Group kernels for Gaussian process metamodels with categorical inputs

Olivier Roustant, Esperan Padonou, Yves Deville, Aloïs Clément, Guillaume Perrin, Jean Giorla, Henry P. Wynn

Group kernels for Gaussian process metamodels with categorical inputs

O. Roustant[1], E. Padonou[1], Y. Deville[2], A. Clément[3], G. Perrin[4], J. Giorla[4], and H. Wynn[5]

[1]Mines Saint-Étienne, Univ. Clermont Auvergne, CNRS, UMR 6158 LIMOS, F–42023 Saint-Étienne, France

[2]AlpeStat, Chambéry, France

[3]CEA/DAM/VA, F–21120, Is-sur-Tille, France

[4]CEA/DAM/DIF, F–91297, Arpajon, France

[5]London School of Economics, England

Abstract

Gaussian processes (GP) are widely used as a metamodel for emulating time-consuming computer codes. We focus on problems involving categorical inputs, with a potentially large number L of levels (typically several tens), partitioned in $G \ll L$ groups of various sizes. Parsimonious covariance functions, or kernels, can then be defined by block covariance matrices $\mathbf{T}$ with constant covariances between pairs of blocks and within blocks. We study the positive definiteness of such matrices to encourage their practical use. The hierarchical group/level structure, equivalent to a nested Bayesian linear model, provides a parameterization of valid block matrices $\mathbf{T}$. The same model can then be used when the assumption within blocks is relaxed, giving a flexible parametric family of valid covariance matrices with constant covariances between pairs of blocks. The positive definiteness of $\mathbf{T}$ is equivalent to the positive definiteness of a smaller matrix of size G, obtained by averaging each block. The model is applied to a problem in nuclear waste analysis, where one of the categorical inputs is atomic number, which has more than 90 levels.

1 Introduction

This research is motivated by the analysis of a time-consuming computer code in nuclear engineering, depending on both continuous and categorical inputs, one of them having more than 90 levels. The final motivation is an inversion problem. However, due to the heavy computational cost, a direct usage of the simulator is hardly possible. A realistic approach is to use a statistical emulator or metamodel. Thus, as a first step, we investigate the metamodelling of such computer code. More precisely, we consider Gaussian process (GP) regression models, also called kriging models ([26], [25]), which have been successfully used in sequential metamodel-based strategies for uncertainty quantification (see e.g. [2]).

Whereas there is a flourishing literature on GP regression, the part concerned with categorical inputs remains quite limited. We refer to [33] for a review. As for continuous inputs, covariance functions or *kernels* are usually built by combination of 1-dimensional ones, most often by multiplication or, more rarely, by addition [4]. The question then comes down to constructing a valid kernel on a finite set, which is a positive semidefinite matrix. Some effort has been spent on parameterization of general covariance matrices [20] and parsimonious parameterizations of smaller classes [19]. Low-rank semidefinite matrices have been used for financial applications (see e.g. [24]), and a variant has been recently introduced and interpreted with latent variables [34]. Some block forms have also been proposed [23], in order to deal with a potential large number of levels. However, their validity (in terms of positive definiteness) was not investigated theoretically. Furthermore, to the best of our knowledge, applications in GP regression are limited to categorical inputs with very few levels, typically less than 5.

The aim of the paper is to investigate the so-called *group kernels* cited in [23], defined by block covariance matrices $\mathbf{T}$ with constant covariances between pairs of blocks and within blocks. The motivations are threefold. First, when the number of levels L is large (such as 90) only the most parsimonious parameterizations are viable. Among them, we can cite ordinal kernels parameterized by a basic transformation (such as the Normal cdf), or the basic compound symmetry (CS) kernel, depending on 2 or 3 parameters. As an extension of CS kernels, group kernels give a flexible alternative, with a number of parameters depending on the number of groups. Second, there are natural situations where levels can be gathered in groups. Chemical elements can be classified by categories, such as metal, gaz, etc.; In mechanics, beams can be classified by their section form. Such information can be easily handled in group kernels. Third, group kernels can be used to question an order assumption which is only partially trustworthy. This is the case of the application where the order given by the atomic number can be trusted between categories of chemical elements but not at the finer resolution of individual chemical elements. Comparing the results obtained with a pure ordinal kernel and a group kernel can help to confirm/deny the order assumption.

Our contributions on group kernels are now listed. We exploit the hierarchy group/level by revisiting a nested Bayesian linear model where the response term is a sum of a group effect and a level effect. The level effects are assumed to sum to zero, which allows recovering negative within-group correlations. This model leads to a parameterization of $\mathbf{T}$ which is automatically positive definite. Interestingly, the assumption on within blocks can be relaxed, and we obtain a parameterization of a wider class of valid group kernels. The positive definiteness condition of $\mathbf{T}$ is also explicited: it is equivalent to the positive definiteness of the smaller covariance matrix obtained by replacing each block by its average.

As mentioned above, this work has some connections with Bayesian linear models as well as linear mixed effect models (see e.g. [16], [28]) in a hierarchical view. Other related works concern hierarchical GPs with a tree structure. For instance, particular forms of group kernels are obtained in multiresolution GP models ([6], [18]). Such models usually assume that children are conditionally independent on the mother. This is not the case in our model, due to the condition that the level effects sum to zero.

The paper is structured as follows. Section 2 gives some background on GP regression with

mixed categorical and continuous inputs. Section 3 presents new findings on group kernels. Section 4 gives some guideline for practical usage. In particular, an algorithm is proposed to recover groups, applicable for a small number of groups. Section 5 illustrates on examples composed of toy functions, simulated data, and a toy problem. Section 6 is devoted to the application which motivated this work. Section 7 gives some conclusions and perspectives for future research.

2 Background and notations

2.1 GPs with continuous and categorical variables

We consider a set of I continuous variables $\mathbf{x}_1, \ldots, \mathbf{x}_I$ defined on a hypercubic domain Δ, and a set of J categorical variables $u_1, \ldots, u_J$ with $L_1, \ldots, L_J$ levels. Without loss of generality, we assume that $\Delta = [0,1]^I$ and that, for each $j = 1, \ldots, J$, the levels of u_j are numbered $1, 2, \ldots, L_j$. We denote $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_I)$, $\mathbf{u} = (u_1, \ldots, u_J)$, and $\mathbf{w} = (\mathbf{x}, \mathbf{u})$.

We consider GP regression models defined on the product space

$$\mathcal{D} = [0,1]^I \times \prod_{j=1}^{J} \{1, \ldots, L_j\},$$

and written as:

$$y_i = \mu(\mathbf{w}^{(i)}) + Z(\mathbf{w}^{(i)}) + \epsilon_i, \qquad i = 1, \ldots, N \tag{1}$$

where μ, Z and ϵ are respectively the trend, the GP part and a noise term. There exist a wide variety of trend functions, as in linear models. Our main focus here is on the centered GP $Z(\mathbf{w})$, characterized by its kernel

$$k : (\mathbf{w}, \mathbf{w}') \mapsto \mathrm{cov}\left(Z(\mathbf{w}), Z(\mathbf{w}')\right).$$

Kernels on $\mathcal{D}$ can be obtained by combining kernels on $[0,1]^I$ and kernels on $\prod_{j=1}^{J}\{1, \ldots, L_j\}$. Standard valid combinations are the product, sum or ANOVA. Thus if k_{cont} denotes a kernel for the continuous variables $\mathbf{x}$, k_{cat} a kernel for the categorical ones $\mathbf{u}$, examples of valid kernels for $\mathbf{w} = (\mathbf{x}, \mathbf{u})$ are written:

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

For conciseness, we will denote by $*$ one of the operations: sum, product or ANOVA. The three formula above can then be summarized by:

$$k(\mathbf{w}, \mathbf{w}') = k_{\mathrm{cont}}(\mathbf{x}, \mathbf{x}') * k_{\mathrm{cat}}(\mathbf{u}, \mathbf{u}') \tag{2}$$

Then, in turn, k_{cont} and k_{cat} can be defined by applying these operations to 1-dimensional kernels. For continuous variables, famous 1-dimensional kernels include squared exponential or Matérn [25]. We denote by $k^i_{\mathrm{cont}}(x_i, x'_i)$ such kernels ($i = 1, \ldots, I$). For a categorical variable, notice that, as a positive semidefinite function on a finite space, a kernel is a positive semidefinite

matrix. We denote by $\mathbf{T}_j$ the matrix of size L_j corresponding to kernels for u_j $(j = 1, \ldots, J)$. Thus, examples of expressions for k_{cont} and k_{cat} are written:

$$k_{\text{cont}}(\mathbf{x}, \mathbf{x}') = k^1_{\text{cont}}(x_1, x'_1) * \cdots * k^I_{\text{cont}}(x_I, x'_I) \quad (3)$$

$$k_{\text{cat}}(\mathbf{u}, \mathbf{u}') = [T_1]_{u_1, u'_1} * \cdots * [T_J]_{u_J, u'_J} \quad (4)$$

The formulation given by equations Eq. (2), (3), (4) is not the most general one, since kernels are not always obtained by combining 1-dimensional ones. Nevertheless, it encompasses the GP models used in the literature of computer experiments with categorical inputs. It generalizes the tensor-product kernels, very often used, and the sum used recently by [4] on the categorical part. It also contains the heteroscedastic case, since the matrices $\mathbf{T}_j$ are not assumed to have a constant diagonal, contrarily to most existing works [33]. This will be useful in the application of Section 6, where the variance of the material is level dependent.

Remark 1 *Combining kernels needs some care to obtain identifiable models. For instance, the product of kernels k_1, k_2 with $k_i(x_i, x'_i) = \sigma_i^2 e^{-|x_i - x'_i|}$ $(i = 1, 2)$, is a kernel depending on only one variance parameter $\sigma^2 := \sigma_1^2 \sigma_2^2$. The GP model is identifiable for this new parameter, but not for the initial parameters σ_1^2, σ_2^2.*

2.2 1-dimensional kernels for categorical variables

We consider here a single categorical variable u with levels $1, \ldots, L$. We recall that a kernel for u is then a L by L positive semidefinite matrix $\mathbf{T}$.

2.2.1 Kernels for ordinal variables

A categorical variable with ordered levels is called ordinal. In this case, the levels can be viewed as a discretization of a continuous variable. Thus a GP Y on $\{1, \ldots, L\}$ can be obtained from a 1-dimensional GP Z on the interval $[0, 1]$ by using a non-decreasing transformation F (also called warping):

$$Y(u) = Z(F(u)).$$

Consequently, the covariance matrix $\mathbf{T}$ can be written:

$$T_{u,u'} = k_Z(F(u), F(u')), \quad u, u' = 1, \ldots, L. \quad (5)$$

When $k_Z(x, x')$ depends on the distance $|x - x'|$, then $T_{u,u'}$ depends on the distance between the levels u, u', distorted by F.

In the general case, F is piecewise-linear and defined by $L - 1$ parameters. However, a parsimonious parameterization may be preferred, based on the cdf of a flexible probability distribution such as the Normal or the Beta. We refer to [17] for examples in regression and to [23] for illustrations in computer experiments.

There is also some flexibility in the choice of the continuous kernel k_Z. The standard Squared-Exponential or Matérn kernels are admissible, but induce positive correlation between levels. In order to allow negative correlations, one may choose, for instance, the cosine correlation kernel on $[0, \alpha)$:

$$k_Z(x, x') = \cos(x - x') \quad (6)$$

where $\alpha \in (0, \pi]$ is a fixed parameter tuning the minimal correlation value. Indeed, Eq. (6) defines a decreasing function of $|x - x'|$ from $[0, \alpha]$ to $[\cos(\alpha), 1]$. It is a valid covariance function obtained by choosing μ as a Dirac non-negative measure in Bochner theorem for real-valued stationary kernels: $k_Z(x, x') = \int \cos(\omega(x - x'))d\mu(\omega)$.

2.2.2 Kernels for nominal variables

For simplicity we present here the homoscedastic case, i.e. when $\mathbf{T}$ has a constant diagonal. It is immediately extended to situations where the variance depends on the level, by considering the correlation matrix.

General parametric covariance matrices There are several parameterizations of positive-definite matrices based on the spectral and Cholesky decompositions. The spectral decomposition of $\mathbf{T}$ is written

$$\mathbf{T} = \mathbf{PDP}^\top \tag{7}$$

where $\mathbf{D}$ is diagonal and $\mathbf{P}$ orthogonal. Standard parameterizations of $\mathbf{P}$ involve the Cayley transform, Eulerian angles, Householder transformations or Givens rotations, as detailed in [12] and [27]. Another general parameterization of $\mathbf{T}$ is provided by the Cholesky decomposition:

$$\mathbf{T} = \mathbf{LL}^\top, \tag{8}$$

where $\mathbf{L}$ is lower triangular. When the variance $T_{u,u}$ does not depend on the level u, the rows of $\mathbf{L}$ have the same norm and represent points on a sphere in $\mathbb{R}^L$. A spherical parameterization of $\mathbf{L}$ is then possible with one variance term and $L(L-1)/2$ angles, representing correlations between levels (see e.g. [20]).

