Group kernels for Gaussian process metamodels with categorical inputs

O. Roustant[a]

Joint work with E. Padonou[b], Y. Deville[c],
A. Clément[d], G. Perrin[d], J. Giorla[d] and H. Wynn[e]

[a] INSA Toulouse – [b] Mines Saint-Étienne – [c] AlpeStat
[d] CEA – [e] London School of Economics

Updated slide show, following talks in the OQUAIDO Chair (funding the project), Isaac Newton Institute (Cambridge, UK), IMT Toulouse and Univ. of Montpellier. Thanks to all the participants for their feedback!

Outline

Outline

Chair in Applied Mathematics OQUAIDO (2016 - 2020)

- Domain : **Computer experiments**
- Position : **Upstream research guided by case-studies**

oquaido.emse.fr

- 6 technological research partners from :
  - Energy : CEA, IFPEN, IRSN, Storengy
  - Transport : Safran
  - Natural risks : BRGM
- 5 academics :
  EMSE, EC Lyon, Univ. of Grenoble, Nice, Toulouse
- 3 experts : J. Garnier (Ecole Polytechnique), D. Ginsbourger (Idiap), Y. Deville (AlpeStat)

- Chair life : PhD supervision, training sessions (maths, software), research invitations (J. Hensmann, T. Santner, H. Wynn), ...

Metamodeling – Computer experiments

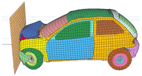
design

Metamodeling – Computer experiments

design → **experiment + measurements** → **meas. output**

Metamodeling – Computer experiments

design → **experiment + measurements** → **meas. output**



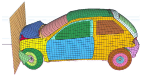
simulator inputs
$\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$

→ **simulator**
physical model, numerical methods

→ **simulator output**
$f_{\text{sim}}(\mathbf{x}^1), f_{\text{sim}}(\mathbf{x}^2), \ldots, f_{\text{sim}}(\mathbf{x}^n)$

Metamodeling – Computer experiments

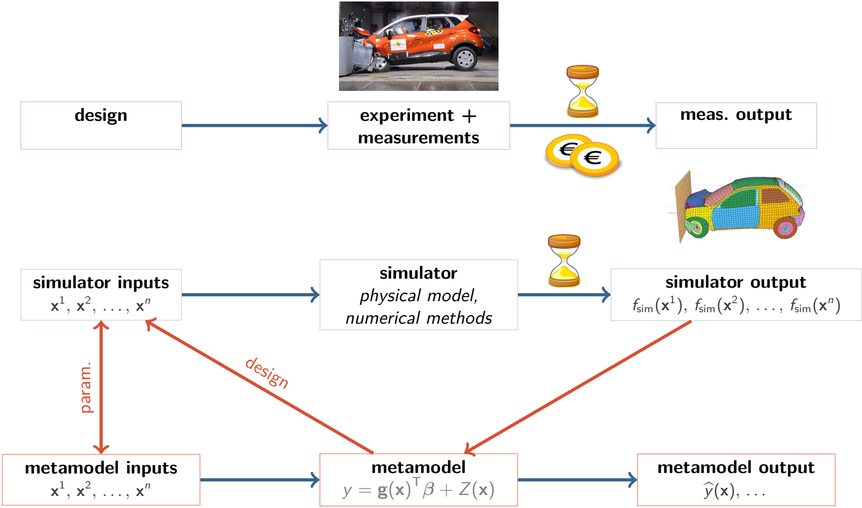
design → **experiment + measurements** → **meas. output**



simulator inputs $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ → **simulator** *physical model, numerical methods* → **simulator output** $f_{\text{sim}}(\mathbf{x}^1), f_{\text{sim}}(\mathbf{x}^2), \ldots, f_{\text{sim}}(\mathbf{x}^n)$

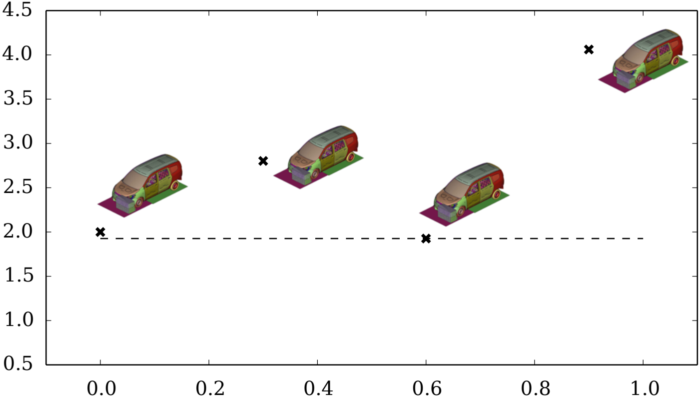
metamodel inputs $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ → **metamodel** $y = \mathbf{g}(\mathbf{x})^\top \beta + Z(\mathbf{x})$ → **metamodel output** $\widehat{y}(\mathbf{x}), \ldots$

Metamodeling – Computer experiments



Metamodeling with Gaussian processes (GP)

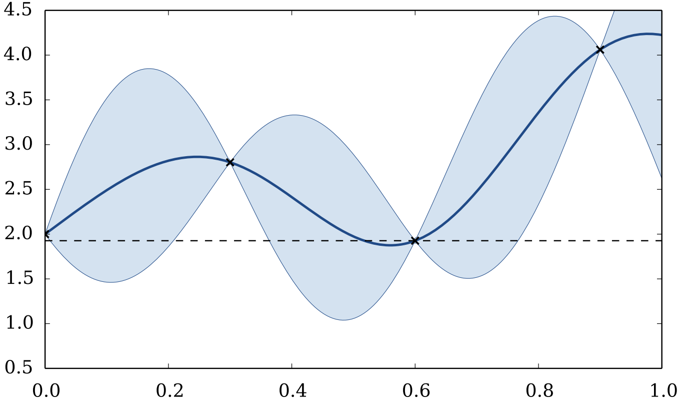
Interpolation of a 1-dimensional function in the context of small data...



Thanks to N. Durrande for the slide !

Metamodeling with Gaussian processes (GP)

Interpolation with GPs : conditional mean and prevision intervals



Gaussian processes

Gaussian processes are stochastic processes (or random fields) s.t. every finite dimensional distribution is Gaussian. $\rightarrow$ Parameterized by two functions

$$Z_{\mathbf{x}} \sim GP(\underbrace{m(\mathbf{x})}_{\textit{trend}}, \underbrace{k(\mathbf{x}, \mathbf{x}')}_{\textit{kernel}})$$

- The trend can be any function.
- The kernel is positive semidefinite :

$$\forall n, \alpha_1, \dots, \alpha_n, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}, \qquad \sum_{i=1}^{n} \alpha_i \alpha_j k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \geq 0.$$

It contains the spatial dependence.

Playing with kernels

A lot of flexibility can be obtained with kernels !

Building a kernel from other ones (basic examples)

Sum, tensor sum	$k_1 + k_2,\ k_1 \oplus k_2$
Product, tensor product	$k_1 \times k_2,\ k_1 \otimes k_2$
ANOVA	$(1 + k_1) \otimes (1 + k_2)$
Warping	$k(\mathbf{x}, \mathbf{x}') = k_1(f(\mathbf{x}), f(\mathbf{x}'))$
...	...

Playing with kernels

A lot of flexibility can be obtained with kernels !

