Department of Statistical Science
School of Management
Clinical Operational Research Unit
Great Ormond Street Hospital For Children NHS Foundation Trust

Integrative Analysis and Imputation of Multiple Data Streams via Deep Gaussian Process

Ali Septiandri, Deyu Ming, F. A. Díaz De la O, Takoua Jendoubi, Samiran Ray

CHIMERA
Engineering and Physical Sciences Research Council

Background

- In ICU settings, data come from multiple sources and are inherently related
- Measurements collected at irregular intervals (informative sampling)—aligning them will result in missing values
- Cannot always get more samples! Some measurements are invasive (Siegal et al., 2023)

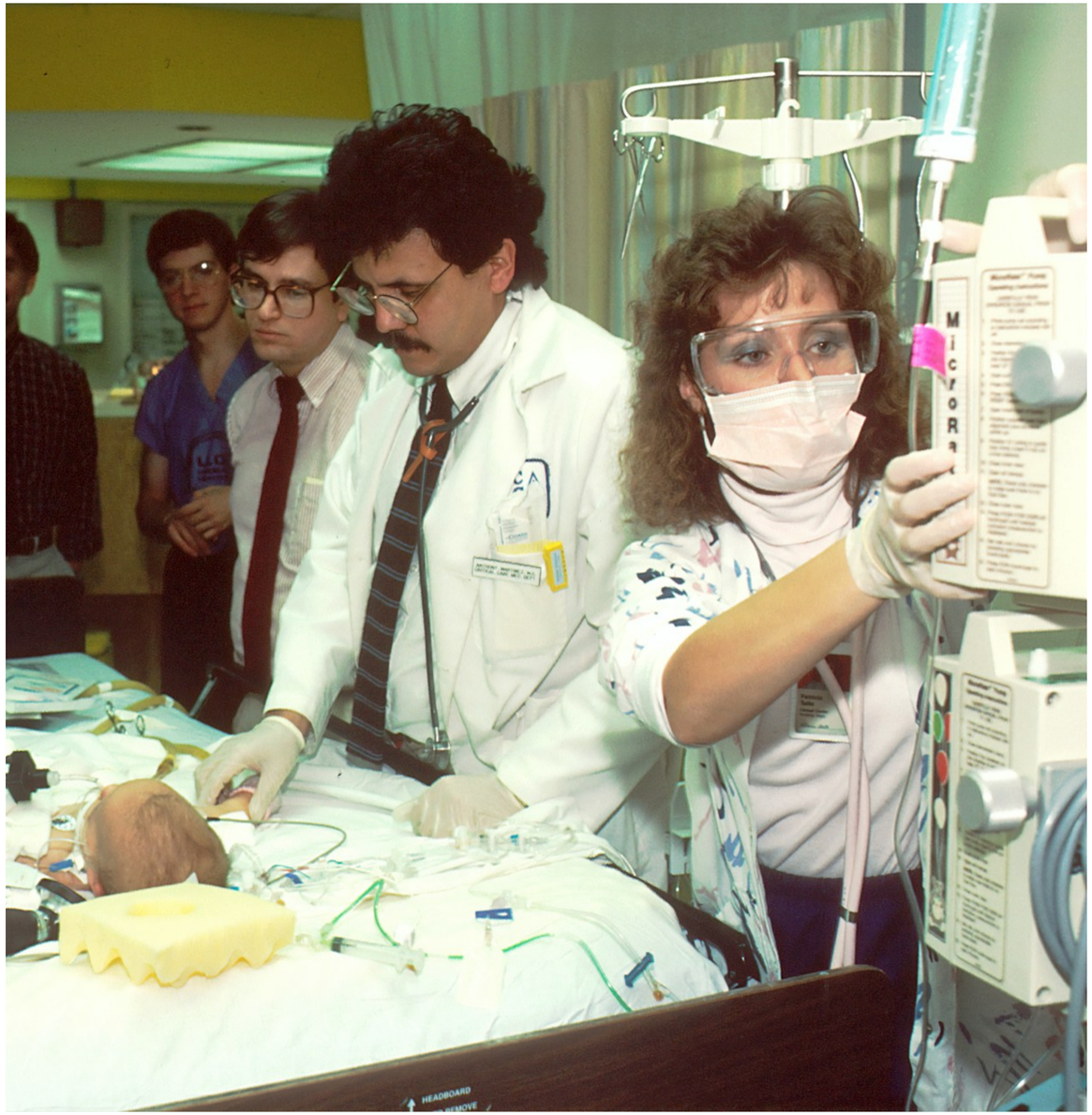
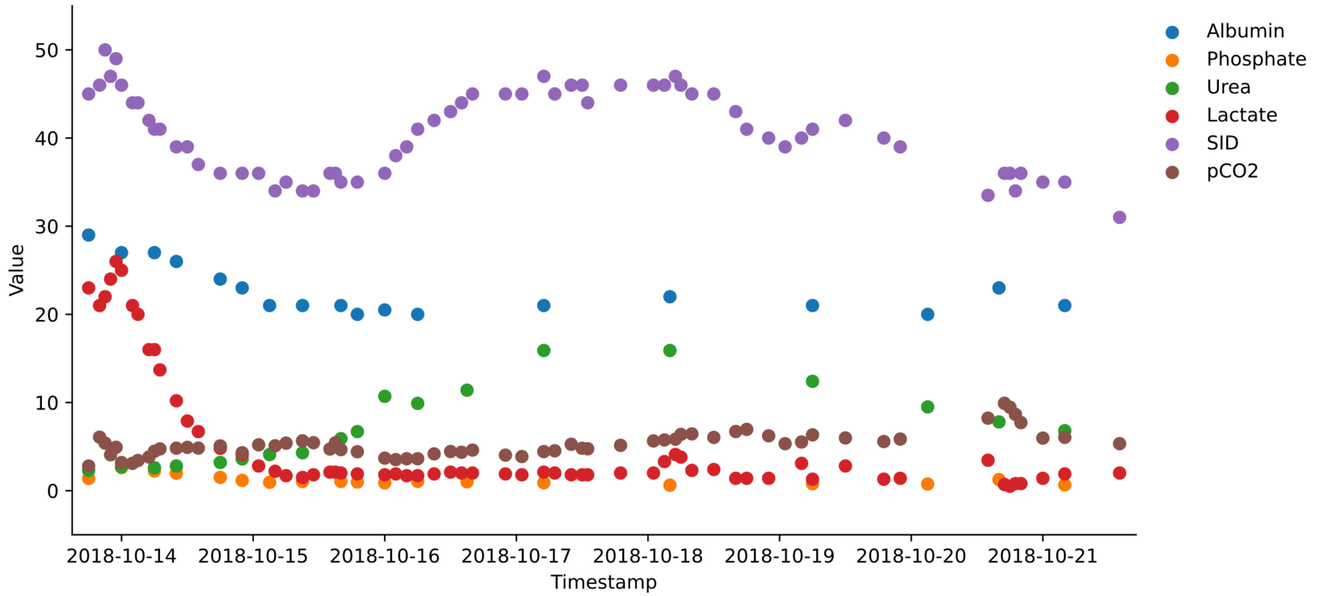


