Non-parametric regression on compositional covariates using Bayesian P-splines

Noname manuscript No.
(will be inserted by the editor)

Non-parametric regression on compositional covariates using Bayesian P-splines

Received: date / Accepted: date

Abstract Methods to perform regression on compositional covariates have recently been proposed using isometric log-ratios (*ilr*) representation of compositional parts. This approach consists of first applying standard regression on ilr coordinates (with no need for either constraint on regression coefficients or the use of Moore Penrose generalized inverse) and second, transforming the estimated *ilr* coefficients into their contrast logratios counterparts which give easy-to-interpret parameters for the linear effect of each compositional part, in relative terms, on the response. In this work we present an extension of this framework, where compositional covariate effects are allowed to be smooth in the *ilr* domain. This is achieved by fitting a smooth function over the multidimensional *ilr* space, using a basis of B-splines and a set of associated spline coefficients. Smoothness is achieved by assuming random walk priors on spline coefficients in a hierarchical Bayesian framework. The proposed methodology is illustrated on a spatial dataset from an ecological survey on a gypsum outcrop located in the Emilia Romagna Region, Italy.

Keywords Compositional data · Bayesian P-splines · Intrinsinc Gaussian Markov Random Fields · Spatial regression · Vegetation cover

1 Introduction

Compositional data consist of vectors whose components refer to proportion of a whole. The most fundamental property of a compositional data set is its *intrinsic* multivariate nature. Compositional data arise in several applied fields, ranging from mineral compositions of rocks to air and water pollution, poll data, economic activities and many others. In most of multivariate analysis, it is the decision of the analyst to treat individual observations of several variables jointly, in order to exploit, understand or modeling their possible mutual

dependence. On the contrary, a composition is multivariate by nature, because the importance of one single component can only be evaluated with respect to the total or with respect to other components. This fundamental feature of compositions gave rise to a specific literature concerning compositional data analysis, whose first comprehensive treatment was given in Aitchison (1986). Such literature is basically concerned with the construction of suitable statistical methods that need to take account of the natural geometry of compositional data: the geometry of the simplex. In fact, a D-part composition belongs to the simplex

$$\mathbb{S}^D = \left\{ (z_1, \ldots, z_d, \ldots, z_D) : z_d > 0, \quad d = 1, \ldots, D; \quad \sum_{d=1}^{D} z_d = c \right\},$$

where c is a positive constant whose value is irrelevant because only the ratios of the parts contain the relevant information.

The case study that motivates the present research refers to an ecological application: in particular, we investigate the relationship between vegetation cover and substrate typologies within a rupicolous basophilic habitat defined as priority by the European Commission (Council Directive 92/43/CEE). Available data consist of post-processed ground photos which provide information about vegetation cover and substrate typology in a fine regular lattice: substrate data are expressed as the proportion of cell grid occupied by each type of substrate. Modeling vegetation cover in terms of substrate typologies is crucial to evaluating substrate suitability, i.e., substrates' natural ability to support vegetation. A thorough discussion of the case study is given in Bruno et al (2014) and Velli (2014). The aim of the case study is to evaluate the effect of the composition of substrate typologies, observed at each pixel, on vegetation coverage: hence, the methodological framework is regression on a compositional covariate.

When dealing with compositional covariates, statistical modeling needs to be coherent with the algebra of the simplex, in order to obtain sensible and interpretable results: to this aim, the analysis starts from some transformation of the covariate. Several contributions have been recently devoted to linear regression on a compositional covariate (Bruno et al, 2014; Hron et al, 2012; Tolosana-Delgado and van den Boogaart, 2011), where several advantages of the *isometric log-ratio* (*ilr*, Egozcue et al (2003)) transformation have been highlighted. In this paper, starting from *ilr*-transformed data, we consider the case of non-parametric regression on a compositional covariate. The *ilr* transformation maps the D-dimensional simplex $\mathbb{S}^D$ into the $(D-1)$-dimensional real space $\mathbb{R}^{D-1}$, thus $(D-1)$-dimensional smoothers are sought. The proposed approach is based on multivariate Bayesian P-splines, i.e. B-splines with roughness penalties, that have been shown to be effective and computationally efficient tools for smoothing multidimensional data (Currie et al, 2006; Eilers et al, 2006). To our knowledge, the only contribution in the framework of non-parametric regression on a compositional covariate is provided by Di Marzio et al (2014), where simplicial kernels are adopted in order to obtain local

constant and local linear non-parametric estimators. The paper is organised as follows. Section 2 discusses linear regression on a compositional covariate, while section 3 deals with non-parametric regression within the Bayesian P-spline framework. In section 4 the aforementioned case study is presented. Some concluding remarks are sketched in section 5.

2 Linear regression on a compositional covariate

For the sake of simplicity, the methodology proposed in sections 2 and 3 considers the case of a linear regression model where y is a real response variable and $\boldsymbol{z} = (z_1, \ldots, z_d, \ldots, z_D)$ is a D-part compositional covariate: this framework is dubbed simplicial-real regression in Di Marzio et al (2014). As will be shown in the application of section 4, all the presented theory can readily be applied to Generalised Linear Models (GLM).
Specifying a linear model with a parameterisation providing sensible interpretation of the compositional parts effect is a non-trivial task. Indeed, untransformed components of the covariate cannot vary freely: hence, if untransformed compositional covariate is used, model parameters can be misleading and suffer of a lack of interpretability; moreover the design matrix of the proportional representation of compositions is singular.
These problems have been tackled following different approaches, all based on some transformation of the compositional covariate: a first advance in the field was proposed in Aitchison and Bacon-Shone (1984), where log-contrast models were introduced by applying the log-ratio transformation to the compositional covariate. According to Tolosana-Delgado and van den Boogaart (2011), the *centered log-ratio* (*clr*) representation of a composition is often convenient in a regression framework because each coefficient can be related to an original component. Nonetheless, the *clr* transformation is not expressed in terms of an orthogonal basis, neither of the simplex nor of the real space and it generates a singular design matrix that requires a sum-to-zero constraint for model estimation. On one hand, this can easily be addressed in a simple linear regression framework, where estimation of the regression coefficients can be obtained by computing the Moore-Penrose inverse. On the other hand, this is not practical in complex hierarchical models because of the computational burden implied by linear constraints (Bruno et al, 2014), and has several disadvantages when performing non-parametric regression, as will be discussed in section 3. In this paper, following the approach proposed in Bruno et al (2014), we adopt the *ilr* transformation which defines an isometry between $\mathbb{S}^D$ and $\mathbb{R}^{D-1}$ and is directly associated with orthogonal coordinates in the simplex. As discussed in Egozcue et al (2003), an orthonormal basis matrix $\boldsymbol{A}$ associated with the coordinate system generated by the *ilr* transformation can be obtained by