Building a kernel from other ones (basic examples)

Sum, tensor sum	$k_1 + k_2,\ k_1 \oplus k_2$
Product, tensor product	$k_1 \times k_2,\ k_1 \otimes k_2$
ANOVA	$(1 + k_1) \otimes (1 + k_2)$
Warping	$k(\mathbf{x}, \mathbf{x}') = k_1(f(\mathbf{x}), f(\mathbf{x}'))$
...	...

Example : $k_1(x, x'; \sigma^2, \ell) = \sigma^2 \exp\left(-\frac{(x-x')^2}{\ell^2}\right)$

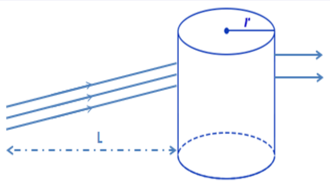
$\rightarrow k_d(\mathbf{x}, \mathbf{x}') = \sigma^2 \prod_{j=1}^{d} k_1(x_j, x_j'; 1, \ell_j) = \sigma^2 \exp\left(-\sum_{j=1}^{d} \frac{(x_j - x_j')^2}{\ell_j^2}\right)$

See other examples in Rasmussen and Williams (2006)... and in this talk !

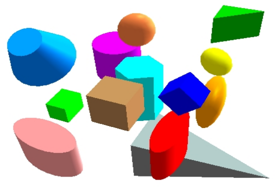
A guiding case-study in nuclear engineering

A particule transport simulator MCNP (Clément, 2016)

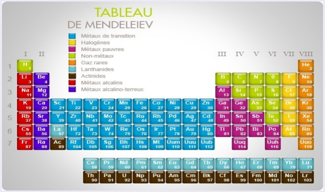
1. Computation using Monte Carlo
2. 4 continuous inputs : L, density, mean width, lateral surface
3. 3 categorical inputs : energy, form, chemical element.



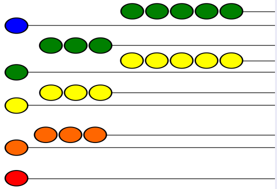
Specific problem : a categorical input with a large number of levels



(a) Form (3 levels)



(b) Atomic number : 94 levels !



(c) Energy (6 levels)

A guiding case-study in nuclear engineering

A 2-stage approach

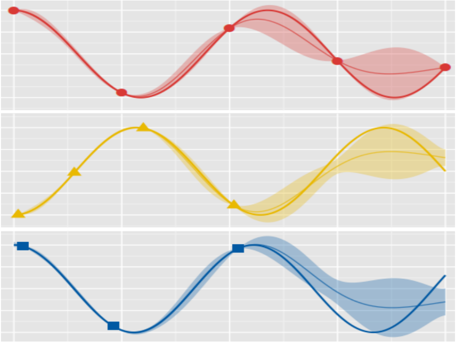
1. GP metamodeling of the computer code
   - This talk !
   - A challenge is the large number of levels (> 90) of one categorical input
   - More details on the preprint, to appear in SIAM/ASA Journal on Uncertainty Quantification
2. Metamodel-based inversion
   - See Clement et al. (2018) on a similar application (continuous inputs)

Outline

GP interpretation when no distance is available

A GP for $(x, u) \in [0, 1] \times \{"red", "yellow", "blue"\}$ can be defined with :

- a kernel on $[0, 1]$, i.e. a **covariance function**
- a kernel on $\{"red", "yellow", "blue"\}$, i.e. a **covariance matrix**
- a valid operation between them, such as *, +, ...



Example : $\mathrm{Cov}(Y(x, "blue"), Y(x', "red")) = k(x, x') \times 0.8$

A formulation - Combination of 1-dimensional kernels

What is a kernel for u_j on $\{1, \dots, m_j\}$?

A positive semidefinite matrix $\mathbf{T}_j$ of size m_j

A formulation - Combination of 1-dimensional kernels

What is a kernel for u_j on $\{1, \dots, m_j\}$?

A positive semidefinite matrix $\mathbf{T}_j$ of size m_j

Combining 1D kernels for $\mathbf{w} = (\mathbf{x}, \mathbf{u})$

Examples of valid operations :

$$\begin{aligned} (\text{Product}) \qquad k(\mathbf{w}, \mathbf{w}') &= k_{\text{cont}}(\mathbf{x}, \mathbf{x}')k_{\text{cat}}(\mathbf{u}, \mathbf{u}') \\ (\text{Sum}) \qquad k(\mathbf{w}, \mathbf{w}') &= k_{\text{cont}}(\mathbf{x}, \mathbf{x}') + k_{\text{cat}}(\mathbf{u}, \mathbf{u}') \\ (\text{ANOVA}) \qquad k(\mathbf{w}, \mathbf{w}') &= (1 + k_{\text{cont}}(\mathbf{x}, \mathbf{x}'))(1 + k_{\text{cat}}(\mathbf{u}, \mathbf{u}')) \end{aligned}$$

A formulation - Combination of 1-dimensional kernels

What is a kernel for u_j on $\{1, \dots, m_j\}$?

A positive semidefinite matrix $\mathbf{T}_j$ of size m_j

Combining 1D kernels for $\mathbf{w} = (\mathbf{x}, \mathbf{u})$

Examples of valid operations :

$$\begin{aligned} \text{(Product)} \quad k(\mathbf{w}, \mathbf{w}') &= k_{\text{cont}}(\mathbf{x}, \mathbf{x}') k_{\text{cat}}(\mathbf{u}, \mathbf{u}') \\ \text{(Sum)} \quad k(\mathbf{w}, \mathbf{w}') &= k_{\text{cont}}(\mathbf{x}, \mathbf{x}') + k_{\text{cat}}(\mathbf{u}, \mathbf{u}') \\ \text{(ANOVA)} \quad k(\mathbf{w}, \mathbf{w}') &= (1 + k_{\text{cont}}(\mathbf{x}, \mathbf{x}'))(1 + k_{\text{cat}}(\mathbf{u}, \mathbf{u}')) \end{aligned}$$

Notice $*$ one of them. Examples of valid kernels for $\mathbf{w}$:

$$k(\mathbf{w}, \mathbf{w}') = k^1_{\text{cont}}(x_1, x_1') * \dots * k^I_{\text{cont}}(x_I, x_I') * [\mathbf{T}_1]_{u_1, u_1'} * \dots * [\mathbf{T}_J]_{u_J, u_J'}$$

A formulation - Combination of 1-dimensional kernels

What is a kernel for u_j on $\{1, \dots, m_j\}$?

A positive semidefinite matrix $\mathbf{T}_j$ of size m_j

Combining 1D kernels for $\mathbf{w} = (\mathbf{x}, \mathbf{u})$

Examples of valid operations :

$$\begin{aligned}
\text{(Product)} \qquad k(\mathbf{w}, \mathbf{w}') &= k_{\text{cont}}(\mathbf{x}, \mathbf{x}')k_{\text{cat}}(\mathbf{u}, \mathbf{u}') \\
\text{(Sum)} \qquad k(\mathbf{w}, \mathbf{w}') &= k_{\text{cont}}(\mathbf{x}, \mathbf{x}') + k_{\text{cat}}(\mathbf{u}, \mathbf{u}') \\
\text{(ANOVA)} \qquad k(\mathbf{w}, \mathbf{w}') &= (1 + k_{\text{cont}}(\mathbf{x}, \mathbf{x}'))(1 + k_{\text{cat}}(\mathbf{u}, \mathbf{u}'))
\end{aligned}$$

Notice $*$ one of them. Examples of valid kernels for $\mathbf{w}$:

$$k(\mathbf{w}, \mathbf{w}') = k^1_{\text{cont}}(x_1, x'_1) * \cdots * k^I_{\text{cont}}(x_I, x'_I) * [\mathbf{T}_1]_{u_1, u'_1} * \cdots * [\mathbf{T}_J]_{u_J, u'_J}$$

Not the most general way, but recovers the usual models of the literature.
$\rightarrow$ Alternatives : Use a d-dim. continuous kernel, use $*_i, *_j$, and so on...