Parsimonious parameterizations The general parameterizations of $\mathbf{T}$ described above require $O(L^2)$ parameters. More parsimonious ones can be used, up to additional model assumptions. Among the simplest forms, the compound symmetry (CS) - often called exchangeable - covariance matrix assumes a common correlation for all levels (see e.g. [19]). The CS matrix with variance v and covariance c is defined by:

$$T_{u,u'} = \begin{cases} v & \text{if } u = u' \\ c & \text{if } u \neq u' \end{cases}, \quad c/v \in (-1/(L-1), 1)\,. \tag{9}$$

This generalizes the kernel obtained by substituting the Gower distance d [9] into the exponential kernel, corresponding to $c/v = e^{-d^2} > 0$.
The CS covariance matrix treats equally all pairs of levels, which is an important limitation, especially when $L \gg 1$. More flexibility is obtained by considering groups of levels. Assume that the L levels of u are partitioned in G groups $\mathcal{G}_1, \ldots, \mathcal{G}_G$ and denote by $g(u)$ the group number corresponding to a level u. Then a desired parameterization of $\mathbf{T}$ is given by the block matrix (see e.g. [23]):

$$T_{u,u'} = \begin{cases} v & \text{if } u = u' \\ c_{g(u),g(u')} & \text{if } u \neq u' \end{cases} \tag{10}$$

where for all $i, j \in \{1, \ldots, G\}$, the terms $c_{i,i}/v$ are within-group correlations, and $c_{i,j}/v$ $(i \neq j)$ are between-group correlations. Notice that additional conditions on the $c_{i,j}$'s are necessary to

ensure that $\mathbf{T}$ is a valid covariance matrix, which is developed in the next section.
Another class of parsimonious kernels is given by low-rank kernels of the form

$$\mathbf{T} = \mathbf{U}\mathbf{U}^\top, \tag{11}$$

where $\mathbf{U}$ is a $L \times q$ matrix, and $q \ll L$. A subset of possible matrices $\mathbf{U}$ can be obtained from the spherical parameterization (see Section 2.2.2) by picking a proper subset of angles [24].
In the same vein, [34] have recently introduced latent variable GP (LVGP), in which a categorical kernel has the form

$$T_{u,u'} = k_q(\mathbf{F}(u), \mathbf{F}(u')) \tag{12}$$

where F is a mapping from $\{1, \dots, L\}$ to the low dimensional space $\mathbb{R}^q$, with $q \ll L$, and k_q is a continuous kernel on $\mathbb{R}^q$. Each of the q coordinates of F is interpreted as an unobserved quantitative feature of the categorical input. When k_q is the usual dot kernel on $\mathbb{R}^q$, we recover the low-rank kernel (11) where $\mathbf{U}$ is the $L \times q$ matrix whose rows are $\mathbf{F}(1)^\top, \dots, \mathbf{F}(L)^\top$. Other standard choices for the continuous kernel k_q such as the Squared-Exponential used in [34], provide in general a full rank matrix, but induce positive correlations between levels. This limitation might be addressed as in Section 2.2.1.
In term of complexity, the low-rank kernel (11) and the latent variable kernel (12) both have $O(qL)$ parameters; The CS kernel (9) has 2, and its extension to groups (10) has $O(G^2/2)$ parameters.

3 Generalized compound symmetry block covariance matrices

We consider the framework of Section 2.2.2 where u denotes a categorical variable whose levels are partitioned in G groups $\mathcal{G}_1, \dots, \mathcal{G}_G$ of various sizes $n_1, \dots, n_G$. Without loss of generality, we assume that $\mathcal{G}_1 = \{1, \dots, n_1\}, \mathcal{G}_2 = \{n_1+1, \dots, n_1+n_2\}, \dots$. We are interested in parsimonious parameterizations of the covariance matrix $\mathbf{T}$, written in block form:

$$\mathbf{T} = \begin{pmatrix} \mathbf{W}_1 & \mathbf{B}_{1,2} & \cdots & \mathbf{B}_{1,G} \\ \mathbf{B}_{2,1} & W_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{B}_{G-1,G} \\ \mathbf{B}_{G,1} & \cdots & \mathbf{B}_{G,G-1} & \mathbf{W}_G \end{pmatrix} \tag{13}$$

where the diagonal blocks $\mathbf{W}_g$ contain within-group covariances, and the off-diagonal blocks $\mathbf{B}_{g,g'}$ are constant matrices containing between-group covariances. We denote:

$$\mathbf{B}_{g,g'} = c_{g,g'}\mathbf{J}_{n_g,n'_g}, \qquad g \neq g' \in \{1, \dots, G\}$$

where $\mathbf{J}_{s,t}$ is the s by t matrix of ones. This means that the between-group covariances only depends on groups (and not on levels).

Although block matrices of the form Eq. (13) may be covariance matrices, they are not positive semidefinite in general. A necessary condition is that all diagonal blocks $\mathbf{W}_g$ are

positive semidefinite. But it is not sufficient. In order to provide a full characterization, we will ask a little more, namely that they remain positive semidefinite when removing the mean:

$$\mathbf{W}_g - \overline{W_g}\mathbf{J}_{n_g} \quad \text{is positive semidefinite, for all} \quad g = 1, \dots, G \tag{14}$$

where $\mathbf{J}_{n_g}$ is matrix of ones of size n_g and $\overline{W_g}$ is the average of $\mathbf{W}_g$ coefficients. This condition will appear naturally in Section 3.3. Notice that valid CS covariance matrices satisfy it. Indeed, if $\mathbf{W}$ is a positive semidefinite matrix with variance v and covariance c, then $\mathbf{W} - \overline{W}\mathbf{J}_n = (v-c)\mathbf{P}$ where $\mathbf{P} = \mathbf{I}_n - n^{-1}\mathbf{J}_n$ verifies $\mathbf{P} = \mathbf{P}\mathbf{P}^\top$, which is positive semidefinite. For this reason, we will call matrices with *Generalized Compound Symmetry (GCS)*, block matrices of the form Eq. (13) verifying Eq. (14). In particular, the class of GCS block matrices contains block matrices of the form Eq. (10).

The rest of the section is organized as follows. Section 3.1 shows how valid CS covariance matrices can be parameterized by a Gaussian model. The correspondence is ensured, thanks to a centering condition on the effects of the levels. Section 3.2 gives material on centered covariance matrices. Section 3.3 contains the main results. It extends the model of Section 3.1 to a GCS block matrix. This gives a proper characterization of positive semidefinite GCS block matrices, as well as a parameterization which automatically fulfills the positive semidefinite conditions. Section 3.4 indicates connections with related works. Finally, in Section 4.1, we collect together the details of our parameterization, for ease of reference.

3.1 A Gaussian model for CS covariance matrices

We first focus on the case of a CS matrix. The following additional notations will be used: for a given integer $L \geq 1$, $\mathbf{I}_L$ is the identity matrix of size L, $\mathbf{1}_L$ is the vector of ones of size L. We denote by

$$\mathbf{\Gamma}_L^{\mathrm{CS}}(v, c) = (v - c)\mathbf{I}_L + c\mathbf{J}_L \tag{15}$$

the CS matrix with a common variance term v and a common covariance term c. It is well-known that $\mathbf{\Gamma}_L^{\mathrm{CS}}(v, c)$ is positive definite if and only if

$$-(L-1)^{-1}v < c < v. \tag{16}$$

For instance, one can check that the eigenvalues of $\mathbf{\Gamma}_L^{\mathrm{CS}}(v, c)$ are $v + (L-1)c$ with multiplicity 1 (eigenvector $\mathbf{1}_L$) and $v - c$ with multiplicity $L - 1$ (eigen-space $\mathbf{1}_L^\perp$). Notice that a CS matrix is positive definite for a range of negative values of its correlation term.

Then we consider the following Gaussian model:

$$\eta_u = \mu + \lambda_u, \qquad u = 1, \dots, L \tag{17}$$

where $\mu \sim \mathcal{N}(0, v_\mu)$ with $v_\mu > 0$, and $\lambda_1, \dots, \lambda_L$ are i.i.d. random variables from $\mathcal{N}(0, v_\lambda)$, with $v_\lambda > 0$, assumed to be independent of μ.

A direct computation shows that the covariance matrix of $\boldsymbol{\eta}$ is the CS covariance matrix $\mathbf{\Gamma}_L^{\mathrm{CS}}(v_\mu + v_\lambda, v_\mu)$. Clearly this characterizes the subclass of positive definite CS covariance matrices $\mathbf{\Gamma}_L^{\mathrm{CS}}(v, c)$ such that c is non-negative. The full parameterization, including negative values of c in the range $(-(L-1)^{-1}v, 0)$, can be obtained by restricting the average of level effects to be zero, as detailed in the next proposition.

Proposition 1 *When $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ are related as in Eq. (17), the covariance of $\boldsymbol{\eta}$ conditional on zero average errors $\overline{\lambda} = 0$ is a CS matrix with variance $v = v_\mu + v_\lambda[1 - 1/L]$ and covariance $c = v_\mu - v_\lambda/L$. Conversely, given a CS covariance matrix $\mathbf{C}$ with variance v and covariance c, there exists a representation Eq. (17) such that $\mathbf{C}$ is the covariance of $\boldsymbol{\eta}$ conditional on zero average errors $\overline{\lambda} = 0$ where $v_\mu = v/L + c[1 - 1/L]$ and $v_\lambda = v - c$.*

3.2 Parameterization of centered covariance matrices

The usage of Model Eq. (17) to describe CS covariance matrices involves Gaussian vectors that sum to zero. This is linked to centered covariance matrices, i.e. covariance matrices $\mathbf{W}^\star$ such that $\overline{W^\star} = 0$, as detailed in the next proposition. We further give a parameterization of centered covariance matrices.

Proposition 2 *Let $\mathbf{W}^\star$ be a covariance matrix of size $L \geq 2$. Then, $\mathbf{W}^\star$ is centered iff there exists a Gaussian vector $\mathbf{z}$ on $\mathbb{R}^L$ such that $\mathbf{W}^\star = \mathrm{cov}(\mathbf{z}|\overline{z} = 0)$. In that case, let $\mathbf{A}$ be a $L \times (L-1)$ matrix whose columns form an orthonormal basis of $\mathbf{1}_L^\perp$. Then $\mathbf{W}^\star$ is written in an unique way*

$$\mathbf{W}^\star = \mathbf{A}\mathbf{M}\mathbf{A}^\top \tag{18}$$

where $\mathbf{M}$ is a covariance matrix of size $L - 1$.
In particular if $\mathbf{W}^\star = v[\mathbf{I}_L - L^{-1}\mathbf{J}_L]$ is a centered CS covariance matrix, then $\mathbf{M} = v\mathbf{I}_{L-1}$, and we can choose $\mathbf{z} \sim \mathcal{N}(0, v\mathbf{I}_L)$.

The choice of $\mathbf{A}$ in Proposition 2 is free, and can be obtained by normalizing the columns of a $L \times (L-1)$ Helmert contrast matrix ([30], §6.2.):

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

3.3 A hierarchical Gaussian model for GCS block covariance matrices

Let us now return to the general case, where the levels of u are partitioned in G groups. It will be convenient to use the hierarchical notation g/u, indicating that u belongs to the group $\mathcal{G}_g$. Then, we consider the following hierarchical Gaussian model:

$$\eta_{g/u} = \mu_g + \lambda_{g/u}, \qquad g = 1, \ldots, G, \quad u \in \mathcal{G}_g \tag{19}$$

where for each g the random variable μ_g represent the *effect of the group g*, and the random variables $\lambda_{g/1}, \ldots, \lambda_{g/n_g}$ represent the *effects of the levels* in this group. We assume that $\boldsymbol{\mu}$ is normal $\mathcal{N}(0, \mathbf{B}^\star)$, and all the $\boldsymbol{\lambda}_{g/.}$ are normal $\mathcal{N}(0, \mathbf{W}_g^\star)$. We also assume that $\boldsymbol{\lambda}_{1/.}, \ldots, \boldsymbol{\lambda}_{G/.}$ are independent, and independent of $\boldsymbol{\mu}$.
Notice that, up to centering conditions on $\boldsymbol{\lambda}_{g/.}$ that will be considered next, μ_g is the mean of group g. Hence, $\mathbf{B}^\star$ is interpreted as the *between group means* covariance. Similarly, $\boldsymbol{\lambda}_{g/.}$ is the

within-group effect around the group mean. This justifies the notations $\mathbf{B}^\star$ and $\mathbf{W}_g^\star$.

As an extension of Proposition 1, the next results show that Eq. (19) gives a one-to-one parameterization of valid GCS block covariance matrices, under the additional assumption that the average of level effects is zero in each group.

Theorem 1 *The covariance matrix of $\boldsymbol{\eta}$ conditional on $\{\overline{\lambda_{g/.}} = 0,\ g = 1, \ldots, G\}$ is a GCS block matrix with, for all $g, g' \in \{1, \ldots, G\}$:*

$$\begin{aligned} \mathbf{W}_g &= B^\star_{g,g} \mathbf{J}_{n_g} + \mathbf{W}^\star_g, \\ \mathbf{B}_{g,g'} &= B^\star_{g,g'} \mathbf{J}_{n_g, n_{g'}}, \end{aligned} \tag{20}$$

where $\mathbf{W}_g^\star$ is a centered positive semidefinite matrix equal to $\mathrm{cov}(\boldsymbol{\lambda}_{g/.} | \overline{\lambda_{g/.}} = 0)$. Conversely, let $\mathbf{T}$ be a positive semidefinite GCS block matrix. Then there exists a representation Eq. (19) such that $\mathbf{T}$ is the covariance of $\boldsymbol{\eta}$ conditional on zero average errors $\overline{\lambda_{g/.}} = 0, (g = 1, \ldots, G)$, with:

$$\begin{aligned} \mathbf{B}^\star &= \widetilde{\mathbf{T}}, \\ \mathrm{cov}(\boldsymbol{\lambda}_{g/.} | \overline{\lambda_{g/.}} = 0) &= \mathbf{W}_g - \overline{W_g} \mathbf{J}_{n_g}, \end{aligned}$$

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

Theorem 2 *Positive semidefinite GCS block matrices with CS diagonal blocks exactly correspond to covariance matrices of $\boldsymbol{\eta}$ in Eq. (19) conditional on the G constraints $\overline{\lambda_{g/.}} = 0$ when $\mathrm{cov}(\boldsymbol{\lambda}_{g/.}) \propto \mathbf{I}_{n_g}$.*

As a by-product, we obtain a simple condition for checking the positive definiteness of GCS block matrices. Interestingly, it only involves a small matrix whose size is the number of groups.