sequential binary partition, namely, the d-th row of $\boldsymbol{A}$ is:

$$\boldsymbol{a}_d = \sqrt{\frac{D-d}{D-d+1}} \left[\underbrace{0,0}_{d-1 \; times} \; , 1, \underbrace{-(D-d)^{-1}, -(D-d)^{-1}}_{D-d \; times} \right], d = 1, \ldots, D \tag{1}$$

The *ilr* transformation leads to *ilr*-transformed covariate $\boldsymbol{x} = ilr(\boldsymbol{z}) = (x_1, \ldots, x_d, \ldots, x_{D-1})$, where the coordinates associated to the basis (1) are:

$$x_d = \sqrt{\frac{D-d}{D-d+1}} \log \frac{z_d}{\sqrt[D-d]{\prod_{j=d+1}^{D} z_j}}, \quad d = 1, \ldots, D. \tag{2}$$

Such transformation gives some important advantages with respect to the *clr* transformation, still providing exactly the same information on the relationship between the response and the compositional covariate.
Let us consider a real response y_i and a vector of *ilr* coordinates $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,D-1})$, $i = 1, \ldots, n$ indexing the observations. The starting point for regression on a compositional covariate is the linear model:

$$y_i = \alpha + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}_l + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma_\epsilon^2), \quad i = 1, \ldots, n \tag{3}$$

where $\boldsymbol{\beta}_l$ is a $(D-1)$-dimensional vector of regression coefficients and α is the intercept. Note that, independently of D, model (3) could be thought as a regression model on *one single* compositional covariate, since what is being explained is just the variation of the response with respect to variation of the compositions of a whole, that in this case is represented by *ilr* coordinates. In practice, the number of parts determines the dimensionality of the domain of the compositional covariate and can generate problems due to the curse of dimensionality.
When parameterisation (3) is adopted, the first coefficient β_{1l} gives information about the relative contribution of the first part to the variation of the response. As an example, a positive coefficient implies that increasing the contribution of the first part of the compositional covariate *with respect to all the other*, has a positive effect on the response: indeed, as can be seen from equation (2), the first *ilr* coordinate involves a comparison between the first component and all the remaining components. Coefficients β_{jl}, $j = 2, \ldots, D-1$, cannot be interpreted analogously. Nonetheless, meaningful regression coefficients for all parts can be obtained switching from *ilr* to *clr* coefficients by means of a simple linear transformation. In fact, *clr* coordinates can be obtained starting from *ilr* coordinates as $clr(\boldsymbol{z}) = ilr(\boldsymbol{z})^{\mathsf{T}} \boldsymbol{A}$: thus, the D-dimensional vector of *clr* regression coefficients can be obtained starting from model (3) as $\widetilde{\boldsymbol{\beta}}_{clr} = \boldsymbol{A}^{\mathsf{T}} \boldsymbol{\beta}$. It is worth noting that coefficients $\widetilde{\boldsymbol{\beta}}_{clr}$ sum to zero by construction and convey all the relevant information concerning the effect of each part on the response variable: in fact, if the proportion of part d increases, then the proportion of at least one of the other parts must decrease. Thus, if increasing the share of the d-th part does have a positive effect on the response

variable, then increasing the share of some other part will necessarily have a negative effect. For this reason, coherently with the most prominent feature of compositional data, *clr* coefficients can be interpreted as the relative effect of each part w.r.t. all the others. Summarizing, regression models based on the *ilr* transformation provide an orthonormal coordinate counterpart to *clr* regression conveying exactly the same information about the relationship between the covariate and the response. The properties of the *ilr* transformation have been proven to be advantageous when estimating complex hierarchical models with no need of introducing linear constraints and reveal extremely useful in the context of non-parametric regression, as discussed in the following section.

3 Non-parametric regression on a compositional covariate

Let us consider the following model

$$y_i = \alpha + f(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2_\epsilon), \quad i = 1, \ldots, n \tag{4}$$

where α is the intercept, $f(\boldsymbol{x}_i)$ is a $(D-1)$-dimensional smooth function of the *ilr* coordinates $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,D-1})^\mathsf{T}$ and ϵ_i is an error term. Model (4), which constitutes the non-parametric generalisation of model (3), basically realizes smoothing of the response data y_i over a compositional covariate $\boldsymbol{z}_i$ by means of its *ilr* coordinates $\boldsymbol{x}_i$.
In general, given data y_i and a covariate x_i, $i = 1, ..., n$, statistical smoothing techniques are such that, for an arbitrary value x_0, an estimate $\hat{f}(x_0)$ is computed by a weighted mean of the data with weights depending on the euclidean distances between x_0 and x_i, for each $i = 1, ..., n$. In our case, distances between realizations of a compositional covariate should be measured in the simplex using *Aitchison* algebra (Aitchison, 1986). However, since the *ilr* transformation is an isomorphism from $\mathbb{S}^D$ to $\mathbb{R}^{D-1}$, i.e. it has the property of being an isometric representation, distances between compositions in the simplex can just be measured by euclidean norms in the *ilr* real domain.

