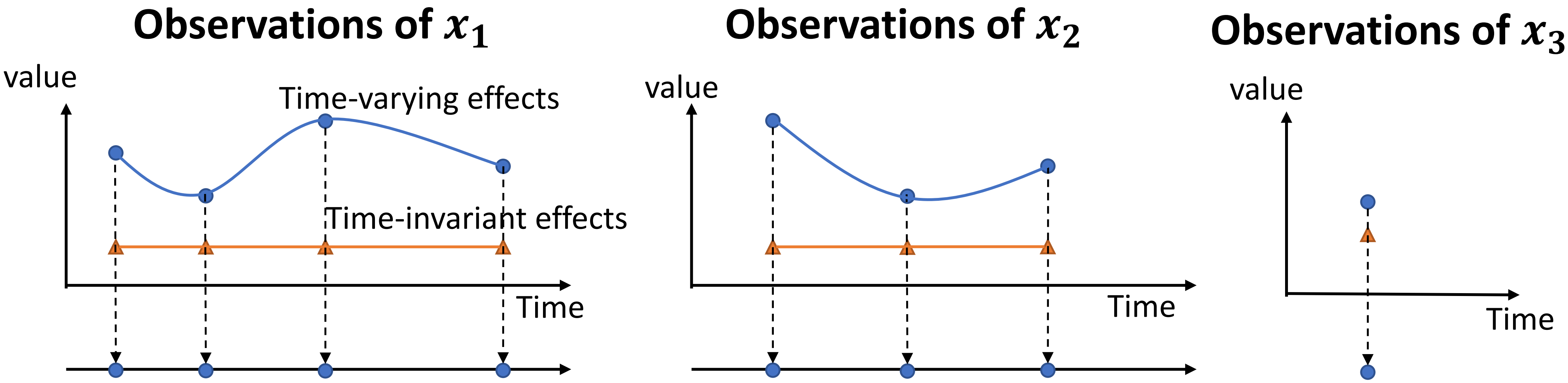
About Me

- 4th-year Ph.D. student at Penn State University
- Research interest:
  - Longitudinal data analysis
  - Causal inference
- E-mail: jul672@psu.edu
- Github: https://github.com/junjieliang672/L-DKGPR.git

Outline

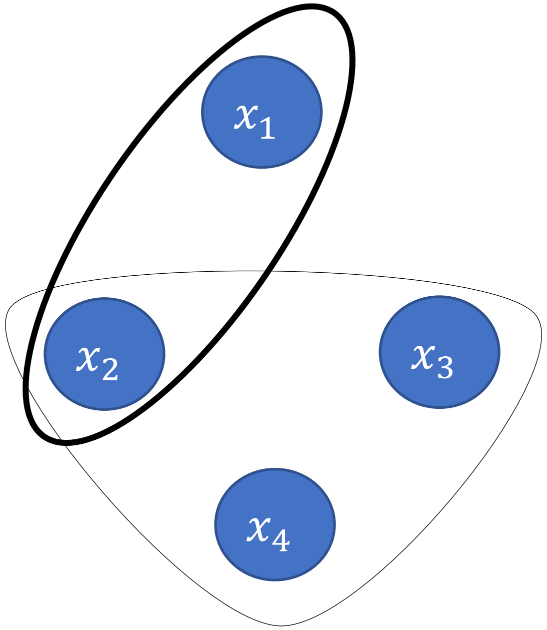
- Background
- Longitudinal Deep Kernel Gaussian Process Regression
- Experiments
- Conclusion

Correlation in Longitudinal Data



Observed outcomes over time are correlated (**Longitudinal Correlation**)

