Mixture of Multivariate Gaussian Processes for Classification of Irregularly Sampled Satellite Image Time-Series

Alexandre Constantin[1*], Mathieu Fauvel[2†] and Stéphane Girard[1†]

[1*]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, 38000, France.
[2]CESBIO, Université de Toulouse, CNES/CNRS/INRAe/IRD/UPS, Toulouse, 31000, France.

*Corresponding author(s). E-mail(s): alexandre.constantin@grenoble-inp.fr;
Contributing authors: mathieu.fauvel@inrae.fr; stephane.girard@inria.fr;
†These authors contributed equally to this work.

Abstract

The classification of irregularly sampled Satellite image time-series (SITS) is investigated in this paper. A multivariate Gaussian process mixture model is proposed to address the irregular sampling, the multivariate nature of the time-series and the scalability to large data-sets. The spectral and temporal correlation is handled using a Kronecker structure on the covariance operator of the Gaussian process. The multivariate Gaussian process mixture model allows both for the classification of time-series and the imputation of missing values. Experimental results on simulated and real SITS data illustrate the importance of taking into account the spectral correlation to ensure a good behavior in terms of classification accuracy and reconstruction errors.

Keywords: Multivariate Gaussian processes, Classification, Multivariate imputation of missing data, Irregular sampling, Satellite image time-series (SITS), Remote sensing

1 Introduction

Satellite images availability has exponentially grown in the last decade. Thanks to free data access policy, optical satellite image time-series (SITS) such as *Landsat* or *Sentinel-2*, offer an unique opportunity to monitor the state and evolution of our living planet. Therefore, SITS have found many applications in ecological monitoring [1, 2], meteorology [3, 4] or agricultural system mapping [5–7], among others.

SITS are characterized by their spatial and spectral resolutions, and their revisit cycle. The spatial resolution corresponds to the size of a pixel on the ground, *e.g.*, a square of 10 meters while the spectral resolution is related to the number of wavelengths collected by the sensor, ranging typically in the visible and near infra-red part of the spectrum [8]. The revisit cycle stands for the time between two acquisitions over the same location: SITS have constant and short (*e.g.* few days) revisit time. Hence, for a given temporal period, a pixel is the collection of spectral measurements made at different times over the same location.

These properties lead to an unprecedented amount of numerical data, for which statistical methods are used to extract meaningful information such as land cover, crops yields ... For a pixel-wise based analysis, the predictor variables are multivariate time-series and the output variables represent the information to be extracted. In the pixel-wise classification setting, spatial independence is often assumed [9] since it drastically reduces the computational load. As an alternative, one can use geostatistics tools (including conditional auto-regressive Gaussian models [10]) to

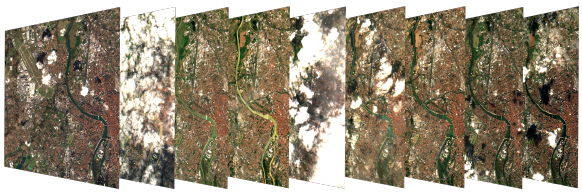

Fig. 1 True color Sentinel-2 satellite image time-series. Data were acquired in 2018 at different time steps over the area of Toulouse, France (images were downloaded from *Theia Land Data Center*: http://www.theia-land.fr/en/presentation/products).

tackle the potential spatial dependence. Temporal and spectral correlations can also be taken into account using various statistical models [11].

However, external random meteorological factors interfere with the availability of the acquired data at the pixel scale. Indeed, as displayed in Fig. 1, shadows and clouds result in missing data in the time-series. Furthermore, orbital trajectory generates an irregular temporal sampling: Even though the acquisition scheme is regular, acquisition days are different for pixels located at different places [12]. As such, each pixel of the SITS has its own size in the temporal domain: Fig. 2 illustrates the irregular temporal sampling on the data under consideration in this paper.

Specific models are thus required to properly analyze such time-series, as described in Section 2. Conventional approaches usually start by resampling the data onto a common temporal grid. In this work, we aim at analyzing irregularly sampled multidimensional SITS without any temporal resampling. In particular, the supervised pixel classification task is considered, *i.e.* the assignment of each pixel of the time-series to a predefined class.

To this end, a mixture of multivariate Gaussian Processes is proposed. A linear dependence model is assumed between the spectral variables leading to a separable covariance function in time and spectral domains. The resulting model provides statistical information on the underlying process for each class (mean and covariance functions) and scales linearly w.r.t. the number of samples. It allows to classify irregularly sampled signals without any temporal resampling and enables, as a by product, to impute missing data.

Section 2 reviews the state-of-the-art on classification with missing data and Gaussian processes. The statistical model is introduced in Section 3 while inference aspects are discussed in Section 4 including the estimation of the model parameters, the supervised classification, and the imputation of missing values.

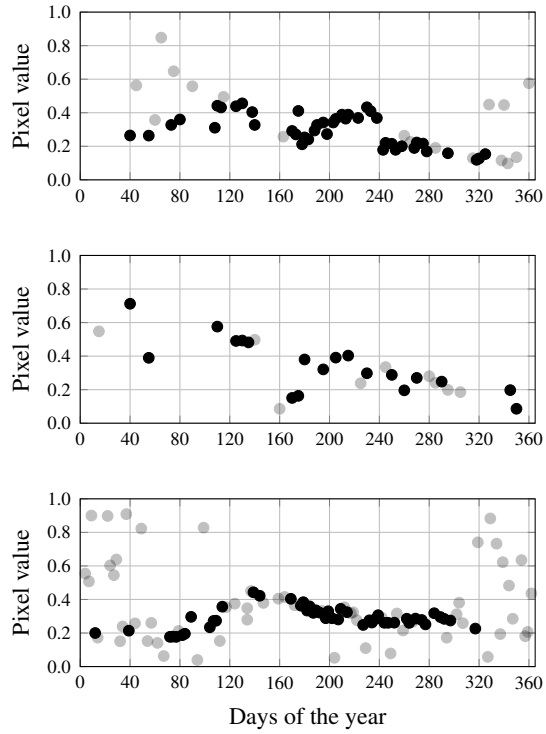

Fig. 2 Illustration of the irregular temporal sampling for the SITS used in this work. Three time-series at different locations for one spectral band are reported: A black dot indicates that the pixel is clear (no shadow or cloud) at the considered time, and a light-gray dot indicates that the pixel has been tagged as clouds or shadows by the data provider.

These statistical procedures are validated on simulated data in Section 5. Section 6 is dedicated to the application of our methodology to the classification of SITS from a Sentinel-2 data-set. Section 7 concludes with a discussion on possible extensions of this work.

2 Related Work

This section briefly reviews state-of-the-art methods for model-based classification, classification dealing with missing values and classification with Gaussian processes.

2.1 Supervised model-based classification

Supervised model-based classification (also referred to as model-based discriminant analysis) starts from a training set of n independent and identically distributed realizations from a random pair $(\mathbf{Y}, Z) \in E \times \{1, \dots, C\}$ and assumes that the conditional distribution of $\mathbf{Y}|Z = c$ belongs to some parametric family: $p(\mathbf{y}|Z = c) = p_c(\mathbf{y};\ \boldsymbol{\theta}_c)$, for all $c \in \{1, \dots, C\}$ and

$\mathbf{y} \in E$, where E is an arbitrary space and $\boldsymbol{\theta}_c$ a set of parameters. Letting $\pi_c = \mathbb{P}(Z = c)$, the marginal distribution of $\mathbf{Y}$ is written as a finite mixture

$$p(\mathbf{y}) = \sum_{c=1}^{C} \pi_c p_c(\mathbf{y};\ \boldsymbol{\theta}_c),$$

whose parameters can be estimated by the maximum likelihood principle. A non-labeled observation can then be classified thanks to the Maximum a posteriori (MAP) criteria:

$$\hat{c} = \underset{c\in\{1,\dots,C\}}{\arg\max}\ p(Z = c|\mathbf{y}) = \underset{c\in\{1,\dots,C\}}{\arg\max}\ \pi_c p_c(\mathbf{y};\ \boldsymbol{\theta}_c),$$

thanks to Bayes' rule. When $E = \mathbb{R}^q$, the multivariate Gaussian distribution is often adopted for $p_c(\mathbf{y};\ \boldsymbol{\theta}_c)$ and gives rise to the well-known Quadratic discriminant analysis (QDA) method. We refer to [13, Section 4.3] for a discussion on the advantages and drawbacks of QDA and for possible extensions. Recent studies extend the model-based classification framework to non-Gaussian distributions such as the skew-normal distribution [14, 15] to deal with asymmetric data, or t-distributions [16, 17] to deal with outliers. We refer to [18, Chapter 9] for an in-depth review. The case $E = \mathbb{R}^q$ also encompasses the situation of discretized time-series on a common grid. Specific models can be then defined, as in [19] for temporal signatures.

If E is discrete, including for example the case of categorical data, extensions focus on the multinomial [20] or the Dirichlet [21] distributions. In the case of ordinal data, other extensions are proposed using a dedicated model of the process generating the data [22]. Finally, when E is more complex, *e.g.* infinite dimensional, non-parametric techniques are used. Kernel methods are probably the most popular non-parametric techniques in this situation [23]. Recall that a kernel is a positive definite function that corresponds to a dot product in a feature space. It allows for the construction of non-linear and non-parametric classifiers on E without computing explicitly the feature space. Kernels can be defined, for instance, on strings [24], graphs [25], vector-valued functions [26, 27], or combinations of several data types [28].

2.2 Classification with missing data

When dealing with remote sensing data, *i.e.* spatial-spectro temporal data such as SITS, handling missing values [29] is a recurrent problem. Classification dealing with missing data occurs when some inputs in the training set are incomplete, *i.e.* the number of available coordinates in $\mathbf{y}$ can be different from one sample to another, see [30–32] for reviews.

Three main approaches can be found in the literature. A first solution is to impute missing values before the classification itself. The pre-processing gives rise to a training set with observations re-sampled on a common grid that can be considered as vectors in a finite space $E = \mathbb{R}^q$, opening the door to classical model-based classification methods. We refer to [33] for a review on imputation techniques. Such two-stage approaches are used on Sentinel-2 SITS where a linear interpolation is applied before performing the classification with a Random Forests classifier [12]. Yet, by applying imputation techniques without any connection to the actual processing, propagated errors from the interpolation may degrade the results.

Alternative solutions are based on functional data analysis [34]. Each observation is interpreted as a sample from a random function. As such, it can be approximated by an expansion on some basis functions. The statistical analysis is then performed on the random vectors of coefficients, see [35] for an application to clustering. Nonparametric smoothing techniques may also adopted, see [36, Chapter 8] for an overview.

Finally, purely non-parametric methods can also be implemented by defining an appropriate dissimilarity measure between samples of varying size. In the context of time-series, Dynamic time warping (DTW) [37] is one of the most popular algorithms. It computes an optimal match between two vectors with different lengths. This map defines a dissimilarity that can be used for comparison in order to cluster samples into multiple groups.

2.3 Classification with Gaussian processes

A recent approach for supervised classification is based on the use of Gaussian processes (GPs) in a Bayesian framework. More specifically, Gaussian processes are used as prior distributions on the regression function linking the label Z to the explanatory variable $\mathbf{X}$. In the binary classification case, the conditional Bernoulli distribution of Z is defined through a logit transformation: $\text{logit}(p(Z = 1|\mathbf{X} = \mathbf{x})) =: f(\mathbf{x})$ where $f(\mathbf{x})$ is a centered Gaussian process. The considered prior Gaussian process is, most of the time, one-dimensional. Extensions to the multi-dimensional case

include the so-called multi-tasks or multi-outputs GP models, see [26, 38]. Finally, some recent works focus on non Gaussian processes such as Student-t processes which have gain attention over the past years [39, 40].

The discrete nature of Z makes the exact inference of model parameters infeasible. To overcome this difficulty, several techniques have been proposed, including the Laplace approximation [41], or through the expectation-propagation algorithm [42]. Such approaches rely on the inversion of a $n \times n$ covariance matrix and thus scale in $O(n^3)$ which makes the inference computationally demanding for large data sets. Scalable GPs were proposed to overcome this vexing effect, using for instance variational inference as in [43] or the Vecchia approximations [44, 45]. We refer to [46] for a review on this topic. In the next Section, we define a mixture of multivariate Gaussian processes which can be used for classification or imputation tasks without resort to approximate inference techniques.

3 Mixture of Multivariate Gaussian processes

The mixture of multivariate Gaussian processes model is introduced in Paragraph 3.1, some associated properties are derived in Paragraph 3.2 and Paragraph 3.3 discusses the link with existing works.

3.1 Model

Let $\mathcal{T}$ be a compact subset of $\mathbb{R}$, we denote by $\mathcal{GP}_1(0, K)$ a continuous univariate centered Gaussian process on $\mathcal{T}$ with covariance function $K : \mathcal{T}^2 \to \mathbb{R}$. Recall that, by definition, $W \sim \mathcal{GP}_1(0, K)$ implies that, for all $(t_1, \ldots, t_q) \in \mathcal{T}^q$, the random vector $(W(t_1), \ldots, W(t_q))^\top$ follows a multivariate centered Gaussian distribution $\mathcal{N}_q(\mathbf{0}, \mathbf{\Sigma})$ such that $\mathbf{\Sigma}_{i,j} = K(t_i, t_j)$, see for instance [47].

For all $p > 0$, let us similarly denote by $\mathcal{IGP}_p(0, K)$ a $p-$ dimensional, independent, centered Gaussian process defined as $\mathbf{W} = (W_1, \ldots, W_p)^\top \sim \mathcal{IGP}_p(0, K)$ if and only if

$$\begin{cases} W_b \sim \mathcal{GP}_1(0, K), \ \forall b \in \{1, \ldots, p\}, \\ W_b \perp\!\!\!\perp W_{b'}, \ \forall b \neq b' \in \{1, \ldots, p\}^2, \end{cases}$$

where $\perp\!\!\!\perp$ stands for independence. The above defined multivariate Gaussian processes are the building blocks to define more general multivariate Gaussian processes denoted by $\mathcal{MGP}_p(\mathbf{m}, K, \mathbf{A})$ where $\mathbf{m} : \mathcal{T} \to \mathbb{R}^p$ is the mean function, $K : \mathcal{T}^2 \to \mathbb{R}$ is the covariance operator and $\mathbf{A}$ a non-singular $p \times p$ matrix: $\mathbf{Y} \sim \mathcal{MGP}_p(\mathbf{m}, K, \mathbf{A})$ if and only if

$$\mathbf{Y} = \mathbf{AW} + \mathbf{m} \text{ with } \mathbf{W} \sim \mathcal{IGP}_p(0, K). \tag{1}$$

Let us remark that model (1) is not identifiable without additional constraints. Indeed, $\mathcal{MGP}_p(\mathbf{m}, K, \mathbf{A})$ and $\mathcal{MGP}_p(\mathbf{m}, \lambda K, \mathbf{A}/\sqrt{\lambda})$ yield the same process for all $\lambda > 0$. This issue is discussed in further details in Section 4, see also the next paragraph for some basic properties of multivariate Gaussian processes defined in (1).