Theorem 3 *Let $\mathbf{T}$ be a GCS block matrix. Then*

(i) $\mathbf{T}$ is positive semidefinite if and only if $\widetilde{\mathbf{T}}$ is positive semidefinite.

(ii) $\mathbf{T}$ is positive definite if and only if $\widetilde{\mathbf{T}}$ is positive definite and the diagonal blocks $\mathbf{W}_g$ are positive definite for all $g = 1, \ldots, G$.

Furthermore, we have

$$\mathbf{T} = \mathbf{X} \widetilde{\mathbf{T}} \mathbf{X}^\top + diag(\mathbf{W}_1 - \overline{W_1} \mathbf{J}_{n_1}, \ldots, \mathbf{W}_G - \overline{W_G} \mathbf{J}_{n_G}) \tag{21}$$

where $\mathbf{X}$ is the $n \times G$ matrix

$$\mathbf{X} := \begin{pmatrix} \mathbf{1}_{n_1} & 0 & \ldots & 0 \\ 0 & \mathbf{1}_{n_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \mathbf{1}_{n_G} \end{pmatrix}.$$

Remark 1 *All the results depend on the conditional distribution* $\boldsymbol{\lambda}_{g/.}|\overline{\lambda_{g/.}} = 0$. *Thus there is some flexibility in the choice of* $\mathbf{W}_g^\star$, *since several matrices* $\mathbf{W}_g^\star$ *can lead to the same conditional covariance matrix* $\mathrm{cov}(\boldsymbol{\lambda}_{g/.}|\overline{\lambda_{g/.}} = 0)$.

Remark 2 (Groups of size 1) *Theorem 1 is still valid for groups of size 1. Indeed if* $n_g = 1$, *then* $(\boldsymbol{\lambda}_{g/.}|\overline{\lambda_{g/.}} = 0)$ *is degenerate and equal to 0. Thus* $\mathbf{W}_g^\star = \mathbf{W}_g - \overline{W_g}\mathbf{J}_1 = 0$ *is positive semidefinite.*

We end this section with an "exclusion property" for groups with strong negative correlation. A positive semidefinite GCS covariance matrix can exhibit negative within-group correlations, but this induces limitations on the between-group correlations. More precisely, the following result shows that if a group has the strongest possible negative within-group correlations, then it must be independent of the others.

Proposition 3 (Exclusion property for groups with minimal correlation) *Let* $\mathbf{T}$ *a GCS covariance matrix, and let* $\mathbf{y}$ *be a centered Gaussian vector such that* $\mathrm{cov}(\mathbf{y}) = \mathbf{T}$. *Let* g *be a group number, and denote by* $\mathbf{y}_g$ *(resp.* $\mathbf{y}_{-g}$*) the subvector extracted from* $\mathbf{y}$ *whose coordinates are (resp. are not) in group* $\mathcal{G}_g$. *Assume that* $\mathbf{W}_g$ *is such that* $\overline{W_g} = 0$. *Then* $\mathbf{y}_g$ *is independent of* $\mathbf{y}_{-g}$.

The condition $\overline{W_g} = 0$ is linked to minimal correlations. Indeed, since $\mathbf{W}_g$ is positive semidefinite, $\overline{W_g} \geq 0$. The limit case $\overline{W_g} = 0$ is obtained when negative terms of $\mathbf{W}_g$ are large enough to compensate positive ones. As an example, if $\mathbf{W}_g$ is a positive semidefinite CS covariance matrix with variance v_g and minimal negative covariance $c_g = -(n_g - 1)^{-1}v_g$, then $\overline{W_g} = 0$.

3.4 Related works

The hierarchical model (19) is similar to the specification of nested factors in the linear model with mixed effects, see e.g. [1] or [16] for a Bayesian interpretation using Gaussian priors for the effects $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}_{g/.}$. The centering constraints $\overline{\lambda_{g/.}} = 0$ are also standard identifiability conditions in such models. Furthermore, the particular case of CS covariance matrices corresponds to the exchangeable assumption of the corresponding random variables. In the framework of linear modelling, Model (19) could typically be used with additional grand mean m and errors $\varepsilon_{g,u}$

$$y_{g,u} = m + \mu_g + \lambda_{g,u} + \varepsilon_{g,u}, \tag{22}$$

the terms μ_g and $\lambda_{g,u}$ representing random effects. More generally, many models (1) with only categorical inputs can be considered as linear models with random effects, namely as ANOVA models. The ML estimation of the hyper-parameters for several covariance structures described above can incidentally be obtained by using a package dedicated to mixed effects, although restricted MLE (RMLE) is often preferred when random effects are present. This can be checked by using the lme4 package ([1]), where MLE can optionally be used.

However, in the general case where continuous inputs are used in the GP term of (1), we get models which are no longer ANCOVA linear models with mixed effects, and which can be called *semi-parametric* due to the flexible dependence of the response on the continuous inputs.

A model with no trend and a product kernel (2) with a stationary k_{cont} can be assessed through its representation

$$y = \sum_i \sum_j \alpha_{ij}\, \beta_i(\mathbf{u})\, \eta_j(\mathbf{x}) + \varepsilon$$

where $\beta_i(\mathbf{u})$ and $\eta_j(\mathbf{x})$ are basis functions related to the kernels k_{cat} and k_{cont} respectively. The smoothness level of the functions $\eta_i(\mathbf{x})$ and that of the resulting response is controlled by the continuous covariance kernel.

As a further difference with linear modelling with mixed effects, our goal is different. In linear modelling, the aim is usually to quantify the effects by estimating their posterior distribution, for instance $(\boldsymbol{\mu}, \boldsymbol{\lambda})|\mathbf{y}$ in the case of (22). On the other hand, we aim at predicting new values of y, which involves choosing a suitable form of covariance matrix on the basis of the available information.

4 Guideline for practical usage

4.1 Choosing a parameterization

The results of the previous sections show that valid GCS block covariance matrices can be parameterized by a family of covariance matrices of smaller sizes. It contains the case where diagonal blocks are CS covariance matrices. The algorithm is summarized below.

1. Generate a covariance matrix $\mathbf{B}^\star$ of size G.
2. For all $g = 1, \ldots, G$,
   If $n_g = 1$, set $\mathbf{W}_g^\star = 0$, else:
   - Generate a covariance matrix $\mathbf{M}_g$ of size $n_g - 1$.
   - Compute a centered matrix $\mathbf{W}_g^\star = \mathbf{A}_g \mathbf{M}_g \mathbf{A}_g^\top$, where $\mathbf{A}_g$ is a n_g by $n_g - 1$ matrix whose columns form an orthonormal basis of $\mathbf{1}_{n_g}^\perp$.
3. For all $1 \leq g < g' \leq G$, compute the within-group blocks $\mathbf{W}_g$ and between-group blocks $\mathbf{B}_{g,g'}$ by Eq. (20).

In steps 1 and 2, the generator covariance matrices $\mathbf{B}^\star$ and $\mathbf{M}_g$ can be general, and obtained by one of the parameterizations of Section 2.2.2. A direct application of Theorem 3 also shows that $\mathbf{T}$ is invertible if and only if $\mathbf{B}^\star$ and the $\mathbf{M}_g$'s are invertible (cf. Appendix for details). Furthermore, some specific form, such as CS matrices, can be chosen. Depending on the number of groups and their sizes, different levels of parsimony can be obtained. Table 1 summarizes some possibilities.
Note that the parameterization is for a general covariance matrix, but not for additional constraint, as for a correlation matrix. In these situations, one can take advantage of the economic condition of Theorem 3: positive semidefiniteness on $\mathbf{T}$ is always equivalent to positive semidefiniteness of the small matrix $\widetilde{\mathbf{T}}$ of size G. This can be handled more easily by non-linear or semidefinite programming algorithms.

Parametric setting		*Resulting form of* $\mathbf{T}$		*Number of parameters*
$\mathbf{M}_g$	$\mathbf{B}^\star$	$\mathbf{W}_g$	$\mathbf{B}_{g,g'}$	
$v_{\lambda_g}\mathbf{I}_{n_g-1}$	$\mathbf{\Gamma}^{\mathrm{CS}}(v_\mu, c_\mu)$	$\mathbf{\Gamma}^{\mathrm{CS}}(v_g, c_g)$	$c_{g,g'} \equiv c_\mu$	$2G+1$
$v_{\lambda_g}\mathbf{I}_{n_g-1}$	General	$\mathbf{\Gamma}^{\mathrm{CS}}(v_g, c_g)$	$c_{g,g'}$	$\frac{G(G+3)}{2}$
General	$\mathbf{\Gamma}^{\mathrm{CS}}(v_\mu, c_\mu)$	General	$c_{g,g'} \equiv c_\mu$	$2+\sum_{g=1}^{G}\frac{n_g(n_g+1)}{2}$
General	General	General	$c_{g,g'}$	$\frac{G(G+1)}{2}+\sum_{g=1}^{G}\frac{n_g(n_g+1)}{2}$

Table 1: Parameterization details for some valid GCS block covariance matrices $\mathbf{T}$.

4.2 Choosing groups

The prediction performance of a GP model defined by a group kernel depends on groups selection. Clearly some expertise in the application field will be helpful for that goal. However, one may be also interested by an automatic procedure. As a first step towards this direction, we suggest the following model-based strategy. We consider the formulation of Section 2, with continuous inputs $\mathbf{x}$ and categorical inputs $\mathbf{u}$, and assume that a group kernel $\mathbf{T}$ has been defined for the first categorical input $u = u_1$.

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

$$d(u,u') = \left(T_{\mathrm{prox}}(u,u) + T_{\mathrm{prox}}(u',u') - 2\,T_{\mathrm{prox}}(u,u')\right)^{1/2}.$$

The distance in Step 2 correspond to the L^2 distance of the underlying centered GP Z corresponding to $\mathbf{T}_{\mathrm{prox}}$: $d(u,u')^2 = \mathrm{E}([Z_u - Z_{u'}]^2) = \mathrm{var}(Z_u - Z'_u)$. We hope that if the initial kernel is well chosen in Step 1, then the classes returned by the clustering algorithm may be a good guess of the groups. For a small number of levels, a general covariance kernel can be used as a proxy. We now discuss more parsimonious kernel choices.

Ordinal variable When the categorical variable is assumed to be ordinal, the initial kernel $\mathbf{T}_{\mathrm{prox}}$ can be chosen as in Section 2.2.1 by a warping. Indeed, the split between groups may correspond to jumps in the warping curve F. Intuitively, the estimated distance $|F(u+1) - F(u)|$ between levels u and $u+1$ may be small when u and $u+1$ belong to a same homogeneous group and larger otherwise. Notice that the power of detection depends on the warping used. For instance, the Normal warping, parameterized by only two parameters, can only detect one split, thus two groups. On the other hand, the general piecewise affine warping, parameterized by the number of levels, may be able to detect many groups. However, it can hardly be applied to a large number of levels (due to estimation issues), and may be subject to overfitting.

Nominal variable For a nominal variable, the initial kernel $\mathbf{T}_{\mathrm{prox}}$ may be chosen as a low-rank kernel. Indeed, observe that if a group of levels is replaced by its center, than the corresponding

GCS block matrix has a rank less or equal to the number of groups. This is visible for instance in Eq. (19) by setting all level effects $\lambda_{g/u}$ to 0: this implies that the columns of the GCS block matrix of Thm. 1 that correspond to a same group are equal. Thus, for a small number of homogeneous groups, we can expect that the GCS block matrix $\mathbf{T}$ is well approximated by a low-rank covariance matrix $\mathbf{T}_{\text{prox}}$ (11). In the same vein, as an extension of low-rank kernels, LVGP kernels (Section 2.2.2) may be a good candidate for $\mathbf{T}_{\text{prox}}$. Indeed, if some levels belong to an homogeneous group, the corresponding latent variables may share some common properties, and conversely (as visible in Eq. 12). Thus the presence of groups should be detected.

5 Examples

This section investigates four toy examples with one categorical input, and one or two continuous input variables. We first gather some common information.
For continuous variables, we have used a Matérn 5/2 kernel, corresponding to a two-times differentiable GP in mean square sense. The kernel mixing continuous and categorical variables is built here by tensor-product, i.e. by choosing a product in Eq. (2).
Parameter estimation is done by maximum likelihood. As the likelihood surface may be multimodal, we have launched several optimizations with different starting points chosen at random in the domain. The optimization algorithm used is COBYLA, from R package `nloptr` [11], which is derivative-free and handles nonlinear inequality and equality constraints. To our experience, it gives more robust results than gradient-based algorithms, and handles missing values returned when matrices are numerically singular for certain parameter choices.

Model accuracy is measured over a test set formed by a regular grid of large size (typically 1000), in terms of Q^2 criterion. The Q^2 criterion has a similar expression than R^2, but is computed on the test set:

$$Q^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{23}$$

where the y_i denote the observations (on the test set), $\bar{y}$ their mean, $\hat{y}_i$ the predictions. It is negative if the model performs worst than the mean, positive otherwise, and tends to 1 when predictions are close to true values. Finally, the process is repeated 100 times, in order to assess the sensitivity of the result to the design.

5.1 Example 1

Consider the deterministic function

$$f(x,u) = \cos\left(7\pi\frac{x}{2} + p(u)\pi - \frac{u}{20}\right)$$

with $x \in [0,1]$, $u \in \{1,\dots,13\}$ and $p(u) = \left(0.4 + \frac{u}{15}\right)\mathbb{1}_{u>9}$. As visible in Fig. 1, there are two groups of curves corresponding to levels $\{1,\dots,9\}$ and $\{10,\dots,13\}$ with strong within-group correlations, and strong negative between-group correlations.