Correlation in Longitudinal Data

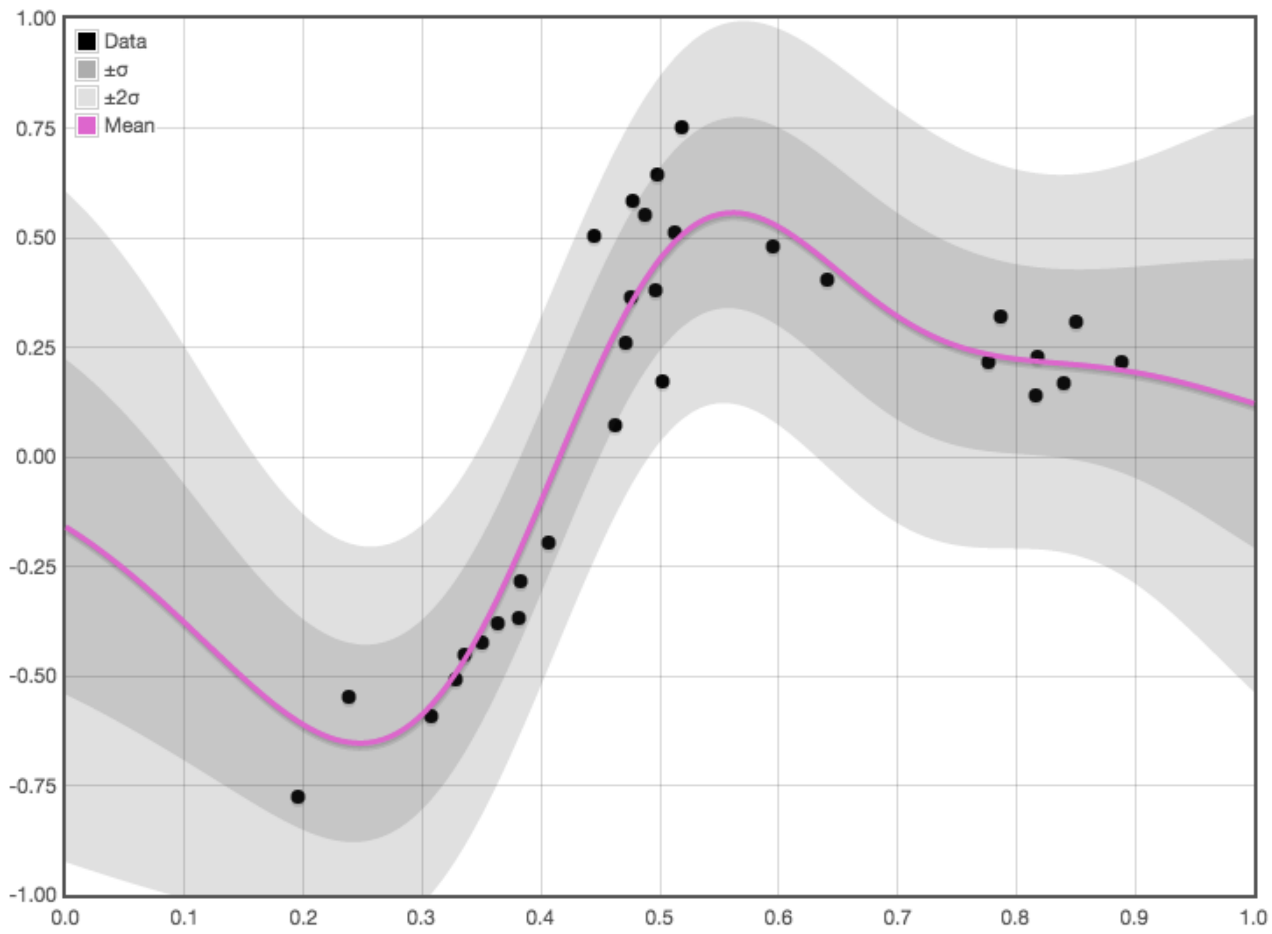


Individuals can also be correlated
(**Cluster Correlation**)

- Correlation can be weak/strong/absent.
- Predictive model needs to account for the complex, unknown data correlation

Gaussian Process

- A distribution over functions
$$f \sim \mathcal{GP}(\mu, k_\gamma)$$
- Kernel function $k_\gamma(\cdot,\cdot)$ describes the correlation between a pair of data.
- Any finite data collections has multivariate Gaussian distribution:
$$(\mathbf{f}|X) \sim N(\mu_X, K_{XX})$$
- Outcome distribution (regression)
$$(\boldsymbol{y}|\mathbf{f}) \sim N(\mathbf{f}, \sigma^2 I)$$



Gaussian Process for longitudinal data

- Pros:
  - Using parametrized kernel to model correlation between observed outcomes. The kernel function provides smooth interpolation between samples, granting GP the ability to cope with irregularly sampled data.
- Cons:
  - Expressive power of GP is dispensed to the choice of kernel. Choosing an appropriate kernel often involves a tedious process of trial and error.
  - Existing GPs for longitudinal data do not scale to thousands of covariates and millions of data points.