The mixture of multivariate Gaussian processes (M2GP) is defined by: Conditionally to $Z = c$,

$$\mathbf{Y} \sim \mathcal{MGP}_p(\mathbf{m}_c, K_c, \mathbf{A}_c), \tag{2}$$

where $\mathbf{m}_c : \mathcal{T} \to \mathbb{R}^p$, $K_c : \mathcal{T}^2 \to \mathbb{R}$ and $\mathbf{A}_c$ is a non-singular $p \times p$ matrix, for all $c \in \{1, \ldots, C\}$. In the context of SITS classification, $\mathbf{Y}$ represents the (unobserved) multidimensional process and p denotes the number of spectral bands. The particular case $\mathbf{A}_c = \mathbf{I}_p$ yields a mixture of independent Gaussian processes (MIGP) whose applications to classification have been investigated in [48].

3.2 First properties

Let $\mathbf{C}$ and $\mathbf{D}$ be two matrices of size $m \times n$ and $p \times q$ respectively. Recall that the Kronecker product $\mathbf{C} \otimes \mathbf{D}$ is the $mp \times nq$ matrix such that

$$\mathbf{C} \otimes \mathbf{D} = \begin{pmatrix} c_{11}\mathbf{D} & \ldots & c_{n1}\mathbf{D} \\ \vdots & \ddots & \vdots \\ c_{m1}\mathbf{D} & \ldots & c_{mn}\mathbf{D} \end{pmatrix}$$

and $\text{vec}(\mathbf{C}) \in \mathbb{R}^{mn}$ is the vector obtained by stacking the n columns of $\mathbf{C}$:

$$\text{vec}(\mathbf{C}) = (c_{11}, \ldots, c_{m1}, c_{12}, \ldots, c_{m2}, \ldots, c_{1n}, \ldots, c_{mn})^\top.$$

The matrix-variate normal distribution $\mathcal{MN}_{p,q}$ [49, 50] is defined for all $p \times q$ random matrix $\mathbf{Y}^\star$ as: $\mathbf{Y}^\star \sim \mathcal{MN}_{p,q}(\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Lambda})$ if and only if

$$\text{vec}(\mathbf{Y}^\star) \sim \mathcal{N}_{pq}(\text{vec}(\mathbf{M}), \mathbf{\Sigma} \otimes \mathbf{\Lambda}), \tag{3}$$

where $\mathbf{M}$ is a $p \times q$ matrix, $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$ are symmetric positive definite matrices of size $q \times q$ and $p \times p$

respectively. We refer to [49] for an early definition of the matrix-variate normal distribution (as well as some of its derivatives) and to [51] for a general account on matrix-variate distributions. The associated density function is defined for all $p \times q$ matrix $\mathbf{y}$ by

$$p(\mathbf{y}) = (2\pi)^{-pq/2} \det(\boldsymbol{\Sigma})^{-p/2} \det(\boldsymbol{\Lambda})^{-q/2} \times \exp\left(-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Lambda}^{-1}(\mathbf{y}-\mathbf{M})\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{M})^\top\right]\right), \quad (4)$$

where $\mathrm{tr}(\cdot)$ denotes the trace operator. The next Proposition establishes that the finite sized marginals of the multivariate Gaussian process (1) can be interpreted as random matrices from a matrix-variate normal distribution.

Proposition 1 *Let $\mathbf{Y} \sim \mathcal{MGP}_p(\mathbf{m}, K, \mathbf{A})$ and introduce $\mathbf{Y}^\star$ the $p \times q$ random matrix defined as $\mathbf{Y}^\star = (\mathbf{Y}(t_1), \ldots, \mathbf{Y}(t_q))$ where $(t_1, \ldots, t_q) \in \mathcal{T}^q$. Then,*

$$\mathbf{Y}^\star \sim \mathcal{MN}_{p,q}(\mathbf{M}, \boldsymbol{\Sigma}, \mathbf{A}\mathbf{A}^\top), \quad (5)$$

where $\mathbf{M} = (\mathbf{m}(t_1), \ldots, \mathbf{m}(t_q))$ and $\boldsymbol{\Sigma}$ is the covariance matrix defined by $\boldsymbol{\Sigma}_{k,\ell} = K(t_k, t_\ell)$ for all $(k, \ell) \in \{1, \ldots, q\}^2$. Equivalently,

$$vec(\mathbf{Y}^\star) \sim \mathcal{N}_{pq}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \otimes \mathbf{A}\mathbf{A}^\top),$$

with $\boldsymbol{\mu} = vec(\mathbf{M})$.

In the SITS framework, $\mathbf{Y}^\star$ represents the observed $q-$dimensional SITS which is a discretized version of $\mathbf{Y}$ at q timestamps. An illustration is provided in Fig. 3 where $\mathcal{T} = [0, 1]$ and $p = q = 10$. Only the first two coordinates are presented. Let $\langle \cdot, \cdot \rangle$ denote the Euclidean scalar product on $\mathbb{R}^p$ and $\|\cdot\|$ be the associated norm. For all non zero vectors $(u, v) \in \mathbb{R}^p \times \mathbb{R}^p$, we also introduce $\cos(u, v) = \langle u, v \rangle / (\|u\| \, \|v\|)$. As a direct consequence of the covariance structure in (5), the correlation ρ between the elements of the random matrix $\mathbf{Y}^\star$ can be derived:

Corollary 1. *Suppose the assumptions of Proposition 1 hold.*

(i) For all $(b, b') \in \{1, \ldots, p\}^2$ and $j \in \{1, \ldots, q\}$, one has

$$\rho(Y^\star_{b,j}, Y^\star_{b',j}) = \cos(\mathbf{a}_b, \mathbf{a}_{b'}),$$

(with $\mathbf{a}_b$ the b^{th} line of $\mathbf{A}$) and is thus independent of $j \in \{1, \ldots, q\}$.

(ii) For all $(j, j') \in \{1, \ldots, q\}^2$ and $b \in \{1, \ldots, p\}$, one has

$$\rho(Y^\star_{b,j}, Y^\star_{b,j'}) = \Sigma_{j,j'} \Big/ \sqrt{\Sigma_{j,j}\Sigma_{j',j'}}, \quad (6)$$

and is thus independent of $b \in \{1, \ldots, p\}$.

It appears that $\mathbf{A}$ tunes the dependence between the lines of $\mathbf{Y}^\star$ (*i.e.* the spectral bands in the SITS context) while $\boldsymbol{\Sigma}$ drives the dependence between the columns (*i.e.* the acquisition times of the SITS).

3.3 Links with existing works

Multivariate Gaussian processes have already been used in the machine learning community without formal definition. In [38], the authors introduced a so-called multi-task Gaussian process where, in our context, each task represents one column from $\mathbf{Y}^\star$. In [40, 52] the authors provided a multivariate Gaussian process for regression. More complex techniques with prior distribution on the mean can be found in [53]. In the latter one, missing values are handled using a matrix which selects the observed timestamps from a larger vector. In contrast, M2GP does not require the introduction of a larger vector to represent all potential timestamps.

The *Linear Model of Coregionalization* (LMC) is commonly used in geostatistics to construct processes to tackle multi-output regression problems [54, 55]. We also refer to [26] for submodels and to [56] for a Bayesian perspective. The LMC can be written similarly to (1) as $\mathbf{Y} = \mathbf{A}\mathbf{W} + \mathbf{m}$ with two differences: $\mathbf{A}$ is a $p \times P$ matrix with $P \geq p$ (not necessarily squared) and the components of $\mathbf{W}$ are not necessarily identically distributed: $W_j \sim \mathcal{GP}(0, K_j)$, $j = 1, \ldots, P$. It is thus clear that a M2GP can be interpreted as a LMC with additional constraints. The latter permits to simplify the covariance structure using a Kronecker product, thus leading to simplified estimation procedures, as discussed in the next section.

Besides, the matrix-variate Gaussian distribution (3) has been widely studied for real-valued vector with fixed length. Applications of matrix-variate normal distribution can be found in different contexts such as electro-encephalography [57] or remote sensing [58]. Let us also mention that, in [59], the same Kronecker product model is used to regularize the estimation of the covariance matrix in high dimension and in [60] to impute missing data.

Finally, a likelihood ratio test is introduced in [61] to check whether the separability of the covariance (5) is adapted to the data in hand. However, this test has not been extended to irregularly sampled time-series.

4 Inference

This section addresses several inference aspects associated with the M2GP model. Consider $\{(\mathbf{Y}^1, Z_1), \ldots, (\mathbf{Y}^n, Z_n)\}$ a set of n random pairs identically distributed from the M2GP model. Clearly, π_c can be estimated by its empirical counterpart $\hat{\pi}_c = n_c/n$ where $n_c = \sum_{i=1}^n \mathbb{I}\{Z_i = c\}$ is the number of samples in class c (and $\mathbb{I}\{\cdot\}$ is the indicator function). Besides, from (2), $\mathbf{Y}^i \sim \mathcal{MGP}_p(\mathbf{m}_c, K_c, \mathbf{A}_c)$ conditionally to $Z_i = c$, for all $i \in \{1, \ldots, n\}$. The unknown quantities to be estimated are $\mathbf{m}_c : \mathcal{T} \to \mathbb{R}^p$, $K_c : \mathcal{T}^2 \to \mathbb{R}$ and the matrix $\mathbf{A}_c$. The use of parametric models for mean and covariance functions is discussed in Subsection 4.1 and the Maximum likelihood estimation (MLE) of all resulting parameters is presented in Subsection 4.2. The associated classification method based on the MAP rule and the imputation of missing values are described in Subsection 4.3 and Subsection 4.4 respectively.

4.1 Parametric mean and covariance functions

Let $J > 0$ and introduce $\{\varphi_1, \ldots, \varphi_J\}$ a subset of J basis functions of $L_2(\mathcal{T})$. For all $b \in \{1, \ldots, p\}$, the bth coordinate $(\mathbf{m}_c(t))_b$ of $\mathbf{m}_c(t)$ is expanded as

$$(\mathbf{m}_c(t))_b = \sum_{j=1}^{J} \alpha_{c,b,j}\, \varphi_j(t), \tag{7}$$

with $t \in \mathcal{T}$, and where $a_{c,b,j}$ is the projection coefficient of $(\mathbf{m}_c(\cdot))_b$ on $\varphi_j(\cdot)$. Denoting by $\boldsymbol{\alpha}_c$ the $p \times J$ matrix defined by:

$$\boldsymbol{\alpha}_c = \begin{pmatrix} \alpha_{c,1,1} & \alpha_{c,1,2} & \ldots & \alpha_{c,1,J} \\ \alpha_{c,2,1} & \ddots & \ldots & \alpha_{c,2,J} \\ \vdots & \vdots & \ddots & \ldots \\ \alpha_{c,p,1} & \ldots & \ldots & \alpha_{c,p,J} \end{pmatrix}$$

and letting $\mathbf{b} : t \in \mathcal{T} \mapsto (\varphi_1(t), \ldots, \varphi_J(t))^\top \in \mathbb{R}^J$, then (7) can rewritten matricially as $\mathbf{m}_c(t) = \boldsymbol{\alpha}_c \mathbf{b}(t)$.

The covariance operator K_c is assumed to belong to a family of symmetric positive-definite kernels [47, Chapter 4]. A typical kernel is the squared exponential kernel (also known as Gaussian or RBF kernel) with an additive white noise covariance function:

$$K_c(t, t'|\boldsymbol{\theta}_c) = \gamma_c^2 \exp\left(-\frac{(t-t')^2}{2h_c^2}\right) + \sigma_c^2 \mathbb{I}\{t = t'\}, \tag{8}$$

where $(t, t') \in \mathcal{T}^2$.The parameters are collected in $\boldsymbol{\theta}_c$ with, in this case, $\boldsymbol{\theta}_c = \{\gamma_c, h_c, \sigma_c\}$.

4.2 Maximum likelihood estimation

Assume each multivariate Gaussian process $\mathbf{Y}^i$ is observed on its own finite grid of distinct q_i timestamps $(t_1^i, \ldots, t_{q_i}^i) \in \mathbb{R}^{q_i}$ and note $\mathbf{Y}^{i,\star} = (\mathbf{Y}^i(t_1^i), \ldots, \mathbf{Y}^i(t_{q_i}^i))^\top$ the associated $p \times q_i$ random matrix. Let us stress that this formalism naturally allows to deal with irregularly sampled SITS since the size of $\mathbf{Y}^{i,\star}$ may depend on i. From Proposition 1, one has that, conditionally to $Z_i = c$,

$$\mathbf{Y}^{i,\star} \sim \mathcal{MN}_{p,q_i}(\boldsymbol{\alpha}_c \mathbf{B}^i, \boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c), \mathbf{A}_c \mathbf{A}_c^\top), \tag{9}$$

where the covariance matrix $\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)$ is defined for all $(j, j') \in \{1, \ldots, q_i\}^2$ by $\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)_{j,j'} = K_c(t_j^i, t_{j'}^i|\boldsymbol{\theta}_c)$ and $\mathbf{B}^i = (\mathbf{b}(t_1^i), \ldots, \mathbf{b}(t_{q_i}^i))$ is a $J \times q_i$ design matrix. Parameters $\{\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c\}$ are estimated by minimizing the negative log-likelihood given hereafter.

Lemma 1. *The negative log-likelihood associated with (9) can be expanded as*

$$\mathcal{L} = \frac{1}{2} \sum_{c=1}^{C} \ell_c(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c \mathbf{A}_c^\top),$$

(up to an additive constant) where $\ell_c(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c \mathbf{A}_c^\top) =$

$$Q_c \log\det(\mathbf{A}_c \mathbf{A}_c^\top) + p \sum_{i|Z_i=c} \log\det \boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c) +$$

$$\operatorname{tr}\left(\sum_{i|Z_i=c} (\mathbf{Y}^{i,\star} - \boldsymbol{\alpha}_c \mathbf{B}^i) \{\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)\}^{-1} (\mathbf{Y}^{i,\star} - \boldsymbol{\alpha}_c \mathbf{B}^i)^\top \{\mathbf{A}_c \mathbf{A}_c^\top\}^{-1} \right),$$

with $Q_c = \sum_{i|Z_i=c} q_i$ *, for all* $c \in \{1, \ldots, C\}$.