We aim at reconstructing g with GP models based on levels grouping. We consider:

- One group (CS covariance matrix);

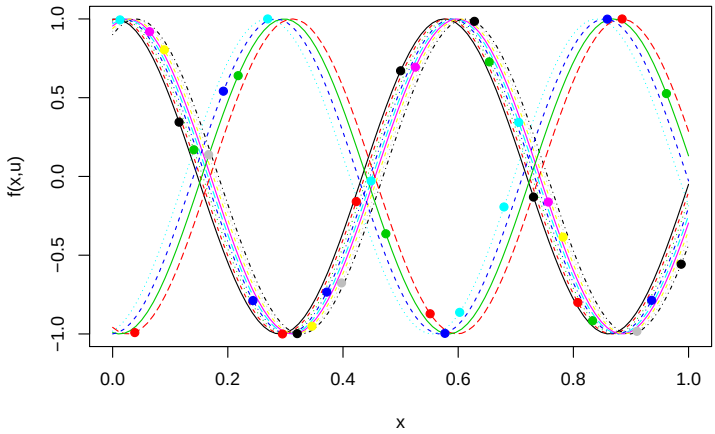


Figure 1: Test function of Example 1. Bullets represent design points.

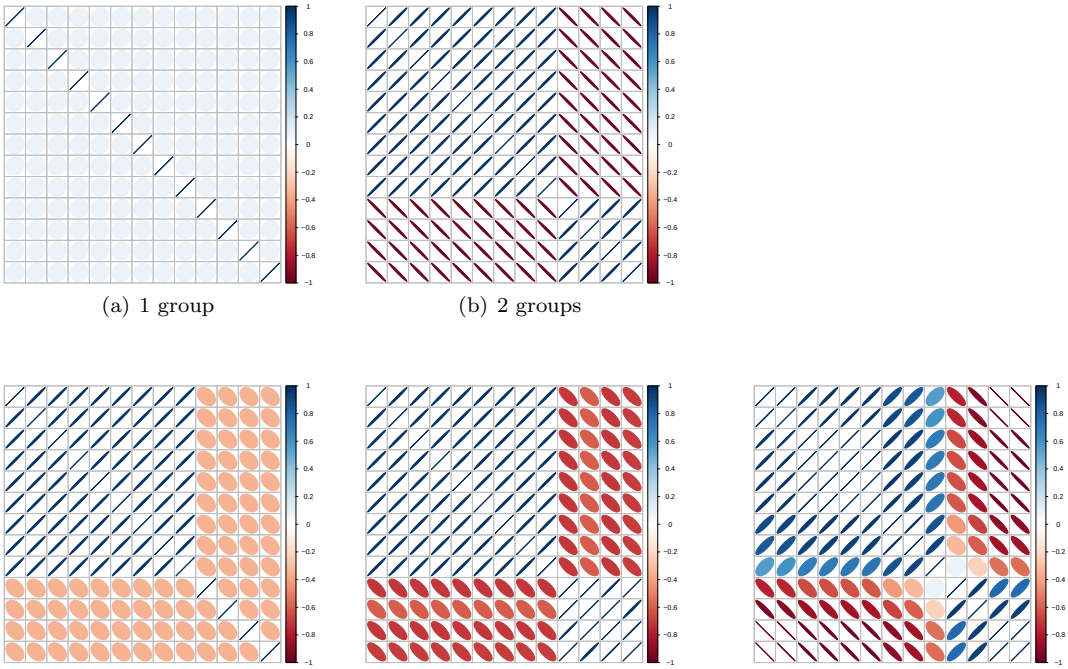
- Two unknown groups. To recover them, we use the strategy presented in Section 4.2 based on an initial low-rank matrix of rank 2 and hierarchical clustering;
- Two groups chosen at random. More precisely, the group numbers of the levels are drawn independently from a Bernoulli distribution.
- Five given groups $\{1,\dots,9\}$, $\{10\}$, $\{11\}$, $\{12\}$, $\{13\}$, with two variants: when the between-group correlation is constant and the general case.

We also compare the results when assuming that u is ordinal. The ordinal kernel is obtained by mapping a cdf F into the cosine kernel k_Z of Equation (6) with $\alpha = \pi$. (see Section 2.2.1). Notice that the choice of k_Z is motivated by recovering negative correlations, and has no link with the sinusoidal form of the curves of the example. We consider two warping choices (Normal, general), and compare the results when the order is known and given by the phase of the sinusoids, or when it is unknown and chosen at random.
Finally, we also consider a low-rank kernel and LVGP.
In order to benefit from the strong link between levels, we use a design that spreads out the points between levels. For instance, the information given by $g(0,1)$ may be useful to estimate $g(x,u)$ at 0 for a different level $u \geq 2$, without computing $g(0,u)$. More precisely, we have used a (random) sliced Latin hypercube design (SLHD) [22] with 3 points by level, for a total budget of 39 points.

Results The estimated correlation parameters are shown in Fig. 2. The interpretation of the correlation plots is as follows. For a given pair of levels $\{u,u'\}$, the ellipse represents the level lines of the pdf of the Gaussian vector with correlation $T_{u,u'}$; In particular strong positive (resp. negative) correlations correspond to thin ellipse with main axis $y = x$ (resp. $y = -x$). The correlation value is also represented with a color scale.

The correlation structure that can be intuited from Fig. 1 is well recovered with two groups

and five groups, with different between-groups correlations. On the opposite, considering only one group or five groups with a common between-group correlation oversimplifies the correlation structure. Finally, the model using an ordinal kernel recovers the two groups of curves, as well as the strong negative correlation between them, which is made possible by the choice of the kernel used in the warping.



(a) 1 group (b) 2 groups

(c) 5 groups (common between-group correlation) (d) 5 groups (general case) (e) Ordinal

Figure 2: Estimated correlation kernel k_{cat}, based on a representative design of experiments (design with median Q^2).

Performance results are shown in Fig. 3. As expected on this example where levels form groups, group kernels outperform the other kernel choices, when correct groups are identified. We can see that the best tradeoff between prediction accuracy and parsimony is obtained with two groups. Notice that the two groups are almost always correctly identified by the automatic algorithm of Section 4.2. This is not surprising since the two groups are very homogeneous, and thus well approximated by a low-rank kernel, as well as strongly negatively correlated, a favorable case for clustering.

When assuming that u is ordinal with a given correct order, the ordinal kernel has good performances as well. The estimated warping (not shown) has a strong jump between levels 9 and 10, corresponding to the two groups, and could be used for group detection.

The results are also informative about misspecification. When groups are chosen at random,

the performances obtained with GCS group kernels are comparable to a simple CS kernel, corresponding to one group. When the order is chosen at random, the performance decreases when compared to the correct order, but remains better than a CS kernel. Actually, an inspection of the estimated warpings reveals that partial orders are detected among levels, a phenomenon enhanced by the warping flexibility. This may explain why the performance does not decrease too much, and is better for a general warping than for the Normal one.

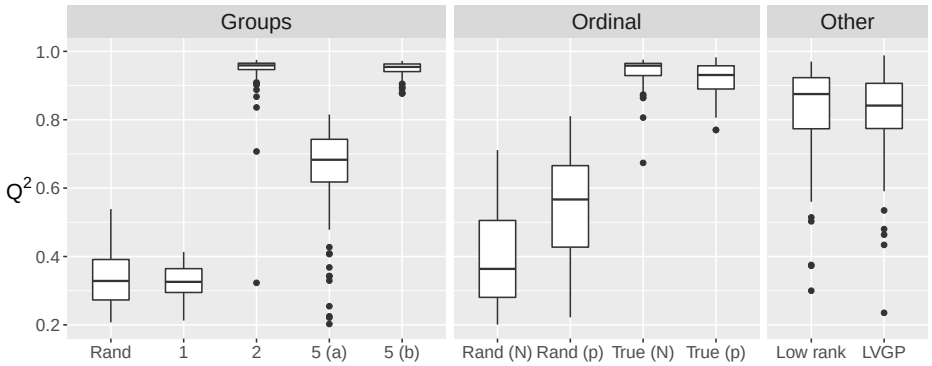


Figure 3: Q^2 of various GP models, based on 100 repetitions of the design. First panel (group kernels): 2 random groups, 1 group (CS structure), 2 estimated groups, 5 given groups (a : common between-group covariance, b : general). Second panel (ordinal kernels): Random order or true order, with two warpings (N: Normal, p: piecewise affine), Third panel (other kernels): Low-rank ($q = 2$), LVGP ($q = 2$). Number of parameters used in each categorical kernel (boxplot order): groups = (4, 2, 4, 3, 12), ordinal = (4, 13, 4, 13), other = (26, 24).

5.2 Example 2

We now provide a second example, in order to illustrate the ability of the hierarchical model (19) to deal with negative within-group correlations. We consider the deterministic function given by:

$$f(x,u) = \begin{cases} (x + 0.01(x - 1/2)^2) \times u/10 & \text{if } u = 1,2,3,4 \\ 0.9\cos(2\pi(x + (u-4)/20)) \times \exp(-x) & \text{if } u = 5,6,7 \\ -0.7\cos(2\pi(x + (u-7)/20)) \times \exp(-x) & \text{if } u = 8,9,10 \end{cases}$$

with $x \in [0,1]$, $u \in \{1,\ldots,10\}$. As visible in Fig. 4, the levels can be split in two groups: a group of almost linear functions (levels $1-4$), and a group of damped sinusoidal functions (levels $5-10$). Within the latter group, there are strong negative correlations between levels $5-7$ and $8-10$. Hence, the levels could also be split into three groups.

In this section, we briefly compare the corresponding GP models:

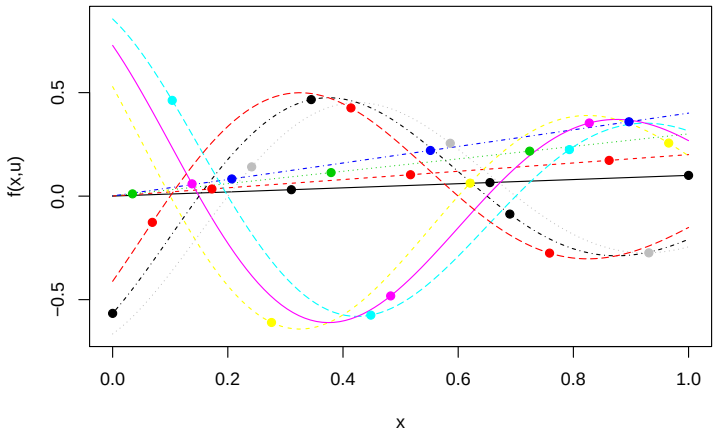


Figure 4: Test function of Example 2. Bullets represent design points.

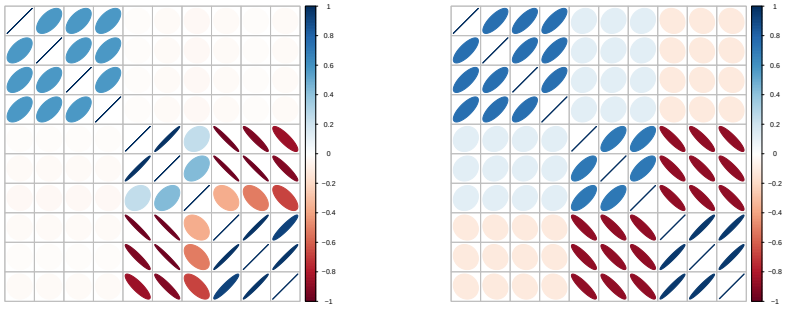
- The first model considers the two groups $\{1, \ldots, 4\}$ and $\{5, \ldots, 10\}$. The within-group structure is CS for the first group (linear functions). But a general structure is chosen for the second one (sinusoids), in order to capture its complex covariance structure.
- The second model is based on the three groups $\{1, \ldots, 4\}$, $\{5, \ldots, 7\}$ and $\{8, \ldots, 10\}$. The within-group structure is CS, and the between-group covariance is general.

For simplicity, we consider a single stratified design of experiments extracted from a sequence of regularly spaced points, with $m = 3$ points per level.

The estimated correlation parameters are shown in Fig. 5. The correlation structure that can be intuited from Fig. 4 is well recovered by the two models. However, in the case of two groups, the estimated between-group correlation is nearly zero. This is an illustration of the exclusion property (Proposition 3). Indeed, due to the strong negative (estimated) correlations within the second group, we have $\overline{W_2} \approx 0$, which induces a small correlation between the other group. In this example, the model with three groups may be more appropriate, which seems confirmed by the larger Q^2 value of 0.94 (compared to 0.88 for two groups). Nevertheless, it is nice to see that, starting from a smaller number of groups, the correlation plot detects the two subgroups of sinusoids.

5.3 Simulated data

In order to have a better intuition of the kind of applications that can be covered by group kernels, we have represented in Figure 6 a couple of sample paths obtained from a GP with a GCS kernel. We have chosen here a strong positive within-group correlation, corresponding to homogeneous groups. This indeed gives a common pattern for the sample paths corresponding to a same group, but also allows some specific behaviours.
That simulation setting provides a sound framework to assess the probability of discovering



(a) 2 groups (b) 3 groups

Figure 5: Estimated correlation kernel k_{cat} for the two GP models.