Outline

- Background
- Longitudinal Deep Kernel Gaussian Process Regression
- Experiments
- Conclusion

Overview of L-DKGPR

- 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

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

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

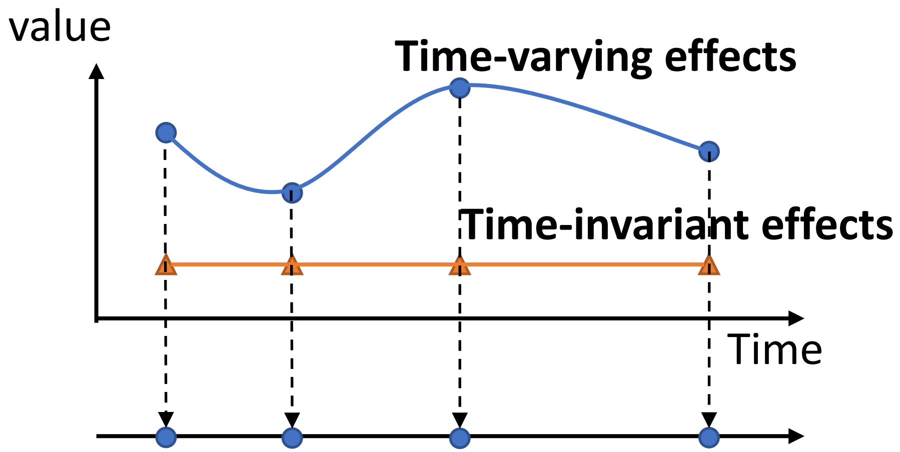
$$f \sim \mathcal{GP}(\mathbf{0}, \boxed{k_\theta})$$

Eliminate the need for kernel searching using **deep kernels**

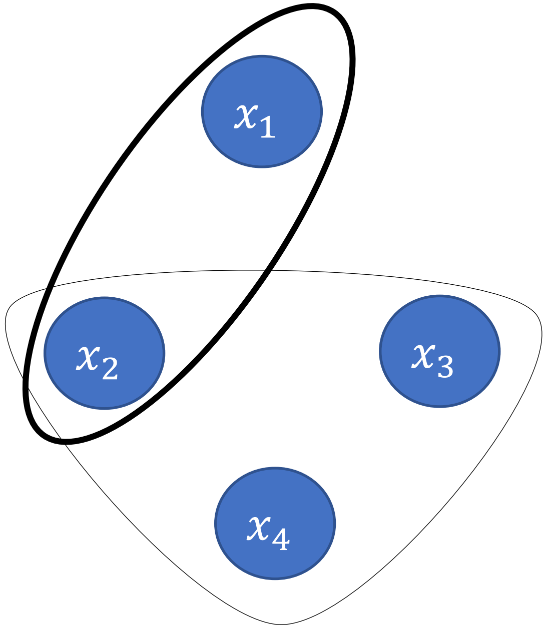
Solve the system with

- Latent space inducing points
- Variational inference

Deep Kernels for Multilevel Correlation



Longitudinal Correlation



Cluster Correlation

$$f = \alpha^{(v)} f^{(v)} + \alpha^{(i)} f^{(i)}$$

$$f^{(v)} \sim \mathcal{GP}\left(\mathbf{0}, \boxed{k_{\gamma}^{(v)}}\right) \rightarrow \text{Kernel for time-varying effects}$$

$$f^{(i)} \sim \mathcal{GP}\left(\mathbf{0}, \boxed{k_{\phi}^{(i)}}\right) \rightarrow \text{Kernel for time-invariant effects}$$