It appears that the likelihood only involves the product of matrices $\mathbf{A}_c \mathbf{A}_c^\top$ and not the matrix $\mathbf{A}_c$ itself. This is a direct consequence of (5): The matrix-variate normal distribution of the sampled process $\mathbf{Y}^\star$ only depends on the above product. The parameters of interest are thus $\boldsymbol{\alpha}_c$, $\boldsymbol{\theta}_c$ and $\mathbf{A}_c \mathbf{A}_c^\top$ and the MLE is obtained by solving C independent optimization problems:

$$(\hat{\boldsymbol{\alpha}}_c, \hat{\boldsymbol{\theta}}_c, \widehat{\mathbf{A}_c \mathbf{A}_c^\top}) = \underset{\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c \mathbf{A}_c^\top}{\arg\min}\ \ell_c(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c \mathbf{A}_c^\top), \tag{10}$$

for all $c \in \{1, \ldots, C\}$. The solution is partially explicit as explained in the next Proposition.

Proposition 2 *Let $c \in \{1, \ldots, C\}$.*
(i) Solutions of (10) satisfy the following two properties. Given $\hat{\boldsymbol{\theta}}_c$, one has:

$$\hat{\boldsymbol{\alpha}}_c = \left[\sum_{i|Z_i=c} \mathbf{Y}^{i,\star}\{\boldsymbol{\Sigma}^{c,i}(\hat{\boldsymbol{\theta}}_c)\}^{-1}(\mathbf{B}^i)^\top\right]\left[\sum_{i|Z_i=c} \mathbf{B}^i\{\boldsymbol{\Sigma}^{c,i}(\hat{\boldsymbol{\theta}}_c)\}^{-1}(\mathbf{B}^i)^\top\right]^{-1} \tag{11}$$

$$\widehat{\mathbf{A}_c\mathbf{A}_c^\top} = \frac{1}{Q_c}\sum_{i|Z_i=c}(\mathbf{Y}^{i,\star} - \hat{\boldsymbol{\alpha}}_c\mathbf{B}^i)\{\boldsymbol{\Sigma}^{c,i}(\hat{\boldsymbol{\theta}}_c)\}^{-1}(\mathbf{Y}^{i,\star} - \hat{\boldsymbol{\alpha}}_c\mathbf{B}^i)^\top. \tag{12}$$

(ii) The partial derivative of the negative log-likelihood (see Lemma 1) w.r.t. the kth coordinate of $\boldsymbol{\theta}_c$ is given by:

$$\frac{\partial \ell_c(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c\mathbf{A}_c^\top)}{\partial(\boldsymbol{\theta}_c)_k} = \sum_{i|Z_i=c} \mathrm{tr}\left(\left[p\{\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)\}^{-1} - \Delta^{c,i}(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c\mathbf{A}_c^\top)\right]\frac{\partial \boldsymbol{\Sigma}^{c,i}}{\partial(\boldsymbol{\theta}_c)_k}(\boldsymbol{\theta}_c)\right), \tag{13}$$

where

$$\Delta^{c,i}(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c\mathbf{A}_c^\top) = \boldsymbol{\beta}^{c,i}(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c)^\top\{\mathbf{A}_c\mathbf{A}_c^\top\}^{-1}\boldsymbol{\beta}^{c,i}(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c),$$
$$\text{with } \boldsymbol{\beta}^{c,i}(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c) = (\mathbf{Y}^{i,\star} - \boldsymbol{\alpha}_c\mathbf{B}^i)\{\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)\}^{-1}.$$

In practice, the computation of the MLE is achieved thanks to an iterative procedure based on (11)–(13), described in Algorithm 1 and discussed in Paragraph 4.5.

4.3 Supervised classification

The objective is to assign a label $\tilde{c} \in \{1, \ldots, C\}$ to a new $p \times q$ random matrix $\tilde{\mathbf{Y}}^\star = (\mathbf{Y}(\tilde{t}_1), \ldots, \mathbf{Y}(\tilde{t}_q))^\top$. We focus on the MAP rule which consists in maximizing w.r.t. c the posterior probability

$$\mathbb{P}(Z = c|\tilde{\mathbf{Y}}^\star) \propto \pi_c p(\tilde{\mathbf{Y}}^\star|Z = c),$$

where $p(\tilde{\mathbf{Y}}^\star|Z = c)$ is matrix-variate normal density defined as

$$-\log p(\tilde{\mathbf{Y}}^\star|Z = c) = \frac{pq}{2}\log(2\pi) + \frac{p}{2}\log\det\left(\tilde{\boldsymbol{\Sigma}}^c(\boldsymbol{\theta}_c)\right) + \frac{q}{2}\log\det\left(\mathbf{A}_c\mathbf{A}_c^\top\right) + \frac{1}{2}\mathrm{tr}\left[\{\mathbf{A}_c\mathbf{A}_c^\top\}^{-1}(\tilde{\mathbf{Y}}^\star - \boldsymbol{\alpha}_c\tilde{\mathbf{B}})\tilde{\boldsymbol{\Sigma}}^c(\boldsymbol{\theta}_c)^{-1}(\tilde{\mathbf{Y}}^\star - \boldsymbol{\alpha}_c\tilde{\mathbf{B}})^\top\right], \tag{14}$$

see the proof of Lemma 1. Here, the covariance matrix $\tilde{\boldsymbol{\Sigma}}^c(\boldsymbol{\theta}_c)$ is defined for all $(j, j') \in \{1, \ldots, q\}^2$ by $\tilde{\boldsymbol{\Sigma}}^c(\boldsymbol{\theta}_c)_{j,j'} = K_c(\tilde{t}_j, \tilde{t}_{j'}|\boldsymbol{\theta}_c)$ and $\tilde{\mathbf{B}} = (\mathbf{b}(\tilde{t}_1), \ldots, \mathbf{b}(\tilde{t}_q))$ is a $J \times q$ design matrix. In practice, all parameters are replaced using their MLE counterparts and $\tilde{c}$ is selected by minimizing the negative log posterior probability, that is:

$$\tilde{c} = \arg\min_c \left\{p\log\det\left(\tilde{\boldsymbol{\Sigma}}^c(\hat{\boldsymbol{\theta}}_c)\right) + q\log\det\left(\widehat{\mathbf{A}_c\mathbf{A}_c^\top}\right) - 2\log(n_c/n) + \mathrm{tr}\left[\left\{\widehat{\mathbf{A}_c\mathbf{A}_c^\top}\right\}^{-1}(\tilde{\mathbf{Y}}^\star - \hat{\boldsymbol{\alpha}}_c\tilde{\mathbf{B}})\tilde{\boldsymbol{\Sigma}}^c(\hat{\boldsymbol{\theta}}_c)^{-1}(\tilde{\mathbf{Y}}^\star - \hat{\boldsymbol{\alpha}}_c\tilde{\mathbf{B}})^\top\right]\right\}.$$

In the SITS framework, the above formula provides a natural way to classify a new multivariate time-series even though it is not observed at the same timestamps as the examples from the training set.

4.4 Imputation of missing values

The next result provides the distribution of the MGP process at time $t^\dagger$ conditionally to its label and to observations at times $t_1, \ldots, t_q$.

Proposition 3 *Assume that, conditionally to $Z = c$, $\mathbf{Y} \sim \mathcal{MGP}_p(\boldsymbol{\alpha}_c\mathbf{b}, K_c, \mathbf{A}_c)$ and introduce $\mathbf{Y}^\star$ the $p \times q$ random matrix defined as $\mathbf{Y}^\star = (\mathbf{Y}(t_1), \ldots, \mathbf{Y}(t_q))$ where $(t_1, \ldots, t_q) \in \mathcal{T}^q$. Let $t^\dagger \in \mathcal{T}$ such that $t^\dagger \neq t_k$ for all $k \in \{1, \ldots, q\}$. Then, conditionally to $Z = c$ and $\mathbf{Y}^\star = \mathbf{y}^\star$,*

$$\mathbf{Y}(t^\dagger) \sim \mathcal{N}_p\left(\boldsymbol{\mu}_c(t^\dagger, \mathbf{y}^\star),\ \boldsymbol{\Lambda}_c(t^\dagger)\right),$$

with

$$\boldsymbol{\mu}_c(t^\dagger, \mathbf{y}^\star) = \boldsymbol{\alpha}_c\mathbf{b}(t^\dagger) + (\mathbf{y}^\star - \boldsymbol{\alpha}_c\mathbf{B})\{\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c)\}^{-1}\mathbf{k}_c(t^\dagger),$$
$$\boldsymbol{\Lambda}_c(t^\dagger) = \left[K_c(t^\dagger, t^\dagger|\boldsymbol{\theta}_c) - \mathbf{k}_c(t^\dagger)^\top\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c)^{-1}\mathbf{k}_c(t^\dagger)\right] \otimes \mathbf{A}_c\mathbf{A}_c^\top,$$

and where $\mathbf{k}_c(t^\dagger) = (K_c(t^\dagger, t_1|\boldsymbol{\theta}_c), \ldots, K_c(t^\dagger, t_q|\boldsymbol{\theta}_c))^\top$. Recall that the covariance matrix $\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c)$ is defined for all $(j, j') \in \{1, \ldots, q\}^2$ by $\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c)_{j,j'} = K_c(t_j, t_{j'}|\boldsymbol{\theta}_c)$ and $\mathbf{B} = (\mathbf{b}(t_1), \ldots, \mathbf{b}(t_q))$ is a $J \times q$ design matrix.

As a consequence, when $\mathbf{Y}(t^\dagger)$ is not observed (but its label is known to be c), this missing value can be imputed by the conditional expectation given in Proposition 3, where the unknown parameters are replaced by their associated MLE:

$$\hat{\mathbf{Y}}_c(t^\dagger) = \hat{\boldsymbol{\alpha}}_c\mathbf{b}(t^\dagger) + (\mathbf{Y}^\star - \hat{\boldsymbol{\alpha}}_c\mathbf{B})\{\boldsymbol{\Sigma}^c(\hat{\boldsymbol{\theta}}_c)\}^{-1}\hat{\mathbf{k}}_c(t^\dagger). \tag{15}$$

This allows for the reconstruction of SITS values at unobserved times. If the label of $\mathbf{Y}^\star$ is unknown, the

distribution of the MGP process at time $t^\dagger$ conditionally to observations at times $t_1, \ldots, t_q$ can still be derived from Proposition 3: conditionally to $\mathbf{Y}^\star = \mathbf{y}^\star$,

$$\mathbf{Y}(t^\dagger) \sim \sum_{c=1}^{C} \tau_c(\mathbf{y}^\star) \mathcal{N}_p\left(\boldsymbol{\mu}_c(t^\dagger, \mathbf{y}^\star),\ \boldsymbol{\Lambda}_c(t^\dagger)\right),$$

with $\tau_c(\mathbf{y}^\star) = \mathbb{P}(Z = c | \mathbf{Y}^\star = \mathbf{y}^\star)$, leading to

$$\begin{aligned}
\boldsymbol{\mu}(t^\dagger, \mathbf{y}^\star) &= \sum_{c=1}^{C} \tau_c(\mathbf{y}^\star) \boldsymbol{\mu}_c(t^\dagger, \mathbf{y}^\star), \\
\boldsymbol{\Lambda}(t^\dagger) &= \sum_{c=1}^{C} \tau_c(\mathbf{y}^\star) \left(\boldsymbol{\Lambda}_c(t^\dagger) + \boldsymbol{\mu}_c(t^\dagger, \mathbf{y}^\star)^\top \boldsymbol{\mu}_c(t^\dagger, \mathbf{y}^\star)\right) \\
&\quad - \boldsymbol{\mu}(t^\dagger, \mathbf{y}^\star)^\top \boldsymbol{\mu}(t^\dagger, \mathbf{y}^\star).
\end{aligned}$$

Thus, when both $\mathbf{Y}(t^\dagger)$ and its label are not observed, $\mathbf{Y}(t^\dagger)$ can be imputed by

$$\hat{\mathbf{Y}}(t^\dagger) = \sum_{c=1}^{C} \widehat{\tau_c}(\mathbf{Y}^\star) \hat{\mathbf{Y}}_c(t^\dagger), \tag{16}$$

where $\hat{\mathbf{Y}}_c(t^\dagger)$ is given in (15), $\mathbf{Y}^\star = (\mathbf{Y}(t_1), \ldots, \mathbf{Y}(t_q))$ and

$$\widehat{\tau_c}(\mathbf{Y}^\star) = \hat{\pi}_c \hat{p}(\mathbf{Y}^\star | Z = c) \Big/ \sum_{k=1}^{C} \hat{\pi}_k \hat{p}(\mathbf{Y}^\star | Z = k),$$

with $\hat{p}(\mathbf{Y}^\star | Z = k)$ the estimated matrix-variate density defined similarly to (14).

4.5 Numerical implementation

The computation of the MLE is implemented as detailed in Algorithm 1 using the results of Proposition 2. To deal with the idenfiability issue mentioned in Paragraph 3.1, $\mathbf{A}_c\mathbf{A}_c^\top$ is normalised by η_c such that $\|\mathbf{A}_c\mathbf{A}_c^\top\|_F = 1$ (where $\|\cdot\|_F$ denotes the Frobenius norm) and each covariance matrix $\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)$ is modified accordingly so that the likelihood remains unaffected (step (d) of Algorithm 1). The gradient step (e) is performed using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, see [62]. More specifically, the L-BFGS-B version is used which allows for box and positivity constraints. As described in [62], the gradient step is obtained by line search and the algorithm stops when either: the objective function (*i.e.* the likelihood) does not change significantly, the (infinite) norm of the projected gradient is sufficiently small or when the maximum number of iterations is reached. Since the objective function is not convex, the optimization process is sensitive to the initialization. In practice, multiple random starts are used and the best solution in terms of negative log-likelihood is retained. Let us highlight that, in practice, steps (a)-(e) are computed for all classes in parallel since the model parameters are decoupled w.r.t. the classes.

Algorithm 1 Computation of MLE of model parameters.