Kernels for ordinal variables

Warping

- When the levels of u are ordered : $1 \leq 2 \leq \cdots \leq L$, define :

$$[\mathbf{T}]_{\ell,\ell'} = k_c(F(\ell), F(\ell')), \quad \ell, \ell' = 1, \dots, L.$$

where $k_c(\mathbf{x}, \mathbf{x}')$ is a continuous kernel, and F is $\uparrow$.
It is the covariance kernel of $Y_{F(\ell)}$ if $Y \sim GP(0, k_c)$.

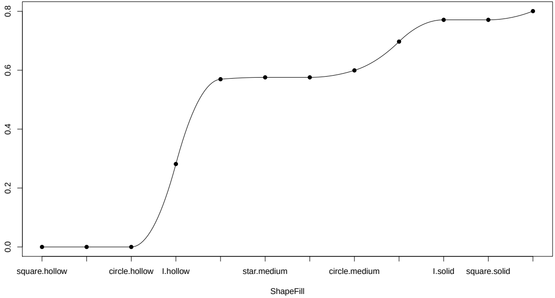


Figure – An example of warping as a spline of degree 2, now available on `kergp`.

Kernels for nominal variables

- **General**
  - Spectral param. $\mathbf{T} = \mathbf{P}\mathbf{D}\mathbf{P}^\top$
  - Spherical param. $\mathbf{T} = \mathbf{L}\mathbf{L}^\top$
- **Compound symmetry** $[\mathbf{T}]_{\ell,\ell'} = \begin{cases} v & \text{if } \ell = \ell' \\ c & \text{if } \ell \neq \ell' \end{cases}$
- **Group kernels**, such as $[\mathbf{T}]_{\ell,\ell'} = \begin{cases} v_g & \text{if } \ell = \ell' \\ c_{g(\ell),g(\ell')} & \text{if } \ell \neq \ell' \end{cases}$
- **Low "rank" approaches** (Rapisarda et al. (2007), Zhang et al. (+2020))

Low-rank	$\mathbf{T} = \mathbf{U}\mathbf{U}^\top,$	with $\mathbf{U} : L \times q$
Latent-variable :	$[\mathbf{T}]_{\ell,\ell'} = k_c(F(\ell), F(\ell')),$	with $F : \{1, \dots, L\} \to \mathbb{R}^q$

Details on low-rank approaches

Interpretation of latent variable kernels (Zhang et al. (+2020))

The underlying Gaussian process for a latent variable kernel is

$$Z(u) = Y(F_1(u), \dots, F_q(u))$$

where $F_1, \dots, F_q$ are mapping from $\{1, \dots, L\} \to \mathbb{R}$, called **"latent variables"**.

- Example : u : type of lubricant, ϕ_1 : viscosity, ϕ_2 : boiling point, ...
- Only the values of the $F_i's$ at $1, \dots, L$ are used : the kernel is parameterized by (a subset of) $F_i(\ell)$, $\ell = 1, \dots, L, \quad i = 1, \dots, q$.

Details on low-rank approaches

Interpretation of latent variable kernels (Zhang et al. (+2020))

The underlying Gaussian process for a latent variable kernel is

$$Z(u) = Y(F_1(u), \ldots, F_q(u))$$

where $F_1, \ldots, F_q$ are mapping from $\{1, \ldots, L\} \to \mathbb{R}$, called **"latent variables"**.

- Example : u : type of lubricant, ϕ_1 : viscosity, ϕ_2 : boiling point, ...
- Only the values of the $F_i's$ at $1, \ldots, L$ are used : the kernel is parameterized by (a subset of) $F_i(\ell)$, $\ell = 1, \ldots, L, \quad i = 1, \ldots, q$.

Links with low-rank kernels

If $k_c(x, x') = \langle x, x' \rangle$ the dot product on $\mathbb{R}^q$, then the latent variable kernel is a low-rank kernel $\mathbf{T} = \mathbf{U}\mathbf{U}^\top$, with $U_{\ell,i} = F_i(\ell)$, for $\ell = 1, \ldots, L, \quad i = 1, \ldots, q$.
$\to$ Latent variables kernels are extending low-rank kernels for general k_c

Outline

Block covariance matrices

- Form considered :

$$\mathbf{T} = \begin{pmatrix} \mathbf{W}_1 & \mathbf{B}_{1,2} & \cdots & \mathbf{B}_{1,G} \\ \mathbf{B}_{2,1} & \mathbf{W}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{B}_{G-1,G} \\ \mathbf{B}_{G,1} & \cdots & \mathbf{B}_{G,G-1} & \mathbf{W}_G \end{pmatrix} \qquad (1)$$

$\mathbf{W}_g$	within-group covariances,	s.t.	$\mathbf{W}_g - \overline{W_g}\mathbf{J}_{n_g,n_g} \succeq 0$
$\mathbf{B}_{g,g'}$	between-group covariances,	with	$\mathbf{B}_{g,g'} \equiv c_{g,g'}$

- Particular case : $\mathbf{W}_g$ is Compound Symmetry (CS)

$$\mathbf{W}_g = \begin{pmatrix} v_g & c_g & \dots & c_g \\ c_g & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & c_g \\ c_g & \dots & c_g & v_g \end{pmatrix}$$

Positive definiteness condition - Algebraic point of view

Theorem 1

For all $\mathbf{T}$ of the form (3),

$$\mathbf{T} \succeq 0 \quad \Longleftrightarrow \quad \widetilde{\mathbf{T}} \succeq 0$$

where $\widetilde{\mathbf{T}}$ is a $G \times G$ matrix obtained by averaging each block.

Positive definiteness condition - Algebraic point of view

Theorem 1

For all $\mathbf{T}$ of the form (3),

$$\mathbf{T} \succeq 0 \quad \Longleftrightarrow \quad \widetilde{\mathbf{T}} \succeq 0$$

where $\widetilde{\mathbf{T}}$ is a $G \times G$ matrix obtained by averaging each block.

Sketch of proof

- $\Rightarrow$ If $\mathbf{T} = \mathrm{Cov}(\mathbf{y})$, then $\widetilde{\mathbf{T}} = \mathrm{Cov}(\overline{\mathbf{y}_1}, \dots, \overline{\mathbf{y}_G})$.

Positive definiteness condition - Algebraic point of view

Theorem 1

For all $\mathbf{T}$ of the form (3),

$$\mathbf{T} \succeq 0 \iff \widetilde{\mathbf{T}} \succeq 0$$

where $\widetilde{\mathbf{T}}$ is a $G \times G$ matrix obtained by averaging each block.