Deep Kernels for Multilevel Correlation

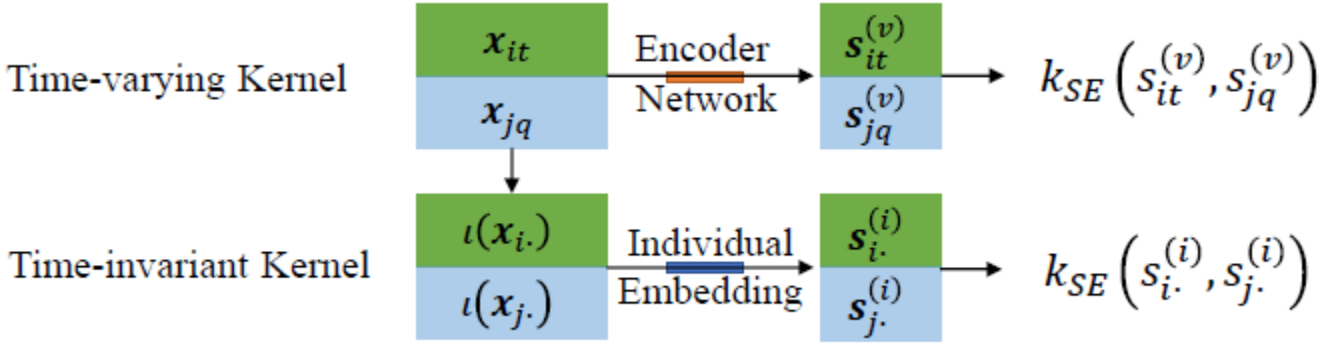
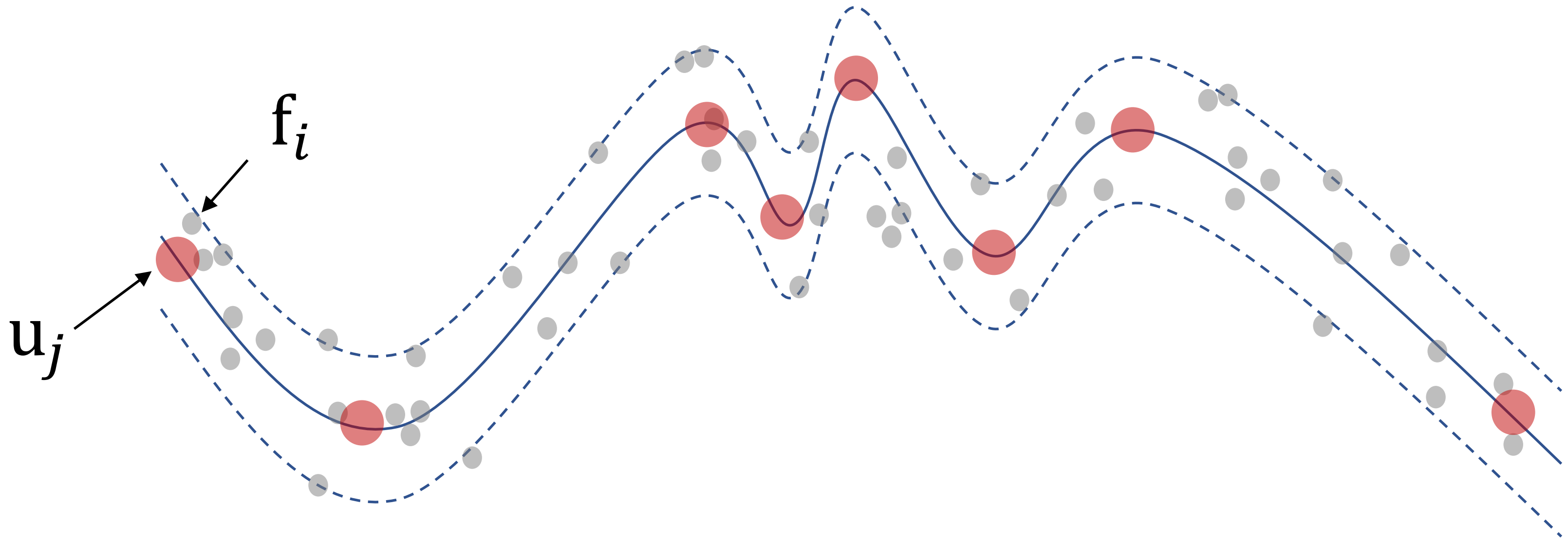


Figure 1: Structure of the deep kernels.