Require:
Sample $\left\{(\mathbf{Y}^{i,\star}, Z_i) \in \mathbb{R}^{p \times q_i} \times \{1, \ldots, C\},\ i = 1, \ldots, n\right\}$
Initialization $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C)$.
Ensure:
MLE $\left(\hat{\boldsymbol{\alpha}}_c, \widehat{\mathbf{A}_c\mathbf{A}_c^\top}, \hat{\boldsymbol{\theta}}_c\right)$, $c = 1, \ldots, C$.
for $c = 1 \to C$ **do**
  while $\ell_c(\boldsymbol{\alpha}_c, \mathbf{A}_c\mathbf{A}_c^\top, \boldsymbol{\theta}_c)$ has converged **do**
    (a) Update $\boldsymbol{\alpha}_c$ using (11)
    (b) Update $\mathbf{A}_c\mathbf{A}_c^\top$ using (12)
    (c) Compute $\eta_c \leftarrow \|\mathbf{A}_c\mathbf{A}_c^\top\|_F$
    (d) Update $\mathbf{A}_c\mathbf{A}_c^\top \leftarrow \mathbf{A}_c\mathbf{A}_c^\top / \eta_c$ and $\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c) \leftarrow \eta_c \boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)$, $i = 1, \ldots, n$
    (e) Update $\boldsymbol{\theta}_c$ with a gradient step using (13)
  end while
end for

The numerical complexity of one iteration for all classes of Algorithm 1 is $O(n(q_\infty^3 + p^3 + J^3))$ where n is the sample size and $q_\infty = \max\{q_i,\ i = 1, \ldots, n\}$. The computation of the MLE thus scales linearly w.r.t. n. In constrast, the cost associated with standard classfication methods based on Gaussian processes is $O((C+1)n^3)$ [47, Algorithm 3.3]. Here, the computation of the MLE only relies on the inversion of $p \times p$ and $q_i \times q_i$ matrices whose sizes do not depend on the sample size.

Let us note that Algorithm 1 can be interpreted as an extension of the so-called Flip-flop method introduced independently by [63, 64]. This latter method is an iterative way to compute the MLE associated with the matrix-variate normal distribution. As such, it is limited to the situation where $q_1 = q_2 = \cdots = q_n$ which only occurs when all Gaussian processes are observed on a common grid. Identifiability issues are discussed in [50] and the method is extended to higher order tensor distributions in [65].

Finally, all the above estimation procedures have been implemented in Python using the Scikit-Lean API, see [66]. The Fourier basis $\{\varphi_1, \ldots, \varphi_J\}$ was chosen to estimate the mean function (see [48] for other bases), while the family of symmetric positive-definite kernels was selected among the *Kernels* class in the Scikit-Learn library.

5 Validation on simulated data

The performance of the inference procedure associated with the M2GP model is illustrated on simulated data.[1] The simulated model is described in Paragraph 5.1. First, the influence of the dependence between coordinates as well as the influence of the number of observation times are investigated in Paragraph 5.2. Second, consequences on the classification and imputation accuracy are discussed in Paragraph 5.3.

5.1 Experimental design

A binary classification problem is considered. Two classes are simulated from a 10-dimensional M2GP model (2) on $\mathcal{T} = [0, 1]$ with 1,000 samples per class leading to $n = 2,000$ and $p = 10$. Let us recall that a class c is completely described by its associated set of parameters $\{\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c\}$. Mean functions are generated following (7) with a Fourier basis of size $J = 11$. Coefficients $\alpha_{c,b,j}$ are simulated independently from a $\mathcal{N}_1(0, 0.02)$ distribution, $c \in \{1, 2\}$, $b \in \{1, \ldots, 10\}$ and $j \in \{1, \ldots, 11\}$, yielding different mean functions $\mathbf{m}_1$ and $\mathbf{m}_2$. The covariance operator is identical for both classes: $K_1(\cdot, \cdot) = K_2(\cdot, \cdot)$. It is defined following (8) as the sum of a RBF kernel and a white noise covariance function. The associated parameters are $\boldsymbol{\theta}_1 = \{\gamma_1, h_1, \sigma_1\} = \{1.5, 150, 0.05\} = \boldsymbol{\theta}_2$. We also set equal covariance matrices, *i.e.* $\mathbf{A}_1 = \mathbf{A}_2$, with

$$\mathbf{A}_1\mathbf{A}_1^\top = \begin{pmatrix} 1 & \beta & \cdots & \beta \\ \beta & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \beta \\ \beta & \cdots & \beta & 1 \end{pmatrix}, \tag{17}$$

so that β tunes the pairwise correlation between the 10 coordinates of the Gaussian processes. In the following, we shall consider $\beta \in \{0, 1/4, 1/2\}$. In practice, M2GP processes are simulated on random grids of varying size $q \in \{10, 20, \ldots, 100\}$, see Fig. 3 for an illustration in the case $q = 10$ and $\beta = 0$.

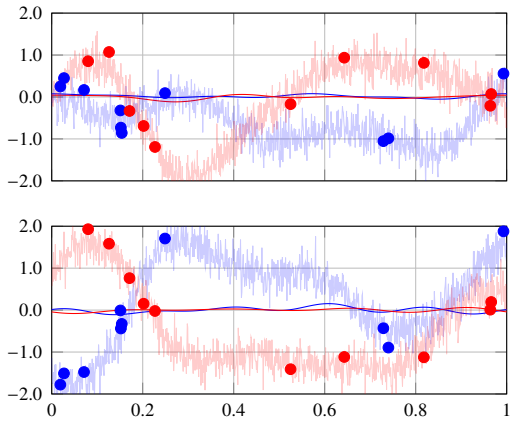

Fig. 3 Two simulated M2GP processes (transparent lines) in dimension $p = 10$ observed at $q = 10$ timestamps (dots), from two classes ($c = 1$: blue, $c = 2$: red). The mean functions are depicted as continuous opaque lines. Top panel: first coordinates, bottom panel: second coordinates (only the first two coordinates p_1 and p_2 are represented).

5.2 Estimation results

All estimation procedures are evaluated on 100 replications of the above described simulation model. First, for all $c \in \{1, 2\}$, the quality of the reconstructed mean $\hat{\mathbf{M}}_c = \hat{\boldsymbol{\alpha}}_c\mathbf{B}$ is measured by the normalized Mean Squared Error (nMSE) defined as:

$$\text{nMSE}(\hat{\mathbf{M}}_c, \mathbf{M}_c) = \frac{\|\mathbf{M}_c - \hat{\mathbf{M}}_c\|_F^2}{\|\mathbf{M}_c - \overline{\mathbf{M}}_c\|_F^2}, \tag{18}$$

where $\overline{\mathbf{M}}_c$ is the empirical mean of the processes in class c. The lower this score is, the better the estimation. An example of reconstructed mean is presented on Fig. 4, for one replication. Second, the quality of the estimation of the covariance structure $\mathbf{A}_c\mathbf{A}_c^\top$ (see 17) by $\widehat{\mathbf{A}_c\mathbf{A}_c^\top}$ is assessed by the cosine score defined as:

$$C(\widehat{\mathbf{A}_c\mathbf{A}_c^\top}, \mathbf{A}_c\mathbf{A}_c^\top) = 1 - \frac{\langle \widehat{\mathbf{A}_c\mathbf{A}_c^\top}, \mathbf{A}_c\mathbf{A}_c^\top \rangle_F}{\|\widehat{\mathbf{A}_c\mathbf{A}_c^\top}\|_F \|\mathbf{A}_c\mathbf{A}_c^\top\|_F}. \tag{19}$$

Let us note that $C(\widehat{\mathbf{A}_c\mathbf{A}_c^\top}, \mathbf{A}_c\mathbf{A}_c^\top) \in [0, 2]$ with $C(\widehat{\mathbf{A}_c\mathbf{A}_c^\top}, \mathbf{A}_c\mathbf{A}_c^\top) = 0$ when $\widehat{\mathbf{A}_c\mathbf{A}_c^\top}$ and $\mathbf{A}_c\mathbf{A}_c^\top$ are proportional. Finally, turning to the estimation of the

[1] The code and a notebook are available at https://gitlab.inria.fr/aconstan/mixture-of-multivariate-gaussian-processes-for-classification-of-irregularly-sampled-satellite-image-time-series

kernel part (8) of the dependence structure, we focus on the estimation accuracy of the length-scale by computing the absolute difference between the true length-scale $h_1 = h_2 = 150$ and its estimated counterpart. The results are averaged over the 100 independent replications and are reported on Fig. 5 for the first class. Similar results are obtained for the second one. It appears that, unsurprisingly, the quality of the estimates increases with the number q of discretization times. At the opposite, the dependence parameter β does not seem to influence significantly the accuracy of the estimation. One can nevertheless note that, as expected, the variability of the estimators increases with β, as the information carried by correlated coordinates decreases. Besides, the estimated length-scales do not depend on β, this may be explained by the separability property exhibited in Corollary 1.

5.3 Classification and imputation results

Here, we focus on the comparison between results associated with M2GP and MIGP models. To assess the classification and imputation performances, $4,000$ samples are generated following the model described in Paragraph 5.1 and then split into two disjoint balanced sets. The first one is used as a training set (of size $n = 2,000$) to estimate model parameters. The second one is used as a test set where the accuracy of the classification and imputation steps associated with the two above methods are compared. The classification performance is assessed thanks to the Overall Accuracy (OA), that is the ratio of the number correctly classified test observations and the total number of test observations, while the nMSE is used for the imputation task. Similarly to (18), we let

$$\text{nMSE}(\hat{\mathbf{Y}}^\star, \mathbf{Y}^\star) = \frac{\|\hat{\mathbf{Y}}^\star - \mathbf{Y}^\star\|_F^2}{\|\mathbf{Y}^\star - \overline{\mathbf{Y}^\star}\|_F^2}, \tag{20}$$

where $\hat{\mathbf{Y}}^\star$ is the imputed discretized process when the class is unknown thanks to (16), given the observed discretized process on q points. $\overline{\mathbf{Y}^\star}$ is the empirical mean of discretized processes in the test set. The above Frobenius norms are computed on a fixed regular grid of $\mathcal{T}$ defined as $\{t_\ell = \ell/100,\ \ell = 1, \ldots, 100\}$. The results are reported in Fig. 6.

It appears that the classification scores associated with M2GP increase with the dependence coefficient β and the number q of discretization times. On the opposite, MIGP scores are decreasing with β, due to the independence assumption. When there is no dependence between coordinates ($\beta = 0$), both methods provide similar classification scores. Unsurprisingly,

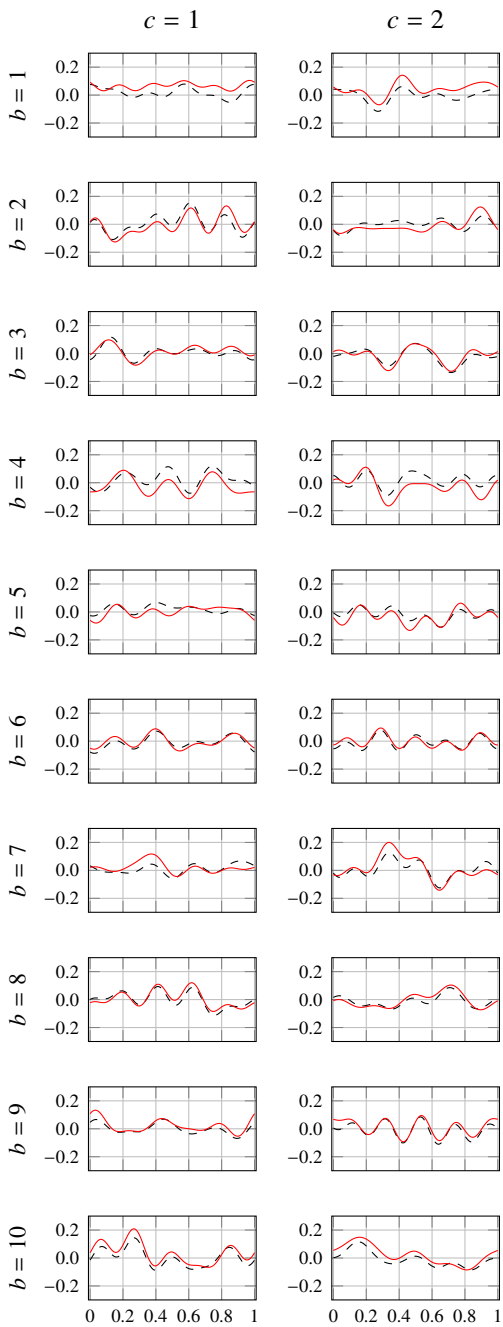

Fig. 4 Estimation of mean functions by M2GP on simulated data for all coordinates $b \in \{1, \ldots, 10\}$, classes $c \in \{1, 2\}$ and $\beta = 0$ on one replication. The dashed line is the true mean, the red line is the estimated GP mean from a discretization on a grid of size $q = 10$.

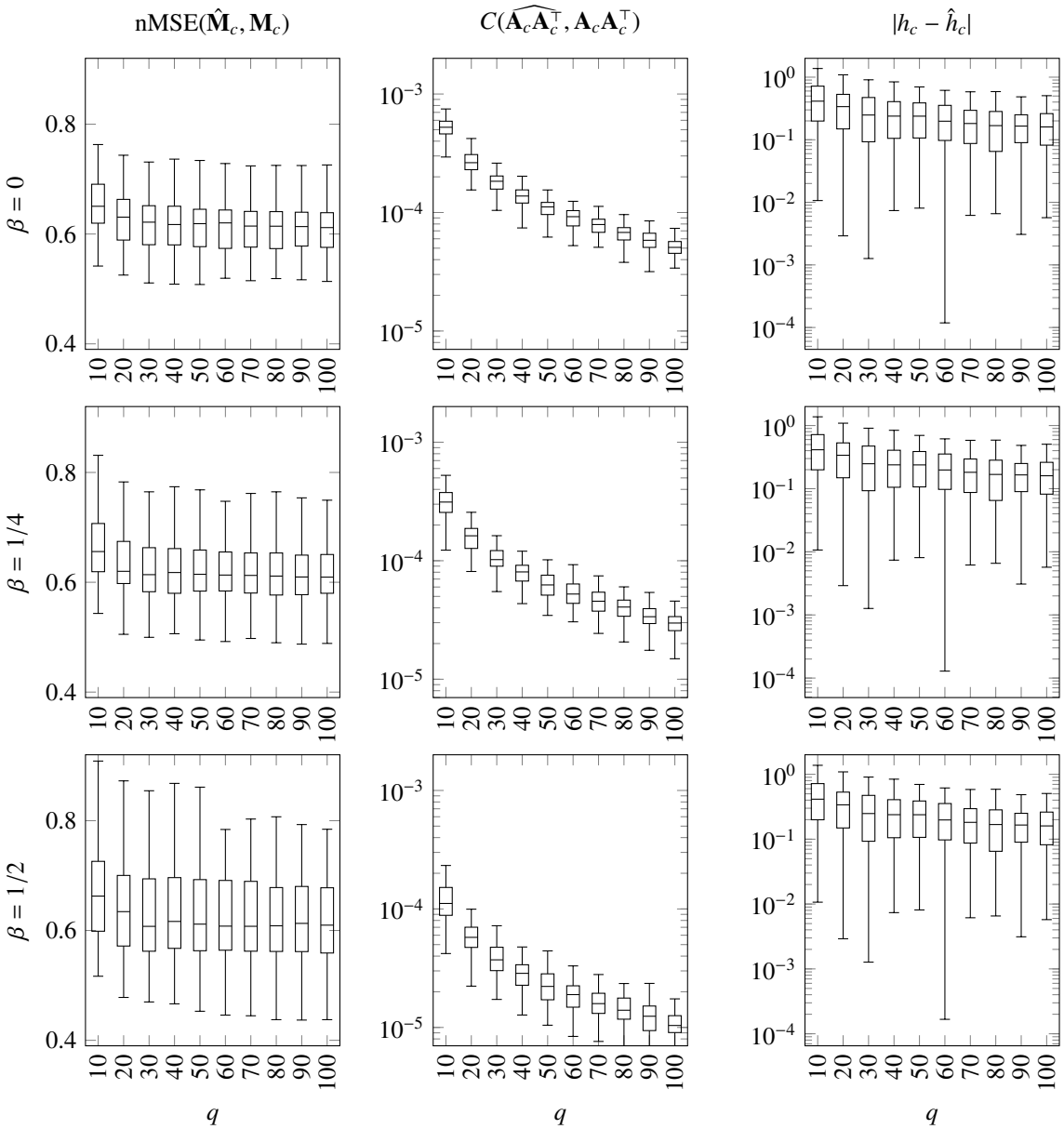

Fig. 5 Estimation of M2GP parameters on simulated data as a function of the number q of discretization times on class $c = 1$. From left to right: normalized mean squared error (18), cosine score (19) and absolute difference of length-scales. From top to bottom: $\beta = 0$, $\beta = 1/4$ and $\beta = 1/2$.

