Additive Gaussian process models for multi-dimensional grid structured data

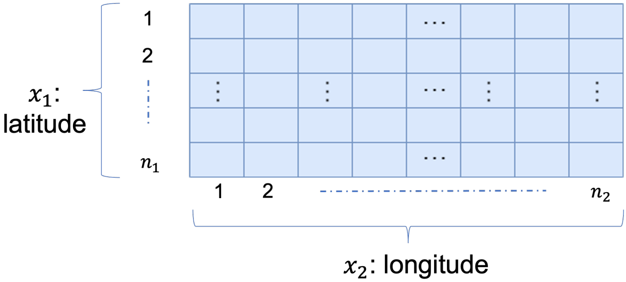
Sahoko Ishida, Wicher Bergsma

Department of Statistics
London School of Economics

19 July 2023

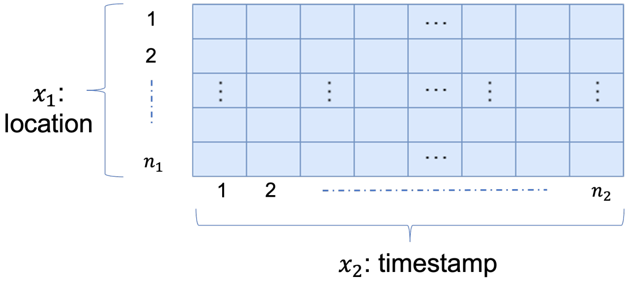
Multi-dimensional grid/panel data

Inputs are on Cartesian grid, e.g.,



Multi-dimensional grid/panel data

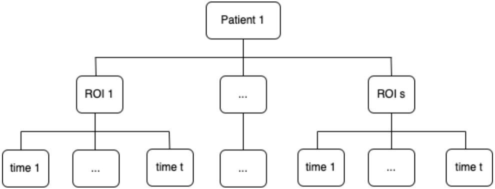
Inputs are on Cartesian grid, e.g.,



And at each grid, we have an observation such as temperature, air-quality levels etc.

Multi-dimensional grid/panel data

Three-dimension example: brain imaging



- Flexible statistical modelling e.g. incorporating spatial and time dependence and its interaction
- Computational efficiency as the number of observations tends to be large

Additive GP models

- For $i = 1, ..., n$, consider a regression model for a response $y_i \in \mathbb{R}$ and two predictors $x_{1i} \in \mathcal{X}_1$ and $x_{2i} \in \mathcal{X}_2$:

$$y_i = f(x_{1i}, x_{2i}) + \epsilon_i$$

  with iid error $\epsilon_i \sim N(0, \sigma^2)$.
- Two model to consider
  - Main effect model

$$f(x_{1i}, x_{2i}) = a + f_1(x_{1i}) + f_2(x_{2i})$$

  - Interaction effect model

$$f(x_{1i}, x_{2i}) = a + f_1(x_{1i}) + f_2(x_{2i}) + f_{12}(x_{1i}, x_{2i})$$

    where a is constant

Statistical modelling through kernels

- Prior for each term given $k_1 : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}$ and $k_1 : \mathcal{X}_2 \times \mathcal{X}_2 \to \mathbb{R}$.

$$a \sim N(0,1), \quad f_1 \sim GP(0,k_1), \quad f_2 \sim GP(0,k_2),$$
$$f_{12} \sim GP(0, k_1 \otimes k_2)$$

- Prior over f: $f \sim GP(0,k)$ where k is defined on input space $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ and given by $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
  - Main effect model

$$k(x,x') = 1 + k_1(x_1,x_1') + k_2(x_2,x_2')$$

  - Interaction effect model

$$k(x,x') = 1 + k_1(x_1,x_1') + k_2(x_2,x_2') + k_1(x_1,x_1')k_2(x_2,x_2')$$

  where $x = (x_1,x_2)^\top \in \mathcal{X}$

Statistical modelling through kernels

Alternatively,

$$\mathbf{f} = (f(x_1), \dots, f(x_n))^\top \sim \mathbf{MVN}(\mathbf{0}, \mathbf{K})$$

where

- Main: $\mathbf{K} = \mathbf{1}_n\mathbf{1}_n^\top + \mathbf{K}_1 + \mathbf{K}_2$
- Interaction: $\mathbf{K} = \mathbf{1}_n\mathbf{1}_n^\top + \mathbf{K}_1 + \mathbf{K}_2 + \mathbf{K}_1 \circ \mathbf{K}_2$

ANOVA decomposition kernel

- With 2 variables, the interaction model is the saturated model with *saturated ANOVA decomposition kernel*

$$k(x, x') = \left(1 + k_1(x_1, x_1')\right)\left(1 + k_2(x_2, x_2')\right)$$

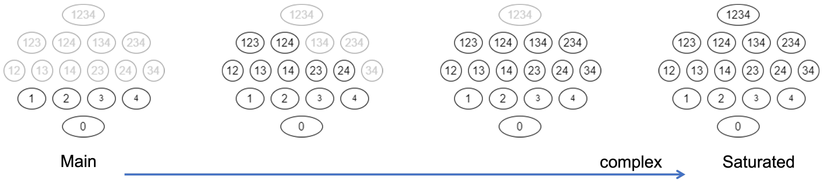
[Wahba, 1990, Gu, 2002] for RKHS and [Stitson et al., 1999] for SVM