Remark 1 Given two D-parts compositions $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{S}^D$, the isometric property implies that

$$d_a(\boldsymbol{w}_1, \boldsymbol{w}_2) = \|ilr(\boldsymbol{w}_1) - ilr(\boldsymbol{w}_2)\| \tag{5}$$

where

$$d_a(\boldsymbol{w}_1, \boldsymbol{w}_2) = \sqrt{\frac{1}{D} \sum_{i<j} \left\{ \log\left(\frac{w_{1i}}{w_{1j}}\right) - \log\left(\frac{w_{2i}}{w_{2j}}\right) \right\}}$$

is the *Aitchison* distance and $\|\cdot\|$ is the euclidean norm.

Property (5) offers a practical solution to manage the difficult task of fitting a D dimensional smooth function on the simplex $\mathbb{S}^D$ by just fitting a $(D-1)$-dimensional smooth function on the unconstrained real space $\mathbb{R}^{D-1}$. This property has been exploited by Di Marzio et al (2014) who use simplicial kernels defined on *ilr* coordinates to perform local linear regression on compositional data.

3.1 P-spline

In this work, we adopt an approach based on the P-spline method proposed by Eilers and Marx (1996), which can easily be adapted to the case of smoothing over a $(D-1)$-dimensional space. The general P-spline method focuses on regression on a basis of B-splines and a difference penalty on the associated spline coefficients. B-splines are required to be equispaced over the covariate domain, i.e. are centred at knots laying on a regular grid, in a way that smoothing is only regulated by penalties (Eilers and Marx, 2010). Typical penalties are made by combining a roughness measure, usually considering first or second order differences at neighbouring knots, and a parameter controlling the balance between smoothness and goodness of fit. In the frequentist approach, estimation is performed via penalized maximum likelihood, conditional on a fixed value for the smoothing parameter. Selection of the smoothing parameter requires cross-validation or grid search procedures, which can be hugely computationally demanding in large datasets or multidimensional settings.
A Bayesian approach has been proposed in Lang and Brezger (2004), using random walk priors on the spline coefficients and an Inverse Gamma prior on the smoothing parameter. The resulting fit accounts for uncertainty of the smoothing parameter estimation.

P-splines have been used as a general tool to model smooth covariate effects in structured additive regression models (Brezger and Lang, 2006). Specifically, to perform surface fitting, the P-spline method has been applied in several contexts involving spatial data, for example in disease mapping (Lee and Durbán, 2009; Goicoa et al, 2012), forestry and ecology (Kneib et al, 2008) and air pollution modelling (Lee and Durbán, 2011). The extension from spatial to multidimensional grid data has been described in Eilers et al (2006) and Currie et al (2006), using classical inferential procedures.

3.2 P-spline smoothing on *ilr* coordinates

In what follows, we show how the P-spline approach can be adapted to the case of non-parametric regression on a compositional covariate, in a hierarchical Bayesian framework. The isometric property of the *ilr* transformation allows the B-spline basis matrix to be built in the unconstrained $\mathbb{R}^{D-1}$ domain. This will produce a smoothing method which is perfectly coherent with the algebra of the simplex. The non-parametric function in equation (4) is specified as

$$f(\boldsymbol{x}_i) = \boldsymbol{B}(\boldsymbol{x}_i)\boldsymbol{\beta}, \tag{6}$$

where $\boldsymbol{B}(x_i)$ is the i-th row entry of a B-spline basis matrix $\boldsymbol{B}_{n\times q}$, containing multivariate B-splines evaluated at the $(D-1)$-dimensional vector $\boldsymbol{x}_i$, and $\boldsymbol{\beta}$ is a q-dimensional vector of spline coefficients.
A practical way to obtain the full basis matrix $\boldsymbol{B}$ is to use tensor products of marginal basis (Wood, 2006) defined over each *ilr* coordinate x_j, $j = 1, ..., D-1$. The steps needed for building $\boldsymbol{B}$ are briefly described in the following. At

the first step, marginal bases $\boldsymbol{B}_j$, $j = 1, ..., D-1$, of dimension $n \times q_j$, are obtained as a collection of q_j univariate B-spline functions centred at equally-spaced knots (laying on a regular knot-grid). In order to define a knot-grid over x_j, let the range $\boldsymbol{r}_j = [min(x_j), max(x_j)]$ be divided into q'_j internal intervals of the same length. A univariate B-spline is a bell-shaped function consisting of $g+1$ polynomial pieces, each piece of degree g (e.g. $g = 3$, cubic). B-splines are non zero over a limited domain spanned by $g+1$ intervals, this giving good numerical properties and stable fitting algorithms. The number of columns q_j is given by $q'_j + g$, while the number of knots required for building the basis is $q'_j + 2g + 1$. Note that, some knots external to the range $\boldsymbol{r}_j$ are needed for any point inside the range to be covered by the same number of B-splines. When all B-splines inside $\boldsymbol{B}_j$ are supported (e.g. when a sufficient amount of data belong to the intervals where 1-dimensional bases are non zero), q_j is also the rank of $\boldsymbol{B}_j$. For all details on computation of a B-spline basis matrix see Eilers and Marx (1996). In practice, the knot-grid can either be specified by choosing q' or simply by defining the locations of regularly spaced knots over $\boldsymbol{r}_j$. Knots over $\boldsymbol{r}_j$ will denoted with vector $\boldsymbol{m}_j$ of length q_j.
At the second step, the tensor product of marginal B-splines is obtained by the kronecker product ($\otimes$):

$$\boldsymbol{B}(\boldsymbol{x}_i) = \boldsymbol{B}_1(x_1) \otimes ... \otimes \boldsymbol{B}_j(x_j) \otimes ... \otimes \boldsymbol{B}_{D-1}(x_{D-1}),$$

where $\boldsymbol{B}(\boldsymbol{x}_i)$ the i-th row of $\boldsymbol{B}$, containing B-splines evaluated at $\boldsymbol{x}_i$. Analogously $\boldsymbol{B}_j(x_j)$ is the row entry of $\boldsymbol{B}_j$ containing B-splines evaluated at x_j. To compute $\boldsymbol{B}$ in practice, the box product operator ($\square$) can be used.