M2GP outperforms MIGP as soon as a dependence occurs.

In terms of reconstruction, both methods feature similar performances, increasing with q. The dependence strength only impacts the variance of the reconstructed processes: The larger β is, the larger the variability.

6 Time-series classification: Application to satellite data

This section is devoted to multivariate SITS classification using the M2GP model. The data were acquired by the Sentinel-2 satellite, and are presented in Paragraph 6.1, with a focus on the irregular temporal sampling. The estimated M2GP parameters are interpreted and discussed in Paragraph 6.2. Finally Paragraph 6.3

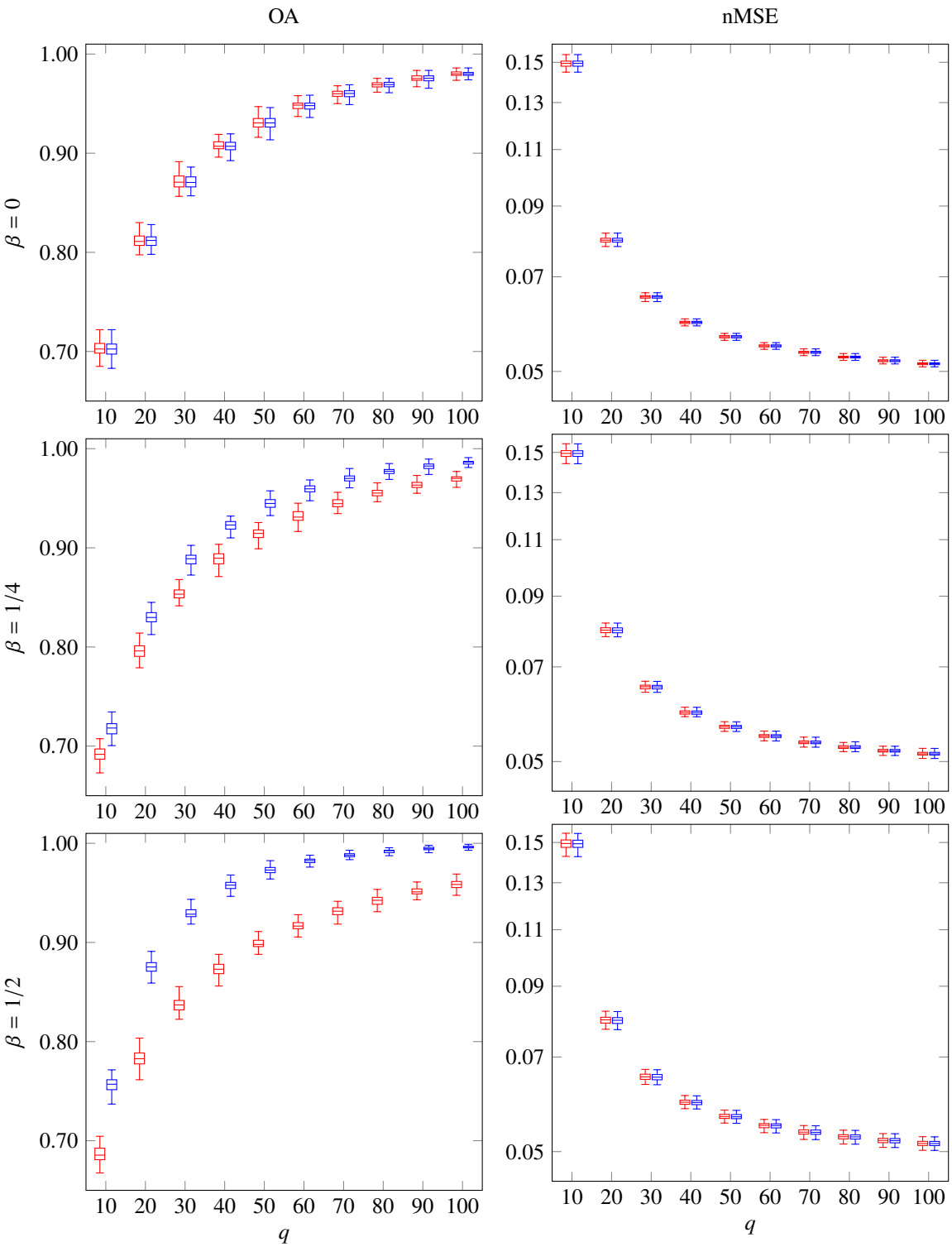

Fig. 6 Classification overall accuracy (OA, left panel) and reconstruction normalized mean-squared error (nMSE, right panel in log scale) boxplots computed on simulated data. Comparison between M2GP (blue) and MIGP (red) results as functions of the number q of discretization times. From top to bottom: $\beta = 0$, $\beta = 1/4$ and $\beta = 1/2$.

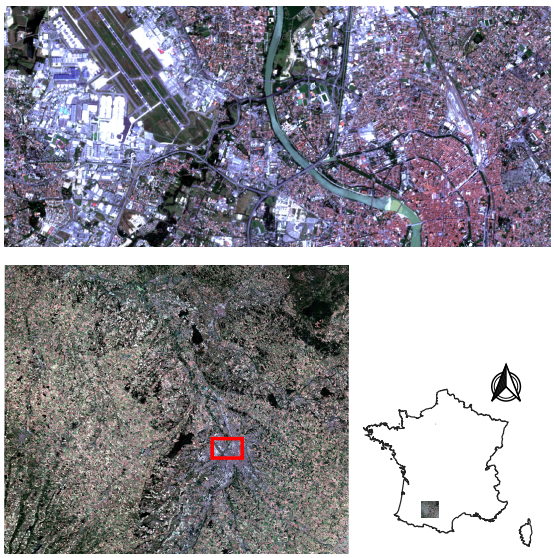

Fig. 7 The study area is located in the south of France (right bottom image). The left bottom image corresponds to the entire area (100 km×100 km) and the upper image is a zoom over the red rectangle (11 km×5 km).

concludes this section with classification results and comparisons to state-of-the-art methods.

6.1 Sentinel-2 satellite image time-series

Since 2016, the Sentinel-2 mission [67] produces massive multispectral images,[2] around 1.6TBytes a day, with a spatial resolution of 10 m/pixel and 13 spectral bands (only 10 bands are used for the analysis). The frequency of revisit is 5 days and clouds as well as shadows are present in the data, at random locations. Most of the clouds and shadows positions are automatically extracted by the data provider. Yet, thin clouds may remain in the data. The selected images cover the area of Toulouse, France (Fig. 7) and all available acquisitions for the year 2018 were used. The image is of spatial size 10,000×10,000 pixels (10,000 km^2). Each extracted time-series i has a dimension of $p = 10$ channels (or bands) and its own number of timestamps q_i. The distribution of the q_is is represented in Fig. 8 for this area in 2018.

Fourteen classes were extracted from national data-bases and 10 pairs of training and validation data-sets are generated independently for the experiments by randomly selecting samples for the training

[2] https://sentinel.esa.int/web/sentinel/missions/sentinel-2/data-products.

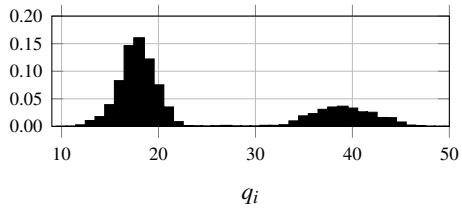

Fig. 8 Normalized histogram of the q_is within the SITS data-set.

and testing sets. Training and testing sets were carefully constructed to avoid spatial dependence between pixels.

Table 1 shows the number of extracted samples for each training and validation set. The number of samples per class is unbalanced but represents the actual proportion of land cover classes in the region.

Table 1 Land cover classes and number of extracted samples n_c per class for each training and validation set.

Class	n_c
Summer crops	40,000
Winter crops	30,000
Broad-leaved forest	10,000
Continuous urban fabric	10,000
Discontinuous urban fabric	10,000
Industrial or commercial units	10,000
Meadow	10,000
Orchards	10,000
Road surfaces	10,000
Vines	10,000
Water bodies	10,000
Woody moorlands	9,972
Coniferous forest	9,957
Natural grasslands	9,939
Total	189,868

6.2 Parameters estimation

M2GP is fitted to the satellite image time-series using the estimators described in Section 4. A Fourier basis is adopted for estimating the means using $J = 19$ functions while the time dependence structure is modeled by a RBF kernel combined with an additive white noise. The influence of the basis and the selection of the dimension J are discussed in the MIGP framework by [48, Fig. 8, and Fig. 1 in the supp. mat.].

Regarding the influence of the hyperparameters initial values, it has been observed in practice good convergence to similar local minima. Yet, some initial configurations might yield poor local minima, in

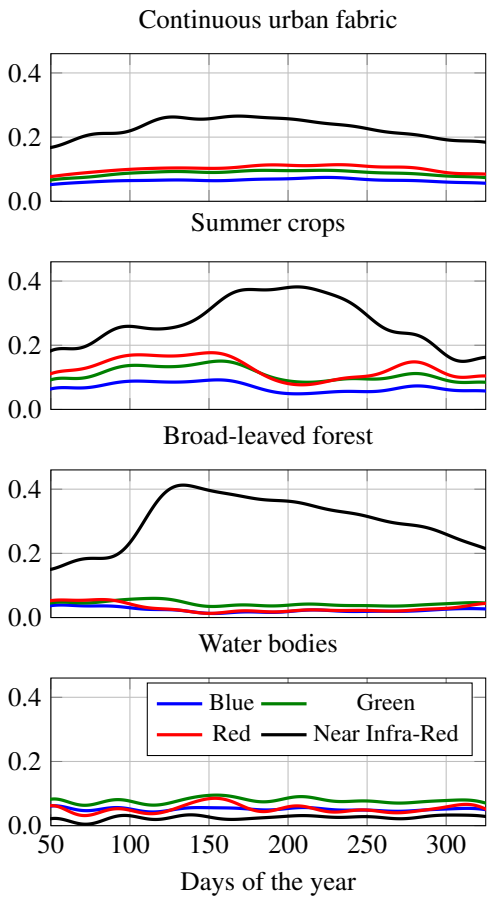

Fig. 9 Estimated means for four channels and four classes (continuous urban fabric, summer crops, broad-leaved forest and water bodies). The horizontal axis represents the days of the year and the vertical axis represents the reflectance value.

particular for which the level of the noise parameter is higher than the level of the RBF kernel (*i.e.*, $\sigma_c^2 \gg \gamma_c^2$ in (8)). In order to prevent such pathological situations, exclusive box constraints for these parameters were used during the optimization (with the L-BFGS-B algorithm).

Estimated mean functions are reported in Fig. 9 for four selected channels: blue, green, red and near infrared (nIR) and four selected classes: continuous urban fabric, summer crops, broad-leaved forest and water bodies. In the context of remote sensing data, nIR is often correlated with the presence or absence of vegetation: Large values of nIR associated with small values of red, indicate that the vegetation is abundant. This behavior is observed in agricultural classes such as summer crops or broad-leaved forest during spring and summer.

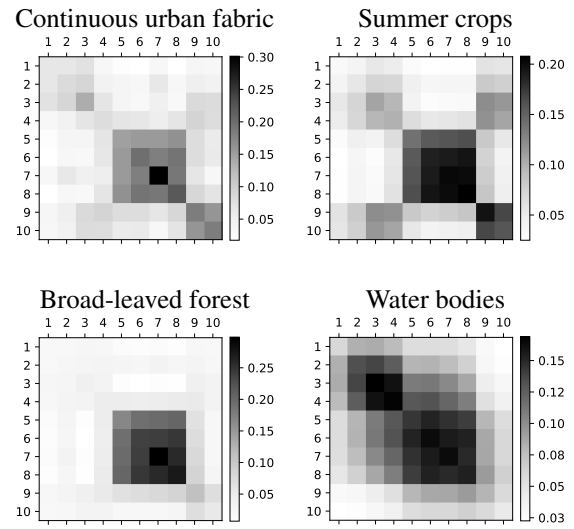

Fig. 10 Estimated covariance matrices $\mathbf{A}_c\mathbf{A}_c^\top$ on four land-cover classes.

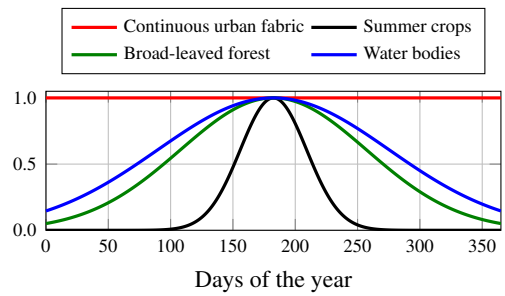

Fig. 11 Normalized RBF kernels (8) centered at day 180: $K(t, 180) = \exp(-0.5(t-180)^2/h_c^2)$ computed on four classes.

The estimated covariance matrices between all 10 channels $\widehat{\mathbf{A}_c\mathbf{A}_c^\top}$ are reported in Fig. 10 for the same classes. Similar covariance matrices have already been observed on mono-temporal Sentinel-2 data, we refer to [68, Fig. 8] for similar results on crops classes.

Finally, the time covariance structure is illustrated on Fig. 11. The estimated RBF kernel on the same four classes is drawn when centered at day 180. The temporal correlation associated with natural elements, such as summer crops or broad-leaved forest, is short since their reflectance evolves along the year (*e.g.* because of the vegetation cycle, or anthropic events). In contrast, man-made materials, such as continuous urban fabrics, exhibit longer temporal correlation because their reflectance does not evolve along time.[3]

[3]This is true when the period of observation is not too long, few years, otherwise the material property might be altered and its reflectance could vary.