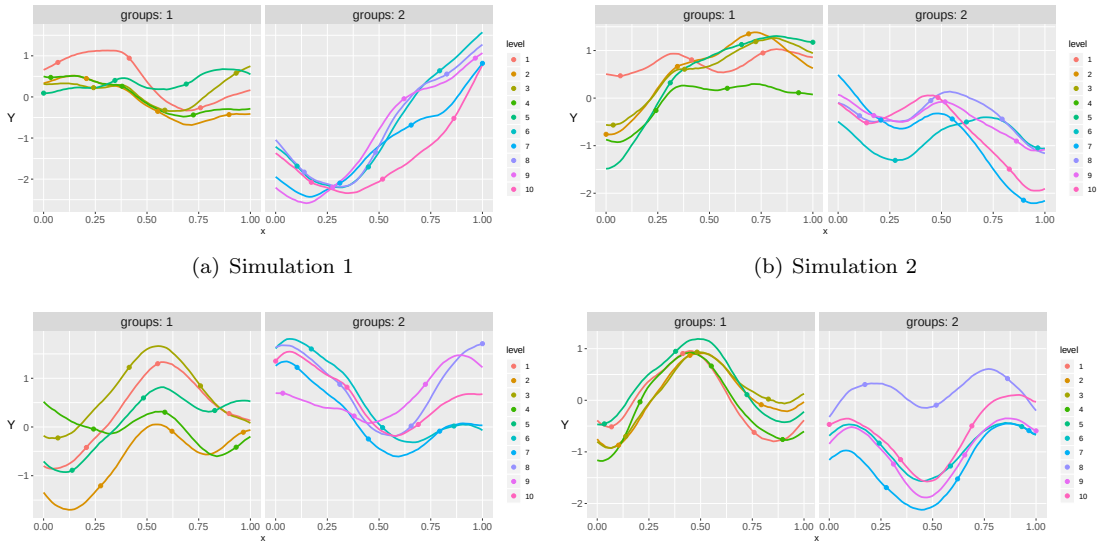
groups, as well as the prediction accuracy of GP models in terms of probability distribution. More precisely, starting from two given groups and given within- and between- group correlations, we simulate 100 sample paths. For each sample path, we extract a training set, formed by a random design of experiments and the corresponding observations. For simplicity, the design is obtained from a fixed stratified design of experiments extracted from a sequence of regularly spaced points, with three points per level; randomness is obtained by drawing uniformly the level corresponding to a strata. The grouping algorithm of Section 4.2 is then run, and various GP models are evaluated on a test set formed by a thin sequence of $[0, 1]$.

Some results are shown on Figure 7, corresponding to three values of between-group correlations ρ_{bet}. On each panel, we observe that the best performance is obtained for group kernels with true groups, which is logical since the simulated paths are based on them. Estimating the groups degrades a little the performance, but is better than using the low-rank kernel on which the algorithm is based. Considering now the difference between panels, a 95% confidence interval for the rate of misclassification is $[12\%, 18\%]$ for $\rho_{\text{bet}} = -0.5$, $[18\%, 25\%]$ for $\rho_{\text{bet}} = 0$ and $[26\%, 32\%]$ for $\rho_{\text{bet}} = 0.5$. This rate corresponds to the expectation of proportion of levels that are not correctly classified by the algorithm. Hence for $\rho_{\text{bet}} = -0.5$, between 1 and 2 levels (over 10) on average are not well classified, whereas between 2 and 3 are not for $\rho_{\text{bet}} = 0.5$. This result is expected since the groups are less separated when the between-group correlation increases. Hence, for $\rho_{\text{bet}} = 0.5$, the performance of group kernels gets close to that of a simple CS kernel, corresponding to one group. Recovering groups is also harder when the within-group correlation decreases since the groups are then less homogeneous.

Finally, notice the poor performance of the ordinal kernel (with order $1, 2, \ldots, 10$), which is simply due to the absence of natural level ordering on these simulated data.

5.4 Toy application: Beam bending problem

To illustrate group kernels, we consider the role of the cross-section shape in the Euler Bernoulli beam bending problem. Indeed, beams exhibit natural shapes (such as circular, rectangular, I-shape) with various filling configurations (e.g. hollow, medium hollow, solid). We consider



(a) Simulation 1

(b) Simulation 2

(c) Simulation 3

(d) Simulation 4

Figure 6: Four simulations of a GP model Z with tensor-product kernel $k(\mathbf{w}, \mathbf{w}') = k_{\text{cont}}(x, x')k_{\text{cat}}(u, u')$, where k_{cont} is a Matérn kernel with lengthscale $\theta = 0.4$ and k_{cat} is a GCS group kernel corresponding to $L = 10$ levels, split in two groups of size 5. The within-group correlation is fixed to 0.8 for the two groups, and the between-group correlation to -0.5. In each panel, the L curves correspond to a sample path of $x \mapsto Z(x, u)$ for $u = 1, \ldots, L$.

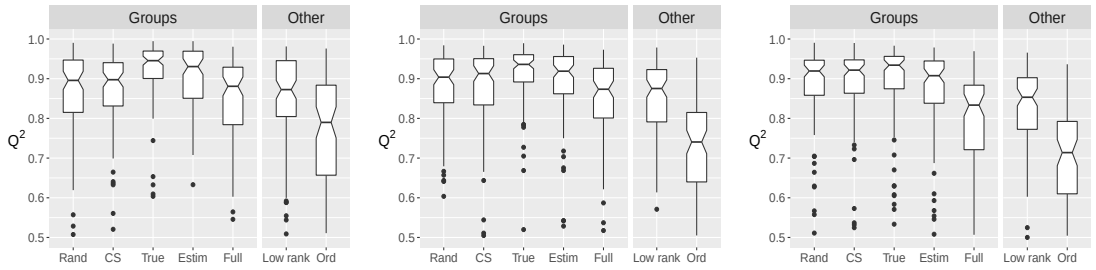
here twelve shapes represented in Figure 8, viewed as the levels of a categorical variable. The aim is to compare GP models for the vertical deflection of a horizontal beam, supposed to be the output of a black-box function. In such configuration where groups can be visually identified, we expect group kernels to be relevant.

Let us now go deeper into details. The beam is supposed to be fixed at one end, and a vertical force is locally applied to the other end. It can then be shown [7] that if the beam length is long enough compared to cross section dimensions, then, under linear elasticity assumption, an approximation of the beam vertical deflection at the free end is given by:

$$y(L, S, \tilde{I}) = \frac{PL^3}{3ES^2\tilde{I}} \tag{24}$$

where:

- L is the horizontal length of the beam,
- E is the Young modulus characterizing the material properties of the beam,
- P is the amplitude of the vertical loading,
- S is the area of the cross-section,
- $\tilde{I} = I/S^2$ is the moment of inertia I normalized by the cross section.



(a) $\rho_{\text{bet}} = -0.5$ (b) $\rho_{\text{bet}} = 0$ (c) $\rho_{\text{bet}} = 0.5$

Figure 7: Probability distributions of Q^2 of various GP models, based on a sample of 100 simulations drawn from a GP with a GCS kernel and random stratified designs. The settings of the simulations are the same as in Figure 6, with three different values of the between-group correlation: $\rho_{\text{bet}} = -0.5, 0, 0.5$. Each panel compares group kernels (2 random groups, 1 group (CS structure), 2 given groups, 2 estimated groups, 10 groups), a low rank kernel ($q = 2$) and an ordinal kernel (order $1, 2, \ldots, 10$) with normal warping. Number of parameters used in each categorical kernel (boxplot order): groups = (4, 2, 4, 4, 45), other = (20, 3).

The Young modulus and applied force are supposed to be constant: $E = 600\text{GPa}$, $P = 600\text{N}$. On the contrary, the beam dimensions can vary. Length L and section S are modeled by two continuous variables varying in the intervals $[10, 20]$ and $[1, 2]$ respectively. The normalized moment of inertia $\tilde{I}$ is modeled by a categorical variable with twelve levels corresponding to the cross-sections of Figure 8. Notice that it is a continuous quantity that can explicitly be derived for all the considered shapes. Indeed, from basic mechanics, we have:

$$\begin{aligned}(\tilde{I}_1, \ldots, \tilde{I}_{12}) = (&0.0833, 0.139, 0.380, 0.0796, 0.133, 0.363,\\ &0.0859, 0.136, 0.360, 0.0922, 0.138, 0.369).\end{aligned} \tag{25}$$

These values form three clusters equal to the groups visible on Figure 8, corresponding to hollow, medium hollow and solid shapes, $\{\tilde{I}_1, \tilde{I}_4, \tilde{I}_7, \tilde{I}_{10}\}$, $\{\tilde{I}_2, \tilde{I}_5, \tilde{I}_8, \tilde{I}_{11}\}$ and $\{\tilde{I}_3, \tilde{I}_6, \tilde{I}_9, \tilde{I}_{12}\}$.
This beam example is usually chosen as a toy example in computer experiments (see e.g. [34]), as the expression (24) provides a challenging non-linear function. However, looking at (24), there is no denying that applying a log transform to the output data would strongly simplify the modeling problem. But if this transform was applied, the model would become linear, and the interest of considering kernel-based prediction models would strongly decrease.
The following GP models are now compared. We underline that $\tilde{I}$ is assumed to be latent, i.e. unknown, otherwise a GP with pure continuous inputs $L, S, \tilde{I}$ will work well.

1. Low-rank approaches (low-rank and latent variable kernels).
2. Group kernel, with two choices:
   - "Shape-based" groups, formed by similarity between the different shapes: $\{\tilde{I}_1, \tilde{I}_2, \tilde{I}_3\}$, $\{\tilde{I}_4, \tilde{I}_5, \tilde{I}_6\}$, $\{\tilde{I}_7, \tilde{I}_8, \tilde{I}_9\}$ and $\{\tilde{I}_{10}, \tilde{I}_{11}, \tilde{I}_{12}\}$.

- "Filling-based" groups, formed by solid, medium hollow and hollow sections: $\{\tilde{I}_1, \tilde{I}_4, \tilde{I}_7, \tilde{I}_{10}\}$, $\{\tilde{I}_2, \tilde{I}_5, \tilde{I}_8, \tilde{I}_{11}\}$ and $\{\tilde{I}_3, \tilde{I}_6, \tilde{I}_9, \tilde{I}_{12}\}$.

3. Ordinal kernel. As $\tilde{I}$ is assumed unknown, we provide the partial order given by the filling-based groups, and sample at random an order inside each group.

The design of experiments is made of a sliced Latin hypercube, with $m = 3$ points per level. For group kernels, we have tried several parameterizations (see Section 4.1), and the best prediction performance was obtained by imposing constraints on the covariance matrices generating the usual GCS kernel with CS diagonal blocks: we chose the identity matrix for $\mathbf{M}_g$, and a homoscedastic symmetric matrix for $\mathbf{B}^\star$. We report only this case.
The results are summarized in Figure 9. We observe that the best prediction performances are obtained by group kernels (with filling-based groups), and low-rank approaches. Group kernels involve much less parameters (4 versus 11), an advantage that will magnify as the number of levels in each group increases. Giving an inappropriate order for group kernels (shape-based groups) degrades the performances, but not worse than for a CS kernel corresponding to a single group. The good score of low-rank techniques is logical since the deflection is explained by the latent continuous variable $\tilde{I}$. The ordinal kernels perform well, especially for a general piecewise affine warping. An inspection of that warping shows that two jumps are visible, corresponding to the filling-based groups (provided as a partial order). On the other hand, the Normal warping is not rich enough to recover this information.

(a) $\tilde{I}_1$ (b) $\tilde{I}_2$ (c) $\tilde{I}_3$ (d) $\tilde{I}_4$ (e) $\tilde{I}_5$ (f) $\tilde{I}_6$ (g) $\tilde{I}_7$ (h) $\tilde{I}_8$ (i) $\tilde{I}_9$ (j) $\tilde{I}_{10}$ (k) $\tilde{I}_{11}$ (l) $\tilde{I}_{12}$

Figure 8: Representation of the shapes considered for the cross-sections. The scale differs from one picture to another, as the areas are supposed to be the same for each cross-section.

6 Application in nuclear engineering

6.1 Position of the problem

As presented in Introduction, this research is originally motivated by the solving of an inverse problem confronting experimental measurements in nuclear engineering and time-consuming numerical simulation. More precisely, this analysis concerns the identification of the mass m of ^{239}Pu that is present in a particular waste container using a non-destructive nuclear detection technique such as the gamma spectrometry [13]. In that case, at each energy level E,

$$m \times \epsilon(E; \mathcal{E}) = y(E), \tag{26}$$

where $y(E)$ is the quantity of interest provided by the gamma transmitter, and $\epsilon(E; \mathcal{E})$ is the attenuation coefficient, which depends on the source environment denoted by $\mathcal{E}$. In practice, only discrete values of E are of interest, corresponding to the natural energy levels of ^{239}Pu:

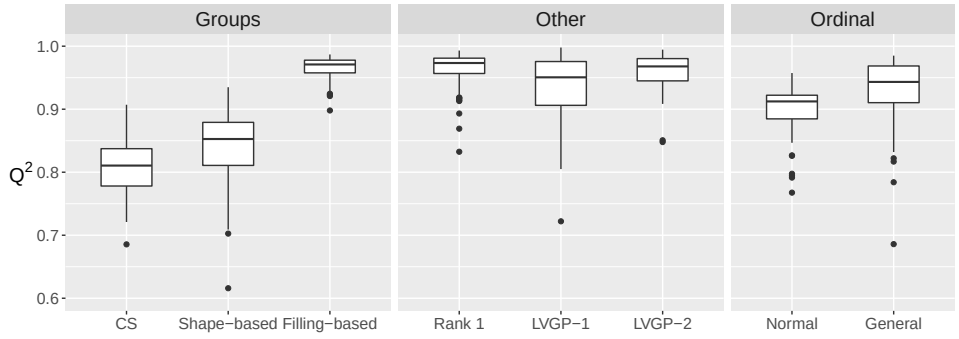


Figure 9: Q^2 of various GP models, based on 100 repetitions of the design. First panel (group kernels): 1 group (CS structure), 4 groups given by the shape of the cross-section, 3 groups given by its filling (hollow to solid). Second panel: low-rank ($q = 1$) and LVGP ($q = 1$, $q = 2$). Third panel (ordinal kernels): Partial order given by cross-section filling, with two warpings (Normal, piecewise affine). Number of parameters used in each categorical kernel (boxplot order): groups = (2, 7, 4), other = (12, 11, 21), ordinal = (4, 13).

$$E \in \{94.66, 129.3, 203.6, 345.0, 375.1, 413.7\} \text{ (keV)}. \tag{27}$$

Then, based on previous studies [10], the real source environment is parameterized by the following input variables:

- An equivalent geometric shape for the nuclear waste: sphere ('sph'), cylinder ('cyl') or parallelepiped ('par').
- An equivalent material for this waste, characterized by its chemical element with atomic number in $\{1, \ldots, 94\}$,
- The bulk density of the waste, in $[0, 1]$,
- The distance of measurement between the container and the measurement device, in $[80, 140]$ (cm),
- The mean width and lateral surfaces (in logarithmic scale) crossed by a gamma ray during the rotation of the object.