Model Inference – Inducing Points



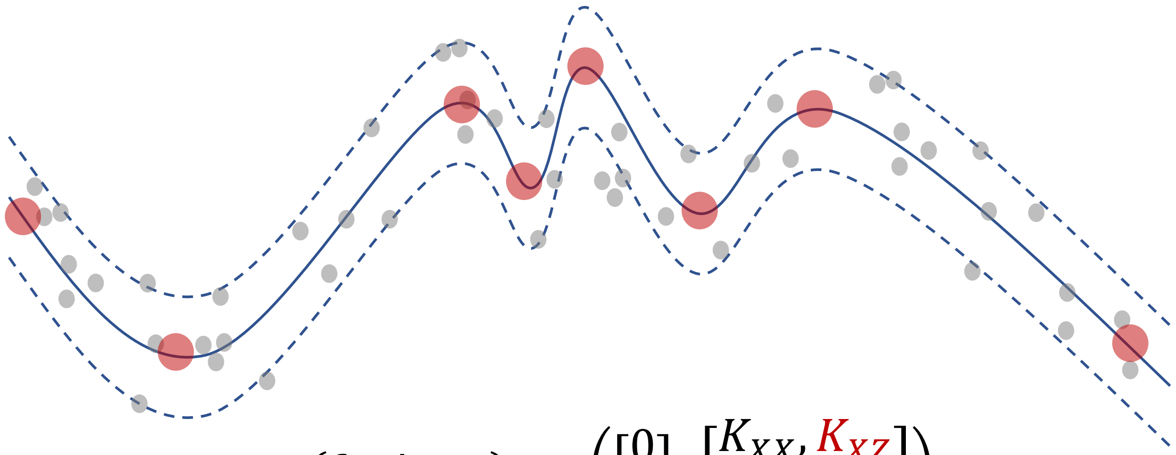
$$(\mathbf{f}|X) \sim N(\mathbf{0}, K_{XX})$$
$$(\mathbf{f}_*|X, X_*, \mathbf{f}) \sim N\left(K_{X_*X}K_{XX}^{-1}\mathbf{f}, K_{X_*X_*} - K_{X_*X}K_{XX}^{-1}K_{X_*X}^{T}\right)$$

$$\mathcal{O}\left(N^3\right)$$

$$(\mathbf{f}, \mathbf{u}|X, Z) \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}K_{XX}, K_{XZ}\\K_{XZ}^{T}, K_{ZZ}\end{bmatrix}\right)$$
$$(\mathbf{f}|X, Z, \mathbf{u}) \sim N(K_{XZ}K_{ZZ}^{-1}\mathbf{u}, K_{XX} - K_{XZ}K_{ZZ}^{-1}K_{XZ}^{T})$$
$$(\mathbf{f}_*|X_*, Z, \mathbf{u}) \sim N\left(K_{X_*Z}K_{ZZ}^{-1}\mathbf{f}, K_{X_*X_*} - K_{X_*X}K_{ZZ}^{-1}K_{X_*Z}^{T}\right)$$

$$\mathcal{O}\left(M^3\right)$$

Model Inference – Latent Space Inducing Points



$$(\mathbf{f}, \mathbf{u} | X, Z) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K_{XX}, K_{XZ} \\ K_{XZ}^{T}, K_{ZZ} \end{bmatrix}\right)$$