6.3 Classification results

In this section, the classification performances of M2GP are compared to state-of-the-art methods. Four competitors are considered: Random forests (RF) [69], Quadratic discriminant analysis (QDA) and High Dimensional Data Analysis (HDDA) [70] which are based on a finite-dimensional Gaussian model, linear Support vector machine (SVM) classifier fitted with a Stochastic Gradient Descent [71], and, finally, Mixture of independent Gaussian processes (MIGP) [48].

The time-series have been resampled on a common temporal grid of size 73 (every 5 days of year 2018) using a linear interpolation for RF, QDA, HDDA and SVM methods since they require a fix vectorial representation of the sample. All the spectral bands have been stacked together to obtain a vector of dimension 73 dates $\times$ 10 spectral bands = 730 features. This strategy was used to produce the first land use and land cover maps in France [12] and is now commonly used at large scale. Yet, for other applications, e.g., changes detection, the interpolation algorithm choice may be more critical than for pixel-wise classification [72].

RF is trained with 100 trees of depth 25, HDDA uses the best model selected by cross-validation among eight possibilities [73] (corresponding to different assumptions on the covariance structure), and QDA is used with a regularized covariance matrix [74], $\tilde{\mathbf{\Sigma}} = (1-\epsilon)\hat{\mathbf{\Sigma}} + \epsilon\mathbf{I}$, where $\epsilon = 10^{-2}$ is selected by cross-validation. Finally, it has been observed on several random starts that the results of Table 2 where not sensitive to the initialization, once proper box constraints were enforced, as discussed in Section 6.2.

The F_1-score is computed to assess numerically the classification accuracy. The F_1 is defined as the harmonic mean of the precision and recall scores [75]. Classification maps are also presented in order to qualitatively evaluate the spatial coherency of the results (despite a spatial pixel-wise independence assumption made by all considered methods).

Means F_1 scores and their standard deviations computed on 10 independent sets are reported in Table 2 for each class as well as the "average F_1 score" computed on all classes. Non-parametric methods (RF and SVM) provide the best classification results in terms of F1-score. Interestingly, the covariance structure assumption of HDDA results in worse results in terms of classification accuracy than QDA. The Kronecker structure and the irregular sampling of M2GP lead to an accuracy similar to QDA but lower than non-parametric techniques. Finally, the MIGP yields the lowest accuracy, highlighting the spectral dependency on Sentinel-2 SITS.

The obtained classification maps are reported in Fig. 12 for three different sites. Large differences are observed in these scenes. For the first column, corresponding to the airport zone, most of the inner vegetations are wrongly classified to natural grasslands with QDA, while RF, SVM and M2GP classify correctly them as meadow. Runway are mostly confused with industrial/commercial units using RF while runways are almost recovered by M2GP. Overall, strong differences between thematic maps are observed, but visual assessment from a mono-date color image is difficult. Yet, without taking into account the spatial dependence, M2GP recovers most of the spatial structure of the image, and the *salt and pepper* classification noise is limited, as for RF and SVM.

7 Discussion

A multivariate Gaussian process model has been introduced for the classification of irregularly sampling satellite image time-series. The multivariate model involves a specific structure of the covariance operator that exploits the data features and also reduces the number of parameters to estimate. Furthermore, the proposed formulation scales linearly w.r.t. the number of samples. Experimental results on simulated and real data sets with irregular sampling show the importance of modeling the dependence between coordinates of the process, in particular for classification accuracy.

Because of the high dimensionality of SITS data, all model-based classifiers investigated in the previous section include some regularization: ridge regularization (QDA) and parsimonious covariance structures (M2GP and HDDA). In the QDA case, the regularization acts as a shrinkage of the eigenvalues associated with the full spectro-temporal covariance matrix (of size 730 in the experiments) and thus robustifies the computation of the inverse. In the M2GP case, the Kronecker structure of the covariance matrix simplifies the processing by inverting two matrices $\mathbf{\Sigma}$ and $\mathbf{A}\mathbf{A}^\top$ of much smaller sizes (less than 50 for the first one and 10 for the second one in our experiments). Interestingly, while QDA and M2GP impose different covariance structures for each class, they reach a similar classification accuracy, significantly higher than

Table 2 Mean F_1 score (mean(%) ± standard deviation) on the 10 independent data-sets.

	Non-parametric		Model-based			
	RF	SVM	QDA	HDDA	MIGP	M2GP
Summer crops	96.8 ± 0.45	95.6 ± 0.81	96.5 ± 0.27	89.0 ± 3.98	90.0 ± 0.83	95.9 ± 0.44
Winter crops	94.0 ± 0.77	93.9 ± 0.66	91.6 ± 0.48	85.2 ± 2.74	80.2 ± 0.83	92.2 ± 0.64
Broad-leaved forest	86.2 ± 2.35	85.3 ± 2.63	77.4 ± 3.91	65.7 ± 10.82	75.7 ± 5.03	81.5 ± 3.10
Cont. urban fabric	58.0 ± 1.55	55.9 ± 2.49	39.8 ± 6.18	12.5 ± 12.78	21.4 ± 3.49	30.9 ± 5.51
Discont. urban fabric	57.3 ± 3.44	40.2 ± 12.61	58.5 ± 1.39	47.2 ± 4.04	42.5 ± 3.17	54.5 ± 0.80
Ind. or commercial units	60.3 ± 1.35	48.3 ± 4.05	31.3 ± 2.14	31.5 ± 8.32	27.4 ± 0.92	38.4 ± 2.34
Meadow	64.8 ± 2.94	63.0 ± 3.17	58.3 ± 4.14	47.6 ± 6.91	43.3 ± 3.80	55.0 ± 4.19
Orchards	81.0 ± 2.64	76.4 ± 3.11	72.9 ± 4.05	62.0 ± 5.48	51.9 ± 5.46	77.6 ± 3.58
Road surfaces	87.1 ± 1.87	78.7 ± 2.79	73.1 ± 1.92	52.1 ± 20.77	54.2 ± 5.79	75.0 ± 2.06
Vines	78.9 ± 6.86	78.5 ± 6.57	71.1 ± 4.35	64.4 ± 10.34	60.9 ± 7.61	71.7 ± 5.18
Water bodies	99.4 ± 0.08	99.3 ± 0.10	98.7 ± 0.35	95.0 ± 2.95	84.9 ± 5.38	96.8 ± 0.84
Woody moorlands	56.6 ± 3.50	56.1 ± 3.85	23.9 ± 7.70	41.6 ± 4.76	14.1 ± 5.52	10.6 ± 12.00
Coniferous forest	86.9 ± 2.76	87.0 ± 2.56	76.6 ± 7.24	74.6 ± 6.13	61.2 ± 5.41	82.4 ± 6.61
Natural grasslands	30.7 ± 16.90	19.4 ± 14.68	29.8 ± 12.88	9.9 ± 10.48	15.4 ± 7.86	20.6 ± 8.46
Average F_1 score	74.2 ± 1.78	69.4 ± 1.78	64.2 ± 1.36	55.6 ± 2.70	51.7 ± 1.63	63.1 ± 1.15

the HDDA one. We thus believe that the M2GP covariance structure is well suited to deal with irregularly sampled SITS while the HDDA one is not supported by such data. Yet, in terms of classification accuracy, the M2GP model is not flexible enough to reach the performance of non-parametric methods, this is the price to pay for interpretability. The Gaussian (and thus unimodal) assumption for the class conditional density might be too limiting. The multi-modality of Sentinel-2 SITS can be addressed thanks to an unsupervised use of M2GP at the price of an increased computational cost. M2GP could be also extended to non-Gaussian processes, *e.g.* Student-t as in [40, 76], or by including mixed effect to account for spatial dependencies [77, 78] with a particular focus on scalability. Extension to non-Gaussian processes would not preserve the LMC framework unless the processes are stable under linear combination, leading to an increased complexity.

Finally, another extension of the proposed model would consist in considering time-series with both irregular temporal and spectral samplings. This would allow for the use of two satellite sources. Indeed, Sentinel-2 satellites are complemented with Sentinel-1 ones which acquire radar data, with a different physical content not affected by clouds.

Acknowledgments. The authors would like to thank S. Iovleff for his support and advices during the design of the model. The authors would also like to thank Y. Tanguy for his help when using the CNES computational resources to run the experiments presented in this paper.

This version of the article has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/s11222-022-10145-8. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms.

Declarations

Funding. This work is supported by the French National Research Agency in the framework of the Investissements d'Avenir program (ANR-15-IDEX-02) and by the Centre National d'Etudes Spatiales (CNES).

Conflict of interest. The authors declare that they have no conflict of interest.

Proofs

Proof of Proposition 1

Let $\mathbf{Y} \sim \mathcal{MGP}_p(\mathbf{m}, K, \mathbf{A})$ and introduce $\mathbf{Y}^\star$ the $p \times q$ random matrix defined as $\mathbf{Y}^\star = (\mathbf{Y}(t_1), \ldots, \mathbf{Y}(t_q))$ where $(t_1, \ldots, t_q) \in \mathcal{T}^q$. From (1), we have $\mathbf{Y} = \mathbf{AW} + \mathbf{m}$ with $\mathbf{W} \sim \mathcal{IGP}_p(0, K)$. Let $\mathbf{W}^\star =$

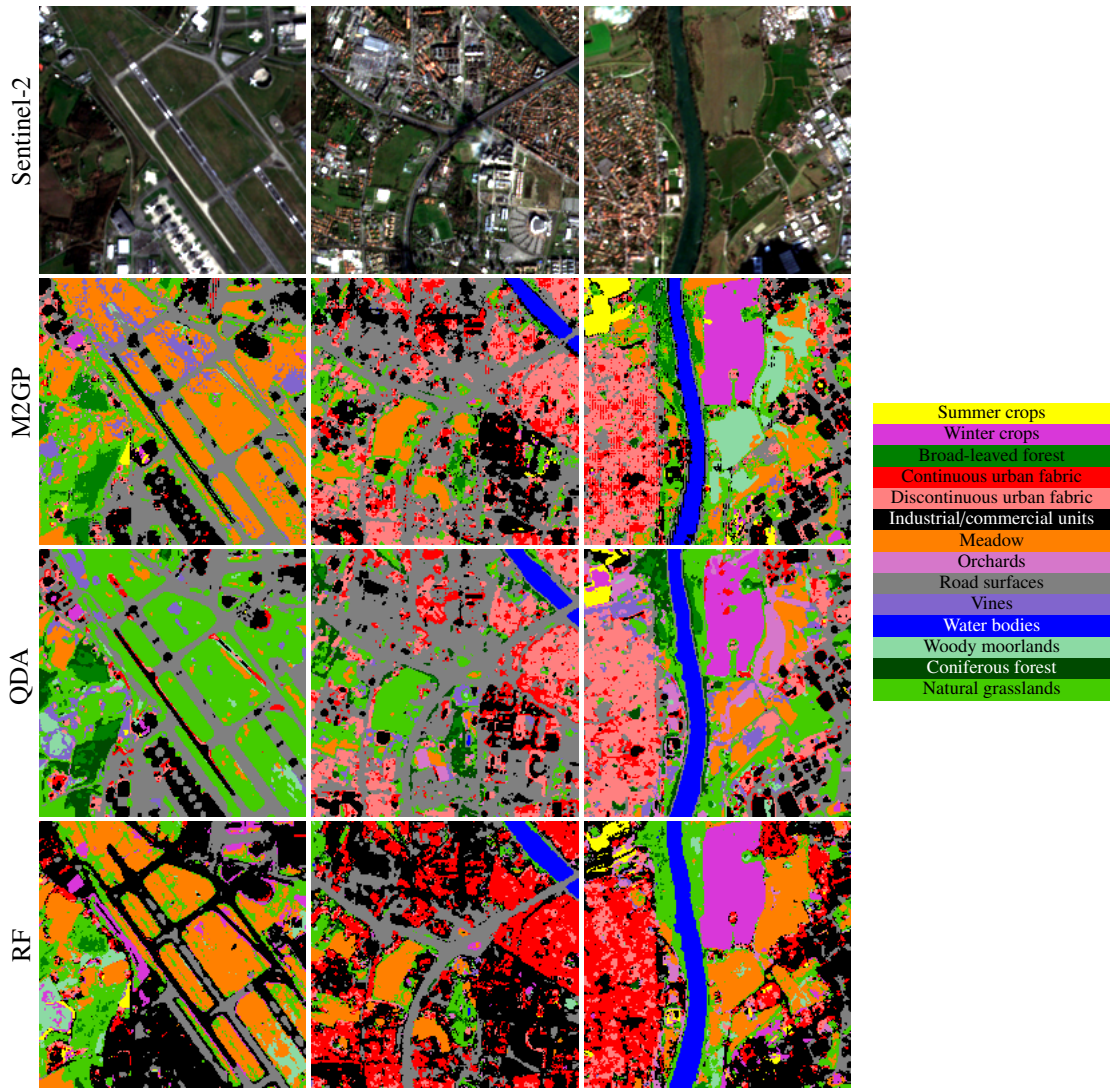

Fig. 12 Three extracts of the classification maps obtained by M2GP and the best methods among model-based and non-parametric families, QDA and RF respectively.

$(\mathbf{W}(t_1), \ldots, \mathbf{W}(t_q))$ be the associated $p \times q$ random matrix. Our first goal is to prove that $\mathbf{W}^\star \sim \mathcal{MN}_{p,q}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I}_p)$ or, equivalently, from (3), to prove that $\mathrm{vec}(\mathbf{W}^\star) \sim \mathcal{N}_{pq}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I}_p)$. To this end, let us consider the random variable

$$S = \sum_{b=1}^{p} \sum_{j=1}^{q} \lambda_{b,j} \mathbf{W}^\star_{b,j},$$

and let us prove that S is a Gaussian random variable pour all $\lambda_{b,j} \in \mathbb{R}$. Clear, one also has

$$S = \sum_{b=1}^{p} S_b, \text{ with } S_b := \sum_{j=1}^{q} \lambda_{b,j} \mathbf{W}^\star_{b,j} = \sum_{j=1}^{q} \lambda_{b,j} \mathbf{W}_b(t_j),$$

where $S_1, \ldots, S_p$ are independent centered Gaussian random variables with variance

$$\mathrm{var}(S_b) = \sum_{j=1}^{q} \sum_{j'=1}^{q} \lambda_{b,j} \lambda_{b,j'} \Sigma_{j,j'}.$$

As a consequence, S is a centered Gaussian random variable with variance

$$\begin{aligned}\text{var}(S) &= \sum_{b=1}^{p} \text{var}(S_b) \\ &= \sum_{b=1}^{p}\sum_{b'=1}^{p}\sum_{j=1}^{q}\sum_{j'=1}^{q} \lambda_{b,j}\lambda_{b,j'}\Sigma_{j,j'}(\mathbf{I}_p)_{b,b'}.\end{aligned}$$

As a conclusion, $\text{vec}(\mathbf{W}^\star) \sim \mathcal{N}_{pq}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}_p)$ and thus $\mathbf{W}^\star \sim \mathcal{MN}_{p,q}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_p)$. Finally, $\mathbf{Y}^\star = \mathbf{A}\mathbf{W}^\star + \mathbf{M} \sim \mathcal{MN}_{p,q}(\mathbf{M}, \boldsymbol{\Sigma}, \mathbf{A}\mathbf{A}^\top)$, see [49, Example 1].

