Additive Gaussian Process Regression on an incomplete grid

Sahoko Ishida

Department of Computer Science

University of Oxford

Wicher Bergsma

Department of Statistics

London School of Economics and Political Science

RSS International Conference 2024, Brighton

GP on multidimensional grid

- Regression model

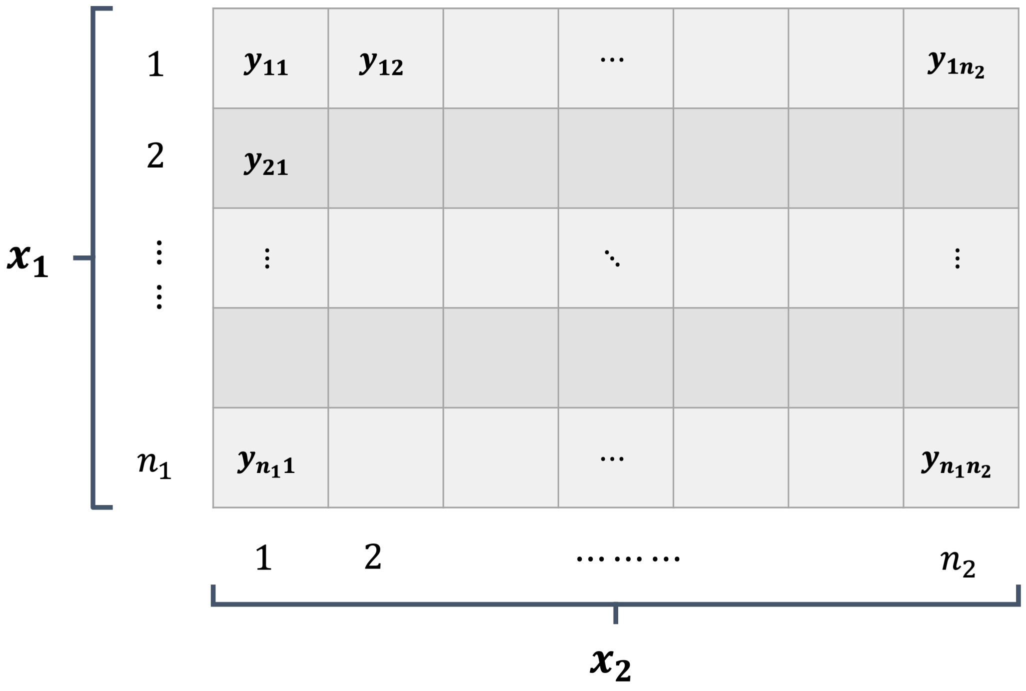
$$y = f(\boldsymbol{x}) + \epsilon,$$

where $f \sim GP(0, k)$ the input $\boldsymbol{x}$ forms Cartesian grid

- Example:
  - Environmental monitoring

$$x = coordinates\ of\ stations \times timestamp$$

$$x = x_1 \times x_2$$



GP on multidimensional grid

- Regression model

$$y = f(\boldsymbol{x}) + \epsilon,$$

where $f \sim GP(0, k)$ the input $\boldsymbol{x}$ forms Cartesian grid

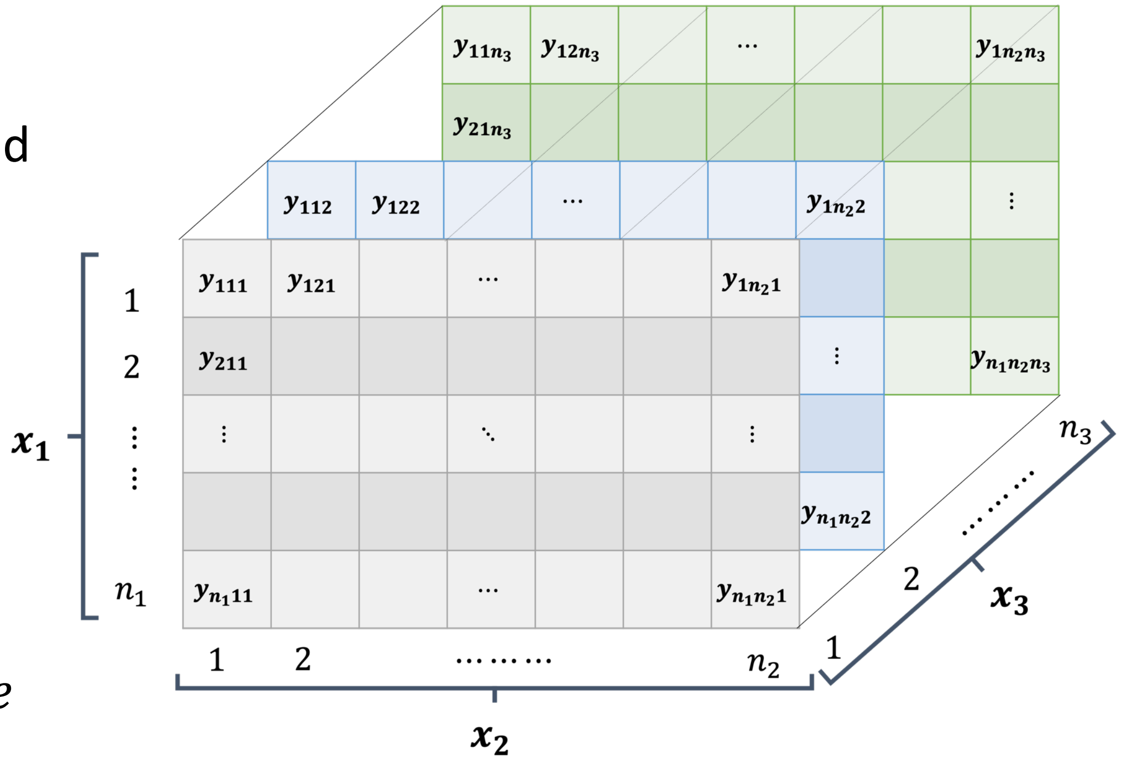
- Example:
  - Environmental monitoring

$x = coordinates\ of\ stations \times timestamp$

  - Brain image

$x = Individual \times location\ in\ the\ brain \times time$

$x = x_1 \times x_2 \times x_3$



GP on non-grid - computation

With 2-dimensional (non-grid) input, for $i = 1, \ldots, n$,

$$y_i = f(x_{si}, x_{ti}) + \epsilon_i$$

Prior: $f \sim GP(0, k)$ with

- $k = k_s \otimes k_t$
- $k = (1 + k_s) \otimes (1 + k_t)$

Alternatively, $\boldsymbol{y} = (y_1 \ldots y_n)^\top \sim MVN(\mathbf{0}, \mathbf{K})$