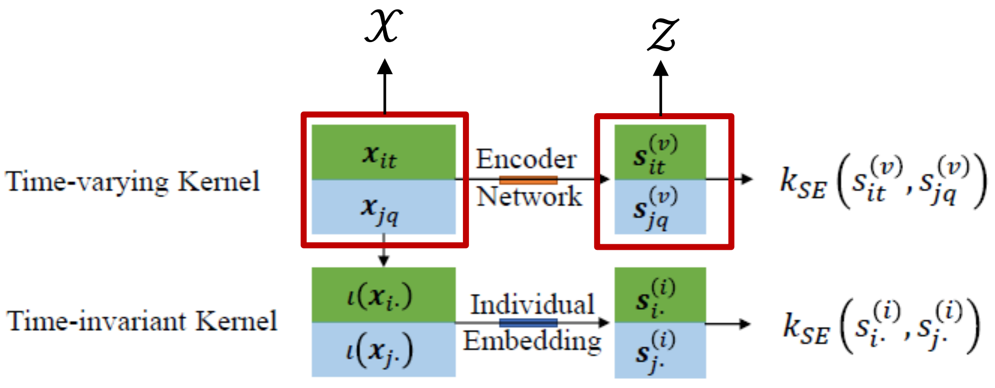


Figure 1: Structure of the deep kernels.

- Z lies in the latent space

$$k_{\gamma}^{(v)}(\boldsymbol{x}, \boldsymbol{z}) = k_{SE}\left(e_{\gamma}(\boldsymbol{x}), \boldsymbol{z}\right)$$
$$k_{\gamma}^{(v)}(\boldsymbol{z}_i, \boldsymbol{z}_j) = k_{SE}\left(\boldsymbol{z}_i, \boldsymbol{z}_j\right)$$

- Z is distinct from all existing X

$$\forall(\boldsymbol{x} \in \mathcal{X}, \boldsymbol{z} \in \mathcal{Z}), \iota(\boldsymbol{x}) \neq \iota(\boldsymbol{z})$$

Model Inference – Variational Inference

$$\Theta^* = \arg\max_{\Theta} \log p(\boldsymbol{y}|X, Z)$$

$$\mathcal{L}_{\text{ELBO}} \stackrel{\text{def}}{=} \mathbb{E}_{q(\mathbf{f},\mathbf{u}|X,Z)}[\log p(\boldsymbol{y}|\mathbf{f})] - \text{KL}\big(q(\mathbf{u}|X,Z)||p(\mathbf{u}|Z)\big)$$

We define the proposal posterior $q(\mathbf{u}|X,Z) = N\big(\boldsymbol{\mu}_q, L_q L_q^T\big)$, to speed up the computation, we follow the deterministic training conditional (DTC) [1].

By simplifying the ELBO, we can:

- Compute the exact ELBO without the need for Monte Carlo sampling.
- Compute the optimal proposal posterior in analytical form.

Algorithm 1: L-DKGPR

Input: Training set $S = \{X, \boldsymbol{y}\}$, latent dimension D_v, D_i, number of inducing points M, gradient-based optimizer and its related hyper-parameters (*i.e.*, learning rate, weight decay, mini-batch size), alternating frequency T.

Initialize the parameters $\Theta = \{\sigma^2, Z, \alpha^{(v)}, \alpha^{(i)}, \gamma, \phi\}$
while *Not converged* **do**
- Update proposal posterior $q(\mathbf{u}|X, Z)$ according to (10) and (12) — **Step1**
- $t = 0$
- **for** $t < T$ **do**
  - Update Θ using the input optimizer.
  - $t = t + 1$

Step2

Outline

- Background
- Longitudinal Deep Kernel Gaussian Process Regression
- Experiments
- Conclusion

Research questions

How does L-DKGPR perform compared to existing models?

- Prediction accuracy on longitudinal regression tasks
- Quality of correlation estimation

How do different model components contribute to L-DKGPR?