Definition 1 Given two matrices $\boldsymbol{P}_{m \times c_1}$ and $\boldsymbol{Q}_{m \times c_2}$, the box product operator is defined as:

$$\boldsymbol{P} \square \boldsymbol{Q} = (\boldsymbol{P} \otimes \boldsymbol{1}_{c_2}^{\mathsf{T}}) \odot (\boldsymbol{1}_{c_1}^{\mathsf{T}} \otimes \boldsymbol{Q})$$

where $\boldsymbol{1}_k$ indicates a column vector of ones with length k and $\odot$ is the element-wise product.

Eventually, the full basis matrix is $\boldsymbol{B} = \boldsymbol{B}_{D-1} \square \boldsymbol{B}_{D-2} \square \ldots \square \boldsymbol{B}_1$ with dimension $n \times q$ where $q = q_1 q_2 \cdot \ldots \cdot q_{D-1}$.
Figure 1 describes the steps leading to the full basis matrix $\boldsymbol{B}$ in the case where $D = 3$. In the upper panels of Figure 1, marginal and bivariate B-splines are displayed to aid visualization of a tensor product operation on two marginal bases, i.e. when we have two *ilr* coordinates x_1 and x_2. In the upper left panel, marginal univariate B-splines are plotted with dashed and solid lines along x_1 and x_2 respectively, with knots indicated by dots. A perspective plot of the bivariate B-splines resulting from the tensor product of the marginal bases is displayed in the upper right panel of Figure 1. In the lower left panel, a flattened version of the perspective plot with the same bivariate B-splines seen from top are shown: contours emphasize the symmetric shape of each B-spline and their regular spacing in $\mathbb{R}^2$. In this panel, knots are represented by dots; note, these are the vertexes of a 2-dimensional regular grid $\boldsymbol{m}$ whose coordinates are $\left(\boldsymbol{m}_1 \otimes \boldsymbol{1}_{q_2}^{\mathsf{T}}, \boldsymbol{1}_{q_1}^{\mathsf{T}} \otimes \boldsymbol{m}_2\right)$. In the lower right panel, a ternary plot

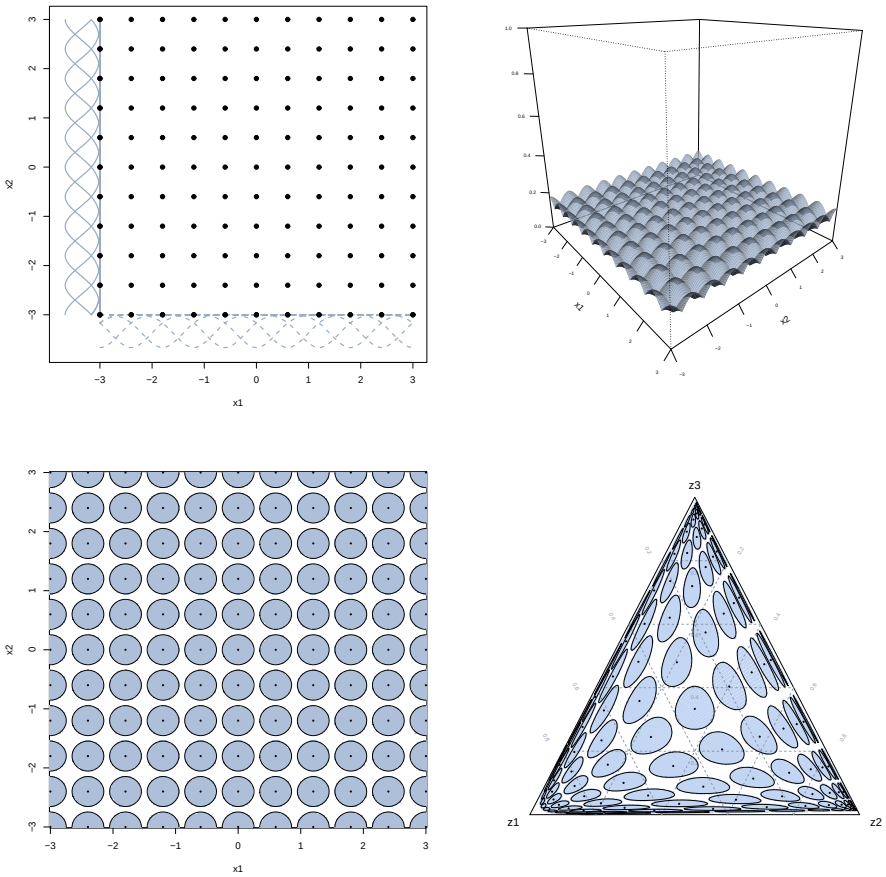

Fig. 1 Bivariate B-spline basis functions over the *ilr* domain, in the case of a 3-part composition. Upper left panel shows marginal basis, with knots identified by dots. Upper right panel and lower left panels display B-splines in a perspective plot and a flat representation, respectively. The lower right panel shows how B-splines appear in the simplex $\mathbb{S}^3$. The shape of the B-splines is highlighted by contours in the last two panels

shows how things appears in the simplex: B-splines have symmetric shapes and are equispaced in terms of *Aitchison* geometry. Ternary diagrams are the most popular tool for compositional data visualization when $D = 3$. A ternary diagram is an equilateral triangle equivalent to the 3-dimensional simplex: a generic 3-parts compositional observation, $\boldsymbol{z} = (z_j)_{j=1,\dots,3}$, will plot at a distance z_j from the opposite side of vertex j. At each vertex of the triangle, the components takes the maximum value of 1.