After normalization, the characteristics of the input space can be summed up in Table 2.

To recapture the notation of the previous sections, let $\mathbf{x}$ and $\mathbf{u}$ be the vectors gathering respectively the continuous and categorical inputs, and $\mathbf{w} = (\mathbf{x}, \mathbf{u})$. For a given value of $\mathbf{w}$, Monte Carlo simulation codes as MCNP [8] can be used to model the measured scene and approach the value of $\epsilon(\mathbf{w}) = \epsilon(E, \mathcal{E})$. The mass m can eventually be searched as the solution of the following optimization problem:

$$(m^\star, \mathbf{w}^\star) = \arg\min_{m,\mathbf{w}} \|\boldsymbol{y}^{\text{obs}} - m \times \boldsymbol{\epsilon}(\mathbf{w})\|, \tag{28}$$

Name of the input	Variation domain
Distance	[0,1]
Density	[0,1]
Width	[0,1]
Surface	[0,1]
Energy	$\{1, 2, 3, 4, 5, 6\}$
Shape	{sph, cyl, par}
Chemical element	$\{1, \ldots, 94\}$

Table 2: Input variables for the application.

where $\|\cdot\|$ is the classical Euclidean norm, $\boldsymbol{\epsilon}(\mathbf{w})$ and $\boldsymbol{y}^{\text{obs}}$ respectively gather the values of ϵ and y at the six values of E that are used for the measurements. To solve Eq. (28), it is therefore necessary to compute $\boldsymbol{\epsilon}$ at a high number of points. However, each evaluation of the MCNP code can be extremely demanding (between several minutes to several hours CPU for one evaluation). Thus, surrogate models have to be introduced to emulate the function $\mathbf{w} \mapsto \epsilon(\mathbf{w})$, which is now investigated in the frame of Gaussian process regression. We refer to [3] for the second step, namely the treatment of the inversion problem.

6.2 Model settings

For pedagogical purpose, a dataset of large size $N = 5076$ has been computed with the MNCP code. The construction of the design of experiments was guided by the categorical inputs, such that each of the $6 \times 3 \times 94 = N/3$ combinations of levels appears 3 times. It was completed by a Latin hypercube of size N to define the values of the four continuous inputs.
From this full dataset, a training set of size $n = 3 \times 94 = 282$ is extracted by selecting at random 3 observations by chemical element. The remaining $N - n$ points serve as test set.

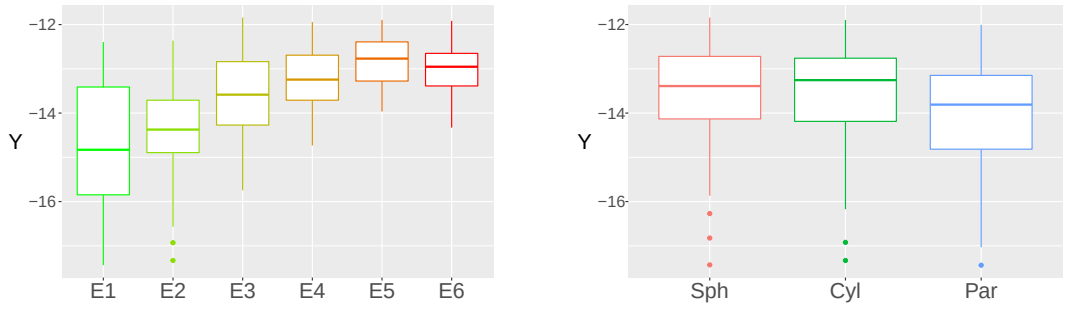


Figure 10: Y in function of the energy (left) and geometric shape (right).

Model settings are now motivated by a graphical analysis. In Fig. 10, the output is displayed in function of the energy and the geometric shape. We observe that successive energy levels correspond to close values. This fact confirms that the energy is ordinal and we use the warped

kernel defined by Eq. (5). The influence of the geometric shape is less obvious, and we have chosen an exchangeable (CS) covariance structure for it.
In Fig. 11, Y is displayed in function of the 94 chemical elements, ordered by atomic number. Two important facts are the high number of levels and heteroscedasticity. For this purpose, the 94 chemical elements are divided into 5 groups, provided by expert knowledge and represented by colors. To obtain the groups, experts in nuclear physics considered families of chemical elements ordered by their atomic number, such as gaz, metals. They also drew attenuation curves and gathered elements showing similar patterns. Although an order can be assumed between groups, it was not clear, a priori, if making this order assumption for chemical elements was reasonable (in particular various phenomena were observed for the heaviest chemical elements). In the following we will compare the results obtained by considering this assumption or not. Due to the large number of levels, only the most parsimonious kernels can be used: ordinal kernels parameterized by a Normal warping or group kernels. The partition suggests to use a group kernel of the form Eq. (13), where the within-group blocks W_g are CS covariance matrices. In order to handle heteroscedasticity, the variance of W_g is assumed to depend on the group number g.

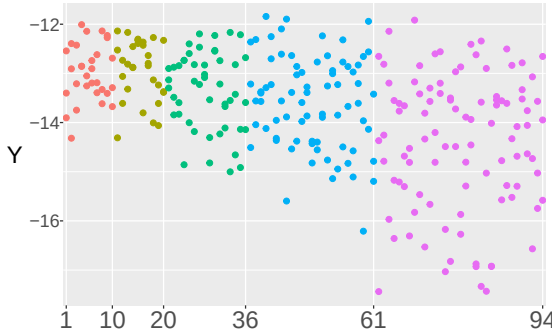


Figure 11: Y in function of chemical elements, ordered by atomic number.

The influence of continuous variables can be observed by panels (not represented), and does not reveal useful information for our purpose. A Matérn 5/2 kernel is set for all continuous inputs, as we expect the output to be a regular function of the continuous inputs. Indeed, for this kernel, the corresponding Gaussian process is two times mean-square differentiable.
For categorical inputs, we have chosen this kernel as well in the warping formula Eq. (5), which seems appropriate since the estimated between-group correlations are clearly positive. Finally, three candidate kernels for $\mathbf{w}$ are obtained by combining the kernels of input variables defined above, by sum, product or ANOVA (see Section 2).

6.3 Results

Following the model settings detailed above, Fig. 12, Panel 3, presents the results obtained with 60 random designs of size n and three operations on kernels. Furthermore, we have implemented three other kernels for the chemical element, in order to compare other model choices for this

categorical input. In the first panel, we grouped all the 94 levels in a single group. In the second one, we kept the 5-group kernel but forced the between-group covariances to have a common value. Finally, in the fourth panel, we considered that the levels were ordered by their atomic number, and used the warped kernel of Eq. (5) with a Normal transform.

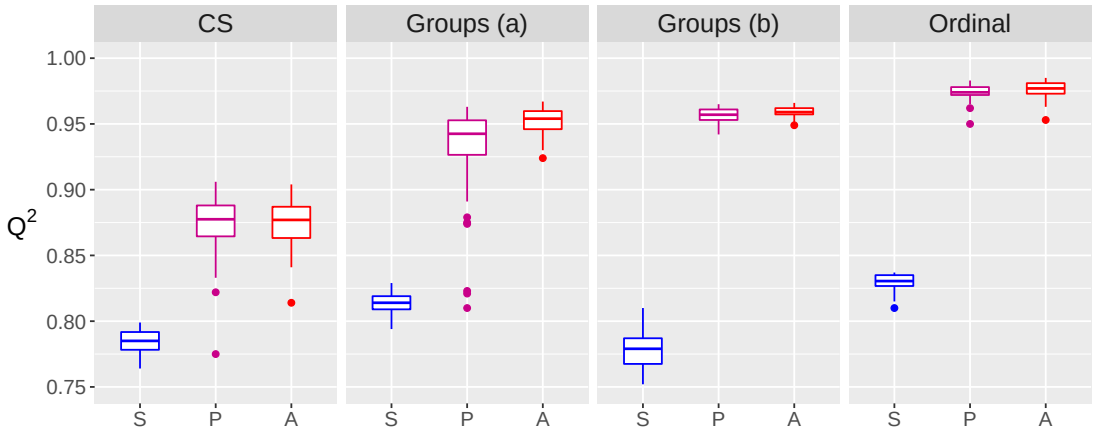


Figure 12: Q^2 of several GP models, based on 60 random designs, corresponding to different model choices for the chemical element. First panel: Single group. Second panel: 5 groups, with a common between-group covariance. Third panel: 5 groups. Fourth panel: Ordered levels. For each panel, three combinations of kernel are tested: sum (S), product (P) or ANOVA (A). Total number of parameters used (panel order): 'prod' = (12, 21, 30, 14), 'sum' = 'prod' + 6, 'anova' = 'prod' + 7.

First, comparing the three operations on kernels, we remark that in all the panels, additive kernels provide the worst results. This suggests the existence of interactions between different inputs of the simulator. Second, the ANOVA combination produces slight improvements, compared to the standard tensor-product, both in terms of accuracy and stability with respect to design choice.

Now, comparing the four panels, we see that gathering the levels in a single group is the least efficient strategy. The 5-group kernel gives very good performances, especially when the between-group covariances vary freely: constraining them to be equal degrades the result. Surprisingly here, the ordinal kernel gives the best performance. Indeed, for this application it was not intuitive to the experts that the chemical elements can be treated as an ordinal variable sorted by its atomic number, since the effect of the order on the output was trusted rather for family of elements (i.e. groups). This is confirmed by the correlation plots of Fig. 14, corresponding to a model with a median Q^2 score. We can see that the estimated correlations between levels seems to decrease as the difference between levels increases, an indication that the levels may be ordered by their atomic number.

We further investigate robustness to group and order misspecification in Figure 13. We can see that choosing five groups at random (with random sizes) gives similar accuracy results than gathering all levels in one group. On the other hand, choosing a wrong order can be more detrimental, either for a random order or when the chemical elements are sorted according to

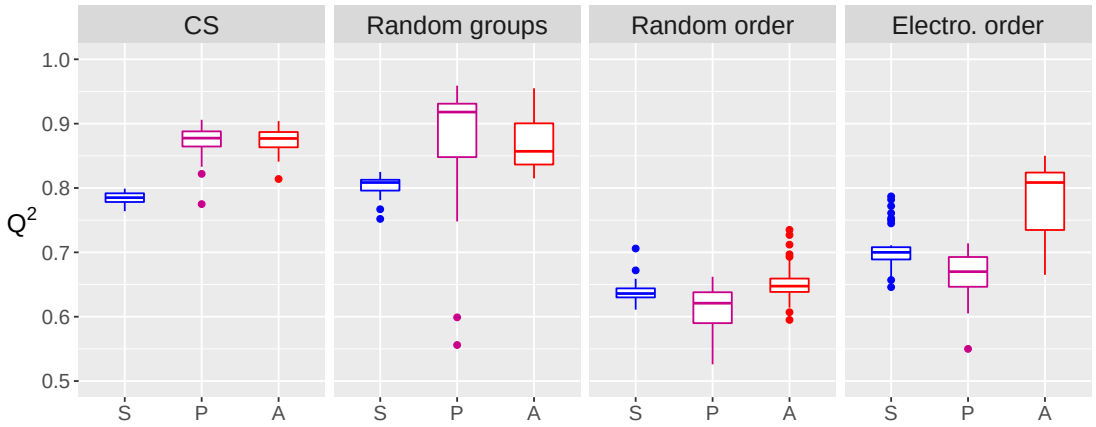


Figure 13: Q^2 of several GP models, based on 60 random designs, corresponding to several misspecified GP models for the chemical element. First panel: Single group. Second panel: 5 groups, chosen at random. Third panel: Ordered levels with a random order. Fourth panel: Ordered levels according to Pauling electronegativity. For each panel, three combinations of kernel are tested: sum (S), product (P) or ANOVA (A).

Pauling electronegativity[1] (see e.g. [21]). Thus in this case, group kernels seem more robust than ordinal kernels to (group / order) misspecification.
Then, we report several post-processing results. First, the estimated transformation of energy levels (Fig. 15, left) is concave and flat near high values, which corresponds to the behaviour observed in Fig. 10 (left panel). In addition, the last three levels lead to similar results (Fig. 15, right). This corresponds to the fact that when the energy is high, the gamma ray almost always crosses the nuclear waste, leading to a high value for the output. Second, the estimated correlation among the sphere, the cylinder and the parallelepiped is very high ($c = 0.9$, Fig. 16). This justifies considering a covariance structure for that categorical input, rather than using three independent GP models for all the three levels.
Finally, although the aim of the paper is to investigate a new class of GP models, it is useful to situate their performances with respect to standard models from machine learning. As an example, we have estimated a random forest (R package `randomForest` [15]), tuned by cross validation (R package `caret` [14]), either by a) treating the atomic number as nominal, or b) as ordinal (by considering its levels as integers). The Q^2 (median on 60 random designs) are respectively 0.61 for a) and 0.89 for b).