Photo by National Cancer Institute on Unsplash

Challenges

- We want to impute missing values…
- but traditional imputation often ignores temporal structure (e.g. MICE) & uncertainty (e.g. deep learning)
- Need for robust, uncertainty-aware imputation in critical care datasets



On uncertainty quantification

- Medical observations are inherently uncertain, coming from measurement errors or the use of surrogate markers → leading to unreliable model predictions (Cabitza et al., 2017)

- Alerts triggered by prediction tools are often not accompanied by a clinically actionable change → *alarm fatigue* (Embi & Leonard, 2012; Umscheid et al., 2015)

Physicochemical model

- In critical care medicine, clinicians monitor pH levels to inform them about the conditions of a patient
- While pH is the primary variable to monitor, other covariates provide information on metabolic status (Gattioni et al., 2017)
- pH can be modelled from strong ion difference (SID), total weak acid, and pCO2 by the Stewart-Fencl approach

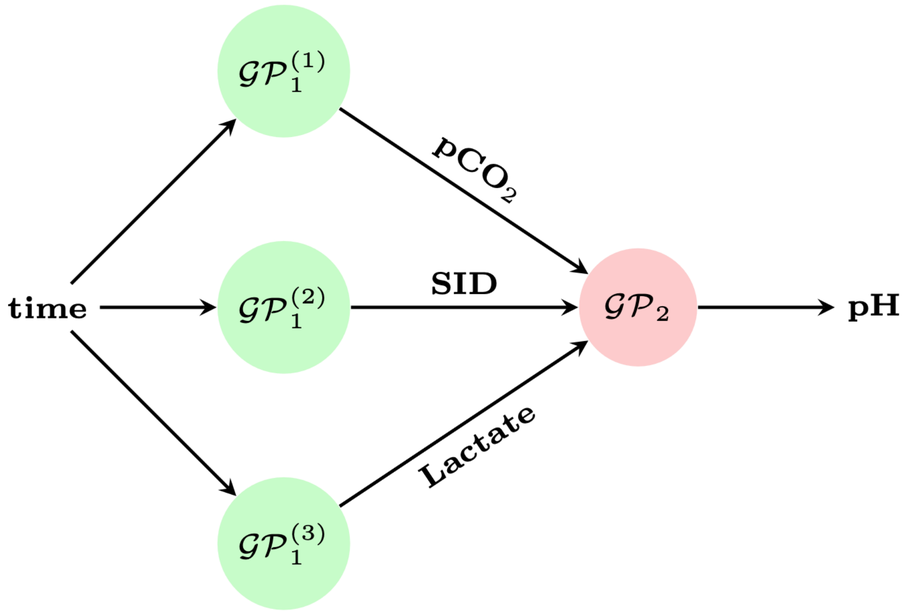
$$[SID] + [H^+] - K_C\frac{pCO_2}{[H^+]} - \frac{K_A A_{TOT}}{K_A + [H^+]} - K_3\frac{K_C pCO_2}{[H^+]^2} - \frac{K_W}{[H^+]} = 0$$

where SID, A_{TOT}, and pCO_2 are independent variables and K_X are constants.

Proposed solution

Deep Gaussian Process with Stochastic Imputation (Ming et al., 2023)

- GPs and Deep GPs are typically used for emulating computationally expensive numerical models
- Integrates longitudinal & cross-sectional information
- Joint modelling for all data streams
- Provides uncertainty quantification for imputed values



Gaussian Processes

$$\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{R}(\mathbf{X}))$$

where $\boldsymbol{\mu} \in \mathbb{R}^N$ is the mean vector, σ^2 is the scale parameter, and $\mathbf{R}(\mathbf{X}) \in \mathbb{R}^{N \times N}$ is the correlation matrix

Cell ij in the matrix $\mathbf{R}(\mathbf{X})$ is specified by $k(\mathbf{X}_{i*}, \mathbf{X}_{j*}) + \eta 1_{\{\mathbf{X}_{i*} = \mathbf{X}_{j*}\}}$, where $k(\cdot, \cdot)$ is a given kernel function with η being the nugget term and $1_{\{\cdot\}}$ being the indicator function

Gaussian Processes

Given a new input position $\mathbf{x_0} \in \mathbb{R}^{1\times D}$, then

$$\mu_0 = \mathbf{r}(\mathbf{x}_0)^T \mathbf{R}(\mathbf{x})^{-1}\mathbf{y}$$

$$\sigma_0^2 = \sigma^2(1 + \eta - \mathbf{r}(\mathbf{x}_0)^T \mathbf{R}(\mathbf{x})^{-1}\mathbf{r}(\mathbf{x}_0))$$

where $\mathbf{r}(\mathbf{x}_0) = [k(\mathbf{x}_0, \mathbf{x}_{1*}), \ldots, k(\mathbf{x}_0, \mathbf{x}_{N*})]^T$

Deep GPs

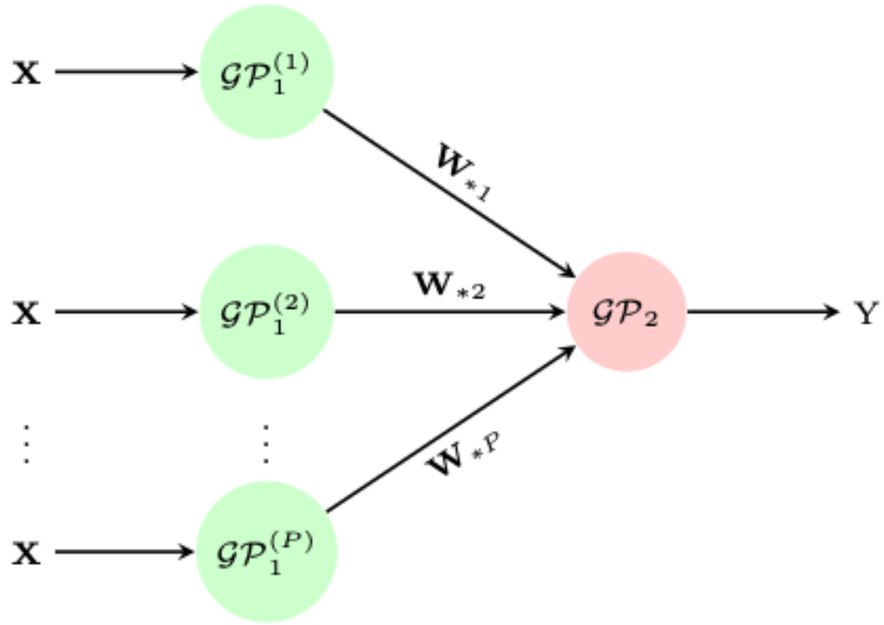
- Consider a GP model with N sets of D-dimensional input ($\mathbf{X} \in \mathbb{R}^{N \times D}$) and produces N sets of P-dimensional output ($\mathbf{W} \in \mathbb{R}^{N \times P}$)
- In the Stewart–Fencl approach, this multi-output GP model can be interpreted as using time as a shared input variable and predicting covariates as outputs
- We can assume that the output $\mathbf{W}$ of this model, i.e. the column vectors $\mathbf{W}_{*p}$, is conditionally independent with respect to $\mathbf{X}$
- We then link the output $\mathbf{W}$ to a second GP model that produces N one-dimensional outputs ($\mathbf{Y} \in \mathbb{R}^{N}$), e.g. to predict pH

Deep GPs

We can see it as a linked GP where, for a new input position $\mathbf{x}_0$, the posterior predictive distribution of the output can be written as

$$p(y_0 \mid \mathbf{x}_0; \mathbf{y}, \mathbf{w}, \mathbf{x}) = \int p(y_0 \mid \mathbf{w}_0; \mathbf{y}, \mathbf{w}, \mathbf{x}) p(\mathbf{w}_0 \mid \mathbf{x}_0; \mathbf{y}, \mathbf{w}, \mathbf{x}) d\mathbf{w}_0$$

$$= \int p(y_0 \mid \mathbf{w}_0; \mathbf{y}, \mathbf{w}) \prod_{p=1}^{P} p(w_{0p} \mid \mathbf{x}_0; \mathbf{w}_p^*, \mathbf{x}) d\mathbf{w}_0$$



Deep GPs

Then the mean and variance become

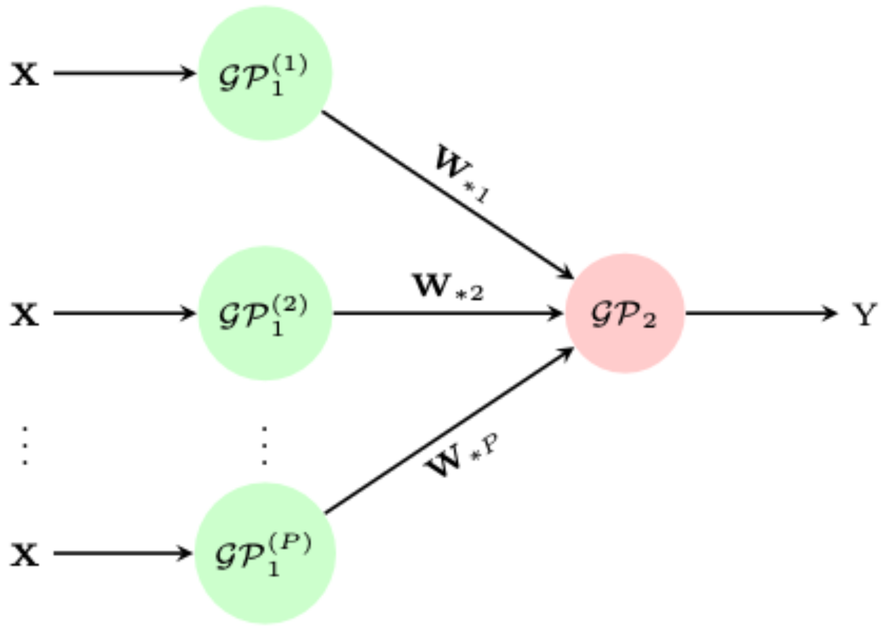
$$\tilde{\mu}_0 = \mathbf{I}(\mathbf{x}_0)^T \mathbf{R}(\mathbf{w})^{-1} \mathbf{y}$$

$$\tilde{\sigma}_0^2 = \mathbf{y}^T \mathbf{R}(\mathbf{w})^{-1} \mathbf{J}(\mathbf{x}_0) \mathbf{R}(\mathbf{w})^{-1} \mathbf{y} - \left[\mathbf{I}(\mathbf{x}_0)^T \mathbf{R}(\mathbf{w})^{-1} \mathbf{y}\right]^2 + \sigma^2 \left(1 + \eta - \text{tr}\left[\mathbf{R}(\mathbf{w})^{-1} \mathbf{J}(\mathbf{x}_0)\right]\right)$$

where $\mathbf{I}(\mathbf{x}_0) \in \mathbb{R}^{N \times 1}$ with its i-th element $I_i = \prod_{p=1}^{P} \mathbb{E}[k_p(W_{0p}(\mathbf{x}_0), w_{ip})]$

and $\mathbf{J}(\mathbf{x}_0) \in \mathbb{R}^{N \times N}$ with its ij-th element $J_{ij} = \prod_{p=1}^{P} \mathbb{E}[k_p(W_{0p}(\mathbf{x}_0), w_{ip}) k_p(W_{0p}(\mathbf{x}_0), w_{jp})]$

Deep GP algorithm



Algorithm 1 Construction of a DGP emulator with the hierarchy in Figure 2

Input: i) Realisations $\mathbf{x}$ and $\mathbf{y}$; ii) A new input position $\mathbf{x}_0$; iii) The number of imputations N.

Output: Mean and variance of $y_0(\mathbf{x}_0)$.

1: **for** $i = 1, \ldots, N$ **do**

2: Given $\mathbf{x}$ and $\mathbf{y}$, draw an imputation $\{\mathbf{w}_{*p,i}\}_{p=1,\ldots,P}$ of the latent output $\{\mathbf{W}_{*p}\}_{p=1,\ldots,P}$ via an Elliptical Slice Sampling [40] update.

3: Construct the LGP emulator $\mathcal{LGP}_i$ with the mean $\tilde{\mu}_{0,i}(\mathbf{x}_0)$ and variance $\tilde{\sigma}^2_{0,i}(\mathbf{x}_0)$, given $\mathbf{x}$, $\mathbf{y}$, and $\{\mathbf{w}_{*p,i}\}$.

4: **end for**

5: Compute the mean $\mu(\mathbf{x}_0)$ and variance $\sigma^2(\mathbf{x}_0)$ of $y_0(\mathbf{x}_0)$ by

$$\mu(\mathbf{x}_0) = \frac{1}{N}\sum_{i=1}^{N} \tilde{\mu}_{0,i}(\mathbf{x}_0),$$

$$\sigma^2(\mathbf{x}_0) = \frac{1}{N}\sum_{i=1}^{N} \left([\tilde{\mu}_{0,i}(\mathbf{x}_0)]^2 + \tilde{\sigma}^2_{0,i}(\mathbf{x}_0) \right) - \mu(\mathbf{x}_0)^2.$$

Numerical experiment

- Data used: Paediatric ICU admissions (n=14)
- Variables: pCO2, SID (Na^+, Cl^-), lactate (weak acid), pH
- Preprocessing: Hourly discretisation, z-score normalisation, masking to simulate missingness
- Benchmarks:
  - Last observation carried forward (LOCF)
  - MICE
  - GP interpolation

Model evaluation

- Four levels of missingness: 10%, 20%, 30%, 40%
- Two evaluation metrics
  - Mean absolute error – imputation accuracy

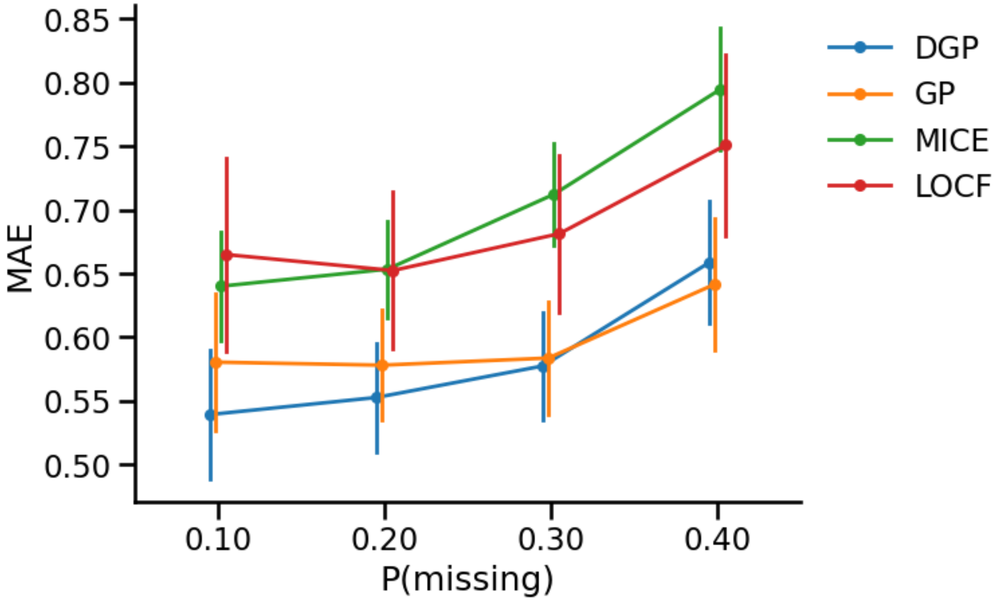
$$\text{MAE} = \frac{1}{N \times D} \sum_{i=1}^{N} \sum_{d \in D} |Y_{id} - \hat{Y}_{id}|$$

  - Negative log likelihood – uncertainty quantification

$$\text{NLL} = -\sum_{i=1}^{N} \log p(Y_i | X_i; \theta)$$

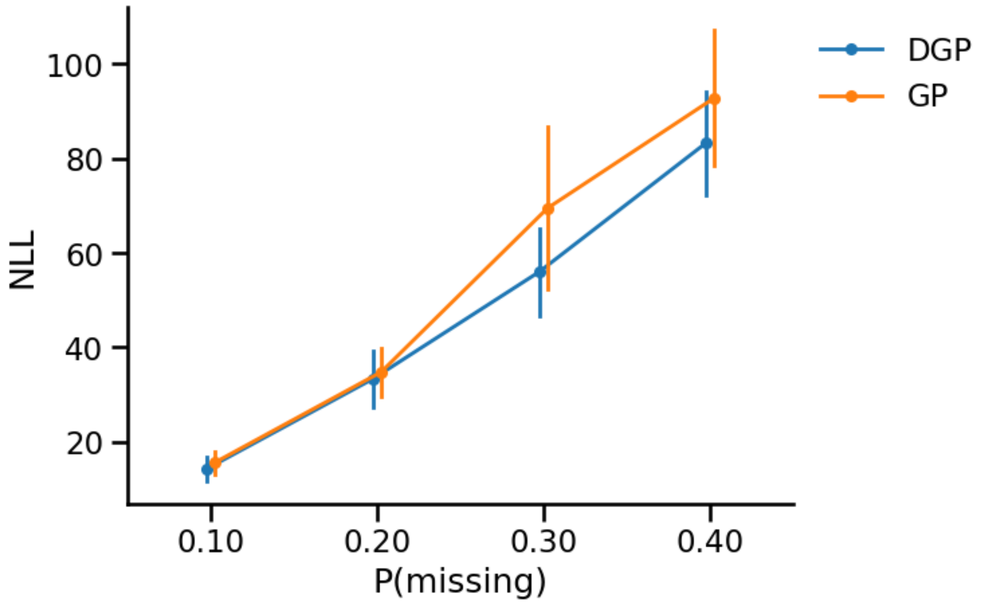
Imputing missing values

- DGP achieved the lowest error rates at 10% to 30% missing values—covering the typical 15–30% missingness in critical care data (Luo et al., 2017)
- As missingness rate increases, longitudinal information is more valuable than cross-sectional information
- DGP combines both → optimal results in lower missingness



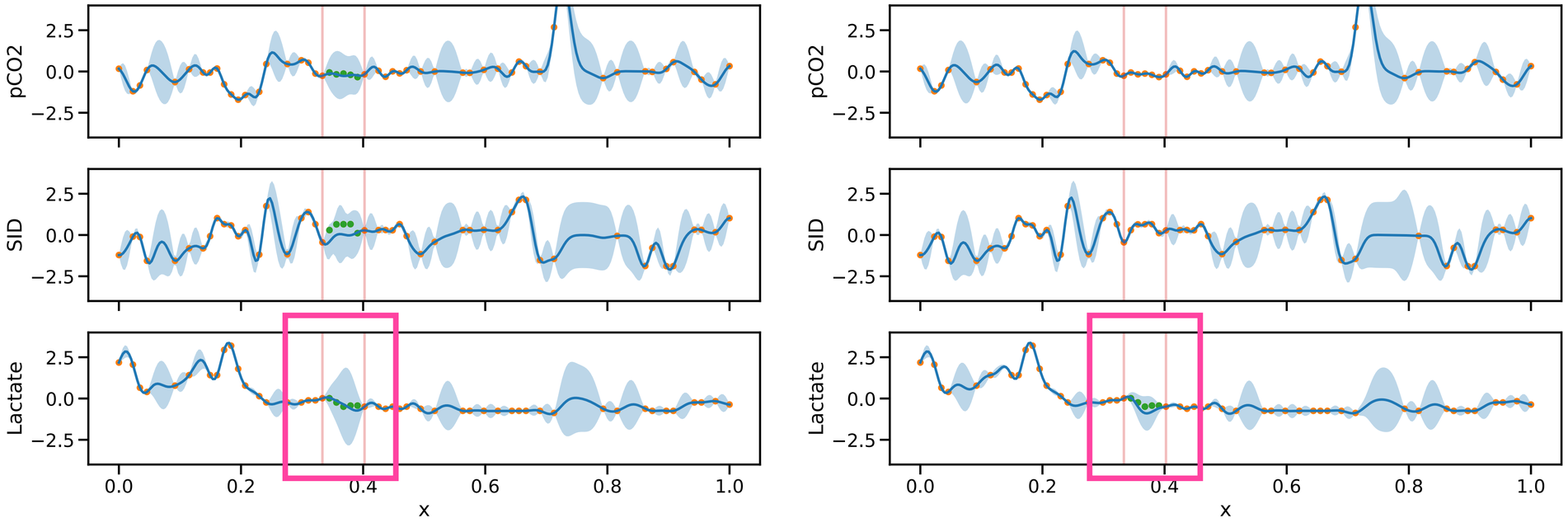
Uncertainty quantification

- DGP performs best when taking into account the uncertainty quantification
- As missingness rate increases, DGP maintains tighter uncertainty bounds than GP
- LOCF was excluded as it does not provide uncertainty quantification



Uncertainty quantification

As the covariates were connected through pH in the output layer using DGP-SI, an observation from one covariate could affect the uncertainty of another covariate where an observation was unavailable.



Implications

- DGP-SI provides reliable, uncertainty-aware imputations to aid clinical decision-making
- Insight into patient status between lab measurements
- Similar problems can be found in human activity recognition from multiple sensors, sleep disorder diagnosis using EEG, and hepatocellular carcinoma (Han et al., 2021)

Limitations & future work

- Computational expense with large datasets → sparse GP (Snelson & Ghahramani, 2007) or GPU parallelisation (Wang et al., 2019)
- Propagated uncertainty which may result in worse performance for predicting pH
- Comparison to deep learning models
- Analysis in higher missingness

Thank you

aliakbars.id/rss25