Proof of Lemma 1

Combining (4) and (9) yields that the density of $\mathbf{Y}^{i,\star}$ conditionally to $Z_i = c$ is given for all $i = 1, \ldots, n$ by

$$\begin{aligned}-\log p_{i,c}(\mathbf{y}) &= \frac{pq_i}{2}\log(2\pi) + \frac{p}{2}\log\det(\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)) \\ &+ \frac{q_i}{2}\log\det(\mathbf{A}_c\mathbf{A}_c^\top) \\ &+ \frac{1}{2}\text{tr}\left[(\mathbf{A}_c\mathbf{A}_c^\top)^{-1}(\mathbf{y} - \boldsymbol{\alpha}_c\mathbf{B}^i)(\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c))^{-1}(\mathbf{y} - \boldsymbol{\alpha}_c\mathbf{B}^i)^\top\right].\end{aligned}$$

The negative log-likelihood can be written as

$$\begin{aligned}\mathcal{L} &= -\sum_{c=1}^{C}\sum_{i|Z_i=c}\log p_{i,c}(\mathbf{Y}^{i,\star}) \\ &:= \frac{1}{2}\sum_{c=1}^{C}\ell_c(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c\mathbf{A}_c^\top) + \frac{p\log(2\pi)}{2}\sum_{c=1}^{C}\sum_{i|Z_i=c} q_i,\end{aligned}$$

with, for all $c = 1, \ldots, C$,

$$\begin{aligned}\ell_c(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c\mathbf{A}_c^\top) &= p\sum_{i|Z_i=c}\log\det(\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)) + \\ &\sum_{i|Z_i=c} q_i\log\det(\mathbf{A}_c\mathbf{A}_c^\top) + \\ &\sum_{i|Z_i=c}\text{tr}\left[(\mathbf{A}_c\mathbf{A}_c^\top)^{-1}(\mathbf{Y}^{i,\star} - \boldsymbol{\alpha}_c\mathbf{B}^i)(\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c))^{-1}(\mathbf{Y}^{i,\star} - \boldsymbol{\alpha}_c\mathbf{B}^i)^\top\right].\end{aligned}$$

The conclusion follows.

Proof of Proposition 2

(i) Let $\boldsymbol{\beta}^{c,i}(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c) = (\mathbf{Y}^{i,\star} - \boldsymbol{\alpha}_c\mathbf{B}^i)\{\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)\}^{-1}$ and consider the differential of $\ell_c(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c\mathbf{A}_c^\top)$ w.r.t. $\boldsymbol{\alpha}_c$:

$$\begin{aligned}&\text{d}\ell_c(\boldsymbol{\alpha}_c) \\ &= -\sum_{i|Z_i=c}\text{tr}\left(\{\mathbf{A}_c\mathbf{A}_c^\top\}^{-1}(\text{d}\boldsymbol{\alpha}_c)\mathbf{B}^i\{\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)\}^{-1}\boldsymbol{\beta}^{c,i}(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c)^\top\right) \\ &\quad -\sum_{i|Z_i=c}\text{tr}\left(\{\mathbf{A}_c\mathbf{A}_c^\top\}^{-1}\boldsymbol{\beta}^{c,i}(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c)(\mathbf{B}^i)^\top(\text{d}\boldsymbol{\alpha}_c)^\top\right) \\ &= -2\sum_{i|Z_i=c}\text{tr}\left(\{\mathbf{A}_c\mathbf{A}_c^\top\}^{-1}\boldsymbol{\beta}^{c,i}(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c)(\mathbf{B}^i)^\top(\text{d}\boldsymbol{\alpha}_c)^\top\right),\end{aligned}$$

by remarking that both terms are equal in view of the properties of the trace operator. Moreover, from Kronecker product properties [79, Theorem 8.12], one has

$$\begin{aligned}&-\frac{1}{2}\text{d}\ell_c(\boldsymbol{\alpha}_c) \\ &= \sum_{i|Z_i=c}\text{vec}(\text{d}\boldsymbol{\alpha}_c)^\top\left(\mathbf{B}^i\{\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)\}^{-1} \otimes \{\mathbf{A}_c\mathbf{A}_c^\top\}^{-1}\right) \\ &\qquad\qquad \times \text{vec}(\mathbf{Y}^{i,\star} - \boldsymbol{\alpha}_c\mathbf{B}^i) \\ &= (\text{dvec}(\boldsymbol{\alpha}_c))^\top\text{vec}\left(\{\mathbf{A}_c\mathbf{A}_c^\top\}^{-1}\sum_{i|Z_i=c}\boldsymbol{\beta}^{c,i}(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c)(\mathbf{B}^i)^\top\right).\end{aligned}$$

Interpreting the above result as a scalar product and using the "broad" definition of matrix derivative defined in [80], if follows:

$$\begin{aligned}&\frac{\partial\ell_c(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c\mathbf{A}_c^\top)}{\partial\boldsymbol{\alpha}_c} = \\ &-2\text{vec}\left(\{\mathbf{A}_c\mathbf{A}_c^\top\}^{-1}\sum_{i|Z_i=c}\boldsymbol{\beta}^{c,i}(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c)(\mathbf{B}^i)^\top\right).\end{aligned}$$

Setting this partial derivative to zero yields

$$\sum_{i|Z_i=c}(\mathbf{Y}^{i,\star} - \boldsymbol{\alpha}_c\mathbf{B}^i)\{\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)\}^{-1}(\mathbf{B}^i)^\top = 0,$$

or equivalently, $\boldsymbol{\alpha}_c =$

$$\left[\sum_{i|Z_i=c}\mathbf{Y}^{i,\star}\{\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)\}^{-1}(\mathbf{B}^i)^\top\right]\left[\sum_{i|Z_i=c}\mathbf{B}^i\{\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)\}^{-1}(\mathbf{B}^i)^\top\right]^{-1},$$

which is the desired result. Second, let us consider the differential of $\ell_c(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c\mathbf{A}_c^\top)$ w.r.t. $\mathbf{A}_c\mathbf{A}_c^\top$:

$$\text{d}\ell_c(\mathbf{A}_c\mathbf{A}_c^\top) = Q_c\text{d}\log\det(\mathbf{A}_c\mathbf{A}_c^\top) + \text{dtr}\left(\mathbf{N}(\boldsymbol{\theta}_c)\{\mathbf{A}_c\mathbf{A}_c^\top\}^{-1}\right),$$

where

$$\mathbf{N}(\boldsymbol{\theta}_c) = \sum_{i|Z_i=c} (\mathbf{Y}^{i,\star} - \boldsymbol{\alpha}_c \mathbf{B}^i)\{\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)\}^{-1}(\mathbf{Y}^{i,\star} - \boldsymbol{\alpha}_c \mathbf{B}^i)^\top.$$

From [79, Example 9.6], the associated partial derivative vanishes for

$$\begin{aligned}\mathbf{A}_c\mathbf{A}_c^\top &= \frac{\mathbf{N}(\boldsymbol{\theta}_c)}{Q_c} \\ &= \frac{1}{Q_c}\sum_{i|Z_i=c} (\mathbf{Y}^{i,\star} - \boldsymbol{\alpha}_c \mathbf{B}^i)\{\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)\}^{-1}(\mathbf{Y}^{i,\star} - \boldsymbol{\alpha}_c \mathbf{B}^i)^\top,\end{aligned}$$

and the result is proved.

(*ii*) Consider the kth coordinate of the gradient of $\ell_c(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c\mathbf{A}_c^\top)$ w.r.t. $\boldsymbol{\theta}$:

$$\begin{aligned}&\frac{\partial \ell_c(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c\mathbf{A}_c^\top)}{\partial \theta_k} \\ &= p\sum_{i|Z_i=c} \frac{\partial}{\partial \theta_k}\log\det(\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)) + \frac{\partial}{\partial \theta_k}\mathrm{tr}\left(\mathbf{N}(\boldsymbol{\theta}_c)\{\mathbf{A}_c\mathbf{A}_c^\top\}^{-1}\right) \\ &= p\sum_{i|Z_i=c} \mathrm{tr}\left(\{\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)\}^{-1}\frac{\partial \boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)}{\partial \theta_k}\right) \\ &- \sum_{i|Z_i=c} \mathrm{tr}\left(\boldsymbol{\beta}^{c,i}(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c)^\top\{\mathbf{A}_c\mathbf{A}_c^\top\}^{-1}\boldsymbol{\beta}^{c,i}(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c)\frac{\partial \boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta})}{\partial \theta_k}\right) \\ &= \sum_{i|Z_i=c} \mathrm{tr}\left(\left[p\{\boldsymbol{\Sigma}^{c,i}(\boldsymbol{\theta}_c)\}^{-1} - \Delta^{c,i}(\boldsymbol{\alpha}_c, \boldsymbol{\theta}_c, \mathbf{A}_c\mathbf{A}_c^\top)\right]\frac{\partial \boldsymbol{\Sigma}^{c,i}}{\partial \theta_k}(\boldsymbol{\theta}_c)\right).\end{aligned}$$

The result is proved.

Proof of Proposition 3

Let $\mathbf{Y}^\star = (\mathbf{Y}(t_1), \ldots, \mathbf{Y}(t_q))$ be a $p \times q$ random matrix where, conditionally to $Z = c$, $\mathbf{Y} \sim \mathcal{MGP}_c(\boldsymbol{\alpha}_c\mathbf{b}, K_c, \mathbf{A}_c)$. Recall that Proposition 1 yields $\mathrm{vec}(\mathbf{Y}^\star) \sim \mathcal{N}_{pq}(\mathrm{vec}(\boldsymbol{\alpha}_c\mathbf{B}), \boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c) \otimes \mathbf{A}_c\mathbf{A}_c^\top)$, where $\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c)$ is defined for all $(j, j') \in \{1, \ldots, q\}^2$ by $\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c)_{j,j'} = K_c(t_j, t_{j'}|\boldsymbol{\theta}_c)$ and $\mathbf{B} = (\mathbf{b}(t_1), \ldots, \mathbf{b}(t_q))$ is a $J \times q$ design matrix. Let $t^\dagger \in \mathcal{T}$ be an unobserved time, *i.e.* $t^\dagger \neq t_k$, for all $k \in \{1, \ldots, q\}$, and $\mathbf{k}_c(t^\dagger) = (K_c(t^\dagger, t_1|\boldsymbol{\theta}_c), \ldots, K_c(t^\dagger, t_q|\boldsymbol{\theta}_c))^\top$. Then, classical properties on conditional Gaussian random vectors (see for instance [81, p. 63]) entail that, conditionally to $Z = c$ and $\mathrm{vec}(\mathbf{Y}^\star) = \mathrm{vec}(\mathbf{y}^\star)$, $\mathbf{Y}(t^\dagger)$ follows the p-variate Gaussian distribution $\mathcal{N}_p(\boldsymbol{\mu}_c(t^\dagger, \mathbf{y}^\star), \boldsymbol{\Lambda}_c(t^\dagger))$ with, on the one hand

$$\begin{aligned}&\boldsymbol{\mu}_c(t^\dagger, \mathbf{y}^\star) - \boldsymbol{\alpha}_c\mathbf{b}(t^\dagger) \\ &= [\mathbf{k}_c(t^\dagger)^\top \otimes \mathbf{A}_c\mathbf{A}_c^\top]\{\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c) \otimes \mathbf{A}_c\mathbf{A}_c^\top\}^{-1}\mathrm{vec}(\mathbf{y}^\star - \boldsymbol{\alpha}_c\mathbf{B}) \\ &= [\mathbf{k}_c(t^\dagger)^\top \otimes \mathbf{A}_c\mathbf{A}_c^\top]\{\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c)^{-1} \otimes (\mathbf{A}_c\mathbf{A}_c^\top)^{-1}\}\mathrm{vec}(\mathbf{y}^\star - \boldsymbol{\alpha}_c\mathbf{B}) \\ &= \left[\{\mathbf{k}_c(t^\dagger)^\top\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c)^{-1}\} \otimes \{(\mathbf{A}_c\mathbf{A}_c^\top)(\mathbf{A}_c\mathbf{A}_c^\top)^{-1}\}\right]\mathrm{vec}(\mathbf{y}^\star - \boldsymbol{\alpha}_c\mathbf{B}) \\ &= \left[\{\mathbf{k}_c(t^\dagger)^\top\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c)^{-1}\} \otimes \mathbf{I}_p\right]\mathrm{vec}(\mathbf{y}^\star - \boldsymbol{\alpha}_c\mathbf{B}) \\ &= \mathrm{vec}\left(\mathbf{I}_p(\mathbf{y}^\star - \hat{\boldsymbol{\alpha}}_c\mathbf{B})\{\mathbf{k}_c(t^\dagger)^\top\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c)^{-1}\}^\top\right) \\ &= (\mathbf{y}^\star - \boldsymbol{\alpha}_c\mathbf{B})\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c)^{-1}\mathbf{k}_c(t^\dagger),\end{aligned}$$

and on the other hand,

$$\begin{aligned}&\boldsymbol{\Lambda}_c(t^\dagger) - K_c(t^\dagger, t^\dagger|\boldsymbol{\theta}_c) \otimes \mathbf{A}_c\mathbf{A}_c^\top \\ &= -[\mathbf{k}_c(t^\dagger)^\top \otimes \mathbf{A}_c\mathbf{A}_c^\top]\left\{\boldsymbol{\Sigma}^c \otimes \mathbf{A}_c\mathbf{A}_c^\top(\boldsymbol{\theta}_c)\right\}^{-1}[\mathbf{k}_c(t^\dagger) \otimes \mathbf{A}_c\mathbf{A}_c^\top] \\ &= -\left[(\mathbf{k}_c(t^\dagger)^\top\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c)^{-1}) \otimes \mathbf{I}_p\right][\mathbf{k}_c(t^\dagger) \otimes \mathbf{A}_c\mathbf{A}_c^\top] \\ &= -\left(\mathbf{k}_c(t^\dagger)^\top\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c)^{-1}\mathbf{k}_c(t^\dagger)\right) \otimes (\mathbf{I}_p\mathbf{A}_c\mathbf{A}_c^\top).\end{aligned}$$

Finally,

$$\boldsymbol{\Lambda}_c(t^\dagger) = \left[K_c(t^\dagger, t^\dagger|\boldsymbol{\theta}_c) - \mathbf{k}_c(t^\dagger)^\top\boldsymbol{\Sigma}^c(\boldsymbol{\theta}_c)^{-1}\mathbf{k}_c(t^\dagger)\right] \otimes \mathbf{A}_c\mathbf{A}_c^\top$$

and the result is proved.