7 Conclusion

In the framework of GP regression with both continuous and categorical inputs, we focus on problems where categorical inputs may have a potentially large number of levels L, partitioned in $G \ll L$ groups of various sizes. We provide new results about parsimonious block covariance

[1] We used the data provided on: https://www.lenntech.com/periodic-chart-elements/electronegativity.htm

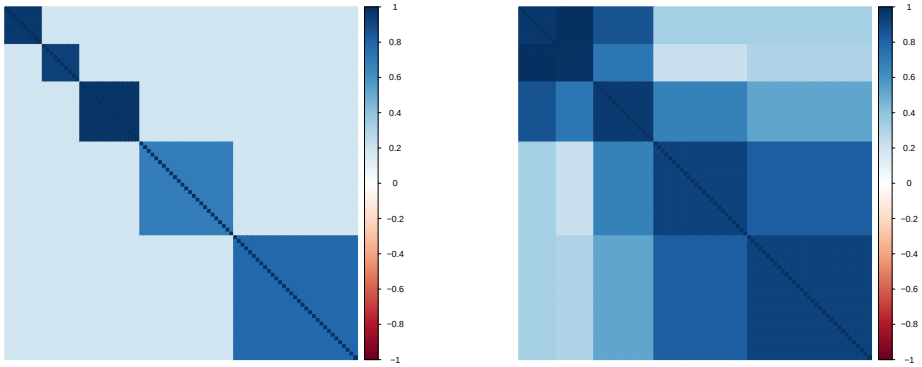


Figure 14: Estimated correlation kernel for the chemical element, with a common between-group parameter (left) or different ones (right).

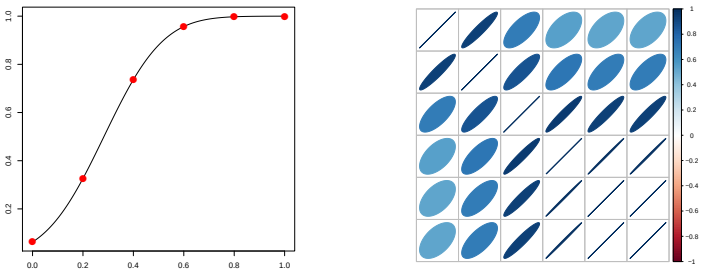


Figure 15: Estimated kernel for the energy: estimated warping (left) and correlation structure (right).

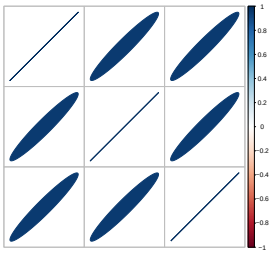


Figure 16: Estimated correlation kernel for the geometric shape.

matrices, defined by a few within- and between-group covariance parameters.

We revisit a two-way nested Bayesian linear model, where the response term is defined as a

sum of a group effect and a level effect. We obtain a flexible parameterization of block covariance matrices which automatically satisfy the positive definiteness conditions. As a particular case, we recover situations where the within-group covariance structures are compound symmetry, with possible negative correlations. Furthermore, we show that the positive definiteness of a given block covariance matrix can be checked by verifying that the small matrix of size G obtained by averaging each block is positive definite. This criterion can be useful if the proposed block matrix has a desirable constraint, such as homoscedasticity, which is not directly handled by the proposed parameterization.

We apply these findings on several toy functions as well as an application in nuclear engineering, with 4 continuous inputs, 3 categorical inputs, one of them having 94 levels corresponding to chemical numbers in Mendeleev's table. In this application, 5 groups were defined by experts. The results, measured in terms of prediction accuracy, outperform those obtained with simpler assumptions, such as gathering all levels into a single group. It is gratifying that our nominal method performs almost as well as using the appropriate order with a warped kernel.

Thanks to these first examples, we can list pros and cons for group kernels. We can cite two main reasons to use group kernels. First, group kernels easily handle information on group of levels, which is sometimes provided by physical knowledge (such as for chemical elements) or due to a latent ordinal variable (e.g. in the beam bending problem). Second, group kernels form the only class of kernels for nominal variables which can be used for a large number of levels, since its complexity only depends on the number of groups. It is thus a useful alternative to ordinal kernels, and can be used to question an order assumption.
About limitations now, we can mention two. First, when there is a true and total order for a categorical variable, we cannot expect group kernels to improve prediction accuracy by partitioning the levels in groups. Second, when the groups are not provided, it may be hard to recover them in full generality, especially when groups are strongly positively correlated, a difficult case for clustering. At least, we have provided a first algorithm toward this direction, designed for a small number of groups, which detected the groups of Example 5.1.

There are several perspectives for this work. First, one future direction is to further investigate a data-driven technique to recover groups of levels, made more difficult when a small number of observations is available. Similarly, if there is an order between levels, can we infer it from the data? Second, the trend of the GP models with mixed continuous and categorical inputs could be made more complex, in the same vein as works on GP models with continuous inputs.

A Proofs

Proof 1 (Proof of Proposition 1) *The vector* $(\boldsymbol{\lambda}, \lambda_1 + \cdots + \lambda_L)$ *is a centered Gaussian vector with covariance matrix*

$$v_\lambda \begin{pmatrix} \mathbf{I}_L & \mathbf{1}_L \\ \mathbf{1}_L^\top & L \end{pmatrix}.$$

Hence the conditional distribution of $\boldsymbol{\lambda}$ *knowing* $\overline{\lambda} = 0$ *is a centered Gaussian vector with covariance matrix*

$$\mathrm{cov}(\boldsymbol{\lambda} \,|\, \overline{\lambda} = 0) = v_\lambda[\mathbf{I}_L - \mathbf{1}_L L^{-1} \mathbf{1}_L^\top] = v_\lambda[\mathbf{I}_L - L^{-1}\mathbf{J}_L].$$

Then, by using the independence between μ *and the* λ_u*'s, we deduce*

$$\begin{aligned} \mathrm{cov}(\boldsymbol{\eta} \,|\, \overline{\lambda} = 0) &= v_\mu \mathbf{J}_L + v_\lambda[\mathbf{I}_L - L^{-1}\mathbf{J}_L] \\ &= v_\lambda \mathbf{I}_L + [v_\mu - L^{-1} v_\lambda]\mathbf{J}_L. \end{aligned}$$

We recognize the CS covariance matrix $\boldsymbol{\Gamma}_L^{\mathrm{CS}}(v, c)$ *with* $v = v_\mu + (1 - L^{-1})v_\lambda$ *and* $c = v_\mu - L^{-1}v_\lambda$*. As a covariance matrix, it is positive semidefinite. Furthermore, we have* $c < v$ *and* $c + (L-1)^{-1}v = v_\mu[1 + (L-1)^{-1}] > 0$*, and the conditions of positive definiteness Eq. (16) are satisfied. Conversely, let* $\mathbf{C}$ *be a positive definite CS matrix* $\boldsymbol{\Gamma}_L^{\mathrm{CS}}(v, c)$*. Then we have* $-(L-1)^{-1}v < c < v$*, and we can define* $v_\mu = L^{-1}[v + (L-1)c]$ *and* $v_\lambda = v - c$*. From the direct sense, we then obtain that the covariance matrix of* $\boldsymbol{\eta} \,|\, \overline{\lambda} = 0$ *is* $\boldsymbol{\Gamma}_L^{\mathrm{CS}}(v, c) = \mathbf{C}$*.*

Proof 2 (Proof of Proposition 2) *The first part of the proposition is obtained by remarking that if* $\mathbf{W}^\star = \mathrm{cov}(\mathbf{z})$*, then* $\overline{W^\star} = \mathrm{var}(\overline{z})$*. Thus, assuming that* $\mathbf{z}$ *is centered,* $\overline{W^\star} = 0$ *is equivalent to* $\overline{z} = 0$ *with probability* 1*.*
For the second part, notice that $\overline{z} = 0$ *means that* $\mathbf{z}$ *is orthogonal to* $\mathbf{1}_L$*. Thus, one can write the expansion of* $\mathbf{z}$ *in the orthonormal basis* $\mathbf{1}_L^\perp$ *defined by* $\mathbf{A}$*. Denoting by* $\mathbf{t}$ *the* $(L-1)$*-vector of coordinates, we have* $\mathbf{z} = \mathbf{A}\mathbf{t}$*. This gives* $\mathbf{W}^\star = \mathrm{cov}(\mathbf{A}\mathbf{t}) = \mathbf{A}\,\mathrm{cov}(\mathbf{t})\mathbf{A}^\top$*, and Eq. (18) follows with* $\mathbf{M} = \mathrm{cov}(\mathbf{t})$*.*
To prove uniqueness, observe that, by definition, $\mathbf{A}^\top\mathbf{A} = \mathbf{I}_{L-1}, \mathbf{A}^\top\mathbf{1}_L = 0$*. Starting from* $\mathbf{W}^\star = \mathbf{A}\mathbf{M}\mathbf{A}^\top$*, and multiplying by* $\mathbf{A}^\top$ *on the left and by* $\mathbf{A}$ *on the right, we get* $\mathbf{M} = \mathbf{A}^\top\mathbf{W}^\star\mathbf{A}$*, showing that* $\mathbf{M}$ *is unique.*
Now, let $\mathbf{W}^\star = v[\mathbf{I}_L - L^{-1}\mathbf{J}_L]$*. Since* $\mathbf{J}_L = \mathbf{1}_L\mathbf{1}_L^\top$*, we obtain*

$$\mathbf{M} = \mathbf{A}^\top\mathbf{W}^\star\mathbf{A} = v[\mathbf{A}^\top\mathbf{A} - L^{-1}(\mathbf{A}^\top\mathbf{1}_L)(\mathbf{1}_L^\top\mathbf{A})] = v\mathbf{I}_{L-1}.$$

As a by-product, notice that resubstituting $\mathbf{M}$ *into* $\mathbf{W}^\star = \mathbf{A}\mathbf{M}\mathbf{A}^\top$ *gives* $\mathbf{A}\mathbf{A}^\top = \mathbf{I}_L - L^{-1}\mathbf{J}_L$*. Finally, if* $\mathbf{z} \sim \mathcal{N}(0, v\mathbf{I}_L)$*, then the properties of conditional Gaussian vectors lead immediately to* $\mathrm{cov}(\mathbf{z}|\overline{z} = 0) = \mathbf{W}^\star$*.*

Proof 3 (Proof of Theorem 1) *The expressions of* $\mathbf{W}_g$ *and* $\mathbf{B}_{g,g'}$ *are obtained directly by using the independence assumptions about* $\boldsymbol{\mu}$ *and the* $\boldsymbol{\lambda}$*'s. Notice that* $\mathbf{W}_g^\star$*, the covariance matrix of* $\boldsymbol{\lambda}_{g/\cdot}$ *knowing* $\overline{\lambda_{g/\cdot}} = 0$*, is centered by Proposition 2. This gives* $\mathbf{W}_g - \overline{W_g}\mathbf{J}_{n_g} = \mathbf{W}_g^\star$*, which is positive semidefinite. Hence* $\mathbf{T}$ *is a GCS block matrix.*

Conversely, let $\mathbf{T}$ *be a positive semidefinite GCS block matrix. Let* $\widetilde{\mathbf{T}}$ *be the matrix obtained from* $\mathbf{T}$ *by averaging each block. Then* $\widetilde{\mathbf{T}}$ *is also a positive semidefinite matrix. Indeed, since* $\mathbf{T}$ *is positive semidefinite, it is the covariance matrix of some vector* $\mathbf{z}$*. Then* $\widetilde{\mathbf{T}}$ *is the covariance matrix of* $\widetilde{\mathbf{z}}$*, the vector obtained from* $\mathbf{z}$ *by averaging by group:* $\widetilde{z}_g = n_g^{-1}\sum_{u \in G_g} z_u$*. Thus there exists a centered Gaussian vector* $(\mu_g)_{1 \le g \le p}$ *whose covariance matrix is*

$$\mathbf{B}^\star = \widetilde{\mathbf{T}}.$$

Now, for $g = 1, \ldots, G$*, define*

$$\mathbf{W}_g^\star = \mathbf{W}_g - B_{g,g}^\star \mathbf{J}_{n_g} = \mathbf{W}_g - \overline{W_g}\mathbf{J}_{n_g}.$$

Observe that $\overline{W_g^\star} = 0$, *and by assumption* $\mathbf{W}_g^\star$ *is positive semidefinite. Hence, from Proposition 2, there exists a centered Gaussian vector* $(\lambda_{g/j})_{1\leq u\leq n_g}$ *such that*

$$\mathbf{W}_g^\star = \mathrm{cov}(\boldsymbol{\lambda}_{g/.}|\overline{\lambda_{g/.}} = 0).$$

We can assume that $\boldsymbol{\lambda}_{1/.}, \ldots, \boldsymbol{\lambda}_{G/.}$ *are independent, and* $\boldsymbol{\mu}$ *and* $\boldsymbol{\lambda}$ *are independent. Finally, we set* $\eta_{g/u} = \mu_g + \lambda_{g/u}$. *By the direct sense and Eq. (20), we obtain that* $\mathbf{T}$ *is the covariance matrix of* $\boldsymbol{\eta}$ *conditional on* $\{\overline{\lambda_{g/.}} = 0,\, g = 1, \ldots, G\}$.