Sketch of proof

- $\Rightarrow$ If $\mathbf{T} = \mathrm{Cov}(\mathbf{y})$, then $\widetilde{\mathbf{T}} = \mathrm{Cov}(\overline{\mathbf{y}_1}, \dots, \overline{\mathbf{y}_G})$.
- $\Leftarrow$

$$\begin{pmatrix} \mathbf{W}_1 & & \left(c_{g,g'}\right) \\ & \ddots & \\ \left(c_{g',g}\right) & & \mathbf{W}_G \end{pmatrix} = \begin{pmatrix} \left(\overline{W_1}\right) & & \left(c_{g,g'}\right) \\ & \ddots & \\ \left(c_{g',g}\right) & & \left(\overline{W_G}\right) \end{pmatrix} + \begin{pmatrix} \mathbf{W}_1 - \left(\overline{W_1}\right) & & (0) \\ & \ddots & \\ (0) & & \mathbf{W}_G - \left(\overline{W_G}\right) \end{pmatrix}$$

Positive definiteness condition - Algebraic point of view

Theorem 1

For all $\mathbf{T}$ of the form (3),

$$\mathbf{T} \succeq 0 \quad \Longleftrightarrow \quad \widetilde{\mathbf{T}} \succeq 0$$

where $\widetilde{\mathbf{T}}$ is a $G \times G$ matrix obtained by averaging each block.

Sketch of proof

- $\Rightarrow$ If $\mathbf{T} = \mathrm{Cov}(\mathbf{y})$, then $\widetilde{\mathbf{T}} = \mathrm{Cov}(\overline{\mathbf{y}_1}, \dots, \overline{\mathbf{y}_G})$.
- $\Leftarrow$

$$\begin{pmatrix} \mathbf{W}_1 & & \left(c_{g,g'}\right) \\ & \ddots & \\ \left(c_{g',g}\right) & & \mathbf{W}_G \end{pmatrix} = \begin{pmatrix} \left(\overline{W_1}\right) & & \left(c_{g,g'}\right) \\ & \ddots & \\ \left(c_{g',g}\right) & & \left(\overline{W_G}\right) \end{pmatrix} + \begin{pmatrix} \mathbf{W}_1 - \left(\overline{W_1}\right) & & (0) \\ & \ddots & \\ (0) & & \mathbf{W}_G - \left(\overline{W_G}\right) \end{pmatrix}$$

Positive definiteness condition - Probabilistic point of view

A hierarchical Gaussian model

$$\eta_{g/\ell} = \mu_g + \lambda_{g/\ell}, \qquad g = 1, \ldots, G, \quad \ell \in \mathcal{G}_g$$

with :

- $\boldsymbol{\mu} \sim \mathcal{N}(0, \mathbf{B}^\star)$ with $\mathbf{B}^\star$ invertible, $\boldsymbol{\lambda}_{g/.} \sim \mathcal{N}(0, \mathbf{W}_g^\star)$, with $\mathbf{W}_g^\star$ invertible.
- $\boldsymbol{\lambda}_{1/.}, \ldots, \boldsymbol{\lambda}_{G/.}, \boldsymbol{\mu}$ are independent.

$\rightarrow$ *response part of a nested two-way ANOVA model, with Gaussian ind. priors*.

Positive definiteness condition - Probabilistic point of view

A hierarchical Gaussian model

$$\eta_{g/\ell} = \mu_g + \lambda_{g/\ell}, \qquad g = 1, \dots, G, \quad \ell \in \mathcal{G}_g$$

with :

- $\boldsymbol{\mu} \sim \mathcal{N}(0, \mathbf{B}^\star)$ with $\mathbf{B}^\star$ invertible, $\boldsymbol{\lambda}_{g/.} \sim \mathcal{N}(0, \mathbf{W}_g^\star)$, with $\mathbf{W}_g^\star$ invertible.
- $\boldsymbol{\lambda}_{1/.}, \dots, \boldsymbol{\lambda}_{G/.}, \boldsymbol{\mu}$ are independent.

$\rightarrow$ *response part of a nested two-way ANOVA model, with Gaussian ind. priors*.

Theorem 2 - Representations of block covariance matrices of the form (3)

$$\mathbf{T} \succeq 0 \iff \mathbf{T} = \mathrm{Cov}(\boldsymbol{\eta} | \overline{\boldsymbol{\lambda}_{1/.}} = \cdots = \overline{\boldsymbol{\lambda}_{G/.}} = 0)$$

with

$$\begin{aligned} \mathbf{W}_g &= B^\star_{g,g} \mathbf{J}_{n_g} + \mathbf{W}_g^\star, \\ \mathbf{B}_{g,g'} &= B^\star_{g,g'} \mathbf{J}_{n_g, n_{g'}}, \end{aligned}$$

where $\mathbf{W}_g^\star = \mathrm{Cov}(\boldsymbol{\lambda}_{g/.} | \overline{\boldsymbol{\lambda}_{g/.}} = 0)$ are centered.

Remarks - CS covariance matrices and negative correlations

- For $G = 1$ this gives a representation of valid CS covariance matrices, including the range of negative correlations.

  Ex for $d = 2$, assume λ_1, λ_2 i.i.d $\mathcal{N}(0, v_\lambda)$, so that we have two parameters v_μ, v_λ, which is the correct number of parameters for CS cov.mat.
  Compare with / without condition $\lambda_1 + \lambda_2 = 0$,

$$\begin{aligned} \eta_1 &= \mu + \lambda_1 \\ \eta_2 &= \mu + \lambda_2 \end{aligned}$$

Remarks - CS covariance matrices and negative correlations

- For $G = 1$ this gives a representation of valid CS covariance matrices, including the range of negative correlations.

  Ex for $d = 2$, assume λ_1, λ_2 i.i.d $\mathcal{N}(0, v_\lambda)$, so that we have two parameters v_μ, v_λ, which is the correct number of parameters for CS cov.mat.
  Compare with / without condition $\lambda_1 + \lambda_2 = 0$,

$$\begin{aligned} \eta_1 &= \mu + \lambda_1 \\ \eta_2 &= \mu + \lambda_2 \end{aligned}$$

- Limitation for groups with strong negative correlation

Proposition (Exclusion property)

If $\mathbf{W}_g$ is a cov. mat. with minimal negative correlation, i.e. $\overline{W_g} = 0$, then $\mathbf{y}_g \perp\!\!\!\perp \mathbf{y}_{-\mathbf{g}}$.

Remarks - CS covariance matrices and negative correlations

- For $G = 1$ this gives a representation of valid CS covariance matrices, including the range of negative correlations.

  Ex for $d = 2$, assume λ_1, λ_2 i.i.d $\mathcal{N}(0, v_\lambda)$, so that we have two parameters v_μ, v_λ, which is the correct number of parameters for CS cov.mat.
  Compare with / without condition $\lambda_1 + \lambda_2 = 0$,

$$\begin{aligned} \eta_1 &= \mu + \lambda_1 \\ \eta_2 &= \mu + \lambda_2 \end{aligned}$$

- Limitation for groups with strong negative correlation

Proposition (Exclusion property)

If $\mathbf{W}_g$ is a cov. mat. with minimal negative correlation, i.e. $\overline{W_g} = 0$, then $\mathbf{y}_g \perp\!\!\!\perp \mathbf{y}_{-\mathbf{g}}$.