3.3 Priors

Intrinsic Gaussian Markov Random Fields (IGMRFs) (Rue and Held, 2005) are widely adopted to model smooth random effects. In what follows $\text{IGMRF}_M(\tau \boldsymbol{K})$ denotes the (multivariate normal) distribution of a M-dimensional random vector with $(M \times M)$-dimensional structure matrix $\boldsymbol{K}$ and precision parameter τ. An IGMRF prior on spline coefficients $\boldsymbol{\beta}$ is specified as:

$$\boldsymbol{\beta} \sim \text{IGMRF}_q(\tau_\beta \boldsymbol{K}_\beta) \tag{7}$$

$$\tau_\beta \sim Gamma(a, b) \tag{8}$$

where the structure matrix $\boldsymbol{K}_\beta$ is a sparse precision matrix, whose non zero pattern describes conditional dependencies between spline coefficients placed in correspondence of the vertexes of the knot-grid. A similar prior was proposed by Lang and Brezger (2004) for Bayesian inference on spline coefficients of a univariate P-spline model. The purpose is to penalize deviations of spline coefficients from a constant (or a linear, depending on the random walk order) unknown level, in a way that a *not too variable* set of $\boldsymbol{\beta}$s is achieved. The degree of variability of the joint posterior distribution of $\boldsymbol{\beta}$ is regulated by the precision parameter τ_β, assuming a global amount of smoothness operating all over the knot-grid. According to the Markov assumption, coefficients $\boldsymbol{\beta}$ are conditionally independent given coefficients at neighbouring knots (e.g. adjacent vertexes). The definition of neighbourhood, determines different smoothing models; see Rue and Held (2005) ch. 3 for a detailed description of IGMRFs on regular lattices. In order to obtain the structure matrix $\boldsymbol{K}_\beta$, we proceed by specifying marginal structures $\boldsymbol{R}_j$, $j = 1, \ldots, D-1$, and combine them by means of sums of Kronecker products. In principle, different marginal structures can be assumed to achieve different smoothing along each dimension. Let $\boldsymbol{R}_j$ be the structure matrix associated to an IGMRF defined on the knots $\boldsymbol{m}_j$. $\boldsymbol{K}_\beta$ turns out to be:

$$\begin{aligned}\boldsymbol{K}_\beta = &\left(\boldsymbol{I}_{q_{D-1}} \otimes ... \otimes \boldsymbol{I}_{q_2} \otimes \boldsymbol{R}_1\right) + \\ &\left(\boldsymbol{I}_{q_{D-1}} \otimes ... \otimes \boldsymbol{I}_{q_3} \otimes \boldsymbol{R}_2 \otimes \boldsymbol{I}_{q_1}\right) + \\ &... \\ &\left(\boldsymbol{R}_{D-1} \otimes \boldsymbol{I}_{q_{D-2}} \otimes ... \otimes \boldsymbol{I}_{q_1}\right)\end{aligned} \tag{9}$$

where $\boldsymbol{I}_q$ is the identity matrix of dimension q. Note that, if $\boldsymbol{R} = \boldsymbol{R}_j$, $j = 1, ..., D-1$, is specified as the structure matrix of a random walk prior of order 1 (RW1) on a one-dimensional grid, i.e. (considering $q_j = 5$ as an example):

$$\boldsymbol{R} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

Table 1 B-spline basis $\boldsymbol{B}$ and structure matrices $\boldsymbol{K}_\beta$ according to the number D of compositional parts

n. parts	$\boldsymbol{B}$	$\boldsymbol{K}_\beta$
2	$\boldsymbol{B}_2 \square \boldsymbol{B}_1$	$(\boldsymbol{I}_{q_2} \otimes \boldsymbol{R}_1) + (\boldsymbol{R}_2 \otimes \boldsymbol{I}_{q_1})$
3	$\boldsymbol{B}_3 \square \boldsymbol{B}_2 \square \boldsymbol{B}_1$	$(\boldsymbol{I}_{q_3} \otimes \boldsymbol{I}_{q_2} \otimes \boldsymbol{R}_1) + (\boldsymbol{I}_{q_3} \otimes \boldsymbol{R}_2 \otimes \boldsymbol{I}_{q_1}) + (\boldsymbol{R}_3 \otimes \boldsymbol{I}_{q_2} \otimes \boldsymbol{I}_{q_1})$
.	.	.
.	.	.
.	.	.
D	$\boldsymbol{B}_{D-1} \square ... \square \boldsymbol{B}_1$	$\left(\boldsymbol{I}_{q_{D-1}} \otimes ... \otimes \boldsymbol{I}_{q_2} \otimes \boldsymbol{R}_1\right) + \left(\boldsymbol{I}_{q_{D-1}} \otimes ... \otimes \boldsymbol{I}_{q_3} \otimes \boldsymbol{R}_2 \otimes \boldsymbol{I}_{q_1}\right) + \left(\boldsymbol{R}_{D-1} \otimes \boldsymbol{I}_{q_{D-2}} \otimes ... \otimes \boldsymbol{I}_{q_1}\right)$

the resulting structure matrix $\boldsymbol{K}_\beta$ corresponds to a RW1 prior on a $D-1$ dimensional regular grid. Table 1 summarises steps for building the full basis $\boldsymbol{B}$ and structure $\boldsymbol{K}_\beta$ for varying D.

4 Application

In this section, we apply non-parametric regression on a compositional covariate in a spatial framework. In particular, we consider data gathered in a study designed within the framework of the priority defined by the European Commission (Council Directive 92/43/CEE), with the aim of investigating habitat "6110* Rupicolous calcareous or basophilic grasslands of the *Alysso-Sedion albi*". The ecological study collected data at several times on an area structured as a regular lattice of dimension 30×30, containing $S = 900$ grid cells. A spatio-temporal analysis of these data is provided in Bruno et al (2014), where linear regression on a compositional covariate was performed tacking account of spatial and temporal correlation.
For the purposes of this paper, we focus on the analysis of a single time; spatio-temporal modelling constitutes one of the possible extensions. A ground-photo of the study area was taken and then processed via GIS algorithms to produce a digital image for vegetation cover and ground composition, considering $D = 4$ substrates: moss, litter, soil and bare rock. Each grid cell in the digital image provides information collected over $N = 100$ pixels. At each grid cell i, we denote with y_i the number of pixels covered by a plant and with $\boldsymbol{z}_s = (z_{i1}, z_{i2}, z_{i3}, z_{i4})$ the proportion of pixels occupied respectively by moss, litter, soil and bare rock. A quick measure of vegetation occupancy probability can be obtained as $\hat{\pi}_i = y_i/N$. A map of the study area reporting vegetation cover is shown in the left panel of Figure 2, where a substantial degree of spatial correlation can be observed. For visualising vegetation cover *vs* the compositional covariate, since $D = 4$, a regular tetrahedron can in principle be used, which is a generalisation of the ternary diagram. A tetrahedron is reported in the right panel of Figure 2 where dots are colored proportionally to vegetation cover observed at each substrate composition. Colors vary smoothly from red (low vegetation) to green (high vegetation). The tetrahedron highlights low vegetation cover in correspondence of compositions where bare rock has a

