Additive interaction modelling with Gaussian process priors

Sahoko Ishida

Department of Statistics
London School of Economics

6 December 2023

Outline

Introduction

Additive interaction modelling with a GP prior

Efficient implementation for large-scale multidimensional grid data

Incomplete grid data

Regression with additive Gaussian process priors

- For a response variable $y_i \in \mathbb{R}$, p-dimensional predictors $x_{li} \in \mathcal{X}_l$ $l = 1, \ldots, p$ and $i = 1, \ldots, n$:

$$y_i = f(x_{1i}, \ldots, x_{pi}) + \epsilon_i \tag{1}$$
$$(\epsilon_1, \ldots, \epsilon_n)^\top \sim N(0, \Sigma)$$

- Assume additive structure on f e.g., for $p = 3$,

$$\begin{aligned} f(x_{1i}, x_{2i}, x_{3i}) = a &+ \underbrace{f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i})}_{\text{main effect}} \\ &+ \underbrace{f_{12}(x_{1i}, x_{2i}) + f_{23}(x_{2i}, x_{3i}) + f_{13}(x_{1i}, x_{3i})}_{\text{two-way interaction effect}} \\ &+ \underbrace{f_{123}(x_{1i}, x_{2i}, x_{3i})}_{\text{three-way interaction effect}} \end{aligned} \tag{2}$$

- Assume $f_j \sim \mathrm{GP}(0, k_j)$ for $j \in \{1, 2, 3, 12, 13, 23, 123\}$.

Challenges and contributions of the thesis

- Large number of terms to consider and parameters to estimate, especially for $I \geq 3$
  - Additive interaction modelling with ANOVA decomposition kernel: Parsimonious specification which makes model fitting, comparison, and interpretation easier
- Implementation of additive GP models for large-scale data Focusing on multi-dimensional grid data and exploiting Kronecker product structure in the model covariance matrix (Kroncker method)
  - Extending the Kronecker method to some cases of the sum of separable kernels, which covers non-saturated interaction models
  - Handling incomplete grid data (Ongoing)

Regression with Gaussian process prior

1D example:

- For $i = 1, ..., n$, consider a regression model for a response $y_i \in \mathbb{R}$ and a predictor $x_i \in \mathcal{X}$:

$$y_i = f(x_i) + \epsilon_i$$

  with iid error $\epsilon_i \sim N(0, \sigma^2)$.

- Prior over f: $f \sim GP(0, k)$ where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called kernel and serves as a covariance function

$$\text{cov}[f(x), f(x')] = k(x, x')$$

- Different kernel leads to different properties of the function f (Linearity, smoothness, etc.)
- Each kernel has some parameters (hyper-parameters) denoted by $\boldsymbol{\theta}$

Regression with Gaussian process prior

- Posterior is also a GP with mean and kernel

$$\bar{m}(x) = \mathbf{k}(x)^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, \qquad x \in \mathcal{X} \quad (3)$$

$$\bar{k}(x, x') = k(x, x') - \mathbf{k}(x)^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}(x'), \qquad x, x' \in \mathcal{X} \quad (4)$$

  where

$$\{\mathbf{K}\}_{1 \le i,j \le n} = k(x_i, x_j)$$

$$\mathbf{k}(x) = (k(x, x_1), \dots, k(x, x_n))^\top$$

- Hyper-parameter estimation
  - Put hyper-prior on $\boldsymbol{\theta}$ and use MCMC, or
  - Optimising log marginal likelihood

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

Outline

Introduction

Additive interaction modelling with a GP prior

Efficient implementation for large-scale multidimensional grid data

Incomplete grid data

Additive interaction modelling with a GP prior

Two variable example

- For $i = 1, \ldots, n$, 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:

$$\begin{aligned} \mathbf{K} &= \mathbf{1}_n \mathbf{1}_n^\top + \mathbf{K}_1 + \mathbf{K}_2 + \mathbf{K}_1 \circ \mathbf{K}_2 \\ &= (\mathbf{1}_n \mathbf{1}_n^\top + \mathbf{K}_1) \circ (\mathbf{1}_n \mathbf{1}_n^\top + \mathbf{K}_2) \end{aligned}$$

ANOVA decomposition kernel

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

$$k(x, x') = \alpha_0^2 \left(1 + k_1(x_1, x_1')\right) \left(1 + k_2(x_2, x_2')\right)$$

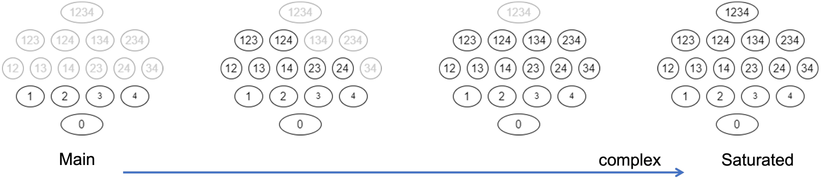
Multiplied by the overall scale parameter α_0^2, so that $a \sim N(0, \alpha_0^2)$.

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

$$k(x, x') = \alpha_0^2 \prod_{l=1}^{d} \left(1 + k_l(x_l, x_l')\right)$$

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

Hierarchical ANOVA decomposition kernel

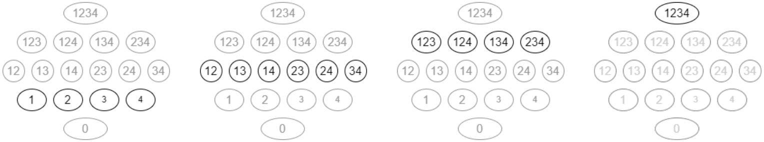


1. Interaction terms – tensor product kernel
2. Interactions included with any main + lower-order interaction terms

Related work

- Functional ANOVA decomposition, Smoothing Spline (SS) ANOVA [Wahba et al., 1995]
  Regression function decomposed in a similar manner as (2), but each term has its own coefficient
- ANOVA kernel for Support Vector Machine [Stitson et al., 1999]

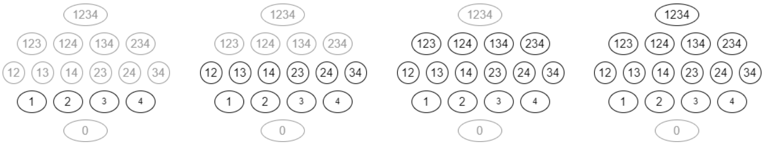
(a) One-way (b) Two-way (c) Three-way (d) Four-way



Related work

- Functional ANOVA decomposition, Smoothing Spline (SS) ANOVA [Wahba et al., 1995]
  Regression function decomposed in a similar manner as (2), but each term has its own coefficient
- Additive Gaussian process models considered in [Duvenaud et al., 2011]

(a) One-way (b) Two-way (c) Three-way (d) Four-way



Additive interaction modelling with a GP prior

Merits

- Hierarchical interaction models give a better fit compared to the model that only accounts for the highest-order interaction
- Parsimonious specification :
  - A smaller number of parameters to estimate compared to classical linear regression or SS ANOVA model.
  - Model selection using log predictive density
- Interpretability: the additive model structure allows for visually interpreting each effect, which is enhanced with k_l being empirically centred.
- Computation: efficient implementation of the proposed model possible for multi-dimensional grid data

Parsimonious specification

Given a set of predictors, all models of any interaction structures share the same set (and number) of parameters

- The different interaction models $\mathcal{M}_k$ can be compared using "plug-in" log marginal likelihood / best fit joint predictive density: $\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}, \mathcal{M}_k)$
- Less costly compared to other criteria, such as
  - Marginal likelihood :

$$p(\mathbf{y}|\mathcal{M}) = \int p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M}_k) p(\boldsymbol{\theta}|\mathcal{M}_k) d\boldsymbol{\theta} \tag{5}$$

  - LOOCV: $\frac{1}{n}\sum_{i=1}^{n} \log p(y_i|\mathbf{y}_{-i}, \mathcal{M}_k)$ where

$$p(y_i|\mathbf{y}_{-i}, \mathcal{M}_k) = \int p(y_i|\boldsymbol{\theta}, \mathcal{M}_k) p(\boldsymbol{\theta}|\mathbf{y}_{-i}, \mathcal{M}_k) d\boldsymbol{\theta}$$

    Does not require fitting the model n times, but some importance sampling procedure needed to approximate the above

Parsimonious specification

- DIC and WAIC are other alternatives but require evaluating $\log p(\mathbf{y}|\boldsymbol{\theta}_s)$ or $\log p(y_i|\boldsymbol{\theta}_s)$ where $\boldsymbol{\theta}_s$ is s-th sample from its posterior distribution.
- A simulation study with 3 variable interaction models show both the best fit predictive density (plug-in marginal likelihood) or marginal likelihood (5) choose the correct model.
- Still requires fitting all candidate models - the model selection is not automated.