- Isolating kernel components
- Solving exact ELBO vs. Monte Carlo estimation

Data sets and Baselines

- Data:
  - Simulated data.
  - Three real-world data sets.
- Baselines:
  - Conventional longitudinal models: GLMM[2]; GEE[3]
  - State-of-the-art longitudinal models: LMLFM[4]; LGPR[5]
  - Gaussian Process models: KISSGP[6]; ODVGP[7]

Regression accuracy

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

Method	LC	MC($C=2$)	MC($C=3$)	MC($C=4$)	MC($C=5$)
L-DKGPR	**86.0±0.2**	**91.3±0.2**	**99.6±0.2**	**99.8±0.2**	**99.8±0.2**
KISSGP	85.9±1.7	-43.4±33.3	-55.5±7.1	-58.2±14.4	-57.2±17.9
ODVGP	82.3±5.2	-1.6±16.9	-14.7±6.5	-13.5±8.4	-6.1±4.4
LGPR	-37.1±19.1	-123.6±162.0	-26.3±43.2	-9.1±14.8	-0.1±5.9
LMLFM	54.7±15.1	-138.3±121.9	-48.3±123.6	22.6±49.0	36.2±41.1
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

Table 2: Regression accuracy R^2 (%) on real-world data sets. We use 'N/A' to denote execution error.

Data sets	N	I	P	L-DKGPR	KISSGP	ODVGP	LGPR	LMLFM	GLMM	GEE
TADPOLE	595	50	24	44.0±5.6	1.2±10.1	9.0±14.1	-261.1±9.0	8.7±5.1	**50.8±5.5**	-11.4±4.8
SWAN	550	50	137	**46.8±4.9**	42.4±4.6	29.0±3.1	-16.6±12.7	38.6±4.2	40.1±7.7	46.4±8.0
GSS	1,500	50	1,553	**19.1±3.7**	12.5±6.3	-7.6±3.3	N/A	15.3±1.4	N/A	-4.6±3.5
TADPOLE	8,771	1,681	24	**64.9±1.4**	0.6±3.9	21.1±1.0	N/A	10.4±0.6	61.9±1.9	17.6±0.7
SWAN	28,405	3,300	137	**52.5±0.4**	20.5±7.6	24.9±21.8	N/A	48.6±2.0	N/A	N/A
GSS	59,599	4,510	1,553	**56.9±0.1**	53.1±0.9	15.4±27.0	N/A	54.8±2.2	N/A	N/A