- With d variables $x = (x_1, \ldots, x_d)^\top$

$$k(x, x') = \prod_{l=1}^{d} \left(1 + k_l(x_l, x_l')\right)$$

Includes 2^d terms : constant term 1, main terms, all interaction terms

Hierarchical ANOVA decomposition kernel



1. Interaction terms – tensor product kernel
2. Interactions included with any main + lower-order interaction terms

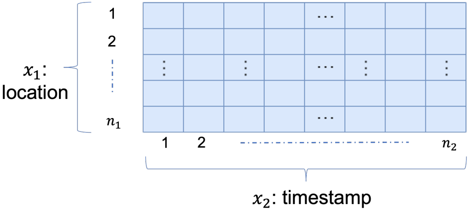
Main constraints

$O(n^3)$ time complexity and $O(n^2)$ memory requirement associated with

1. Inverse of Covariance matrix and its multiplication with a vector $\mathbf{v}$
$$\left(\mathbf{K} + \sigma^2 \mathbf{I}_n\right)^{-1} \mathbf{v}$$
2. Log determinant
$$\log \left|\mathbf{K} + \sigma^2 \mathbf{I}_n\right|$$

Kronecker products in Covariance matrix

When we have multi-dimensional grid data, Kronecker product structure in $\mathbf{K}$ enables efficient evaluation of the above.



- Interaction effect model (saturated):

$$\mathbf{K} = (\mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top + \mathbf{K}_1) \otimes (\mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top + \mathbf{K}_2)$$

- Main effect model:

$$\mathbf{K} = \mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top \otimes \mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top + \mathbf{K}_1 \otimes \mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top + \mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top \otimes \mathbf{K}_2$$

Kronecker products in Covariance matrix

- Existing literature on Kronecker approach in GP handles a limited number of models (separable kernel) including
  - a saturated model
  - a model with only the highest interaction
- Our contribution: flexible with any hierarchical ANOVA kernel

Efficient implementation using Kronecker products

Main goal: Decomposition of Gram matrix

$$\mathbf{K} = (\mathbf{Q}_1 \otimes \mathbf{Q}_2)\mathbf{D}(\mathbf{Q}_1 \otimes \mathbf{Q}_2)^\top$$

where $\mathbf{Q}_l$ is orthonormal, and $\mathbf{D}$ is diagonal with all non-negative diagonal elements

1. $$\left(\mathbf{K} + \sigma^2 \mathbf{I}_n\right)^{-1} \mathbf{v} = (\mathbf{Q}_1 \otimes \mathbf{Q}_2)(\mathbf{D} + \sigma^2 \mathbf{I})^{-1}(\mathbf{Q}_1 \otimes \mathbf{Q}_2)^\top \mathbf{v}$$

2. $$\log |\mathbf{K} + \sigma^2 \mathbf{I}_n| = \sum_i \log \mathbf{D}_{ii} + \sigma^2$$

Time complexity: $O(\sum n_l^3)$ or $O(n \sum n_l)$, memory: $O(\sum n_l^2)$

Eigendecomposition of $\mathbf{K}$

Separable kernel

$$\begin{aligned}\mathbf{K} &= \tilde{\mathbf{K}}_1 \otimes \tilde{\mathbf{K}}_2 \\ &= (\mathbf{Q}_1 \boldsymbol{\Lambda}_1 \mathbf{Q}_1^\top) \otimes (\mathbf{Q}_2 \boldsymbol{\Lambda}_2 \mathbf{Q}_2^\top) \\ &= (\mathbf{Q}_1 \otimes \mathbf{Q}_2)(\boldsymbol{\Lambda}_1 \otimes \boldsymbol{\Lambda}_2)(\mathbf{Q}_1 \otimes \mathbf{Q}_2)^\top\end{aligned}$$

e.g.$\tilde{\mathbf{K}}_l = \mathbf{1}_{n_l}\mathbf{1}_{n_l}^\top + \mathbf{K}_l$

Eigendecomposition of $\mathbf{K}$

Non-separable kernel such as

$$\mathbf{K} = \mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top \otimes \mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top + \mathbf{K}_1 \otimes \mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top + \mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top \otimes \mathbf{K}_2$$

Each term consists of Kronecker product of $\mathbf{1}_{n_l}\mathbf{1}_{n_l}^\top$ and $\mathbf{K}_l$.