References

[1] Li, C., Wulf, H., Schmid, B., He, J.-S., Schaepman, M.E.: Estimating Plant Traits of Alpine Grasslands on the Qinghai-Tibetan Plateau Using Remote Sensing. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **11**(7), 2263–2275 (2018)

[2] Fauvel, M., Lopes, M., Dubo, T., Rivers-Moore, J., Frison, P.-L., Gross, N., Ouin, A.: Prediction of plant diversity in grasslands using Sentinel-1 and -2 satellite image time series. Remote Sensing of Environment **237**, 111536 (2020)

[3] Liu, X., Gopal, V., Kalagnanam, J.: A spatio-temporal modeling framework for weather radar image data in tropical Southeast Asia. The Annals of Applied Statistics **12**(1), 378–407 (2018)

[4] Bertolacci, M., Cripps, E., Rosen, O., Lau, J.W., Cripps, S.: Climate inference on daily rainfall across the Australian continent, 1876-2015. The

Annals of Applied Statistics **13**(2), 683–712 (2019)

[5] Useya, J., Chen, S.: Comparative Performance Evaluation of Pixel-Level and Decision-Level Data Fusion of Landsat 8 OLI, Landsat 7 ETM+ and Sentinel-2 MSI for Crop Ensemble Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **11**(11), 4441–4451 (2018)

[6] Moeini Rad, A., Ashourloo, D., Salehi Shahrabi, H., Nematollahi, H.: Developing an Automatic Phenology-Based Algorithm for Rice Detection Using Sentinel-2 Time-Series Data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(5), 1471–1481 (2019)

[7] Feng, S., Zhao, J., Liu, T., Zhang, H., Zhang, Z., Guo, X.: Crop Type Identification and Mapping Using Machine Learning Algorithms and Sentinel-2 Time Series Data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(9), 3295–3306 (2019)

[8] Manolakis, D.G., Lockwood, R.B., Cooley, T.W.: Hyperspectral Imaging Remote Sensing: Physics, Sensors, and Algorithms. Cambridge University Press, Cambridge (2016)

[9] Landgrebe, D.A.: Signal Theory Methods in Multispectral Remote Sensing. Wiley, Newark, NJ (2005)

[10] Pettitt, A.N., Weir, I.S., Hart, A.G.: A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. Statistics and Computing **12**(4), 353–367 (2002)

[11] Lopes, M., Fauvel, M., Ouin, A., Girard, S.: Spectro-temporal heterogeneity measures from dense high spatial resolution satellite image time series: Application to grassland species diversity estimation. Remote Sensing **9**(10) (2017)

[12] Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I.: Operational high resolution land cover map production at the country scale using Satellite Image Time Series. Remote Sensing **9**(1) (2017)

[13] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, And Prediction. Springer, New-York (2009)

[14] Theodossiou, P.: Financial Data and the Skewed Generalized T Distribution. Management Science **44**(12-Part-1), 1650–1661 (1998)

[15] Chamroukhi, F.: Skew t mixture of experts. Neurocomputing **266**, 390–408 (2017)

[16] Andrews, J.L., McNicholas, P.D.: Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. Statistics and Computing **22**(5), 1021–1029 (2012)

[17] Murray, P.M., Browne, R.P., McNicholas, P.D.: A mixture of SDB skew-t factor analyzers. Econometrics and Statistics **3**(C), 160–168 (2017)

[18] Bouveyron, C., Celeux, G., Murphy, T.B., Raftery, A.E.: Model-Based Clustering and Classification for Data Science: With Applications in R. Cambridge University Press, Cambridge (2019)

[19] Povinelli, R.J., Johnson, M.T., Lindgren, A.C., Jinjin Ye: Time series classification using Gaussian mixture models of reconstructed phase spaces. IEEE Transactions on Knowledge and Data Engineering **16**(6), 779–783 (2004)

[20] Celeux, G., Govaert, G.: Clustering criteria for discrete data and latent class models. Journal of Classification **8**(2), 157–176 (1991)

[21] Bouguila, N., Ziou, D., Vaillancourt, J.: Novel Mixtures Based on the Dirichlet Distribution: Application to Data and Image Classification. In: Perner, P., Rosenfeld, A. (eds.) Machine Learning and Data Mining in Pattern Recognition. Lecture Notes in Computer Science, pp. 172–181. Springer, Berlin, Heidelberg (2003)

[22] Biernacki, C., Jacques, J.: Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. Statistics and Computing **26**(5), 929–943 (2016)

[23] Hofmann, T., Schölkopf, B., Smola, A.J.: Kernel methods in machine learning. The Annals of Statistics **36**(3), 1171–1220 (2008)

[24] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C.: Text classification using string kernels. The Journal of Machine Learning Research **2**, 419–444 (2002)

[25] Kriege, N.M., Johansson, F.D., Morris, C.: A survey on graph kernels. Applied Network Science **5**(1), 1–42 (2020)

[26] Alvarez, M.A., Rosasco, L., Lawrence, N.D.: Kernels for vector-valued functions: A review. Foundations and Trends ® in Machine Learning **4**(3), 195–266 (2012)

[27] Flaxman, S., Chirico, M., Pereira, P., Loeffler, C.: Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ "Real-Time Crime Forecasting Challenge". The Annals of Applied Statistics **13**(4), 2564–2585 (2019)

[28] Bouveyron, C., Fauvel, M., Girard, S.: Kernel discriminant analysis and clustering with parsimonious Gaussian process models. Statistics and Computing **25**(6), 1143–1162 (2015)

[29] Shen, H., Li, X., Cheng, Q., Zeng, C., Yang, G., Li, H., Zhang, L.: Missing Information Reconstruction of Remote Sensing Data: A Technical Review. IEEE Geoscience and Remote Sensing Magazine **3**(3), 61–85 (2015)

[30] Schafer, J.L.: Analysis of Incomplete Multivariate Data. Chapman and Hall/CRC, New-York (1997)

[31] García-Laencina, P.J., Sancho-Gómez, J.-L., Figueiras-Vidal, A.R.: Pattern classification with missing data: a review. Neural Computing and Applications **19**(2), 263–282 (2010)

[32] Bagnall, A., Lines, J., Bostrom, A., Large, J., Keogh, E.: The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery **31**(3), 606–660 (2017)

[33] Lin, W.-C., Tsai, C.-F.: Missing value imputation: a review and analysis of the literature (2006-2017). Artificial Intelligence Review **53**(2), 1487–1509 (2020)

[34] Ramsay, J., Silverman, B.W.: Functional Data Analysis. Springer Series in Statistics. Springer, New-York (2005)

[35] Schmutz, A., Jacques, J., Bouveyron, C., Chèze, L., Martin, P.: Clustering multivariate functional data in group-specific functional subspaces. Computational Statistics **35**(3), 1101–1131 (2020)

[36] Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis: Theory And Practice. Springer Series in Statistics. Springer, New-York (2006)

[37] Chouakria, A.D., Nagabhushan, P.N.: Adaptive dissimilarity index for measuring time series proximity. Advances in Data Analysis and Classification **1**(1), 5–21 (2007)

[38] Bonilla, E.V., Chai, K., Williams, C.: Multi-task Gaussian Process Prediction. Advances in Neural Information Processing Systems **20**, 153–160 (2007)

[39] Shah, A., Wilson, A., Ghahramani, Z.: Student-t Processes as Alternatives to Gaussian Processes. In: Artificial Intelligence and Statistics, pp. 877–885. PMLR, Reykjavik, Iceland (2014)

[40] Chen, Z., Wang, B., Gorban, A.N.: Multivariate Gaussian and Student-t process regression for multi-output prediction. Neural Computing and Applications **32**(8), 3005–3028 (2020)

[41] Hartmann, M., Vanhatalo, J.: Laplace approximation and natural gradient for Gaussian process regression with heteroscedastic Student-t model. Statistics and Computing **29**(4), 753–773 (2019)

[42] Nickisch, H., Rasmussen, C.E.: Approximations for binary Gaussian process classification. Journal of Machine Learning Research **9**(67), 2035–2078 (2008)

[43] Hensman, J., Matthews, A., Ghahramani, Z.: Scalable variational Gaussian process classification. In: Lebanon, G., Vishwanathan, S.V.N.

(eds.) Artificial Intelligence and Statistics. PMLR, vol. 38, pp. 351–360. San Diego, California, USA (2015)

[44] Vecchia, A.V.: Estimation and model identification for continuous spatial processes. Journal of the Royal Statistical Society: Series B **50**(2), 297–312 (1988)

[45] Katzfuss, M., Guinness, J.: A general framework for Vecchia approximations of Gaussian processes. Statistical Science **36**(1), 124–141 (2021)

[46] Liu, H., Ong, Y.-S., Shen, X., Cai, J.: When Gaussian process meets big data: A review of scalable GPs. IEEE Transactions on Neural Networks and Learning Systems **31**(11), 4405–4423 (2020)

[47] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)

[48] Constantin, A., Fauvel, M., Girard, S.: Joint supervised classification and reconstruction of irregularly sampled satellite image times series. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2022)

[49] Dawid, A.P.: Some matrix-variate distribution theory: Notational considerations and a Bayesian application. Biometrika **68**(1), 265–274 (1981)

[50] Srivastava, M.S., von Rosen, T., von Rosen, D.: Models with a Kronecker product covariance structure: estimation and testing. Mathematical Methods of Statistics **17**(4), 357–370 (2008)

[51] Gupta, A.K., Nagar, D.K.: Matrix Variate Distributions. Chapman and Hall/CRC, New-York (1999)

[52] Shi, J.Q., Murray-Smith, R., Titterington, D.M.: Hierarchical Gaussian process mixtures for regression. Statistics and Computing **15**(1), 31–41 (2005)

[53] Ren, Q., Banerjee, S.: Hierarchical factor models for large spatially misaligned data: A low-rank predictive process approach. Biometrics **69**(1), 19–30 (2013)

[54] Goovaerts, P.: Geostatistics for Natural Resources Evaluation. Oxford University Press, USA (1997). Publisher: Cambridge University Press

[55] Goulard, M.: Inference in a coregionalization model. In: Geostatistics, vol. 4, pp. 397–408. Springer, Dordrecht (1989)

[56] Zhang, L., Banerjee, S.: Spatial factor modeling: a Bayesian matrix-normal approach for misaligned data. Biometrics, 1–14 (2021)

[57] Spinnato, J., Roubaud, M., Burle, B., Torrésani, B.: Finding EEG space-time-scale localized features using matrix-based penalized discriminant analysis. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6004–6008 (2014)

[58] Glanz, H., Carvalho, L.: An expectation-maximization algorithm for the matrix normal distribution with an application in remote sensing. Journal of Multivariate Analysis **167**, 31–48 (2018)

[59] Mahanta, M.S., Aghaei, A.S., Plataniotis, K.N.: Regularized LDA based on separable scatter matrices for classification of spatio-spectral EEG patterns. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1237–1241 (2013)

[60] Allen, G.I., Tibshirani, R.: Transposable regularized covariance models with an application to missing data imputation. The Annals of Applied Statistics **4**(2), 764–790 (2010)

[61] Lu, N., Zimmerman, D.L.: The likelihood ratio test for a separable covariance matrix. Statistics & Probability Letters **73**(4), 449–457 (2005)

[62] Zhu, C., Byrd, R.H., Lu, P., Nocedal, J.: Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Transactions on Mathematical Software **23**(4), 550–560 (1997)

[63] Mardia, K.V., Goodall, C.R.: Spatial-temporal analysis of multivariate environmental monitoring data. In: Multivariate Environmental

Statistics vol. 6, pp. 347–385. Elsevier, North-Holland, New-York (1993)

[64] Dutilleul, P.: The MLE algorithm for the matrix normal distribution. Journal of Statistical Computation and Simulation **64**(2), 105–123 (1999)

[65] Manceur, A.M., Dutilleul, P.: Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. Journal of Computational and Applied Mathematics **239**, 37–49 (2013)

[66] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G.: API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pp. 108–122 (2013)

[67] Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P.: Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sensing of Environment **120**, 25–36 (2012)

[68] Wang, B., Jia, K., Liang, S., Xie, X., Wei, X., Zhao, X., Yao, Y., Zhang, X.: Assessment of Sentinel-2 MSI spectral band reflectances for estimating fractional vegetation cover. Remote Sensing **10**(12), 1927 (2018)

[69] Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (2001)

[70] Bouveyron, C., Girard, S., Schmid, C.: High-Dimensional Discriminant Analysis. Communications in Statistics - Theory and Methods **36**(14), 2607–2623 (2007)

[71] Zhang, T.: Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms. In: Icml 2004: Proceedings of the Twenty-First International Conference on Machine Learning. Omnipress, pp. 919–926 (2004)

[72] Holloway-Brown, J., Helmstedt, K.J., Mengersen, K.L.: Interpolating missing land cover data using stochastic spatial random forests for improved change detection. Remote Sensing in Ecology and Conservation **7**(4), 649–665 (2021)

[73] Bergé, L., Bouveyron, C., Girard, S.: HDclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data. Journal of Statistical Software **46**(6), 1–29 (2012)

[74] Friedman, J.H.: Regularized Discriminant Analysis. Journal of the American Statistical Association **84**(405), 165–175 (1989)

[75] Tharwat, A.: Classification assessment methods. Applied Computing and Informatics **17**(1), 168–192 (2021)

[76] Shah, A., Wilson, A., Ghahramani, Z.: Student-t Processes as Alternatives to Gaussian Processes. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, vol. 33, pp. 877–885. PMLR, Reykjavik, Iceland (2014)

[77] Stroup, W.W.: Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. CRC Press, Boca Raton (2012)

[78] Cressie, N.A.C.: Statistics for Spatial Data. John Wiley & Sons, Ltd, New-York (1993)

[79] Schott, J.R.: Matrix Analysis for Statistics. Wiley Series in Probability and Statistics. Wiley, New Jersey (2016)

[80] Magnus, J.R.: On the concept of matrix derivative. Journal of Multivariate Analysis **101**(9), 2200–2206 (2010)

[81] Bilodeau, M., Brenner, D.: Theory of Multivariate Statistics. Springer, New-York (2008)