Proof 4 (Proof of Theorem 2) *Let* $\mathbf{T}$ *be a positive semidefinite GCS block matrix with CS diagonal blocks. Then the diagonal CS matrices are positive semidefinite, leading to* $v_g - c_g \geq 0$. *Thus,*

$$\mathbf{W}_g - \overline{W_g}\mathbf{J}_{n_g} = (v_g - c_g)(\mathbf{I}_{n_g} - n_g^{-1}\mathbf{J}_{n_g})$$

is a positive semidefinite CS matrix. Hence, by Theorem 1, $\mathbf{T}$ *is obtained from Model Eq. (19), with* $\mathrm{cov}(\boldsymbol{\lambda}_{g/.}|\overline{\lambda_{g/.}} = 0) = (v_g - c_g)(\mathbf{I}_{n_g} - n_g^{-1}\mathbf{J}_{n_g})$. *By Proposition 2 (last part), we can choose* $\mathbf{W}_g^\star = v_{\lambda_g}\mathbf{I}_{n_g}$, *with* $v_{\lambda_g} = v_g - c_g \geq 0$.
Conversely, if $\mathbf{W}_g^\star = v_{\lambda,g}\mathbf{I}_{n_g}$, *then by Proposition 2,* $\mathbf{W}_g^\star = \mathrm{cov}(\boldsymbol{\lambda}_{g/.}|\overline{\lambda_{g/.}} = 0)$ *is a CS covariance matrix. The result follows by Theorem 1.*

Proof 5 (Proof of Theorem 3) *First observe that Equation (21) is straightforward.*
The direct sense of (i) has already been derived in the proof of Theorem 1: $\widetilde{\mathbf{T}}$ *is the covariance matrix of* $\widetilde{\mathbf{z}}$. *Conversely, inspecting that proof, we see that if* $\widetilde{\mathbf{T}}$ *is positive semidefinite, then* $\mathbf{T}$ *admits the representation Eq. (19). Thus* $\mathbf{T}$ *is a covariance matrix, and positive semidefinite. This can be also proved from Eq. (21): the two terms of the right-hand side are positive semidefinite, thus so is their sum.*
Now consider (ii). If $\mathbf{T}$ *is positive definite, then its diagonal blocks* $\mathbf{W}_g$ *are all positive definite. Furthermore, by (i),* $\widetilde{\mathbf{T}}$ *is positive semidefinite. It if were singular, there would exist a non-zero vector* $\boldsymbol{\gamma}$ *such that* $\boldsymbol{\gamma}^\top\widetilde{\mathbf{z}} = 0$ *with probability* 1. *This gives a non-trivial linear combination of* $\mathbf{z}$ *which is equal to zero, and* $\mathbf{T}$ *would not be positive definite. Thus* $\widetilde{\mathbf{T}}$ *is positive definite.*

Conversely, let us assume that $\widetilde{\mathbf{T}}$ *and all* $\mathbf{W}_g$*'s are positive definite. We will need the following Lemma:*

lemma 1 *Let* $\mathbf{F}$ *be a centered covariance matrix of size* n, *with rank* $n-1$. *If for some vector* $\boldsymbol{\beta}$, *we have* $\boldsymbol{\beta}^\top\mathbf{F}\boldsymbol{\beta} = 0$, *then* $\boldsymbol{\beta}$ *is a constant vector.*

Proof 6 *A symmetric square root* $\mathbf{L}$ *with* $\mathbf{L}^2 = \mathbf{F}$ *has also rank* $n-1$, *hence the nullspace of* $\mathbf{L}$ *is one-dimensional. Since* $\mathbf{F}$ *is centered, we have* $\mathbf{1}_n^\top\mathbf{F}\mathbf{1}_n = 0$. *Thus* $\mathbf{L}\mathbf{1}_n = 0$. *Similarly,* $\mathbf{L}\boldsymbol{\beta} = 0$. *The result follows.*

Now, let $\boldsymbol{\beta}$ *be a vector of length* $\sum_g n_g$ *such that* $\boldsymbol{\beta}^\top\mathbf{T}\boldsymbol{\beta} = 0$. *Equation (21) gives*

$$\boldsymbol{\beta}^\top\mathbf{T}\boldsymbol{\beta} =: U + V$$

where U *and* V *are obtained from each term in Eq. (21) by left-multiplying by* $\boldsymbol{\beta}^\top$ *and right-multiplying by* $\boldsymbol{\beta}$. *Since* U *and* V *are non-negative, we must have* $U = V = 0$.

- *$V = 0$ implies that all subvectors $\boldsymbol{\beta}_g$ corresponding to groups g verify $\boldsymbol{\beta}_g^\top [\mathbf{W}_g - \overline{W_g}\mathbf{J}_{n_g}]\boldsymbol{\beta}_g = 0$. This implies that they are all constant vectors. Indeed, since (by assumption) $\mathbf{W}_g$ has rang n_g and $\overline{W_g}\mathbf{J}_{n_g}$ has rank 1, the centered matrix $\mathbf{F}_g := \mathbf{W}_g - \overline{W_g}\mathbf{J}_{n_g}$ must have rank $n_g - 1$, and the result is given by the lemma.*

- *$U = 0$ gives $\mathbf{X}^\top \boldsymbol{\beta} = 0$ by positive definiteness of $\widetilde{\mathbf{T}}$. Now, $\mathbf{X}^\top \boldsymbol{\beta}$ is a vector of length G whose component g is the sum of $\boldsymbol{\beta}_g$ coefficients.*

Gathering the conclusions of the two items gives $\boldsymbol{\beta} = 0$. Finally $\mathbf{T}$ is positive definite.

Remark *Notice that for (ii) we needed to add the condition that $\mathbf{W}_g$ is positive definite for all $g = 1, \ldots, G$. However, adding an equivalent condition for (i), namely that $\mathbf{W}_g$ is positive semidefinite, was not necessary. Indeed, it is a consequence of the fact that $\mathbf{W}_g - \overline{W_g}\mathbf{J}_{n_g}$ is positive semidefinite and that $\widetilde{\mathbf{T}}$ is positive semidefinite, which implies $\widetilde{T}_{g,g} = \overline{W_g} \geq 0$.*

Proof 7 (Proof of Proposition 3) *By assumption, $\widetilde{T}_{g,g} = \overline{W_g} = 0$. Now by Theorem 3, $\widetilde{\mathbf{T}}$ is positive semidefinite. As its diagonal term $\widetilde{T}_{g,g}$ is zero, it implies that all the terms on the same row are zero. Hence for all $g' \neq g$: $\widetilde{T}_{g,g'} = 0$. As $\mathbf{T}$ is a GCS covariance matrix, its off-diagonal blocks are constant, and thus, $T_{g,g'} = \widetilde{T}_{g,g'} = 0$, which proves that $\mathbf{y}_g$ and $\mathbf{y}_{-g}$ are non correlated. The result follows by Gaussianity of $\mathbf{y}$.*

Proof 8 (Proof of the assertion of Section 4.1) *We claim that $\mathbf{T}$ is invertible iff $\mathbf{B}^\star$ and all the $\mathbf{M}_g$'s are invertible. Indeed, from Theorem 3, $\mathbf{T}$ is invertible iff $\widetilde{\mathbf{T}}$ and all the $\mathbf{W}_g$'s are invertible. Now, from Theorem 1, $\widetilde{\mathbf{T}} = \mathbf{B}^\star$, and*

$$\mathbf{W}_g = [\mathbf{B}^\star]_{g,g}\, \mathbf{J}_{n_g} + \mathbf{A}_g \mathbf{M}_g \mathbf{A}_g^\top,$$

where $\mathbf{A}_g$ is a n_g by $n_g - 1$ matrix whose colums vectors are an orthonormal basis of $\mathbf{1}_{n_g}^\perp$. Left multiplying by $\mathbf{A}_g^\top$ and right multiplying by $\mathbf{A}_g$, and remarking that $\mathbf{J}_{n_g} = \mathbf{1}_{n_g}\mathbf{1}_{n_g}^\top$, we obtain: $\mathbf{A}_g^\top \mathbf{W}_g \mathbf{A}_g = \mathbf{M}_g$. It is now clear that $\mathbf{W}_g$ is invertible iff $\mathbf{M}_g$ is.

Software and acknowledgements

Implementations have been done with the R packages `mixgp`, `kergp` [5] and `LVGP` [29]. Illustrations use `ggplot2` [32] and `corrplot` [31]. We thank two reviewers and an associate editor for their useful remarks that led to substantial improvement on the initial version. This research was conducted within the frame of the Chair in Applied Mathematics OQUAIDO, gathering partners in technological research (BRGM, CEA, IFPEN, IRSN, Safran, Storengy) and academia (CNRS, Ecole Centrale de Lyon, Mines Saint-Etienne, University of Grenoble, University of Nice, University of Toulouse) around advanced methods for Computer Experiments. The authors thank the participants for fruitful discussions, and especially F. Gamboa and J.M. Azaïs. They are also grateful to the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme UNQ when work on this paper was undertaken (EPSRC grant number EP/K032208/1).

References

[1] D. Bates, M. Mächler, B. Bolker, and S. Walker, *Fitting Linear Mixed-Effects Models using lme4*, arXiv e-prints, (2014), arXiv:1406.5823, p. arXiv:1406.5823, https://arxiv.org/abs/1406.5823.

[2] C. Chevalier, J. Bect, D. Ginsbourger, E. Vazquez, V. Picheny, and Y. Richet, *Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set*, Technometrics, 56 (2014), pp. 455–465.

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

[4] X. Deng, C. D. Lin, K.-W. Liu, and R. K. Rowe, *Additive Gaussian process for computer models with qualitative and quantitative factors*, Technometrics, 59 (2017), pp. 283–292.

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

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

[7] J. Gere and S. Timoshenko, *Mechanics of Materials*, PWS Publishing Company, 1997.

[8] J. T. Goorley, M. Fensin, and G. McKinney, *MCNP6 users manual, Version 1.0*, May 2013.

[9] J. C. Gower, *Euclidean distance geometry*, The Mathematical Scientist, 7 (1982), pp. 1–14.

[10] N. Guillot, *Quantification gamma de radionucléides par modélisation équivalente*, PhD thesis, Université Blaise Pascal - Clermont-Ferrand II, France, 2015.

[11] S. G. Johnson, *The NLopt nonlinear-optimization package*, 2014, http://ab-initio.mit.edu/nlopt. R package version 1.0.4.

[12] A. Khuri and I. Good, *The parameterization of orthogonal matrices : a review mainly for statisticians : review paper*, South African Statistical Journal, 23 (1989), pp. 231–250.

[13] G. F. Knoll, *Germanium Gamma-ray detectors*, vol. 3, John Wiley & Sons, 2010.

[14] M. Kuhn., *caret: Classification and Regression Training*, 2017, https://CRAN.R-project.org/package=caret. R package version 6.0-77.

[15] A. Liaw and M. Wiener, *Classification and regression by randomforest*, R News, 2 (2002), pp. 18–22, http://CRAN.R-project.org/doc/Rnews/.

[16] D. Lindley and A. Smith, *Bayes estimate for the linear model (with discussion) part 1*, Journal of the Royal Statistical Society, Series B, 34 (1972), pp. 1–41.

[17] P. McCullagh, *Regression models for ordinal data*, Journal of the Royal Statistical Society, Series B, 42 (1980), pp. 109–142.

[18] S. Park and S. Choi, *Hierarchical Gaussian process regression*, in Proceedings of 2nd Asian Conference on Machine Learning, M. Sugiyama and Q. Yang, eds., vol. 13 of Proceedings of Machine Learning Research, 2010, pp. 95–110.

[19] J. Pinheiro and D. Bates, *Mixed-effects models in S and S-PLUS*, Statistics and Computing, Springer New York, 2009.

[20] J. C. Pinheiro and D. M. Bates, *Unconstrained parametrizations for variance-covariance matrices*, Statistics and Computing, 6 (1996), pp. 289–296.

[21] H. O. Pritchard and H. A. Skinner, *The concept of electronegativity*, Chemical Reviews, 55 (1955), pp. 745–786.

[22] P. Z. G. Qian, *Sliced Latin hypercube designs*, Journal of the American Statistical Association, 107 (2012), pp. 393–399.

[23] P. Z. G. Qian, H. C. F. Wu, and J. Wu, *Gaussian process models for computer experiments with qualitative and quantitative factors*, tech. report, Department of statistics, University of Wisconsin, 2007.

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

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

[26] J. Sacks, W. Welch, T. Mitchell, and H. Wynn, *Design and analysis of computer experiments*, Statistical Science, (1989), pp. 409–435.

[27] R. Shepard, S. R. Brozell, and G. Gidofalvi, *The representation and parametrization of orthogonal matrices*, The Journal of Physical Chemistry A, 119 (2015), pp. 7924–7939.

[28] A. Smith, *Bayes estimates in one-way and two-way models*, Biometrika, 60 (1973), pp. 319–329.

[29] S. Tao, Y. Zhang, D. W. Apley, and W. Chen, *LVGP: Latent Variable Gaussian Process Modeling with Qualitative and Quantitative Input Variables*, 2019, https://CRAN.R-project.org/package=LVGP. R package version 2.1.5.

[30] W. N. Venables and B. D. Ripley, *Modern applied statistics with S*, Springer, 4 ed., 2002.

[31] T. Wei and V. Simko, *corrplot: Visualization of a correlation matrix*, 2016. R package version 0.77.

[32] H. Wickham, *ggplot2: Elegant graphics for data analysis*, Springer-Verlag New York, 2009, http://ggplot2.org.

[33] Y. Zhang and W. I. Notz, *Computer experiments with qualitative and quantitative variables: A review and reexamination*, Quality Engineering, 27 (2015), pp. 2–13.

[34] Y. Zhang, S. Tao, W. Chen, and D. Apley, *A latent variable approach to Gaussian process modeling with qualitative and quantitative factors*, tech. report, Northwestern University, 2018.