Sketch of proof
If $\overline{W_g} = 0$, then $\widetilde{T}_{g,g} = 0$.
Since $\widetilde{\mathbf{T}}$ is p.s.d., we must have $0 = \widetilde{T}_{g,g'} = c_{g,g'}$ for all $g' \neq g$.

Remarks - Centered covariance matrices

Centered matrices $\mathbf{W}_g^\star$ can be parameterized.

Let $\mathbf{A}$ be a $L \times (L-1)$ matrix whose columns form an orthonormal basis of $\mathbf{1}_L^\perp$.
A centered matrix $\mathbf{W}^\star$ is written in an unique way

$$\mathbf{W}^\star = \mathbf{AMA}^\top \tag{2}$$

where $\mathbf{M}$ is a covariance matrix of size $L-1$.

As an example, $\mathbf{A}$ can be obtained by normalizing the columns of a Helmert contrast matrix (Venables and Ripley (2002), §6.2.) :

$$\begin{bmatrix} -1 & -1 & -1 & \cdots & -1 \\ 1 & -1 & -1 & \cdots & -1 \\ 0 & 2 & -1 & \cdots & -1 \\ \vdots & & & \ddots & \vdots \\ \vdots & 0 & 3 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -1 \\ 0 & 0 & \cdots & 0 & L-1 \end{bmatrix}$$

Parameterization of block covariance matrices

Form of $\mathbf{T}$		Parametric setting		Number of parameters
$\mathbf{W}_g$	$\mathbf{B}_{g,g'}$	$\mathbf{M}_g$	$\mathbf{B}^\star$	
CS	$c_{g,g'} \equiv c$	$\propto \mathbf{I}_{n_g-1}$	CS	$G+2$
CS	$c_{g,g'}$	$\propto \mathbf{I}_{n_g-1}$	General	$\frac{G(G+3)}{2}$
General	$c_{g,g'} \equiv c$	General	CS	$2+\sum_{g=1}^{G} \frac{n_g(n_g+1)}{2}$
General	$c_{g,g'}$	General	General	$\frac{G(G+1)}{2}+\sum_{g=1}^{G} \frac{n_g(n_g+1)}{2}$

Reminder :

$$\mathbf{W}_g = B^\star_{g,g} \mathbf{J}_{n_g} + \mathbf{A}_g \mathbf{M}_g \mathbf{A}_g^\top$$
$$\mathbf{B}_{g,g'} = B^\star_{g,g'} \mathbf{J}_{n_g,n_{g'}}$$

Outline

Group kernels

- Form considered :

$$\mathbf{T} = \begin{pmatrix} \mathbf{W}_1 & (c_{1,2}) & \cdots & (c_{1,G}) \\ (c_{1,2}) & \mathbf{W}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & (c_{G-1,G}) \\ (c_{1,G}) & \cdots & (c_{G-1,G}) & \mathbf{W}_G \end{pmatrix} \quad (3)$$

constant between-group covariances

$\mathbf{W}_g$ within-group covariances, s.t. $\mathbf{W}_g - \overline{W_g}\mathbf{J}_{n_g,n_g} \succeq 0$

- Particular case : $\mathbf{W}_g$ is exchangeable, i.e. $\mathbf{W}_g = \begin{pmatrix} v_g & c_g & \dots & c_g \\ c_g & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & c_g \\ c_g & \dots & c_g & v_g \end{pmatrix}$

Group kernels

- Form considered :

$$\mathbf{T} = \begin{pmatrix} \mathbf{W}_1 & (c_{1,2}) & \cdots & (c_{1,G}) \\ (c_{1,2}) & \mathbf{W}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & (c_{G-1,G}) \\ (c_{1,G}) & \cdots & (c_{G-1,G}) & \mathbf{W}_G \end{pmatrix} \qquad (3)$$

constant between-group covariances

$\mathbf{W}_g$ within-group covariances, s.t. $\mathbf{W}_g - \overline{W_g}\mathbf{J}_{n_g,n_g} \succeq 0$

- Particular case : $\mathbf{W}_g$ is exchangeable, i.e. $\mathbf{W}_g = \begin{pmatrix} v_g & c_g & \dots & c_g \\ c_g & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & c_g \\ c_g & \dots & c_g & v_g \end{pmatrix}$

→ If groups are perfectly homogeneous ($c_g = v_g$), then $\mathbf{T}$ has rank G

A first algorithm for group selection

A model-based algorithm

1. Estimate a first GP model for $(\mathbf{x}, \mathbf{u})$ by replacing $\mathbf{T}$ by a proxy kernel $\mathbf{T}_{\mathrm{prox}}$
2. Apply a clustering algorithm on levels, using the L^2 distance given by $\mathbf{T}_{\mathrm{prox}}$

$$\begin{aligned} d(\ell, \ell')^2 &= \mathrm{E}([Z_u - Z_{u'}]^2) \\ &= \mathbf{T}_{\mathrm{prox}}(\ell, \ell) + \mathbf{T}_{\mathrm{prox}}(\ell', \ell') - 2\,\mathbf{T}_{\mathrm{prox}}(\ell, \ell') \end{aligned}$$

A first algorithm for group selection

A model-based algorithm

1. Estimate a first GP model for $(\mathbf{x}, \mathbf{u})$ by replacing $\mathbf{T}$ by a proxy kernel $\mathbf{T}_{\text{prox}}$
2. Apply a clustering algorithm on levels, using the L^2 distance given by $\mathbf{T}_{\text{prox}}$

$$\begin{aligned} d(\ell, \ell')^2 &= \mathrm{E}([Z_u - Z_{u'}]^2) \\ &= \mathbf{T}_{\text{prox}}(\ell, \ell) + \mathbf{T}_{\text{prox}}(\ell', \ell') - 2\,\mathbf{T}_{\text{prox}}(\ell, \ell') \end{aligned}$$

Choice of $\mathbf{T}_{\text{prox}}$ and scope of applicability

- *If there are few homogeneous groups*, a group kernel should be well approx. by a low rank kernel
  $\to$ Choose $\mathbf{T}_{\text{prox}}$ as a low-rank kernel (of rank $\geq G$)

A first algorithm for group selection

A model-based algorithm

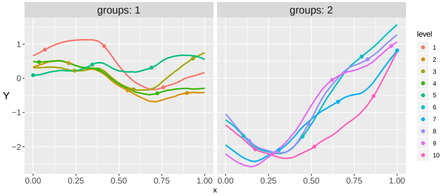
1. Estimate a first GP model for $(\mathbf{x}, \mathbf{u})$ by replacing $\mathbf{T}$ by a proxy kernel $\mathbf{T}_{\mathrm{prox}}$
2. Apply a clustering algorithm on levels, using the L^2 distance given by $\mathbf{T}_{\mathrm{prox}}$

$$\begin{aligned} d(\ell, \ell')^2 &= \mathrm{E}([Z_u - Z_{u'}]^2) \\ &= \mathbf{T}_{\mathrm{prox}}(\ell, \ell) + \mathbf{T}_{\mathrm{prox}}(\ell', \ell') - 2\,\mathbf{T}_{\mathrm{prox}}(\ell, \ell') \end{aligned}$$

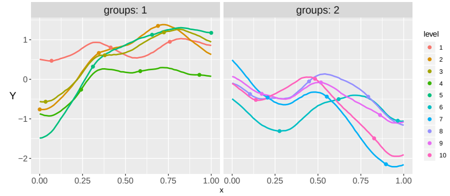
Choice of $\mathbf{T}_{\mathrm{prox}}$ and scope of applicability

- *If there are few homogeneous groups*, a group kernel should be well approx. by a low rank kernel
  $\rightarrow$ Choose $\mathbf{T}_{\mathrm{prox}}$ as a low-rank kernel (of rank $\geq G$)
- *If groups are homogeneous and levels are ordered*, they should be visible as jumps in the warping function
  $\rightarrow$ Choose $\mathbf{T}_{\mathrm{prox}}$ as a warped kernel (with degrees of freedom $\geq G$)

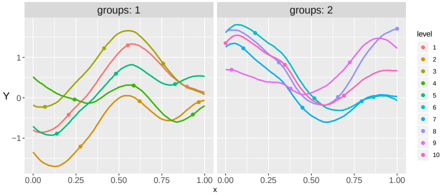
Performance assessment of the group selection algorithm



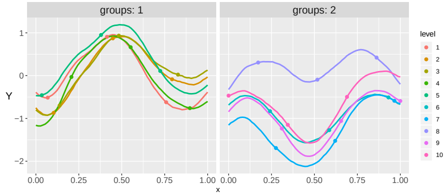
(a) Simulation 1



(b) Simulation 2



(c) Simulation 3



(d) Simulation 4

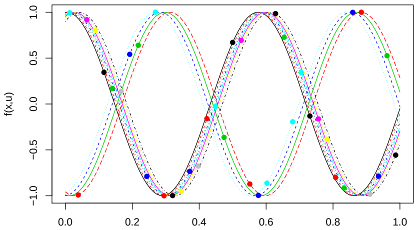
Figure – Four simulations of a GP model Z with tensor-product kernel $k(\mathbf{w}, \mathbf{w}') = k_{\mathrm{cont}}(x, x')k_{\mathrm{cat}}(u, u')$. k_{cont} : Matérn kernel with lengthscale $\theta = 0.4$. k_{cat} : GCS group kernel (10 levels, 2 groups of same size). Within-group correlation : 0.8. Between-group correlation : -0.5.

Performance assessment of the group selection algorithm

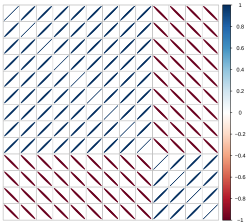
ρ_{bet}	-0.5	0	0.5
Misclassif. rate (95% conf. int)	[12%, 18%]	[18%, 25%]	[26%, 32%]

- For the example of previous slide ($\rho_{\mathrm{bet}} = -0.5$), the misclassif. rate is $\approx$ 15%.
  $\rightarrow$ All the 10 levels but 1 (or 2) are correctly classified
- Misclassification decreases when groups are less separated ($\rho_{\mathrm{bet}} \uparrow$)

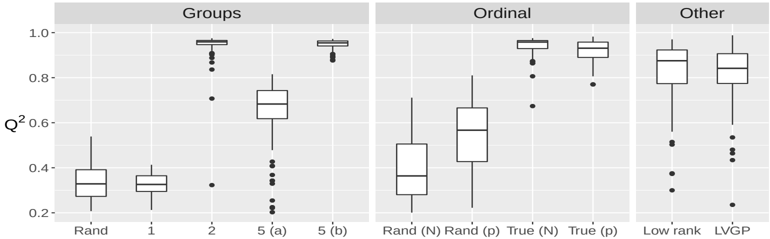
Results on a simple toy example



(a) The example (3 design points per level)

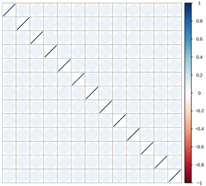


(b) $\hat{\mathbf{T}}$ (median Q^2)

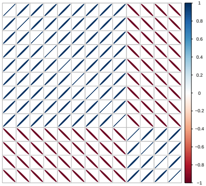


(c) Q^2. Nb. param. : groups = (4, 2, 4, 3, 12), ordinal = (4, 13, 4, 13), other = (26, 24)

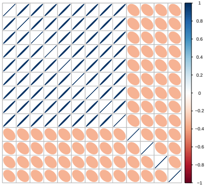
A toy example



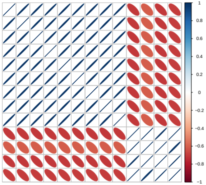
(d) 1 group



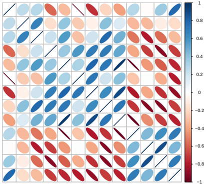
(e) 2 groups



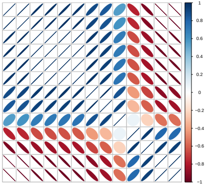
(f) 5 groups (common between-group corr.)



(g) 5 groups (general)



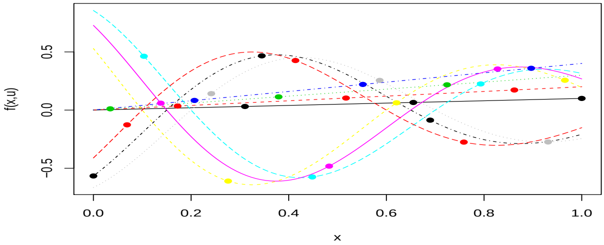
(h) 13 groups

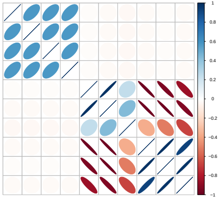


(i) Ordinal

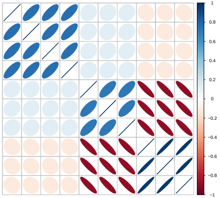
Figure – Estimated correlation kernel k_{cat}, for a design with median Q^2.

A second toy example, with negative within-group correlations





(a) 2 groups



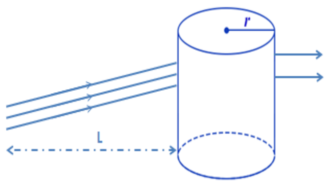
(b) 3 groups

Outline

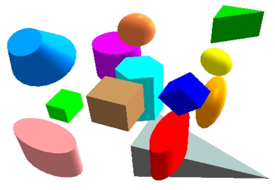
A guiding case-study in nuclear engineering

A particule transport simulator MCNP (Clément, 2016)

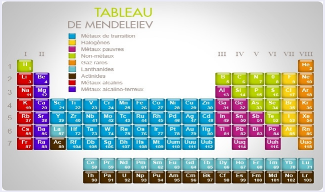
1. Computation using Monte Carlo
2. 4 continuous inputs : L, density, mean width, lateral surface
3. 3 categorical inputs : energy, form, chemical element.



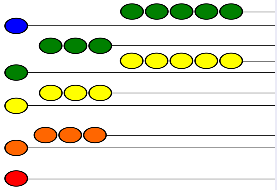
Specific problem : a categorical input with a large number of levels



(c) Form (3 levels)



(d) Atomic number : 94 levels !