Correlation estimation

Ground truth

L-DKGPR

Baselines

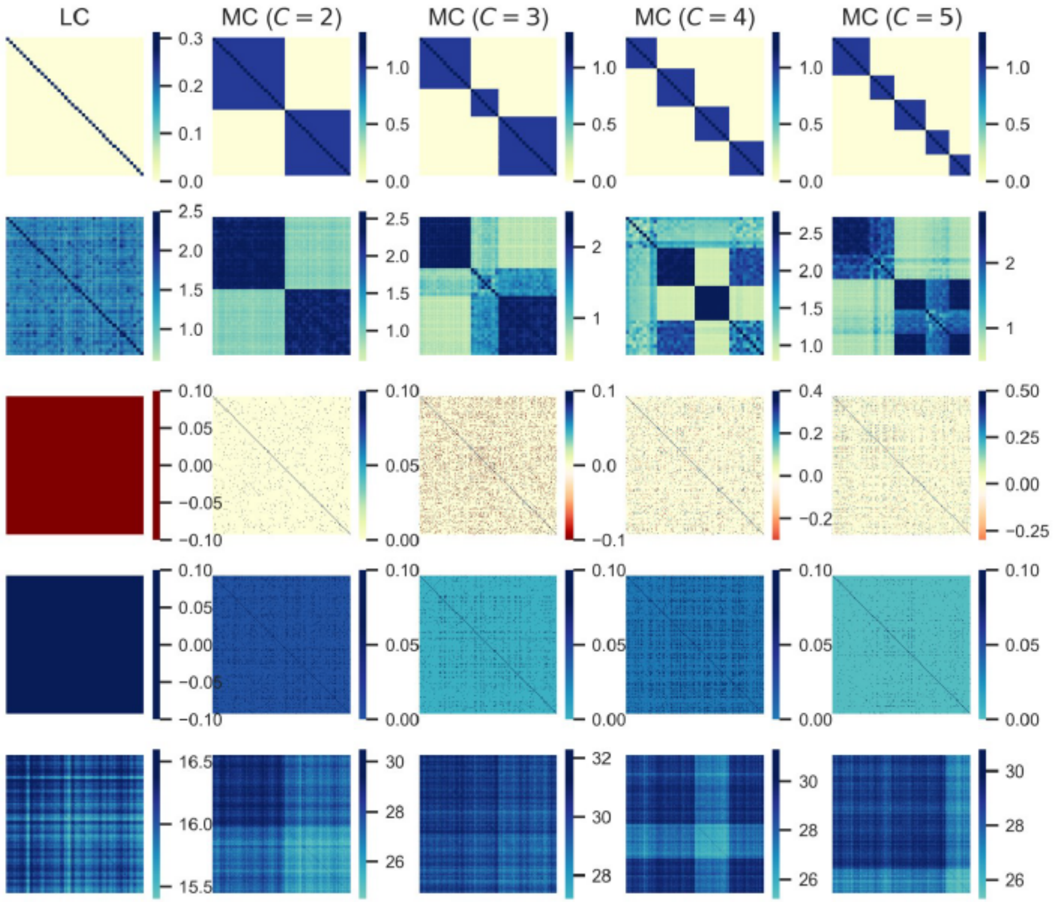


Figure 2: Outcome correlation estimated by all GP methods on simulated data.

Isolating kernel components

Table 3: Effect on the regression accuracy R^2 (%) of different components of L-DKGPR

Data sets	L-DKGPR	L-DKGPR-v	L-DKGPR-i	L-RBF-GPR
TADPOLE	**64.9±1.4**	13.2±1.1	56.3±1.3	55.5±2.4
SWAN	**52.5±0.4**	29.0±3.2	16.7± 2.4	5.4±1.6
GSS	**56.9±0.1**	56.2±0.1	-0.2±0.2	-14.1±0.4
		Time-varying kernel only	**Time-invariant kernel only**	**Non-deep kernel**

Advantage of solving the exact ELBO

Table 4: Effect of solving L-DKGPR using Algorithm 1 vs. Monte Carlo sampling.

Data sets	Solver	M	Iterations	R^2 (%)
SWAN	Alg. 1	10	300	**52.5±0.4**
	Sampling	10	300	3.1±0.2
	Sampling	128	3,000	51.4±0.4
GSS	Alg. 1	10	300	**56.9±0.1**
	Sampling	10	300	4.5±0.1
	Sampling	128	3,000	55.6±0.1

13x Inducing points **10x Iterations**

Summary

- We proposed L-DKGPR, a novel GP with deep kernel specifically designed to cope with longitudinal data that exhibits complex, unknown correlation.
- We improved the scalability of existing GP using two key techniques: (i) latent space inducing points; (ii) variational inference.
- With extensive experiments using both simulated and real-world data, we demonstrated the superior performance of L-DKGPR over state-of-the-art models.

Thank you!

- E-mail: jul672@psu.edu
- Github: https://github.com/junjieliang672/L-DKGPR.git

Reference

[1] Seeger, M., Williams, C., & Lawrence, N. (2003). Fast forward selection to speed up sparse Gaussian process regression (No. CONF).

[2] Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1).

[3] Inan, G., & Wang, L. (2017). PGEE: An R Package for Analysis of Longitudinal Data with High-Dimensional Covariates. R J., 9(1), 393.

[4] Liang, J., Xu, D., Sun, Y., & Honavar, V. (2020, April). LMLFM: Longitudinal Multi-Level Factorization Machine. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 04, pp. 4811-4818).

[5] Timonen, J., Mannerström, H., Vehtari, A., & Lähdesmäki, H. (2019). An interpretable probabilistic machine learning method for heterogeneous longitudinal studies. arXiv preprint arXiv:1912.03549.

[6] Wilson, A. G., Hu, Z., Salakhutdinov, R. R., & Xing, E. P. (2016). Stochastic variational deep kernel learning. Advances in Neural Information Processing Systems, 29, 2586-2594.

[7] Salimbeni, H., Cheng, C. A., Boots, B., & Deisenroth, M. (2018). Orthogonally decoupled variational gaussian processes. In Advances in neural information processing systems (pp. 8711-8720).