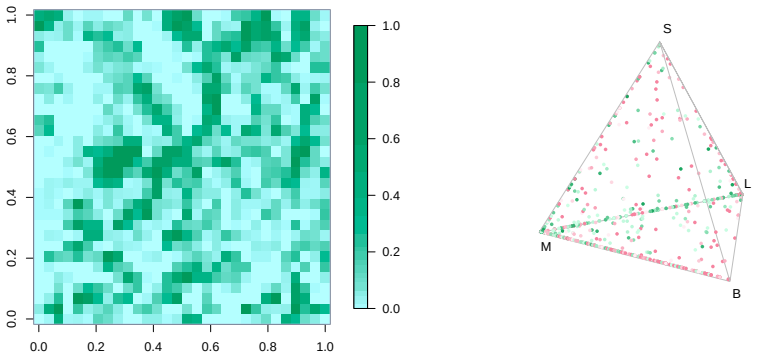

Fig. 2 Left panel: map of vegetation cover. Right panel: vegetation cover *vs* substrate composition; green (red) colors correspond to high (low) vegetation cover.

high relative weight, while high vegetation cover is observed when moss and litter are prevalent.
The aim of the hierarchical model proposed in what follows is to fit a smooth surface over the simplex identified by the tetrahedron in the right panel of Figure 2; a smooth surface is used to describe substrate suitability for vegetation cover. At the same time, spatial correlation observed in the left panel of Figure 2 is taken into account in the model. The linear and non-parametric alternatives discussed in sections 2 and 3 are compared in order to understand the meaning and emphasize the merits of P-spline smoothing in the context of regression on a compositional covariate.

4.1 Model

We propose a Bayesian hierarchical model which specifies a smooth relationship between vegetation occupancy probability and substrate composition by using the methodology proposed in section 3. Conditionally on vegetation occupancy probability π_i, counts y_i are assumed to follow a Binomial distribution, i.e., the binary response at each pixel within a grid cell is considered as the realization of an independent trial:

$$y_i|\pi_i, N \sim Binomial(\pi_i, N) \qquad i = 1, \ldots, S.$$

To model the vegetation occupancy probability as a function of covariates, we adopt the *probit* link; the linear predictor is specified as follows:

$$\Phi^{-1}(\pi_i) = \alpha + f(\boldsymbol{x}_i) + \theta_i,$$

where α is the intercept and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_S)$ is a spatially structured vector of random effects modelled as an IGMRF:

$$\boldsymbol{\theta} \sim \text{IGMRF}_S(\tau_\theta \boldsymbol{K}_\theta).$$

$\boldsymbol{K}_\theta$ is specified as the structure matrix of a Random Walk of order 1 on the regular lattice formed by the grid cells.
The term $f(\boldsymbol{x}_i)$ captures the relationship between substrate composition and vegetation cover. According to the linear model, model hierarchy can be completed as

$$f(\boldsymbol{x}_i) = \boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta}_l \qquad \boldsymbol{\beta}_l \sim N(0, 1000).$$

For specifying the P-spline smoothing model, we assume:

$$\begin{aligned} f(\boldsymbol{x}_i) &= \boldsymbol{B}(x_i)\boldsymbol{\beta} \\ \boldsymbol{\beta} &\sim \text{IGMRF}_S(\tau_\beta \boldsymbol{K}_\beta) \\ \tau_\beta &\sim Gamma(a, b) \end{aligned} \tag{10}$$

using a B-spline basis $\boldsymbol{B}$ built as tensor products of marginal bases $\boldsymbol{B}_j$, $j = 1, 2, 3$, as described in section 3. Structure matrix $\boldsymbol{K}_\beta$ is specified as in equation (9). In building $\boldsymbol{K}_\beta$, the same difference matrix $\boldsymbol{R}_j$ has been used for each *ilr* coordinate. This is coherent with the assumption of a global τ_β regulating the degree of smoothing all over the simplex. As regards knot-grid resolution, $q_j = 7$ equispaced knots have been selected on each *ilr* coordinate, giving a $7^3 = 343$-dimensional structure matrix $\boldsymbol{K}_\beta$ and a (900×343)-dimensional full basis matrix $\boldsymbol{B}$. As it is very often the case in multidimensional smoothing using B-splines, $rank(\boldsymbol{B}) < q$, since extreme knots are not supported. Nevertheless, prior (10) allows model identifiability by borrowing strength between adjacent coefficients.

The Markov property of IGMRF models implies sparseness of the precision matrix, which allows fast computations, making multidimensional P-spline smoothing feasible. Two alternative strategies are currently very popular for approximating the joint posterior distribution: MCMC sampling and Integrated Nested Laplace Approximations (INLA, Rue and Martino 2009). Computations have been performed using both methods, codes are available upon request from the corresponding author. Note that, to allow model identifiability, the smooth surface $\boldsymbol{B}\boldsymbol{\beta}$ must be estimated under the constraint $\mathbf{1}_q^\mathsf{T} \boldsymbol{B}\boldsymbol{\beta} = 0$, i.e. a sum to zero constraint is imposed on the smooth surface: this can easily be implemented both in INLA and MCMC.