- $\mathbf{K} = \mathbf{K}_s \circ \mathbf{K}_t$
- $\mathbf{K} = (\boldsymbol{J}_n + \mathbf{K}_s) \circ (\boldsymbol{J}_n + \mathbf{K}_t)$

Note: $\mathbf{K}_s / \mathbf{K}_t$ is a $n \times n$ matrix and $\boldsymbol{J}_n = \mathbf{1}_{\boldsymbol{n}} \mathbf{1}_{\boldsymbol{n}}^\top$

Main bottleneck – the Gram matrix $\mathbf{K}$

1. Inverse of Covariance matrix and its multiplication with a vector $\mathbf{v}$

$$(\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{v}$$

2. Log determinant

$$\log |\mathbf{K} + \sigma^2 \mathbf{I}_n|$$

$O(n^3)$ operations and $O(n^2)$ storage

GP on multidimensional grid - computation

With 2-dimensional **grid**, for $i = 1, \dots, n_1$ and $j = 1, \dots, n_2$

$$y_{ij} = f\left(x_{si}, x_{tj}\right) + \epsilon_{ij}$$

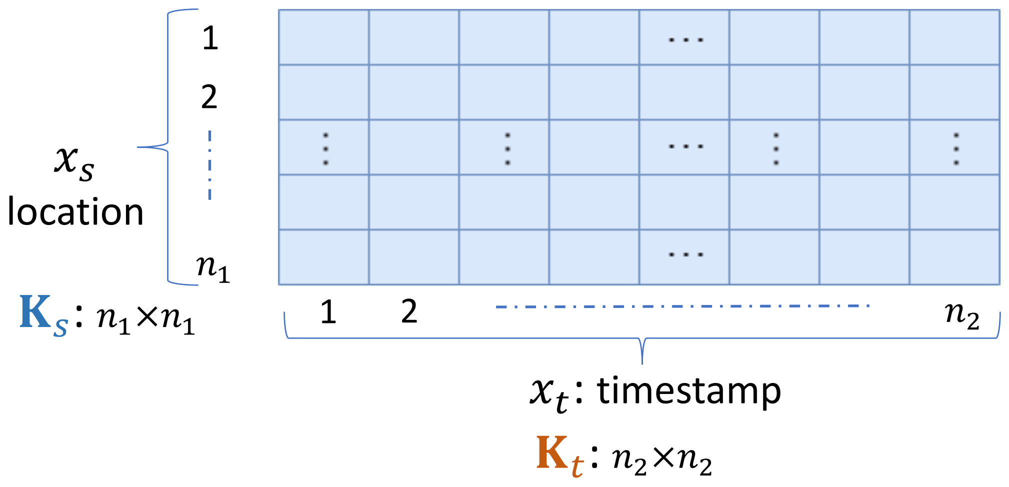
Prior: $f \sim GP(0, k)$ with

- $k = k_s \otimes k_t$
- $k = (1 + k_s) \otimes (1 + k_t)$

Alternatively, $\boldsymbol{y} = (y_1 \dots y_n)^\top \sim MVN(\mathbf{0}, \mathbf{K})$

- $\mathbf{K} = \mathbf{K}_s \otimes \mathbf{K}_t$
- $\mathbf{K} = (\boldsymbol{J}_{n_1} + \mathbf{K}_s) \otimes (\boldsymbol{J}_{n_2} + \mathbf{K}_t)$

Main bottleneck – the Gram matrix $\mathbf{K}$



$$\begin{aligned}\mathbf{K} &= \mathbf{K}_s \otimes \mathbf{K}_t \\ &= \mathbf{Q}_s \boldsymbol{\Lambda}_s \mathbf{Q}_s^\top \otimes \mathbf{Q}_t \boldsymbol{\Lambda}_t \mathbf{Q}_t^\top \\ &= (\mathbf{Q}_s \otimes \mathbf{Q}_t)(\boldsymbol{\Lambda}_s \otimes \boldsymbol{\Lambda}_t)(\mathbf{Q}_s \otimes \mathbf{Q}_t)^\top\end{aligned}$$

GP on multidimensional grid - computation

Main bottleneck – the Gram matrix $\mathbf{K}$

1. Inverse of Covariance matrix and its multiplication with a vector $\mathbf{v}$

$$(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{v} = \underbrace{(\mathbf{Q}_s\otimes\mathbf{Q}_t)(\boldsymbol{\Lambda}_s\otimes\boldsymbol{\Lambda}_t+\sigma^2\mathbf{I}_n)^{-1}\overbrace{(\mathbf{Q}_s\otimes\mathbf{Q}_t)^\top\mathbf{v}}^{\text{Mat-vec multiplication, } O(n(n_1+n_2))}}_{\text{Eigendecomposition, } O(n_1^3+n_2^3)}$$

2. Log determinant

$$\log|\mathbf{K}+\sigma^2\mathbf{I}_n| = \sum_{i,j}\log(\lambda_{si}\lambda_{tj}+\sigma^2)$$

$O(\max(\sum n_l^3, n\sum n_l))$ operations

$O(\sum n_l^2)$ storage

GP on multidimensional grid – additive GP?

3-dimensional case and Functional ANOVA

$$f(x_1, x_2, x_3) = a + \underbrace{f_1(x_1) + f_2(x_2) + f_3(x_3)}_{\text{Main effect}}$$

$$\underbrace{+f_{12}(x_1, x_2) + f_{13}(x_1, x_3) + f_{23}(x_2, x_3)}_{\text{Two-way interaction}}$$

$$\underbrace{+\, f_{123}(x_1, x_2, x_3\,)}_{\text{Three-way interaction}}$$

$$k_{ANOVA}(\mathbf{x}, \mathbf{x}') = \alpha_0 \prod_{l=1}^{3} (1 + \alpha_l k_l(x_l, x_l'))$$

$$\mathbf{K}_{ANOVA} = \alpha_0 \otimes_{l=1}^{3} (J_{n_l} + \alpha_l \mathbf{K}_l)$$

If only main effects or some of the interaction effects are appropriate, we have **a sum of Kronecker product** in the Gram matrix – e.g. main effect only and let $\alpha_0 = 1$

$$\mathbf{K} = J_n + \mathbf{K}_1 \otimes J_{n_2} \otimes J_{n_3} + J_1 \otimes \mathbf{K}_2 \otimes J_{n_3} + J_1 \otimes J_{n_2} \otimes \mathbf{K}_3$$

1. <u>Computational efficiency</u> – does the Kronecker trick still work?
2. <u>Identifiability</u> –the constant term and the functions are not identifiable

GP on multidimensional grid – additive GP?

- E.g. To achieve sum to zero constraint on e.g. f_1 or f_{12} i.e. to achieve $\sum_i f_1(x_{1i}) = 0$ or $\sum_i f_{12}(x_{1i}, x_2) = 0 \ \forall x_2 \in \mathcal{X}_2$, we constraint the kernel k_1
  - Centring:

$$\tilde{k}_1(x_1, x_1') = k(x_1, x_1') - \frac{1}{n}\sum_j k\left(x_1, x_{1j}\right) - \frac{1}{n}\sum_i k(x_{1i}, x'_1) + \frac{1}{n^2}\sum_{ij} k(x_{1i}, x_{1j})$$

  - Lu et al.(2022) – Additive Gaussian Process Revisited

$$\tilde{k}_1(x_1, x_1') = k(x_1, x_1') - \frac{\sum_j k\left(x_1, x_{1j}\right) \sum_i k\left(x_{1i}, x'_1\right)}{\sum_{ij} k(x_{1i}, x_{1j})}$$

GP on multidimensional grid – additive GP?

The corresponding Gram matrix:

- $\widetilde{\mathbf{K}_1} = (\mathbb{I} - \frac{1}{\mathrm{n}_1} J_{n_1})\mathbf{K}_1(\mathbb{I} - \frac{1}{\mathrm{n}_1} J_{n_1})$
- $\widetilde{\mathbf{K}_1} = \mathbf{K}_1 - \frac{\mathbf{K}_1 \mathbf{1}\mathbf{1}^\top \mathbf{K}_1}{\mathbf{1}^\top \mathbf{K}_1 \mathbf{1}}$

1. For both cases, at least one eigenvalue is zero
2. Eigenvectors corresponding to non-zero eigenvalues are all orthogonal to $\mathbf{1}$
3. Given $\widetilde{\mathbf{K}_1} = Q_1 \Lambda_1 Q_1^\top$, the matrix J_{n_1} can be decomposed using the same orthonormal matrix Q_1

$$J_{n_1} = Q_1 \mathrm{A}_1 Q_1^\top$$

GP on multidimensional grid – additive GP?

$$\mathbf{K} = \otimes_{l=1}^{3} J_{n_l} + \overbrace{\mathbf{K}_1}^{Q_1\Lambda_1 Q_1^\top} \otimes \overbrace{J_{n_2}}^{Q_2\Lambda_2 Q_2^\top} \otimes J_{n_3} + \overbrace{J_1}^{Q_1\Lambda_1 Q_1^\top} \otimes \overbrace{\mathbf{K}_2}^{Q_2\Lambda_2 Q_2^\top} \otimes J_{n_3} + J_1 \otimes J_{n_2} \otimes \mathbf{K}_3$$

$$= \otimes_{l=1}^{3} Q_l \Lambda_l Q_l^\top + Q_1\Lambda_1 Q_1^\top \otimes Q_2\Lambda_2 Q_2^\top \otimes Q_3 A_3 Q_3^\top$$

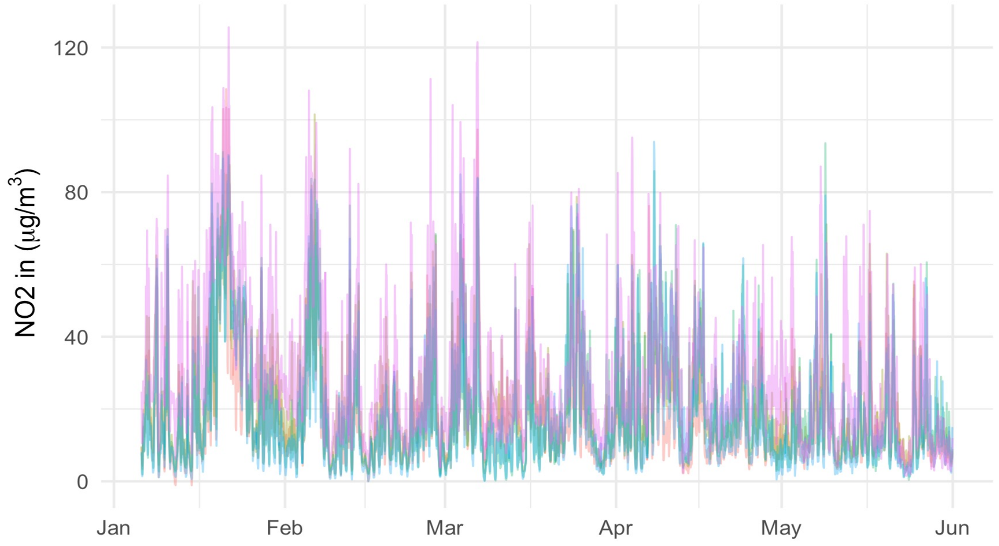
$$+ Q_1 \Lambda_1 Q_1^\top \otimes Q_2 \Lambda_2 Q_2^\top \otimes Q_3 A_3 Q_3^\top + \cdots$$

$$= (\otimes_{l=1}^{3} Q_l)\underbrace{(\otimes_{l=1}^{3} A_l + \Lambda_1 \otimes A_2 \otimes A_3 + \cdots)}_{\text{Diagonal}}(\otimes_{l=1}^{3} Q_l^\top)$$

Application

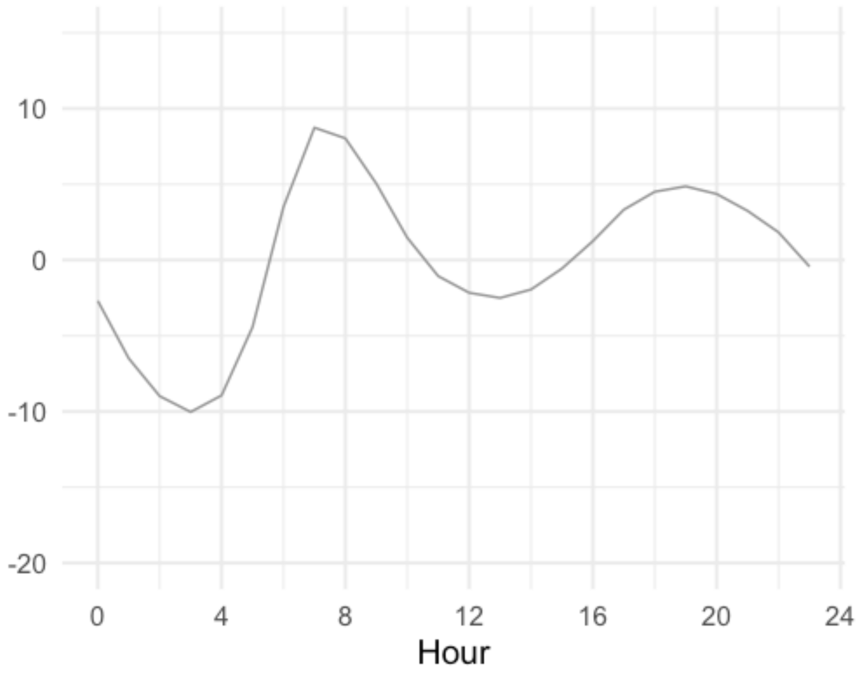
Example with NO_2 in London

- 59 Monitoring stations, x_1: coordinates
- 147days in early/mid 2020, x_2: days
- Hourly measured, x_3: hour of the day
- Total number of observation > 200,000



https://www.londonair.org.uk/

- MCMC (HMC, Stan) takes 10-15 minutes
- Maximum marginal likelihood estimation of scale parameters - convergence in a few seconds

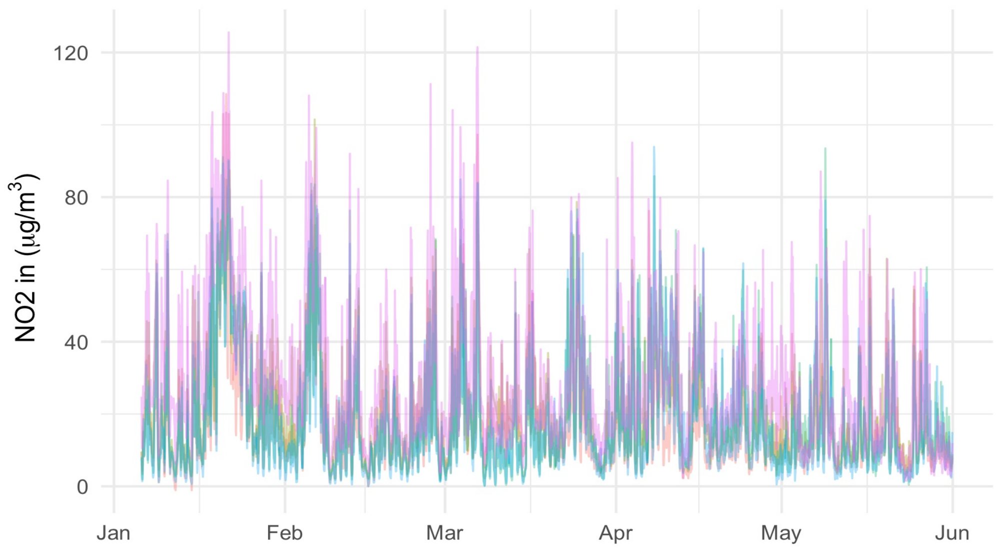


the daily NO2 pattern

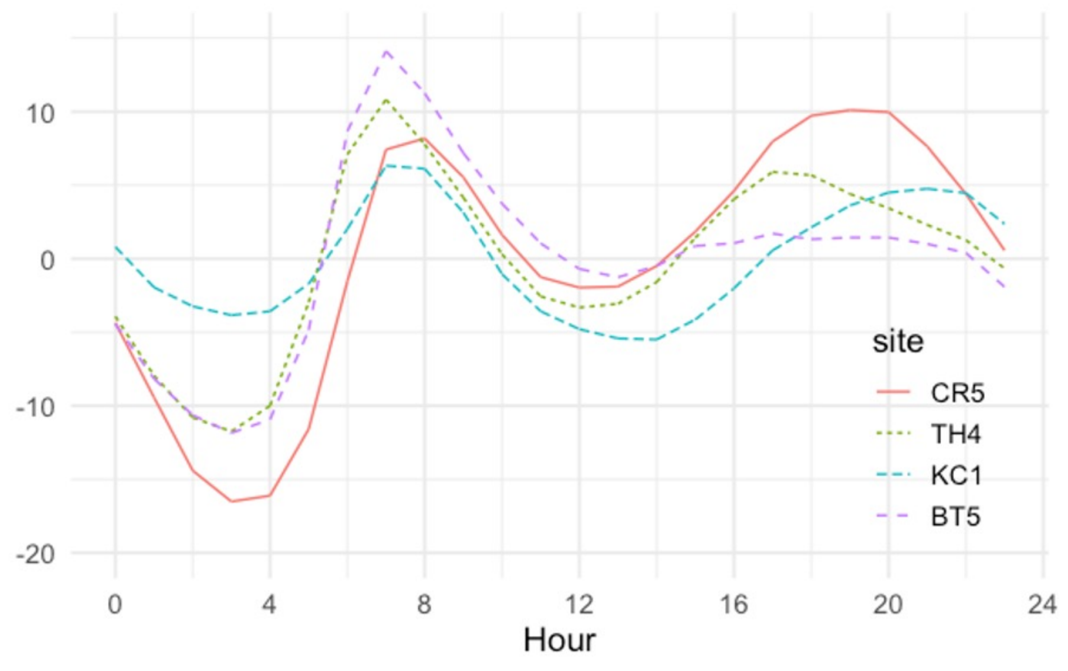
Application

Example with NO_2 in London

- 59 Monitoring stations, x_1: coordinates
- 147days in early/mid 2020, x_2: days
- Hourly measured, x_3: hour of the day
- Total number of observation > 200,000



- MCMC (HMC, Stan) takes 10-15 minutes
- Maximum marginal likelihood estimation of scale parameters - convergence in a few seconds

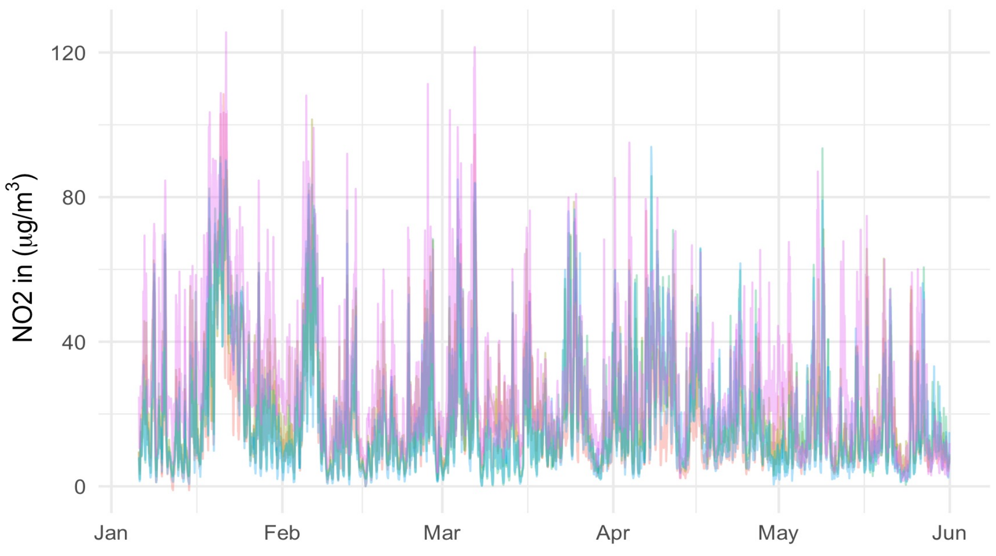


the daily NO2 pattern

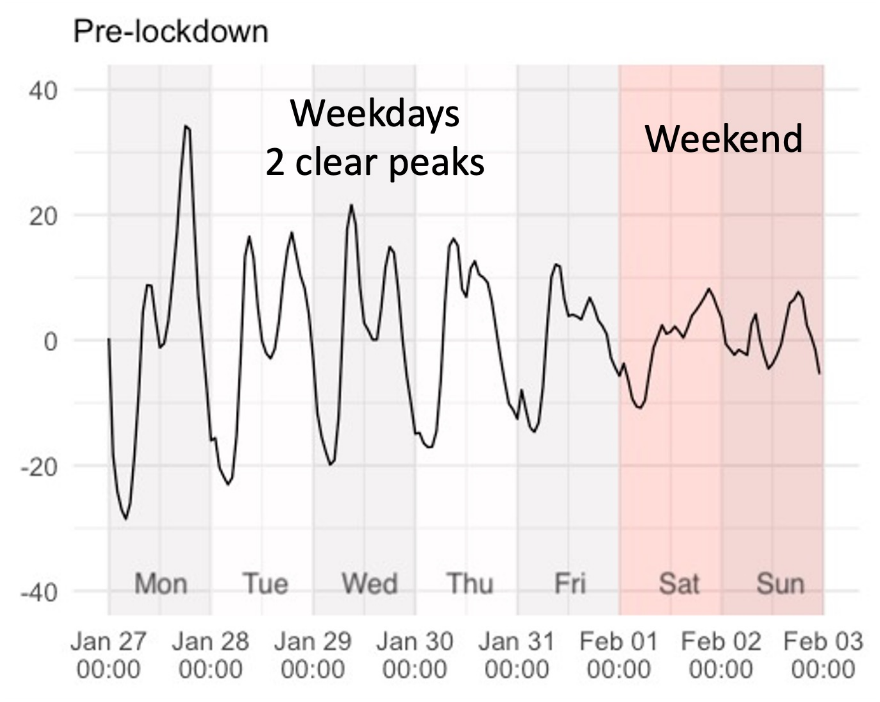
Application

Example with NO_2 in London

- 59 Monitoring stations, x_1: coordinates
- 147days in early/mid 2020, x_2: days
- Hourly measured, x_3: hour of the day
- Total number of observation > 200,000



- MCMC (HMC, Stan) takes 10-15 minutes
- Maximum marginal likelihood estimation of scale parameters - convergence in a few seconds

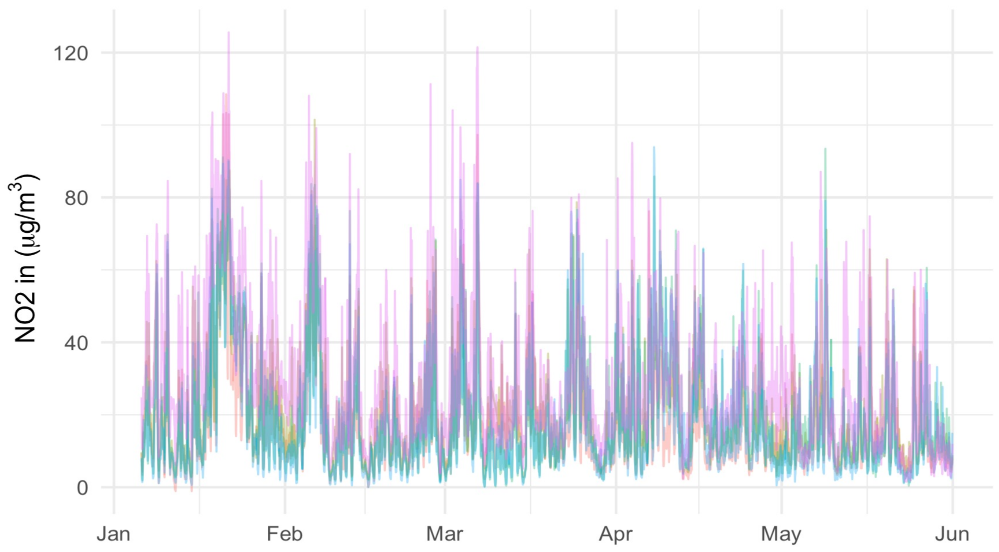


the daily NO2 pattern

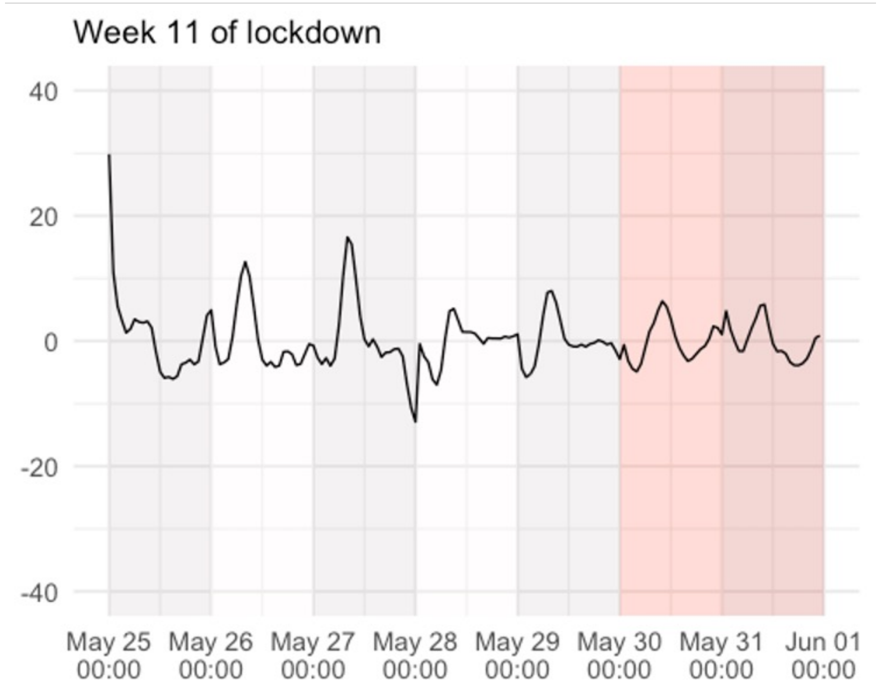
Application

Example with NO_2 in London

- 59 Monitoring stations, x_1: coordinates
- 147days in early/mid 2020, x_2: days
- Hourly measured, x_3: hour of the day
- Total number of observation > 200,000

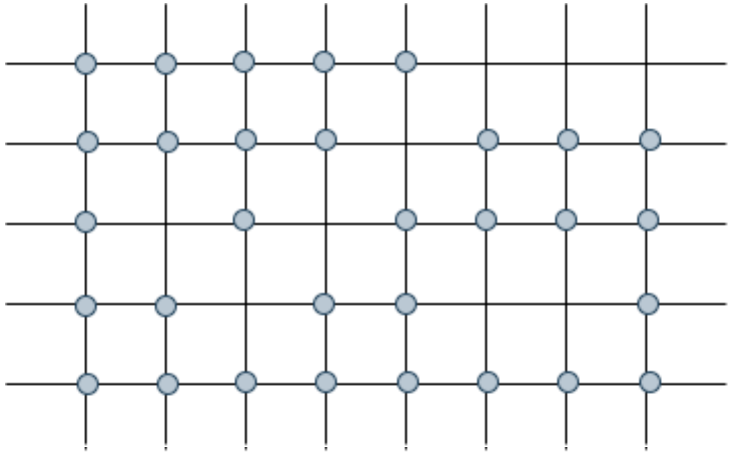


- MCMC (HMC, Stan) takes 10-15 minutes
- Maximum marginal likelihood estimation of scale parameters - convergence in a few seconds



the daily NO2 pattern

GP on incomplete multidimensional grid

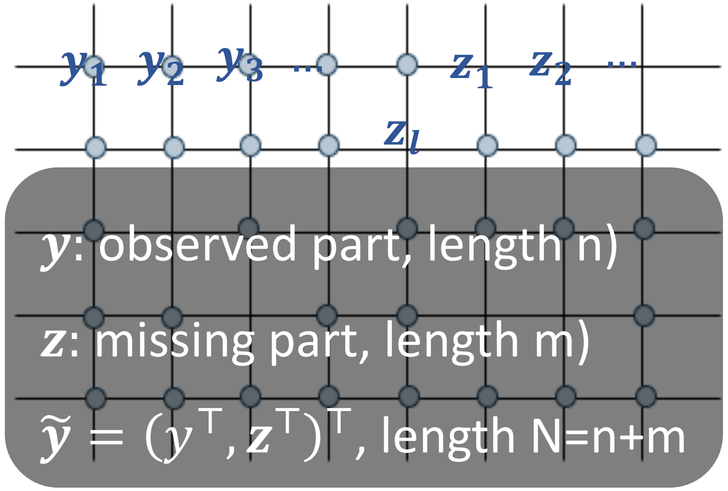


Is Missingness mechanism MCAR/MAR?

→ Yes

Complete case analysis

GP on incomplete multidimensional grid



Complete case analysis

$\mathbf{K}_{NN}$: Evaluated at complete grid

$\mathbf{K}_{nn}$: Evaluated at input corresponding to the observed part $\boldsymbol{y}$

No Kronecker product structure

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = -\frac{1}{2}(\mathbf{y}^\top(\mathbf{K}_{\mathbf{nn}} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y} + \log|\mathbf{K}_{nn} + \sigma^2\mathbf{I}_n|) + c$$

Approximation to complete case analysis

- Gilboa et al.(2013) / Flaxman et al. (2015)

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}) = -\frac{1}{2}\big(\underbrace{\mathbf{y}^\top(\mathbf{K_{nn}} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y}}_{\text{Term 1}} + \underbrace{\log|\mathbf{K}_{\boldsymbol{nn}} + \sigma^2\mathbf{I}_n|}_{\text{Term 2}}\big)$$

$\boldsymbol{y}$: observed part, length n)

$\boldsymbol{z}$: missing part, length m)

$\tilde{\boldsymbol{y}} = (y^\top, \boldsymbol{z}^\top)^\top$, length N=n+m

- Term 1: fill $\boldsymbol{z}$ with "imaginary" observations and

$$\tilde{\boldsymbol{y}}^\top(\mathbf{K_{NN}} + \mathbf{D})^{-1}\tilde{\boldsymbol{y}} \;\rightarrow\; \text{term 1}, \qquad \text{as} \quad \mathrm{w} \rightarrow 0$$

where

$$\mathbf{D} = \begin{bmatrix} \sigma^2\mathbf{I_n} & \mathbf{0}_{\boldsymbol{nm}} \\ \mathbf{0}_{\boldsymbol{mn}} & w^{-\mathbf{1}}\,\mathbf{I}_{\boldsymbol{m}} \end{bmatrix}$$

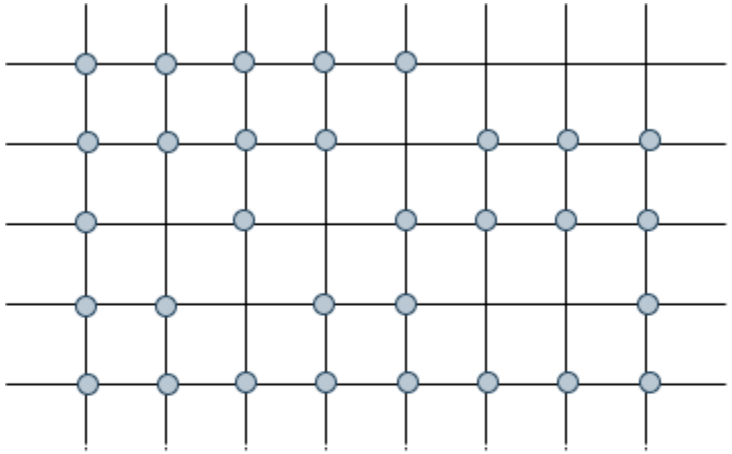
Can be computed using Conjugate Gradient decent algorithm and with $O(JN\sum n_l)$

- Term 2

$$\log|\mathbf{K}_{\boldsymbol{nn}} + \sigma^2\mathbf{I}_n| \approx \sum_{i=1}^{n} \log\big(\tilde{\lambda}_i + \sigma^2\big)$$

where $\tilde{\lambda}_i = \frac{n}{N}\lambda_i^{(N)}$ and $\lambda_1^{(N)}, \ldots, \lambda_n^{(N)}$ are the n largest eigenvalues of $\mathbf{K_{NN}}$

GP on incomplete multidimensional grid



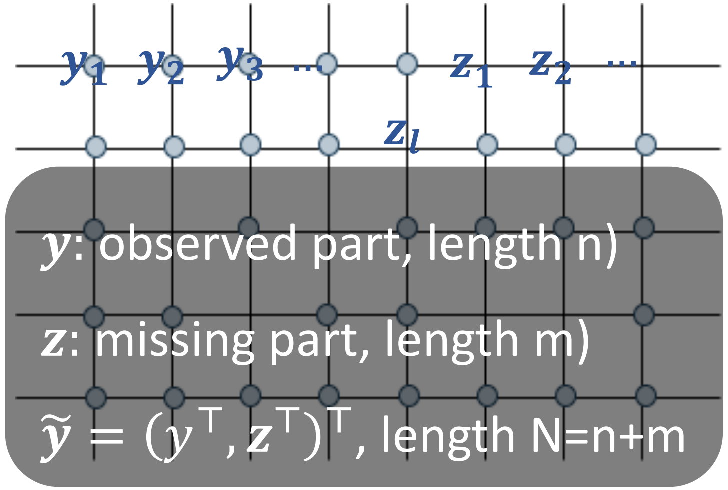
Is Missingness mechanism MCAR/MAR?

$\longrightarrow$ No

- Monitoring devices are more likely to fail at higher(lower) end, or some values may be censored
- Repeated measurement of mental health status – no entries when symptoms are worse

- Possible to model probability of missing/observed and incorporate it when fitting model using the complete case analysis approximation
- Some occasions, partial knowledge on missing part z is available (cut-off, interval)

Stochastic EM algorithm



<u>Stochastic EM with Gibbs Sampling</u>

At each step, sample from $z_1 | \boldsymbol{y}, \boldsymbol{z}_{-1}^{t-1}, \theta^{t-1}$…

Univariate problem

- easy to take into accont missingness mechanism
- Mean and variance can be computed by rank 2 update of Gram matrix $\mathbf{K}_{NN}$

EM algorithm

$$Q(\theta|\theta^{t-1}) = \int \log p(\tilde{\boldsymbol{y}}|\theta) p(\boldsymbol{z}|\boldsymbol{y}, \theta^{t-1}) d\boldsymbol{z}$$

- Directly evaluating $Q(\theta|\theta^{t-1})$ is costly
- MCEM / stochastic (Approximation) EM possible, but sampling from $p(\boldsymbol{z}|\boldsymbol{y}, \theta^{t-1})$ faces multiple challenges, especially with some constraints reflecting missingness mechanism

Sampling from $p(\boldsymbol{z}|\boldsymbol{y}, \theta^{t-1})$

$\boldsymbol{y}$: observed part, length n)

$\boldsymbol{z}$: missing part, length m)

$\widetilde{\boldsymbol{y}} = (y^\top, \boldsymbol{z}^\top)^\top$, length N=n+m

Conditional distribution $(\boldsymbol{z}|\boldsymbol{y}, \theta^{t-1})$

$$p(\boldsymbol{z}|\boldsymbol{y}, \theta^{t-1}) = MVN\big(\mu(\theta^{t-1}), \Sigma(\theta^{t-1})\big)$$

where

$$\mu(\theta^{t-1}) = \mathbf{K}_{mn}(\mathbf{K}_{\mathbf{nn}} + \sigma^2\mathbf{I}_n)^{-1}\boldsymbol{y}$$

$$\Sigma(\theta^{t-1}) = \mathbf{K}_{mm} - \mathbf{K}_{mn}(\mathbf{K}_{\mathbf{nn}} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{K}_{nm}$$

- Both $(\mathbf{K}_{\mathbf{nn}} + \sigma^2\mathbf{I}_n)^{-1}\boldsymbol{y}$ and $(\mathbf{K}_{\mathbf{nn}} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{K}_{nm}$ can be computed using CG decent
- Takes takes $O(\frac{m(m+1)}{2}JN\sum n_l)$ to compute $m\times m$ covariance matrix

Sampling from $p(\boldsymbol{z}|\boldsymbol{y}, \theta^{t-1})$- Gibbs sampling

$\boldsymbol{y}$: observed part, length n)

$\boldsymbol{z}$: missing part, length m)

$\tilde{\boldsymbol{y}} = (\boldsymbol{y}^\top, \boldsymbol{z}^\top)^\top$, length N=n+m

At t-th iteration,

- Sample from $z_1^{(t)}|\boldsymbol{y}, \boldsymbol{z}_{-1}^{(t-1)}, \theta^{t-1} \sim N\left(\mu_1^{(t)}, v_1^{(t)}\right)$ where

$$\mu_1^{(t)} = \boldsymbol{\alpha}\left(x_1^{(ms)}\right)^\top \boldsymbol{y}$$

$$v_1^{(t)} = k(x_1^{(ms)}, x_1^{(ms)}) - \boldsymbol{\alpha}\left(x_1^{(ms)}\right)^\top \mathbf{k}_{N-x_1^{(ms)}}(x_1^{(ms)})$$

And $\alpha\left(x_1^{(ms)}\right) = \left(\mathbf{K}_{N-x_1^{(ms)}, N-x_1^{(ms)}} + \sigma^2 \mathbf{I}_{n-1}\right)^{-1} \mathbf{k}_{N-x_1^{(ms)}}(x_1^{(ms)})$

- $\alpha\left(x_1^{(ms)}\right)$ can be computed by rank 2 update of $\left(\mathbf{K}_{N,N} + \sigma^2 \mathbf{I}_N\right)^{-1} \mathbf{k}_N(x_1^{(ms)})$

- Sample from $z_2^{(t)}|\boldsymbol{y}, z_1^{(t)}, \boldsymbol{z}_{-(1,2)}^{(t-1)}, \theta^{t-1} \sim N\left(\mu_2^{(t)}, v_2^{(t)}\right)$

- …

Stochastic EM with Gibbs sampling

$n = \prod n_l$: l ength of the observed $\boldsymbol{y}$

m : length of the missing $\boldsymbol{z}$

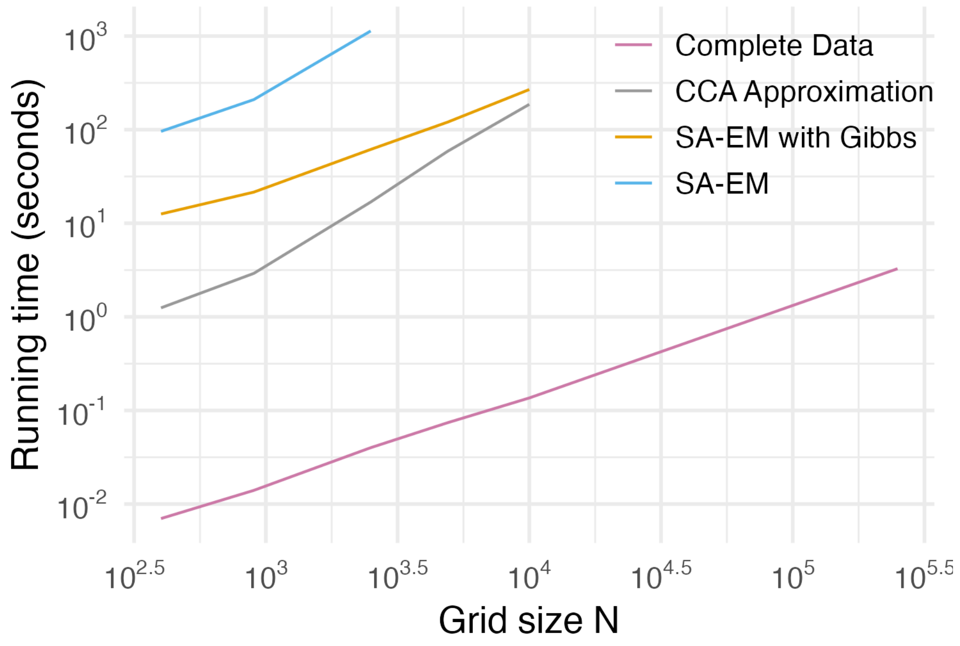
$N = n + m$: length of $\widetilde{\boldsymbol{y}} = (y^\top, \boldsymbol{z}^\top)^\top$,

B: # of iterations for CG decent

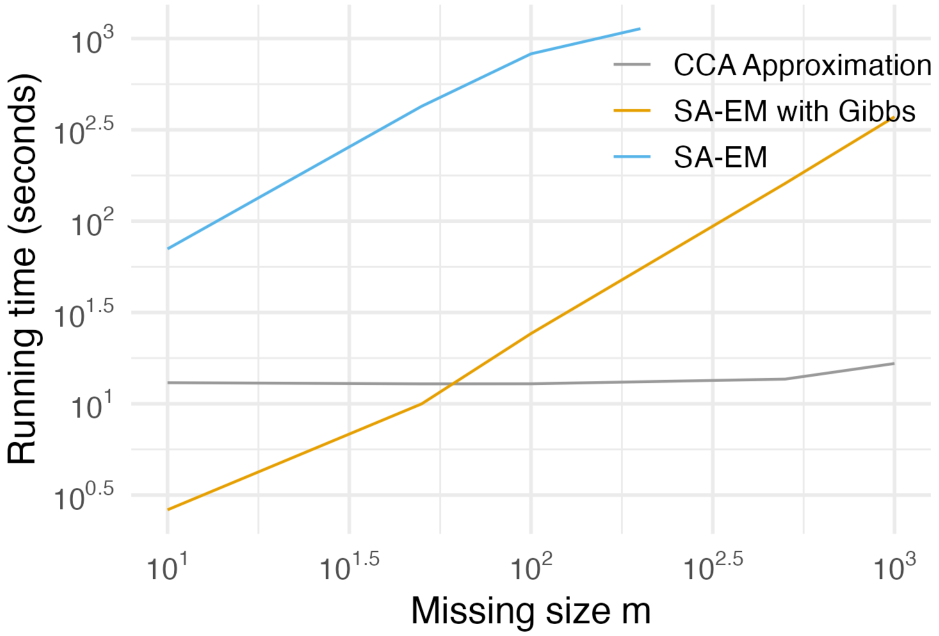
Merits

- Efficiency: $mN\sum n_l$ operations instead of
  - EM without Gibbs: $\frac{m(m+1)}{2}BN\sum n_l + O(m^3)$
  - Complete case analysis approximation: $BN\sum n_l$
- Incorporating some missingness mechanism e.g., $z > c$ for some constant c can be ensured in the sampling step.

Simulation – Computation time



$m = 200$



$N = 50{\times}50$

Summary and future work

	Complete case analysis	Missing value imputation with EM + Gibbs
+	Model fitting and posterior mean fast to compute (appoximation using conjugate gradient decent available)	Wider missing not at random scenarios can be handled
-	• Missingness mechanism that can be incorporated is limited, could lead to bias • Sampling from posterior scales badly with m	Scalability for large N and m still under investigation

- More realistic missingness mechanism and application to real world data
- Modifying the stochastic EM algorithm
- Gibbs + HMC (MH) for full MCMC, similar to De Oliveira(2005) but on the grid

Reference

De Oliveira, Victor. "Bayesian inference and prediction of Gaussian random fields based on censored data." *Journal of Computational and Graphical Statistics* 14, no. 1 (2005): 95-115.

Flaxman, Seth, Andrew Wilson, Daniel Neill, Hannes Nickisch, and Alex Smola. "Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods." In *International conference on machine learning*, pp. 607-616. PMLR, 2015.

Gilboa, Elad, Yunus Saatçi, and John P. Cunningham. "Scaling multidimensional inference for structured Gaussian processes." *IEEE transactions on pattern analysis and machine intelligence* 37, no. 2 (2013): 424-436.

Ishida, Sahoko, and Wicher Bergsma. "Efficient and Interpretable Additive Gaussian Process Regression and Application to Analysis of Hourly-recorded NO2 Concentrations in London." *arXiv preprint arXiv:2305.07073* (2023).

Lu, Xiaoyu, Alexis Boukouvalas, and James Hensman. "Additive gaussian processes revisited." In *International Conference on Machine Learning*, pp. 14358-14383. PMLR, 2022.