Eigendecomposition of $\mathbf{K}$

If each $\mathbf{K}_l$ is centered using centering matrix $\mathbf{C} = \mathbf{I}_{n_l} - \frac{1}{n_l}\mathbf{1}_{n_l}\mathbf{1}_{n_l}^\top$

- it has at least 1 zero eigenvalues, and
- all eigenvectors corresponding to non-zero (and positive) eigenvalues are orthogonal to $\mathbf{1}_{n_l}$

Eigendecomposition

- $\mathbf{K}_l = \mathbf{Q}_l \boldsymbol{\Lambda}_l \mathbf{Q}_l^\top$ with

$$\boldsymbol{\Lambda}_l = \mathrm{diag}(0, \lambda_2, \ldots, \lambda_{n_l})$$

$$\mathbf{Q}_l = \begin{bmatrix} \frac{1}{\sqrt{n_l}}\mathbf{1}_{n_l} & \mathbf{q}_2 & \ldots & \mathbf{q}_{n_l} \end{bmatrix}$$

- $\mathbf{1}_{n_l}\mathbf{1}_{n_l}^\top = \mathbf{Q}_l \mathbf{A}_l \mathbf{Q}_l^\top$ with

$$\mathbf{A}_l = \mathrm{diag}(n_l, 0, \ldots, 0)$$

Eigendecomposition of $\mathbf{K}$

For centered $\mathbf{K}_1$ and $\mathbf{K}_2$,

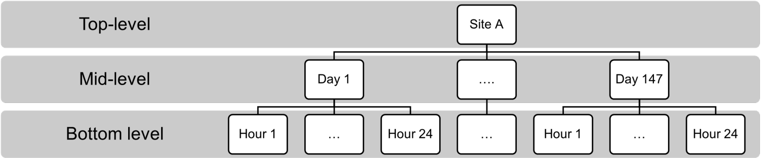
$$\begin{aligned}\mathbf{K} &= \overbrace{\mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top}^{\mathbf{Q}_1\mathbf{A}_1\mathbf{Q}_1^\top} \otimes \overbrace{\mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top}^{\mathbf{Q}_2\mathbf{A}_2\mathbf{Q}_2^\top} + \overbrace{\mathbf{K}_1}^{\mathbf{Q}_1\boldsymbol{\Lambda}_1\mathbf{Q}_1^\top} \otimes \mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top + \mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top \otimes \overbrace{\mathbf{K}_2}^{\mathbf{Q}_2\boldsymbol{\Lambda}_2\mathbf{Q}_2^\top} \\ &= (\mathbf{Q}_1 \otimes \mathbf{Q}_2)\underbrace{(\mathbf{A}_1 \otimes \mathbf{A}_2 + \boldsymbol{\Lambda}_1 \otimes \mathbf{A}_2 + \mathbf{A}_1 \otimes \boldsymbol{\Lambda}_2)}_{\text{diagonal}}(\mathbf{Q}_1 \otimes \mathbf{Q}_2)^\top\end{aligned}$$

Centring also has advantage in terms of identifiability and interpretability

Application to hourly-recorded air-quality monitoring data

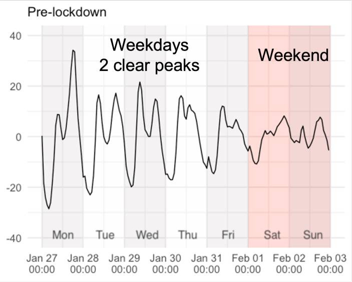
[Ishida and Bergsma, 2023]

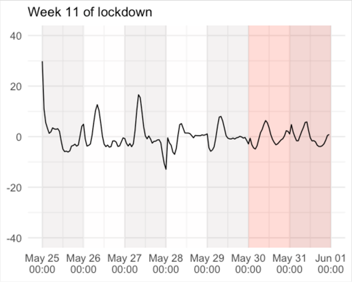
- NO_2 concentrations in London during from January 2020 to May 2020 (for a period of 147 days covering the first lockdown) collected from 59 monitoring stations
- Sample size $> 200,000$
- 3 dimensional grid structure



Application to hourly-recorded air-quality monitoring data

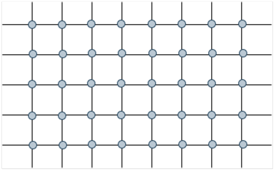
- Saturated model with three-way interaction effect was the best fit
- Under 20 minutes for MCMC sampling (Stan, 200+400 samples)
- A few seconds for marginal likelihood optimisation

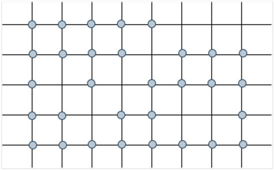




Extensions

- Incomplete grid





  - Possible to handle with MC-EM algorithm with Gibbs sampling
  - Iterative algorithm for missing value imputation (grid completion) and hyper-parameter estimation
- Additive Kronecker products naturally extends to models for multivariate response

References

Gu, C. (2002).
Smoothing spline ANOVA models, volume 297.
Springer.

Ishida, S. and Bergsma, W. (2023).
Efficient and interpretable additive gaussian process regression and application to analysis of hourly-recorded NO_2 concentrations in london.
arXiv preprint arXiv:2305.07073.

Stitson, M., Gammerman, A., Vapnik, V., Vovk, V., Watkins, C., and Weston, J. (1999).
Support vector regression with anova decomposition kernels.
Advances in kernel methods—Support vector learning, pages 285–292.

Wahba, G. (1990).
Spline models for observational data.
Society for Industrial and Applied Mathematics, Philadelphia.