4.2 Results

Comparison of the fitting performances in terms of Deviance Information Criterion (DIC, Spiegelhalter et al (2002)) shows a slightly better performance of the non-parametric model (DIC = 4343) with respect to the linear model (DIC = 4352). Models not including the spatial random effect $\boldsymbol{\theta}$, show poor fitting performances both in the linear and non-parametric case, delivering DIC

Table 2 Posterior summaries for the regression coefficients - linear model

parameter	mean	0.025quant	0.975quant	Substrate	$\widetilde{\boldsymbol{\beta}}_{clr}$
α	-1.8042	-1.9202	-1.6901	Moss (M)	0.082
β_{l1}	0.0944	0.0391	0.1498	Litter (L)	0.107
β_{l2}	0.1644	0.1181	0.2109	Soil (S)	-0.048
β_{l3}	0.0659	0.0163	0.1154	Bare rock (B)	-0.141

values respectively equal to 16324 and 16157: hence, the explanatory variable does not capture all the spatial variation of the response variable. In table 2, some posterior summaries for the linear model parameters are reported. Posterior means of the *ilr* coefficients $\boldsymbol{\beta}_l$ are all statistically significant, showing a significant effect of substrate composition on vegetation cover. As discussed in section 2, while β_{l1} can be interpreted as the effect of the first part with respect to all the others, coefficients β_{l2} and β_{l3} do not have a sensible interpretation since they are not directly related to compositional parts. For this reason, the last column of the table reports *clr* coefficients obtained as $\widetilde{\boldsymbol{\beta}}_{clr} = \boldsymbol{A}^{\mathsf{T}}\boldsymbol{\beta}_{\boldsymbol{l}}$ where $\boldsymbol{A}$ is the orthonormal basis defined in equation (1): these coefficients sum to zero and can readily be interpreted as relative substrate suitability measures. It can be seen that moss and litter show positive relative suitability, on the contrary soil and bare rock show negative relative suitability, with litter being the most suitable substrate and bare rock being the less suitable one.

Building appropriate graphics to visualise the effect of a compositional covariate when $D = 4$ is a non trivial task. A practical approach is to focus on the relative magnitudes of three components (sub-compositions) conditioning on the remaining one. This is the approach adopted in Figure 3, where, in each row, ternary diagrams conditional on a single part are plotted: from top to bottom, diagrams are conditional on moss (M), litter (L), soil (S) and bare rock (B), respectively. Ternary diagrams in the first column report the effect of our compositional covariate in the linear model case; while in the second and third columns, ternaries display the same compositional effect in the smooth model case. In the first two column panels, the conditioning value is equal to the average of the conditioning part over the study region: for example, panels (1.1) and (1.2) are conditional on the moss average, i.e. $M = 0.47$. In the third column panels, we choose an arbitrary conditioning value equal to 0.7. In each ternary diagram, dark green (dark red) colors are associated to substrate compositions that are more suitable (less suitable) for vegetation cover, *conditionally* on the chosen value for the conditioning component.

The patterns observed in the first column of Figure 3 are just a graphical display of the inference that can be drawn by *clr* coefficients: high suitability (green) values are in the neighbourhood of vertexes associated to moss (M) and litter (L), while low suitability (red) values are in the neighbourhood of vertexes associated to soil (S) and bare rock (B). The patterns observed in the first column ternaries, display how a linear relationship appears in the simplex: it is worth to notice that, when adopting the linear model, the same

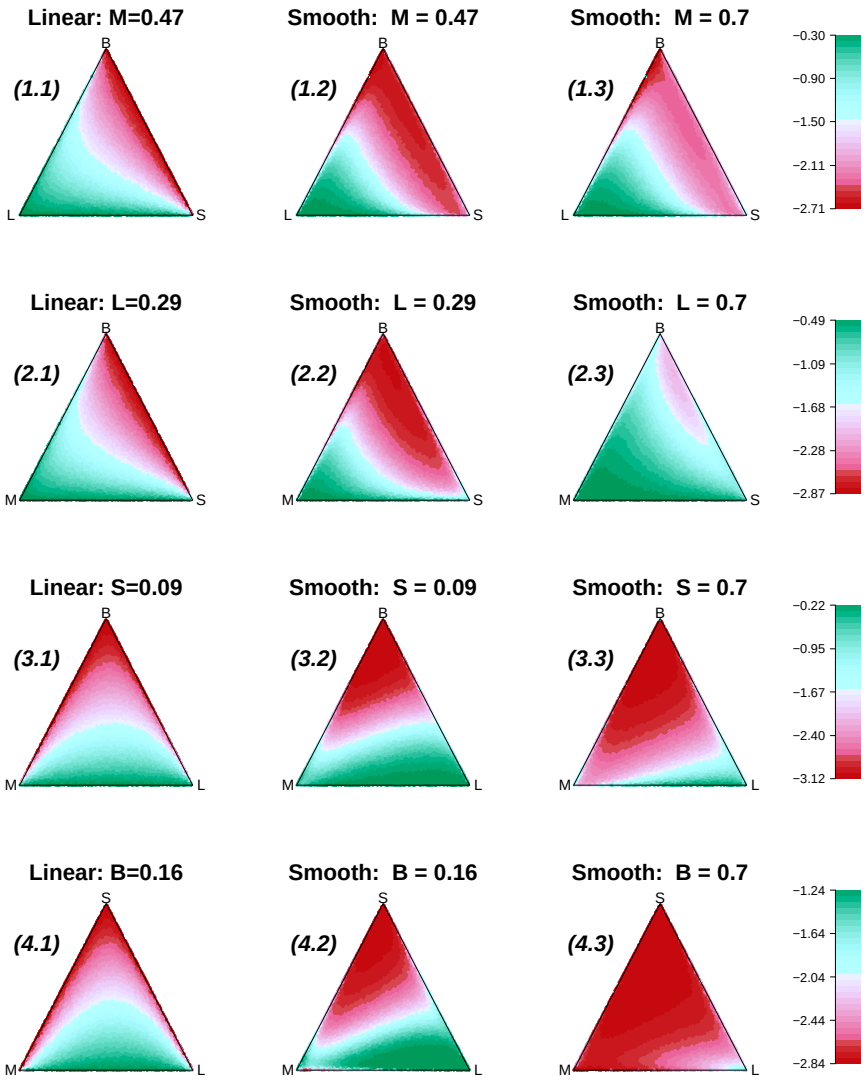

Fig. 3 Ternary diagrams showing the substrate composition effect on vegetation cover (substrate suitability). In rows 1-4, ternaries are plotted conditionally on moss (M), litter (L), soil (S), and bare rock (B) respectively. First column panels show ternaries referred to the linear model while second and third column panels show ternaries referred to the P-spline model