Interpretability

The result can be interpreted by plotting the posterior mean

- Posterior mean decomposition: for additive models with $f = \sum_j f_j$ and priorss $f_j \sim GP(0, k_j)$

$$\bar{m}_j(\mathbf{x}_j) = \mathbf{k}_j(\mathbf{x}_j)^\top(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, \quad \mathbf{x}_j \in \mathcal{X}_j$$

  for $j \in J$ where e.g. $J = \{0, 1, 2, 3, ..., 12, 13, 23, ...\}$

- To interpret the two-way interaction (e.g.,between x_1 and x_2) effect, plot

$$\bar{m}_1(x_1) + \bar{m}_{12}(x_1, x_2^*)$$

  as function of x_1, at different value of x_2^*

- The same principle applies to higher-order interactions
- Possible to intuitively understand the effect of lower-order interaction (including the main effect) if kernels are centred.

Interpretability

Centring of kernels

- Any p.d. kernel can be centred by

$$k_{cent}(x, x) = k(x, x') - \mathbb{E}[k(x, X')] - \mathbb{E}[k(X, x')] + \mathbb{E}[k(X, X')]$$

  where $X, X' \sim P$.
- Empirical centring using centring matrix $\mathbf{C} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$

$$\mathbf{K}^{(c)} = \mathbf{CKC}$$

  - All columns and rows sum to zero
  - Ensures $\sum f(x_i) = 0$
- For a linear kernel $k(x, x') = x^\top x'$, or, $\mathbf{K} = \mathbf{XX}^\top$, it is equivalent to centring the covariates by $\mathbf{X}_{cent} = \mathbf{CX}$

Interpretability

- When kernels are centred, each mean function sums to zero over each input, e.g.,

$$\sum_{i=1}^{n} \bar{m}_1(x_{1i}) = 0, \quad \sum_{i=1}^{n} \bar{m}_{12}(x_1, x_{2i}) = 0.$$

- The lower-order interaction can be seen as the averaged effect

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \{\bar{m}_1(x_1) + \bar{m}_{12}(x_1, x_{2i})\} &= \bar{m}_1(x_1) + \underbrace{\sum_{i=1}^{n} \bar{m}_{12}(x_1, x_{2i})}_{=0} \\ &= \bar{m}_1(x_1) \end{aligned}$$

Intepretability

Example with cattle growth longitudinal data

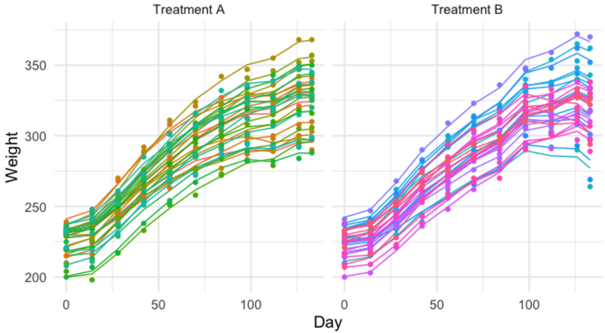


Figure: The observed and fitted growth curve over 133 days of 60 cattle by treatment group

Intepretability

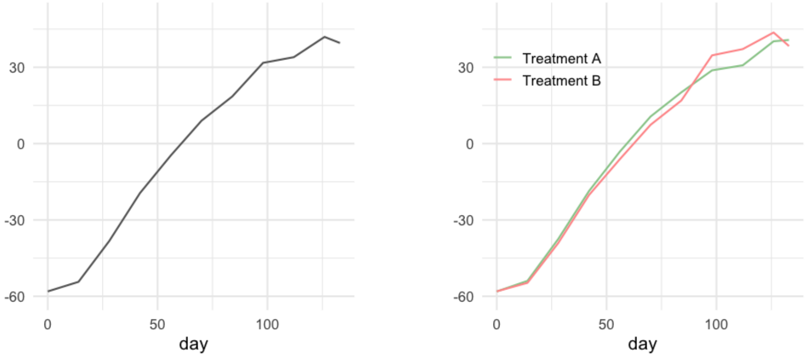
Three-way interaction model:

$$y = f(day, id, group) + \epsilon$$

where

$$\begin{aligned} f(day, group, id) = a &+ f_1(day) + f_2(group) + f_3(id) \\ &+ f_{12}(day, group) + f_{13}(day, id) + f_{23}(group, id) \\ &+ f_{123}(day, group, id) \end{aligned}$$

Intepretability



(a) $\bar{m}_1(day)$ (b) $\bar{m}_1(day) + \bar{m}_{12}(day, group)$

Figure: Average centred growth curve

Outline

Introduction

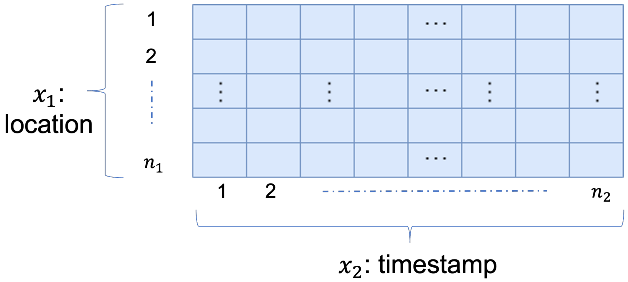
Additive interaction modelling with a GP prior

Efficient implementation for large-scale multidimensional grid data

Incomplete grid data

Multi-dimensional grid/panel data

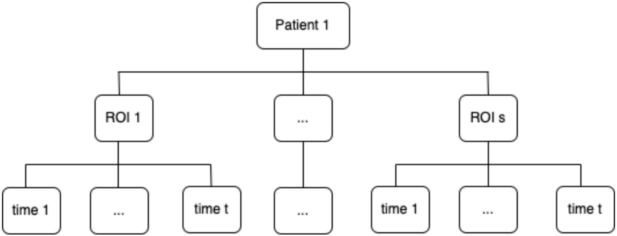
Inputs are on Cartesian grid, e.g.,



- At each grid, we have an observation such as temperature, air-quality levels, etc.
- The grid needs not be equispaced
- Tensor time series

Multi-dimensional grid/panel data

Three-dimension example: brain imaging



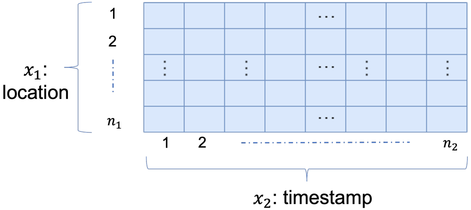
Main constraints

$O(n^3)$ 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):

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

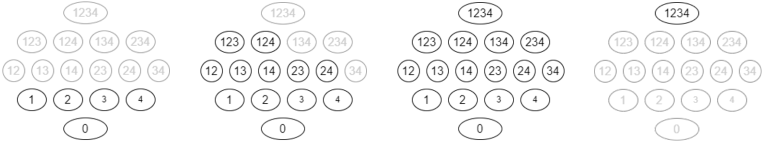
- Main effect model:

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

Kronecker products in Covariance matrix

- Existing literature on the Kronecker approach in GP handles a limited number of models (separable kernel), including
  - a saturated model
  - a model with only the highest interaction

(a) Main (b) Hierarchical (c) Saturated (d) Tensor



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

   Note $(\mathbf{Q}_1 \otimes \mathbf{Q}_2)^\top \mathbf{v} = \mathrm{vec}(\mathbf{Q}_2^\top \mathbf{V} \mathbf{Q}_1)$ where $\mathbf{V} = \mathrm{vec}^{-1}(\mathbf{v})$

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 $\mathbf{K}$

A special case of the sum of separable kernels 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 $\mathbf{K}_l$.
- Do they share the same orthonormal basis?

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 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

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

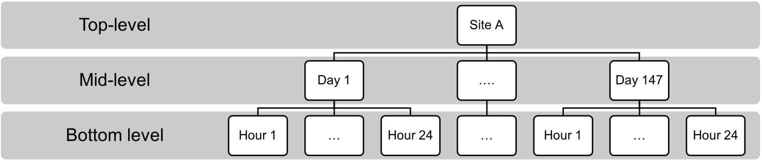
Eigendecomposition of $\mathbf{K}$

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

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

Application to hourly-recorded air-quality monitoring data

- 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

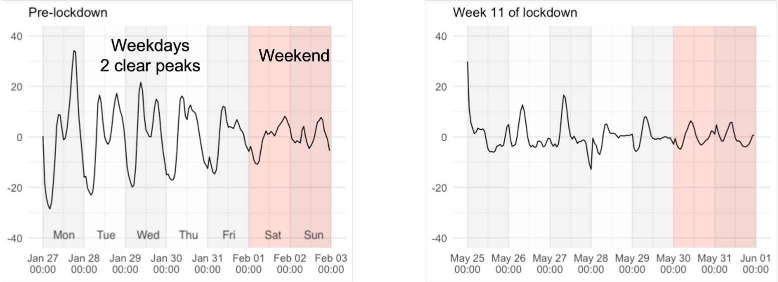


Figure: Plot of $\bar{m}_3(\textit{hour of the day}) + \bar{m}_{13}(\textit{hour of the day, day number})$

Other scalable approaches

- Toeplitz method: similar to Kronecker's as it exploits the data structure
  - The input has to be uni-dimensional and equispaced.
  - Only stationary kernel can be used

  so that the Gram matrix is constant along its diagonal
- Sparse GP with inducing points of length $m < n$, then the costly matrix inversion and matrix-vector multiplication involve these inducing points only.
  - Approximation method while Kronecker method is exact
  - How to choose inducing points?
- Combination of sparse GP with Kronecker method by imposing grid structure in inducing point [Wilson and Nickisch, 2015]

Extensions

Adding random effect on each level to relax iid error assumption, e.g., error term $e_{ij} = u_i + v_j + \epsilon_{ij}$ where $u_i \sim N(0, \sigma_u^2)$ and $v_j \sim N(0, \sigma_v^2)$

$$(e_{11}, e_{12}, \ldots, e_{n_1 n_2})^\top \sim N(\mathbf{0}, \Sigma)$$

where

$$\Sigma = \sigma_u^2 \mathbf{I}_{n_1} \otimes \mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top + \sigma_v^2 \mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top \otimes \mathbf{I}_{n_2} + \sigma^2 \mathbf{I}_{n1} \otimes \mathbf{I}_{n_2}$$

The same orthonormal matrices $\mathbf{Q}_l$ can be used for the decomposition, given $\mathbf{K}_l$ is centred.

Extensions

Incorporating $p \ll n$ dimensional cross-level covariates denoted by $\mathbf{z}_{ij}$

$$y_{ij} = \mathbf{z}_{ij}^{\top} \boldsymbol{\beta} + f(x_{1i}, x_{2j}) + \epsilon_{ij}$$

with $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{B})$. Then the model covariance matrix is

$$\mathbf{Z}\mathbf{B}\mathbf{Z}^{\top} + \mathbf{K} + \sigma^2 \mathbf{I}_n$$

and the inverse (and matrix-vector multiplication) and determinant can still be computed in $O(pn \sum n_l)$ detail

- If the effect of z interacts with x, this is not the case

Limitations

- Forecasting:
  kernels are centred using the observed $\mathbf{x}_1, \ldots, \mathbf{x}_n$, not suited when the main aim is forecasting.
- Kernel sum and product at one level:
  if the base kernel k_l consists of multiple kernels e.g. $k_l = 1 + k_{l1} + k_{l2}$ or $k_l = 1 + k_{l1} + k_{l2} + k_{l1} \otimes k_{l2}$, not all interaction models can be handled within the proposed framework.
- Incomplete grid:
  most repeated measurements and longitudinal data are with missing values

Outline

Introduction

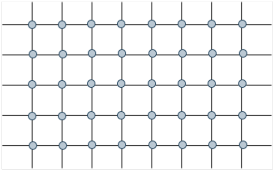
Additive interaction modelling with a GP prior

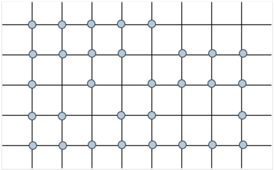
Efficient implementation for large-scale multidimensional grid data

Incomplete grid data

Extention to incomplete grid

- Incomplete grid





  - The work of [Gilboa et al., 2013] addresses this issue, but it is an approximation to a complete case analysis; hence does not work well the cases where the missingness is not at random.
  - Possible to handle with stochastic EM algorithm with Gibbs sampling

Approximation to complete case analysis

Some notations

- $\mathbf{y}_{obs}$ (length n):the observed part
- $\mathbf{y}_{ms}$ (length m): the missing part of the response
- $\tilde{\mathbf{y}} = (\mathbf{y}_{obs}^\top, \mathbf{y}_{ms}^\top)^\top$ which is of length $N = n + m$

Similar notation for $\mathbf{X}_{obs}, \mathbf{X}_{ms}$ and $\mathbf{X}$ for the input. To evaluate

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

- Term 1: fill $\mathbf{y}_{ms}$ with "imaginary" observations and

$$\tilde{\mathbf{y}}^\top(\mathbf{K}_{NN} + \sigma^2\mathbf{D})^{-1}\tilde{\mathbf{y}} \to \text{term 1} \quad \text{as} \quad w \to 0$$

  where

$$\mathbf{D} = \begin{pmatrix} \sigma^2\mathbf{I}_n & \mathbf{0}_{nm} \\ \mathbf{0}_{nm}^\top & w^{-1}\mathbf{I}_m \end{pmatrix}.$$

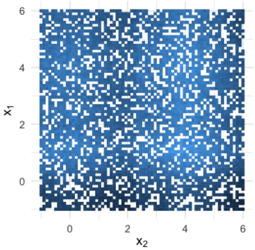
Approximation to complete case analysis

- Term 2 can be approximated by

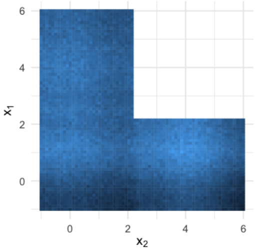
$$\log|\mathbf{K}_{nn} + \sigma^2 \mathbf{I}_n| \approx \sum_{i=1}^{n} \log(\tilde{\lambda}_i^n + \sigma^2)$$

  where $\tilde{\lambda}_i^n = \frac{n}{N}\lambda_i^N$ for $i = 1, \ldots, n$, and $\lambda_1^N, \ldots, \lambda_n^N$ are the n largest eigenvalues of the Gram matrix $\mathbf{K}_{NN}$
- Similar procedure for computing posterior mean and covariance of $\mathbf{y}_{ms}|\mathbf{y}_{obs}$

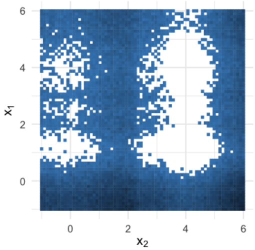
Approximation to complete case analysis



(a) MCAR



(b) MAR



(c) MNAR

Figure: Three missing data mechanisms for the synthetic data with the grid size 70×70 and the missing proportion 30%.

simulation

EM for MNAR

References I

Duvenaud, D. K., Nickisch, H., and Rasmussen, C. (2011).
Additive Gaussian processes.
Advances in neural information processing systems, 24.

Gilboa, E., Saatçi, Y., and Cunningham, J. P. (2013).
Scaling multidimensional inference for structured gaussian processes.
IEEE transactions on pattern analysis and machine intelligence, 37(2):424–436.

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.

References II

Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995).
Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological study of diabetic retinopathy: the 1994 neyman memorial lecture.
The Annals of Statistics, 23(6):1865–1895.

Wilson, A. and Nickisch, H. (2015).
Kernel interpolation for scalable structured gaussian processes (kiss-gp).
In *International conference on machine learning*, pages 1775–1784. PMLR.

Incorporating cross-level covariates

Let

$$\tilde{\mathbf{K}} = \mathbf{Z}\mathbf{B}\mathbf{Z}^\top + \underbrace{\mathbf{K} + \sigma^2 \mathbf{I}_n}_{\mathbf{K}_\sigma}$$

Using Woodbury matrix identity and matrix determinant lemma, we have

$$\tilde{\mathbf{K}}^{-1} = \mathbf{K}_\sigma^{-1} - \mathbf{K}_\sigma^{-1}\mathbf{Z}(\mathbf{B}^{-1} + \mathbf{Z}^\top \mathbf{K}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{K}^{-1}$$

$$\log|\tilde{\mathbf{K}}| = \log|\mathbf{B}^{-1} + \mathbf{Z}^\top \mathbf{K}^{-1}\mathbf{Z}| + \log|\mathbf{K}_\sigma| + \log|\mathbf{B}|$$

back

Simulation study

	MCAR			MAR			MNAR		
	10%	20%	30%	10%	20%	30%	10%	20%	30%
$\bar{\sigma}$	1.5	1.49	1.49	1.5	1.5	1.5	1.45	1.43	1.42
RMSE-σ	0.02	0.02	0.023	0.018	0.017	0.019	0.051	0.074	0.085
RMSE-f	0.16	0.17	0.18	0.17	0.19	0.22	0.73	0.89	1.01
Time(s)	138	146	141	111	110	104	155	147	141

Table: RMSEs for the parameters and for missing grid. Running time is measured in seconds. The synthetic data with 70×70 grid size. For each scenario, the experiment is repeated 20 times.

EM algorithm for incomplete grid with missing-not-at-random cases

- Objective function for EM algorithm

$$Q(\theta|\theta^{t-1}) = \int \log p(\mathbf{y}_{obs}, \mathbf{y}_{ms}|\theta) p(\mathbf{y}_{ms}|\mathbf{y}_{obs}, \theta^{t-1}) d\mathbf{y}_{ms}$$

- Directly evaluating above is costly, especially for large m.
- Numerical approximation can be used, but sampling from $p(\mathbf{y}_{ms}|\mathbf{y}_{obs}, \theta^{t-1})$ iss another challenge.

Stochastic EM algorithm with Gibbs sampling

The conditional distribution

$$p(\mathbf{y}_{ms}|\mathbf{y}_{obs}, \theta^{t-1}) = \mathbf{MVN}(\boldsymbol{\mu}(\theta^{t-1}), \Sigma(\theta^{t-1}))$$

where

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

- To take advantage of the d-dimensional grid structure $(\mathbf{K}_{nn} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{K}_{nm}$ can be replaced by $(\mathbf{K}_{NN} + \mathbf{D})^{-1}\mathbf{K}_{Nm}$ and computed using conjugate gradient (CG) descent algorithm
- This takes $O(\frac{m(m+1)}{2} JN \sum_{l=1}^{d} n_l)$ where J is the number of iterations needed for the CG descent algorithms.

Stochastic EM algorithm with Gibbs sampling

Sampling from a univariate normal distribution

- At t-th iteration,
  1. Sample $y^t_{ms(1)}|\mathbf{y}_{obs}, y^{t-1}_{ms(2)}, y^{t-1}_{ms(3)}, \ldots$ from $N(\mu^t_{(1)}, \sigma^t_{(1)})$ where

     $$\mu^t_{(1)} = \boldsymbol{\alpha}^{t\top}_{ms(1)}\tilde{\mathbf{y}}_{-ms(1)}$$
     $$\sigma^t_{(1)} = k(x_{ms(1)}, x_{ms(1)}) - \boldsymbol{\alpha}^{t\top}_{ms(1)}\mathbf{k}(x_{ms(1)})$$

     where $\tilde{\mathbf{y}}_{-ms(1)} = (\mathbf{y}_{obs}, y^{t-1}_{ms(2)}, y^{t-1}_{ms(3)}, \ldots)$ and

     $$\boldsymbol{\alpha}^t_{ms(1)} = (\mathbf{K}_{N-x_{ms(1)}, N-x_{ms(1)}} + \sigma^2\mathbf{I}_{N-1})^{-1}\mathbf{k}(x_{ms(1)})$$

     can be computed efficiently using a rank 2 update of $(\mathbf{K}_{NN} + \sigma^2\mathbf{I}_N)^{-1}$.
  2. Sample $y^t_{ms(2)}|\mathbf{y}_{obs}, y^t_{ms(1)}, y^{t-1}_{ms(3)}, \ldots$ from $N(\mu^t_{(1)}, \sigma^t_{(1)})$
  3. $\vdots$

Stochastic EM with Gibbs sampling

Merits

- Efficiency: $O(4mN\sum_{l=1}^{d} n_l)$ instead of $O(\frac{m(m+1)}{2}JN\sum_{l=1}^{d} n_l)$
  Generally $4m << \frac{m(m+1)}{2}J$
- Incorporating missingness mechanism e.g., $y_{ms(j)} > c$ for some constant c can be ensured in the sampling step.