(e) Energy (6 levels)

Settings

Full dataset ($N = 5076$)

- Simulator runs from a stratified sampling w.r.t. categorical inputs
  $\rightarrow$ *3 points for each of the* $6 \times 3 \times 94 = 1692$ *combinations of levels*
- Latin hypercube of size N for the continuous inputs

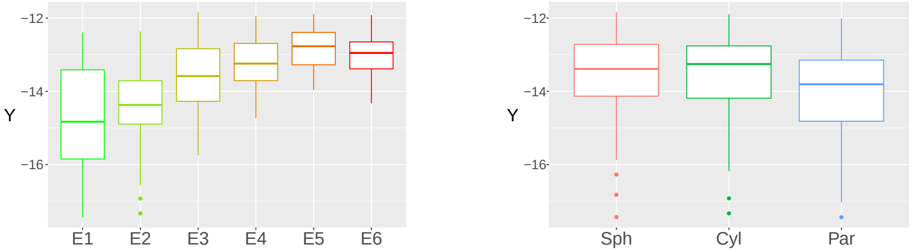
Design of experiments ($n = 282$)

Obtained from the full dataset by stratified sampling w.r.t. 'chemical element'
$\rightarrow$ *3 points for each of the* 94 *levels*

Test set ($N - n$)

Remaining data set

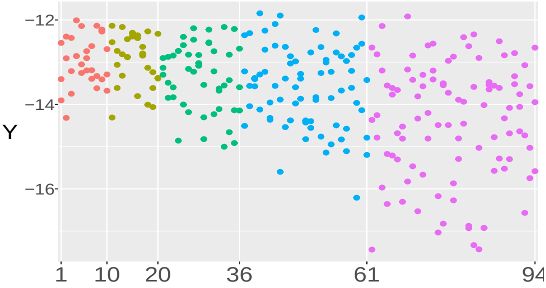
Exploratory analysis - Variables 'Energy' & 'Geometry shape'



Modelling choices :

- 'Energy' : ordinal variable
- 'Geometry shape' : levels seem approx. exchangeable $\rightarrow$ CS cov. matrix

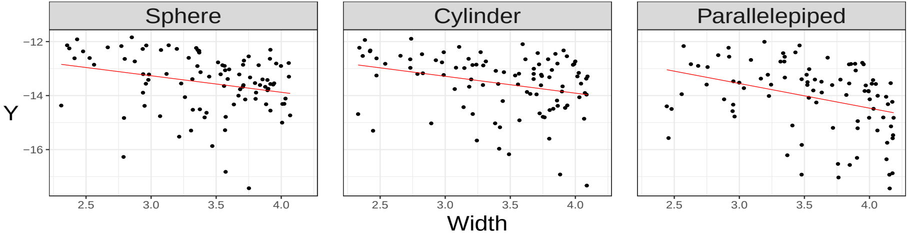
Exploratory analysis – Variable 'Chemical Element' (94 levels)

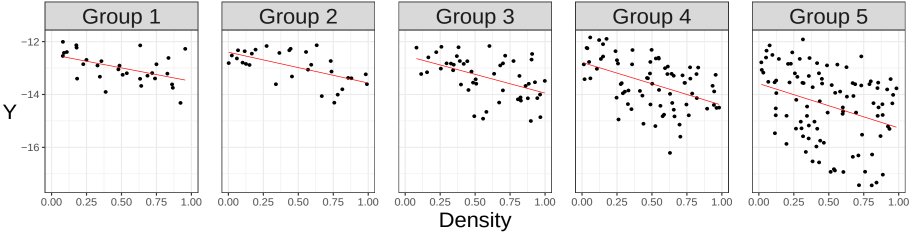


Modelling choice :

- Make the variance depend on the group number

Exploratory analysis – Continuous variables





Prediction accuracy

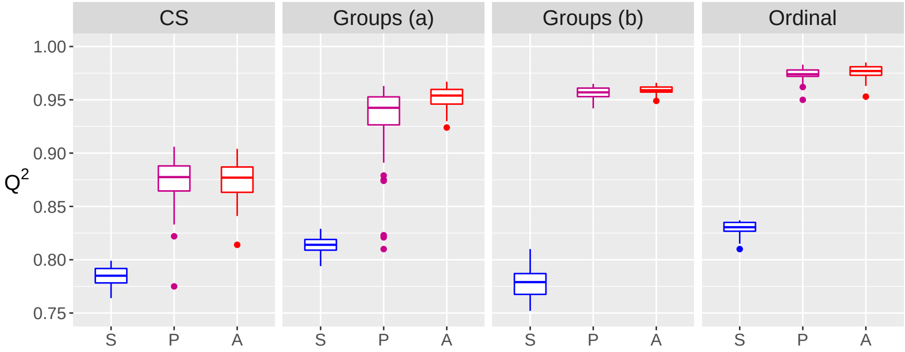
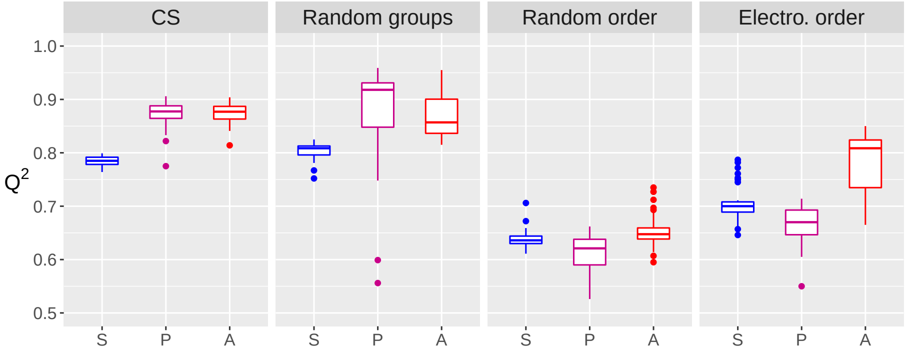


Figure – Q2 of several GP models (in %), based on 60 random designs ($n = 282$).
Operation used : sum, product, ANOVA
Nb of param : 'prod' $= (12, 21, 30, 14)$, 'add' $=$ 'prod'$+6$, 'anova' $=$ 'prod'$+7$

The nominal approach with groups confirms the atomic order as a right order

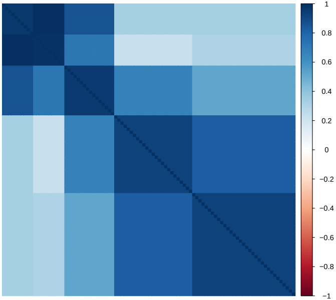
Robustness to group / order misspecification



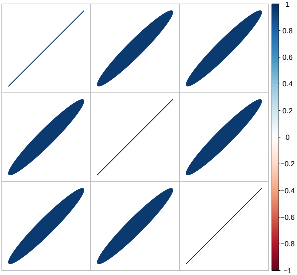
Remarks

- Choosing groups at random is here equivalent to considering 1 group
- Choosing ordering at random can be more detrimental !
- Low-rank approaches are intractable (with the general param. of F)

Some results – Estimated correlations between levels of categorical variables

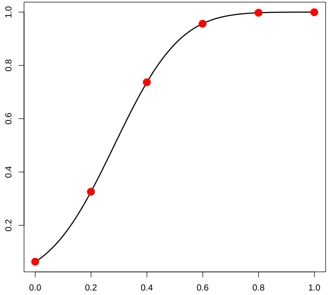


(a) Chemical element



(b) Geometric shape

Some results – Estimated correlations between levels of categorical variables



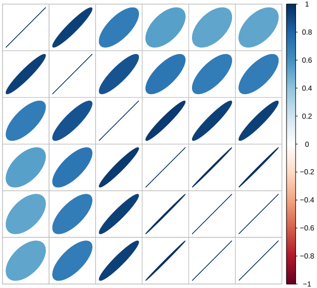
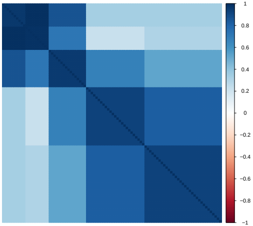
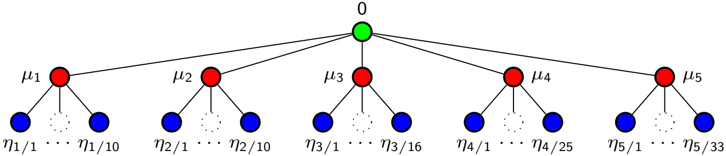


Figure – Estimated kernel for the energy : warping (left) and correlation structure (right).

Towards trees

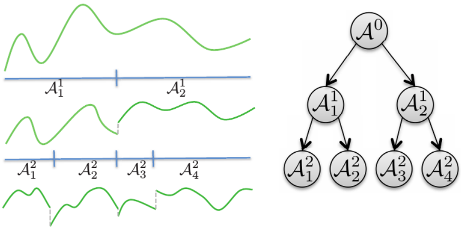




More on hierarchical GPs

- Wavelet kernels (Amato et al., 2006)
- Treed Gaussian processes (Gramacy, 2007)
- Lattice Kriging (Nychka et al., 2015)
- Multiresolution GPs (Fox and Dunson, 2012)
- Hierarchical GPs (Park and Choi, 2010)
- ...

Remark : In these models, the children ("details in subareas") are independent conditionaly on the mother ("trend").
This was not the case before since children sum to 0 (cond. on the mother).



Source : Fox and Dunson (2012), Figure 2.

Outline

Conclusions

General comments for GPs with categorical inputs

- A kernel on a discrete space is a positive semidefinite matrix
- Build kernels from old : product, sum, ANOVA, warping, ...
- Heteroscedasticity / level can be handled directly

Conclusions

General comments for GPs with categorical inputs

- A kernel on a discrete space is a positive semidefinite matrix
- Build kernels from old : product, sum, ANOVA, warping, ...
- Heteroscedasticity / level can be handled directly

Group covariance functions, for levels grouping

1. Prop 1 (Algebra) : Checking PSD only involves the number of groups
   $\rightarrow$ Check if the block average matrix is PSD

Conclusions

General comments for GPs with categorical inputs

- A kernel on a discrete space is a positive semidefinite matrix
- Build kernels from old : product, sum, ANOVA, warping, ...
- Heteroscedasticity / level can be handled directly

Group covariance functions, for levels grouping

1. Prop 1 (Algebra) : Checking PSD only involves the number of groups
   $\rightarrow$ Check if the block average matrix is PSD
2. Prop 2 (Proba) : Hierarchical model, where level effects sum to 0
   $\rightarrow$ The whole range of negative correlations for CS is recovered
   $\rightarrow$ Children are not independent conditionaly on father $\neq$ tree GPs

Conclusions

General comments for GPs with categorical inputs

- A kernel on a discrete space is a positive semidefinite matrix
- Build kernels from old : product, sum, ANOVA, warping, ...
- Heteroscedasticity / level can be handled directly

Group covariance functions, for levels grouping

1. Prop 1 (Algebra) : Checking PSD only involves the number of groups
   $\rightarrow$ Check if the block average matrix is PSD
2. Prop 2 (Proba) : Hierarchical model, where level effects sum to 0
   $\rightarrow$ The whole range of negative correlations for CS is recovered
   $\rightarrow$ Children are not independent conditionaly on father $\neq$ tree GPs
3. Prop 3 (Param) : Parameterization
   $\rightarrow$ Avoids SD programming

Conclusions

General comments for GPs with categorical inputs

- A kernel on a discrete space is a positive semidefinite matrix
- Build kernels from old : product, sum, ANOVA, warping, ...
- Heteroscedasticity / level can be handled directly

Group covariance functions, for levels grouping

1. Prop 1 (Algebra) : Checking PSD only involves the number of groups
   $\rightarrow$ Check if the block average matrix is PSD
2. Prop 2 (Proba) : Hierarchical model, where level effects sum to 0
   $\rightarrow$ The whole range of negative correlations for CS is recovered
   $\rightarrow$ Children are not independent conditionaly on father $\neq$ tree GPs
3. Prop 3 (Param) : Parameterization
   $\rightarrow$ Avoids SD programming

Software implementation

R packages kergp (Deville et al., 2018) (available on CRAN)
and mixgp (Padonou, 2016) (internal).

Open questions and perspectives

Modelling

- How to include trend in models ?
- Nominal inputs : How to group levels ?
- Ordinal inputs : How to order levels ?

Open questions and perspectives

Modelling

- How to include trend in models ?
- Nominal inputs : How to group levels ?
- Ordinal inputs : How to order levels ?

Operational goals

- How to design optimizers that deal with discrete & continuous inputs ?


THANK
YOU

Références I

U. Amato, A. Antoniadis, and M. Pensky. Wavelet kernel penalized estimation for non-equispaced design regression. *Statistics and Computing*, 16(1) :37–55, 2006.

A. Clément. Stochastic Approach for Nuclear Materials Quantification applied on Waste Packages. In *International Nuclear Materials Management Annual Meeting*, Atlanta,USA, 2016.

A. Clement, N. Saurel, and G. Perrin. Stochastic approach for radionuclides quantification. *EPJ Web of Conferences*, 170 :06002, 2018. URL https://doi.org/10.1051/epjconf/201817006002.

Y. Deville, D. Ginsbourger, and O. Roustant. *kergp : Gaussian process laboratory*, 2018. URL https://CRAN.R-project.org/package=kergp. Contributors : N. Durrande. R package version 0.4.0.

E. Fox and D. B. Dunson. Multiresolution Gaussian processes. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 737–745. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4682-multiresolution-Gaussian-processes.pdf.

R. B. Gramacy. An R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian process models. *Journal of Statistical Software*, 19(9) :1–46, 2007.

D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain. A multiresolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2) :579–599, 2015.

Références II

E. Padonou. *mixgp : Kriging models for mixed data*, 2016. R package version 0.1.

S. Park and S. Choi. Hierarchical Gaussian process regression. In M. Sugiyama and Q. Yang, editors, *Proceedings of 2nd Asian Conference on Machine Learning*, volume 13 of *Proceedings of Machine Learning Research*, pages 95–110, 2010.

F. Rapisarda, D. Brigo, and F. Mercurio. Parameterizing correlations : a geometric interpretation. *IMA Journal of Management Mathematics*, 18(1) :55–73, 01 2007.

C. Rasmussen and C. Williams. *Gaussian processes for machine learning*. The MIT Press, 2006.

W. N. Venables and B. D. Ripley. *Modern applied statistics with S*. Springer, 4 edition, 2002.

Y. Zhang, S. Tao, W. Chen, and D. Apley. A latent variable approach to Gaussian process modeling with qualitative and quantitative factors. *to appear in Technometrics*, +2020. URL https://doi.org/10.1080/00401706.2019.1638834.