pattern is observed independently of the value of the conditioning part, since it is determined by constant *clr* coefficients. On the other hand, the local nature of the smooth model allows the shape of the relationship to change according to the value of the conditioning part, as can be seen by comparing the second and third columns of Figure 3. As a matter of fact, joint smoothing performed by Bayesian P-splines, incorporates the effect of interaction among parts, allowing a far more flexible and detailed description of the relationship between compositions and the response, with the drawback that no synthetic parameters are available to describe such relationship.

5 Conclusions

In this work, a non-parametric approach to regression on compositional covariates is developed using Bayesian P-spline. The availability of fast Bayesian computation (Rue et al, 2009) makes Bayesian P-splines appealing for multidimensional smoothing with a low-rank B-spline basis, which is the approach assumed in this paper. Furthermore, we believe particular features of P-splines give advantages in terms of modelling and computations which can be important in multidimensional data settings (such as when estimating compositional covariate effects). Firstly, the rank of a B-spline basis matrix is much lower than sample size (low-rank smoother). Secondly, B-splines basis functions are local, i.e. non zero over a limited domain, this allowing sparse matrix computation and numerical stability. Thirdly, penalized regression with B-splines centred at equispaced knots work well in sparse data contexts, when, for instance, covariate data are sparsely scattered over $\mathbb{R}^{D-1}$; in such situations the combined use of equispaced knots with difference penalties on coefficients allows to smoothly interpolate between gaps. Data sparseness over the compositional domain is a feature of our application analysed in section 4 but is also frequent in spatial statistics, especially in geostatistics.

The methodology is illustrated on a case study concerning a habitat included in the framework of the priority defined by the European Commission, where investigating the relationship between vegetation cover and substrate typology can be helpful in quantifying substrate suitability. The approach proposed in this paper allows a flexible and detailed description of the relationship between compositions and the response, but it can be extended in several directions. Future work includes computationally feasible extensions to the spatio-temporal case and careful evaluation of the prior assumptions on the spline coefficients and smoothing parameter.

References

Aitchison J (1986) The Statistical Analysis of Compositional Data. Chapman and Hall

Aitchison J, Bacon-Shone J (1984) Log contrast models for experiments with mixtures. Biometrika 71(2):pp. 323–330, URL http://www.jstor.org/stable/2336249

Brezger A, Lang S (2006) Generalized structured additive regression based on bayesian p-splines. Computational Statistics and Data Analysis 50(4):967–991, DOI http://dx.doi.org/10.1016/j.csda.2004.10.011

Bruno F, Greco F, Ventrucci M (2014) Spatio-temporal regression on compositional covariates: modeling vegetation in a gypsum outcrop. Environmental and Ecological Statistics pp 1–19, DOI 10.1007/s10651-014-0305-4, URL http://dx.doi.org/10.1007/s10651-014-0305-4

Currie I, Durbán M, Eilers P (2006) Generalized linear array models with applications to multidimensional smoothing. J R Statist Soc B 68:259–280

Di Marzio M, Panzera A, Venieri C (2014) Non-parametric regression for compositional data. Statistical Modelling DOI 10.1177/1471082X14535522

Egozcue J, Pawlowsky-Glahn V, Mateu-Figueras G, Barcel-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Mathematical Geology 35(3):279–300, DOI 10.1023/A:1023818214614

Eilers P, Marx B (1996) Flexible smoothing with b-splines and penalties. Statistical Science 11:89–121

Eilers P, Marx B (2010) Splines, knots, and penalties. Wiley Interdisciplinary Reviews: Computational Statistics 2:637–653

Eilers P, Currie I, Durbán M (2006) Fast and compact smoothing on large multidimensional grids. Computational Statistics & Data Analysis 5:61–76

Goicoa T, Ugarte M, Etxeberria J, Militino A (2012) Comparing car and p-spline models in spatial disease mapping. Environmental Ecological Statistics 19:537–599

Hron K, Filzmoser P, Thompson K (2012) Linear regression with compositional explanatory variables. Journal of Applied Statistics 39(5):1115–1128, DOI 10.1080/02664763.2011.644268

Kneib T, Muller J, Hothorn T (2008) Spatial smoothing techniques for the assessment of habitat suitability. Environmental Ecological Statistics 15:343–364

Lang S, Brezger A (2004) Bayesian p-splines. Journal of Computational and Graphical Statistics 13:183–212

Lee D, Durbán M (2009) Smooth-car mixed models for spatial count data. Computational Statistics and data Analysis 53:2968–2977

Lee DJ, Durbán M (2011) P-spline anova-type interaction models for spatio-temporal smoothing. Statistical Modelling 11(1):49–69, DOI 10.1177/1471082X1001100104

Rue H, Held L (2005) Gaussian Markov Random Fields. Chapman and Hall/CRC

Rue H, Martino S, Chopin N (2009) Approximate bayesian inference for latent gaussian models using inte- grated nested laplace approximations (with discussion). Journal of the Royal Statistical Society, Series B 71(2):319–392

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. Journal of the Royal Statistical So-

ciety: Series B (Statistical Methodology) 64(4):583–639, DOI 10.1111/1467-9868.00353

Tolosana-Delgado R, van den Boogaart KG (2011) Linear models with compositions in r. Compositional Data Analysis: Theory and Applications pp 356–371

Velli A (2014) Relationships between plant diversity and environmental heterogeneity in rupicolous grasslands on gypsum. The case study of Alysso-Sedion albi (Habitat 6110). PhD Dissertation, University of Bologna

Wood S (2006) Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC