ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

SCIENZE STATISTICHE

Ciclo XXXVII

Settore Concorsuale: : 13/D1 - STATISTICA

Settore Scientifico Disciplinare: SECS-S/01 - STATISTICA

Variance Partitioning Priors for Latent Gaussian Models

Presentata da: **Luisa Ferrari**

Coordinatore Dottorato

Prof.ssa Angela Montanari

Supervisore

Prof. Massimo Ventrucci

Esame finale anno 2025

Abstract

Variance partitioning priors have recently been proposed in the context of Bayesian hierarchical models. They are defined as those priors that make use of a reparametrization of the variance parameters into a total variance and a set of proportions. This work proposes a standardization procedure to accommodate variance partitioning priors into a large class of models, namely latent Gaussian models. The standardization procedure to be applied on the model effects guarantees an intuitive interpretation for the new parameters. The procedure acknowledges how the interpretation of variance contributions as intended by the user can differ between fixed and random effects. Particular attention is given to the special class of intrinsic Gaussian Markov random fields, which are popularly used to model spatial and temporal correlation. The benefits of the proposal are validated through simulations, which have confirmed the practical relevance of the standardization procedure. The importance of the contribution lies in the possibility of fully exploiting prior information through variance partitioning priors, which is particularly beneficial to those applications and fields that require complex modelling structures. This advantage is exemplified in the context of species distribution models used in ecology, which are usually composed by different fixed and random effects.

Contents

List of Symbols

Y	random variable with Exponential family distribution of parameters $\eta, \boldsymbol{\psi}$
η	linear predictor defined as $g(E[Y\|\eta, \boldsymbol{\psi}])$ for a given link function $g(\cdot)$
$\boldsymbol{\psi}$	additional likelihood parameters of Y
μ	intercept of the linear predictor
$\boldsymbol{X}$	random vector of covariates $X_1, X_2, ...$
$\pi(\boldsymbol{x})$	joint probability distribution of $\boldsymbol{X}$
$\mathcal{X}$	support of covariate X
$\boldsymbol{D}(\cdot)$	column vector called *basis* of K functions $D_1(\cdot), ..., D_K(\cdot)$
$\boldsymbol{u}$	random vector called *coefficients* with a multivariate Gaussian distribution
$\boldsymbol{\mu}$	mean vector of a multivariate Gaussian distribution
$\boldsymbol{P}$	precision matrix of a multivariate Gaussian distribution
σ^2	scale parameter of a Gaussian distribution
$\boldsymbol{Q}$	scaled precision matrix, i.e. $\boldsymbol{P}/\sigma^2$ (for IGMRFs, called structure matrix)
$\boldsymbol{M}^*$	generalized inverse of matrix $\boldsymbol{M}$
$\|\boldsymbol{M}\|^*$	generalized determinant of matrix $\boldsymbol{M}$
$\boldsymbol{\Sigma}$	generalized inverse of the scaled precision matrix, i.e. $\boldsymbol{\Sigma} = \boldsymbol{Q}^*$
$\boldsymbol{S}$	matrix spanning the null space of $\boldsymbol{Q}$, i.e. $\boldsymbol{QS} = \boldsymbol{0}$
$\boldsymbol{S}_{(d)}$	Vandermonde matrix of degree d
$\boldsymbol{S}_{(d)}(x)$	row vector of $d+1$ functions $x^0, ..., x^d$
$Var_{X,\boldsymbol{u}}[\cdot\|\sigma^2]$	joint variance with respect to X and $\boldsymbol{u}$ conditional on σ
$Var_{\boldsymbol{u}}[\cdot\|\sigma^2, X]$	variance with respect to $\boldsymbol{u}$ conditional on σ and X
$Var_X[\cdot\|\boldsymbol{u}]$	variance with respect to X conditional on $\boldsymbol{u}$
s^2	finite-population variance $Var_X[f(X)\|\boldsymbol{u}]$
$E[s^2]$	expected value of the finite-population variance with respect to $\boldsymbol{u}\|\sigma$
$\boldsymbol{a}$	column vector containing entries $E_X[D_1(X)], ..., E_X[D_K(X)]$
C	expectation-based scaling constant $E_X\{Var_{\boldsymbol{u}}[f(X)\|\sigma^2 = 1, X]\}$
$GM_X[\cdot]$	geometric mean for a random variable defined as $\exp\{E_X[\log(\cdot)]\}$
σ^2_{ref}	geometric mean-based scaling constant $GM_X\{Var_{\boldsymbol{u}}[f(X)\|\sigma^2 = 1, X]\}$
$\int \boldsymbol{M}(x)\,\mathrm{d}x$	component-wise integration of the matrix of functions $\boldsymbol{M}(x)$
$\widetilde{\boldsymbol{S}}$	modified null space matrix equal to $\int_{\mathcal{X}} \boldsymbol{D}(x) \cdot \boldsymbol{S}_{(d-1)}(x) \cdot \pi(x)\,\mathrm{d}x \in \mathbb{R}^{K \times d}$
$\widetilde{\boldsymbol{Q}}$	modified precision matrix such that $\widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{S}} = \widetilde{\boldsymbol{0}}$
$\boldsymbol{B}(\cdot)$	cubic B-Spline basis on equidistant knots on the interval $[m, M]$
$\otimes$	Kronecker product
$\text{IG}(\alpha, \beta)$	Inverse-Gamma distribution with shape α and rate β
$\text{PC}_b(\delta)$	Penalized Complexity prior with base model b and hyperparameter δ

Chapter 1

Introduction

Bayesian hierarchical models are an extremely popular modelling approach, thanks to their flexibility and the recent computational innovations. Popular subclasses are Latent Gaussian Models (LGMs, see Rue, Martino, and Chopin 2009). These models commonly include a mix of fixed and random effects, often in the form of Intrinsic Gaussian Markov Random Fields (IGMRFs, see Rue and Held 2005), which are for example used to model spatio-temporal correlation (Fahrmeir, Kneib, and Lang 2004). The problem of prior specification for these complex models remains an open question, especially with respect to variance parameters, whose estimation can be quite sensitive to prior assumptions. Traditionally, priors on variance parameters are often chosen to be weakly informative, to reflect the idea that limited prior knowledge about their value is available.

Although it is true that there is usually little information about the variance parameters, experts usually have at least some intuition about the relative importance of different effects. This is the idea behind the Hierarchical Decomposition (HD) priors framework proposed by Fuglstad et al. 2020, which aims to leverage this underlying knowledge through the design of an appropriate joint prior on the variance parameters of a model. This goal is achieved through a reparametrization of the original parameters into a total variance and a set of proportions. Proportion parameters are more intuitive for users, as they directly indicate the relative importance of effects on the total variance of the linear predictor on a (0-1) interval, thereby facilitating prior specification. Other works have proposed the same reparametrization for different purposes, namely variable selection (Zhang et al. 2022, Aguilar and Bürkner 2023). We term this reparametrization Variance Partitioning (VP) and define VP priors all those joint priors on the original variance parameters built using this technique (Franco-Villoria, Ventrucci, and Rue 2022).

It is desirable to extend VP priors to a general LGM setting and be able to

elicit priors on the variance contribution of fixed and random effects simultaneously. However, this extension is challenging. VP parameters are only meaningful if the original variance parameters can be interpreted as the variance contribution of the corresponding effects (which we denote as *intuitive interpretation* that the user has about the variance parameters). Variance parameters in an LGM do not always have such an intuitive interpretation. For example, the variance of an IGMRF effect (e.g. a conditional autoregressive model for spatial data) cannot be interpreted directly as the contribution of the corresponding effect, since IGMRFs are improper models (Sørbye and Rue 2014). The problem also arises for proper models, for which the effects must be scaled such that the variance parameters match their intuitive interpretation. This is not a trivial problem, as it requires a formal definition of what we mean by variance contribution of an effect. Existing research has addressed the issue of interpretability offering various definitions of variance contribution and multiple approaches for ensuring this requirement. Nevertheless, the problem has only been investigated for specific subsets of effects within the broader context of LGMs. This limitation is particularly significant as it precludes the broader applicability of VP priors, which are otherwise both interpretable and competitive with respect to state-of-the-art alternatives.

The goal of this thesis is to explore the challenges that need to be addressed in order to correctly implement VP priors in LGMs. Chapter 2 is a preliminary chapter that reviews the fundamental concepts, necessary for a full understanding of the remainder of the thesis. In Chapter 3, we introduce a formal definition of variance contribution and derive the conditions under which the variance of a given effect (either fixed or random) fulfils the definition. This leads us to propose a novel standardization procedure that must be applied to each effect of an LGM in order to accommodate VP priors. Our procedure has been inspired by the work of Sørbye and Rue 2014 on IGMRFs. Multiple examples are illustrated, with particular emphasis on IGMRFs, and simulations are carried out to investigate whether the theory proposed translates into tangible practical benefits. Chapter 4 showcases the advantages of VP priors through an application in the field of ecology, specifically in the context of species distribution models (SDMs, see Ovaskainen et al. 2017). Chapter 5 concludes the thesis, summarizing the main contributions and possible future research lines.

Chapter 2

Preliminaries

In this chapter, we review the foundational concepts that are useful for a better understanding of the rest of the thesis. We start by describing Latent Gaussian models (LGMs), Gaussian Markov Random Fields (GMRFs), and Intrinsic Gaussian Markov Random Fields (IGMRFs). Later on, we briefly introduce the class of Penalized Complexity (PC) priors. Finally, we review the literature on variance partitioning priors, made up by two main branches: Hierarchical Decomposition (HD) priors and R2D2 priors. Readers who are already familiar with these concepts may skip this preliminary chapter.

2.1 Latent Gaussian Models

Latent Gaussian models are a subclass of Bayesian Hierarchical models (BHMs). The class of BHMs is a very popular statistical tool which allows the specification of very complex and flexible models for a given response (Gelman et al. 2013, Wakefield et al. 2013, Hrafnkelsson 2023). The specification of such models is carried out using (at least) three hierarchical levels. First, there is a response level at which a distribution is assumed on the n response observations $\boldsymbol{y} = [y_1, ..., y_n]^T$, conditional on some latent parameters $\boldsymbol{\alpha}$ and hyperparameters $\boldsymbol{\gamma}$. Secondly, there is the latent level in which the distribution of the latent parameters is specified conditional on the hyperparameters $\boldsymbol{\gamma}$. Finally, the hyperparameters $\boldsymbol{\gamma}$ represent the last level and must be assigned a prior distribution (hyperparameter level). A BHM can therefore be fully specified by the joint distribution of its $\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\gamma}$:

$$\pi(\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\gamma}) = \pi(\boldsymbol{y}|\boldsymbol{\alpha},\boldsymbol{\gamma})\pi(\boldsymbol{\alpha}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma}). \tag{2.1}$$

Due to their flexibility and the efficiency of their computational estimation, BHMs have become extremely popular in many areas of study, e.g. disease mapping, (Lawson 2018), environmental sciences (Banerjee, Carlin, and Gelfand 2003), engineering (Hrafnkelsson 2023), social sciences (Gelman and Hill 2007).

The class of models adhering to Equation 2.1 is exceptionally broad, encompassing the vast majority of Bayesian models employed in practical applications. Nevertheless, supplementary assumptions are frequently introduced to facilitate inference and prediction. Latent Gaussian models (LGMs) constitute a particularly prevalent subclass of BHMs, where the latent model $\pi(\boldsymbol{\alpha}|\boldsymbol{\gamma})$ is constrained to follow a Gaussian distribution. We here define LGMs under additional assumptions that are virtually universally adopted in real-world applications. In particular, we focus here on the subclass of LGMs in which the response is linked to the latent parameters only through a generalized linear model (see Bayesian LGMs with a univariate link function in Hrafnkelsson 2023).

We formally define here the class of Latent Gaussian model through the following specification where $\boldsymbol{\gamma} = [\boldsymbol{\psi}, \boldsymbol{\sigma}]$.

- **Response model**. Response observations are assumed to be conditionally independent given $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$:

$$\pi(\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\psi}) = \prod_{i=1}^{n} \pi(y_i|\boldsymbol{\alpha}, \boldsymbol{\psi}).$$

  In particular, $g^{-1}(\eta_i) = E[y_i|\eta_i, \boldsymbol{\psi}]$ where η_i is a linear predictor in the form:

$$\eta_i = \mu + \sum_{p=1}^{P} x_{ip}\beta_p + \sum_{r=1}^{R} f_r(z_{ir})$$

  and $x_1, ..., x_P$ and $z_1, ..., z_R$ are covariates. The functions $f_r(\cdot)$ are always defined in practice using a basis $\boldsymbol{D}_r(\cdot)$ and a set of latent random coefficients $\boldsymbol{u}_r$:

$$f_r(Z_r) = \boldsymbol{D}_r^T(Z_r)\boldsymbol{u}_r \tag{2.2}$$

  where the basis $\boldsymbol{D}_r(\cdot) = [D_{r,1}(\cdot), ..., D_{r,K_r}(\cdot)]^T$ is a column vector of K_r known basis functions, and the vector $\boldsymbol{u}_r$ is of dimension $K_r \times 1$. Hence, η_i can be

rewritten as a linear combination of the latent parameters $\boldsymbol{\alpha} = [\mu, \boldsymbol{\beta}, \boldsymbol{u}_1, ..., \boldsymbol{u}_R]$:

$$\eta_i = \mu + \sum_{p=1}^{P} x_{ip}\beta_p + \sum_{r=1}^{R} \boldsymbol{D}_r^T(z_{ir})\boldsymbol{u}_r. \tag{2.3}$$

- **Latent model**. The latent parameters $\mu, \boldsymbol{\beta}, \boldsymbol{u}_1, ..., \boldsymbol{u}_R$ are all assigned independent Gaussian distributions conditional on hyperparameters, i.e. $\mu \perp\!\!\!\perp \boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{u}_1 \perp\!\!\!\perp ... \perp\!\!\!\perp \boldsymbol{u}_R | \boldsymbol{\sigma}$. As a whole, the latent parameter vector $\boldsymbol{\alpha}$ will follow a multivariate Gaussian distribution, i.e. a Gaussian Markov random field (see Section 2.2). Both the mean parameter $\boldsymbol{\mu}_\alpha$ and the precision matrix $\boldsymbol{P}_\alpha$ of the Gaussian distribution can depend on the hyperparameter vector $\boldsymbol{\sigma}$ so that they are denoted as $\boldsymbol{\mu}_\alpha(\boldsymbol{\sigma})$ and $\boldsymbol{P}_\alpha(\boldsymbol{\sigma})$:

  $$\boldsymbol{\alpha}|\boldsymbol{\sigma} \sim N(\boldsymbol{\mu}_\alpha(\boldsymbol{\sigma}), \boldsymbol{P}_\alpha^*(\boldsymbol{\sigma}))$$

  where $\boldsymbol{M}^*$ represents the generalized inverse of the matrix $\boldsymbol{M}$.

  Thanks to the conditional independence assumption, the precision matrix $\boldsymbol{P}_\alpha(\boldsymbol{\sigma})$ will be sparse, which is computationally convenient. Moreover, Gaussianity ensures that $\boldsymbol{\eta} = [\eta_1, ..., \eta_n]^T$ is also a Gaussian field, conditional on $\boldsymbol{\sigma}$.

- **Hyperparameter model**. The vector of hyperparameters $\boldsymbol{\gamma} = [\boldsymbol{\psi}, \boldsymbol{\sigma}]$ must be assigned a prior. No restrictions are imposed on the $\pi(\boldsymbol{\gamma})$ density.

The joint posterior distribution for the unknown parameters of an LGM is found to be:

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\gamma}|\boldsymbol{y}) \propto \prod_{i=1}^{n} \pi(y_i|\boldsymbol{\alpha}, \boldsymbol{\gamma}) |\boldsymbol{P}_\alpha(\boldsymbol{\gamma})|^{1/2} \exp\left\{-\frac{1}{2}\left[\boldsymbol{\alpha} - \boldsymbol{\mu}_{\boldsymbol{\alpha}}(\boldsymbol{\gamma})\right]^T \boldsymbol{P}_\alpha(\boldsymbol{\gamma})\left[\boldsymbol{\alpha} - \boldsymbol{\mu}_{\boldsymbol{\alpha}}(\boldsymbol{\gamma})\right]\right\} \pi(\boldsymbol{\gamma}). \tag{2.4}$$

In general, it is not possible to obtain a closed form solution for this posterior and simulation-based algorithms or approximation methods are usually needed for the computation of Equation 2.4. Along with MCMC methods, LGMs can be nicely fitted using the increasingly popular methodology proposed by Rue, Martino, and Chopin 2009, which introduced the Integrated Nested Laplace approximation (INLA). A package for the implementation of INLA is available in R and will be used for posterior inference for the rest of the thesis.

2.2 Gaussian Markov Random Fields (GMRFs)

As described in the previous section, Gaussian Markov Random Fields (GMRFs) are used as priors on the $\boldsymbol{\alpha}$, especially for the $\boldsymbol{u}_r$ coefficient vectors of LGMs. GMRFs have been thoroughly investigated by Rue and Held 2005, who defined them as follows.

Definition 2.1 (GMRF)**.** *A random vector $\boldsymbol{u} = [u_1, ..., u_n]^T \in \mathbb{R}^n$ is called a Gaussian Markov Random Field with respect to an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, ..., n\}$ is the set of nodes and $\mathcal{E}$ is the set of edges, with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{P}$, where $\boldsymbol{P}$ is a symmetric, positive definite matrix, if and only if its density has the form:*

$$\pi(\boldsymbol{u}) = (2\pi)^{-n/2}|\boldsymbol{P}|^{1/2}\exp\left(-\frac{1}{2}(\boldsymbol{u}-\boldsymbol{\mu})^T\boldsymbol{P}(\boldsymbol{u}-\boldsymbol{\mu})\right)$$

and

$$P_{ij} \neq 0 \implies \{i, j\} \in \mathcal{E} \quad i \neq j.$$

In other words, a GMRF is a random vector that follows a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{P}$. There is complete correspondence between the precision matrix $\boldsymbol{P}$ and the undirected graph $\mathcal{G}$.

Although technically the precision matrix could be dense, the properties of a GMRF are nicer when $\boldsymbol{P}$ is sparse, so that Rue and Held 2005 focused on this case. Among such properties, Gaussianity ensures that the conditional independence structure on the vector is fully summarized by $\boldsymbol{P}$, or equivalently by the graph $\mathcal{G}$: for example, the conditional distribution of x_i given the remaining $\boldsymbol{u}_{-i} = [u_1, ..., u_{i-1}, u_{i+1}, ..., u_n]^T$ only depends on u_j such that $P_{ij} \neq 0$, which correspond to the first-order neighbours of u_i on the graph $\mathcal{G}$ (local Markov property). Thanks to this and other Markovian properties of a GMRF, we can conclude that its joint probability distribution is uniquely defined by the collection of its full conditionals (Lemma 2.3 of Rue and Held 2005):

$$\pi(\boldsymbol{u}) = \prod_{i=1}^{n} \pi(u_i|\boldsymbol{u}_{-i}).$$

Rue and Held 2005 exploited these properties to derive computationlly efficient algorithms for sampling a GMRF, either unconditionally, conditionally or subject to linear constraints. These algorithms make use of the Cholesky factorization of the

precision matrix, which computes a lower triangle matrix $\boldsymbol{L}$ such that $\boldsymbol{P} = \boldsymbol{L}\boldsymbol{L}^T$. This factorization also provides a fast computation of the density function.

2.3 Intrinsic Gaussian Markov Random Fields (IGMRFs)

Another important class of effects consists of intrinsic GMRFs (IGMRFs), which are popularly used in application, mostly to capture spatial or temporal correlation structures. IGMRFs are improper GMRFs, which are defined by Rue and Held 2005 as follows.

Definition 2.2 (Improper GMRF)**.** *A random vector* $\boldsymbol{u} = [u_1, ..., u_n]^T \in \mathbb{R}^n$ *is called an improper GMRF of rank* $n - k$ *with parameters* $\boldsymbol{\mu}$ *and* $\boldsymbol{P}$*, where* $\boldsymbol{P}$ *is a symmetric, positive semi-definite matrix with rank* $n-k$*, if its density has the form:*

$$\pi(\boldsymbol{u}) = (2\pi)^{-(n-k)/2}(|\boldsymbol{P}|^*)^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{u}-\boldsymbol{\mu})^T\boldsymbol{P}(\boldsymbol{u}-\boldsymbol{\mu})\right) \tag{2.5}$$

and

$$P_{ij} \neq 0 \implies \{i,j\} \in \mathcal{E} \quad i \neq j.$$

Note that $|\boldsymbol{M}|^*$ denotes the generalized determinant of matrix $\boldsymbol{M}$. Equation 2.5 is a proper density for the GMRF $\boldsymbol{u}$ conditional on k linear constraints $\boldsymbol{S}^T\boldsymbol{u} = \boldsymbol{e}$ where $\boldsymbol{S}$ is a matrix of dimension $n \times k$, such that $\boldsymbol{P}\boldsymbol{S} = \boldsymbol{0}$, i.e. $\boldsymbol{S}$ is the null space of $\boldsymbol{P}$. Therefore, an improper GMRF can be viewed as a GMRF under linear constraints where all vectors $\boldsymbol{u} \in \mathbb{R}^n$ can be realizations and not only those $\boldsymbol{u}$ respecting the constraints. Because $\boldsymbol{P}$ is singular, the density in Equation 2.5 is invariant to the addition to $\boldsymbol{u}$ of any vector $\boldsymbol{u}^{(0)} = \boldsymbol{S}\boldsymbol{v}$ for any choice of $\boldsymbol{v} \in \mathbb{R}^k$.

Among the large class of improper GMRFs, there are some important cases that are popularly used in practice. These include cases in which the null space of the rank-deficient matrix $\boldsymbol{P}$ has a well known structure. Based on such requirement, Rue and Held 2005 defined for example IGMRFs of first order.

Definition 2.3 (IGMRF of first order)**.** *An IGMRF of first order is an improper GMRF of rank* $n-1$ *where* $\boldsymbol{P}\boldsymbol{1}=\boldsymbol{0}$*.*

From the definition, the full conditional of an IGMRF has the following expec-

tation:

$$E[u_i|\boldsymbol{u}_{-i}] = \sum_{j:j\sim i} \frac{P_{ij}}{P_{ii}} u_j \qquad i = 1, ..., n$$

which means that the conditional mean of an entry is the weighted average of its neighbours' values but does not involve an overall level μ. As such, an IGMRF of first order describes the behaviour of deviations from the mean without the need to directly specify it.

A first-order random walk can be proven to be an IGMRF of order 1 and can be called in fact an IGMRF on regular locations on the line. A first-order random walk is defined assuming independent increments:

$$u_{i+1} - u_i \overset{iid}{\sim} N(0, \sigma^2). \tag{2.6}$$

Since the process is defined on differences, it requires a starting condition on the value on u_1, or alternatively on any other value.

This specification leads to the following full conditionals:

$$u_i|\boldsymbol{u}_{-i} \sim N\left(\frac{u_{i-1} + u_{i+1}}{2}, \frac{\sigma^2}{2}\right) \qquad i = 1, ..., n.$$

The joint distribution of $\boldsymbol{u}$ has the following precision matrix $\boldsymbol{P} = \dfrac{\boldsymbol{Q}}{\sigma^2}$ where:

$$\boldsymbol{Q} = \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & ... & ... & ... & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix} \tag{2.7}$$

Since $\boldsymbol{P}\mathbf{1}=\mathbf{0}$, the first-order random walk is clearly an IGMRF of order 1.

Another type of first-order IGMRF is the intrinsic conditional autoregressive model (ICAR) on a lattice, either regular or irregular (Besag and Kooperberg 1995). This model is usually specified on the basis of the adjacency matrix $\boldsymbol{W}$ based on the lattice, where $\boldsymbol{W}_{ij} = 1$ if the regions of the lattice i and j are first-order neighbours. The precision matrix is then defined as $\boldsymbol{P} = \dfrac{1}{\sigma^2}[\boldsymbol{G} - \boldsymbol{W}]$, where $\boldsymbol{G}$ is a diagonal matrix with $G_{ii} = \sum_{j=1}^{n} W_{ij}$. This definition implies that $\boldsymbol{Q}\mathbf{1}=\mathbf{0}$.

The full conditionals for the ICAR model are:

$$u_i|\boldsymbol{u_{-i}} \sim N\left(\frac{1}{n_i}\sum_{j:j\sim i} u_j, \frac{\sigma^2}{n_i}\right) \qquad i = 1, ..., n$$

where $j \sim i$ denotes that region i and j are first-order neighbours and n_i denotes the total number of first-order neighbours of region i. A weighted version can be created simply replacing $\boldsymbol{W}$ with any other symmetric matrix of positive weights.

IGMRFs of higher order are similarly defined, i.e. on the basis of the null space of $\boldsymbol{P}$. For example, we can define IGMRFs of d^{th}-order on the line.

Definition 2.4 (IGMRF of d^{th}-order on the line)**.** *An IGMRF of order d is an improper IGMRF of rank $n - d$ with $\boldsymbol{P}\boldsymbol{S}_{(d-1)} = \boldsymbol{0}$, where $\boldsymbol{S}_{(d-1)}$ is a Vandermonde matrix (Hoffman and Kunze 1971) of degree $d - 1$.*

An example of IGMRF of order 2 on the line is the second-order random walk, which defines independent increments for the second-order differences:

$$u_{i+2} - 2u_{i+1} + u_i \overset{iid}{\sim} N(0, \sigma^2).$$

Again, the process requires a starting condition to be realized: specifically, an initial value for u_1 and u_2 must be specified (or any other couple of neighbours).

The joint distribution of $\boldsymbol{u}$ has the following precision matrix $\boldsymbol{P} = \dfrac{\boldsymbol{Q}}{\sigma^2}$ where:

$$\boldsymbol{Q} = \begin{bmatrix}
1 & -2 & 1 & & & & & & \\
-2 & 5 & -4 & 1 & & & & & \\
1 & -4 & 6 & -4 & 1 & & & & \\
 & 1 & -4 & 6 & -4 & 1 & & & \\
 & & ... & ... & ... & ... & ... & & \\
 & & & 1 & -4 & 6 & -4 & 1 & \\
 & & & & 1 & -4 & 6 & -4 & 1 \\
 & & & & & 1 & -4 & 5 & -2 \\
 & & & & & & 1 & -2 & 1
\end{bmatrix}. \tag{2.8}$$

One popular use of second-order random walks is in the definition of P-Spline effects (Eilers and Marx 1996, Lang and Brezger 2004), which are commonly used for smoothing. P-Splines are defined as functions $f(x) = \boldsymbol{B}^T(x)\boldsymbol{u}$ of a continuous covariate x through a basis $\boldsymbol{B}(x) = [B_1(x), ..., B_K(x)]^T$, made up by equidistant cubic B-Splines (Figure 2.1).

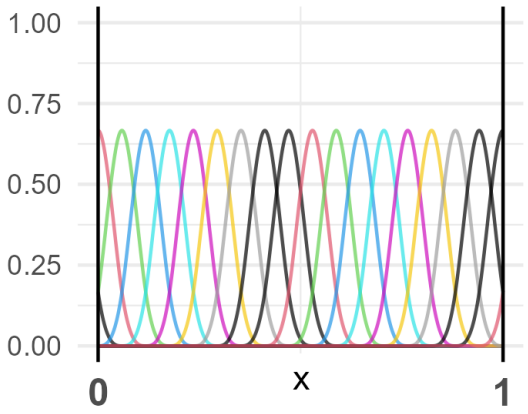

Figure 2.1: 20 cubic equidistant B-Spline functions on the support $x \in [0, 1]$. Each colored line represents a $B_k(x)$ function of the basis $\boldsymbol{B}(x)$.

The set of coefficients $\boldsymbol{u}$ is specified as a second-order random walk process: this choice corresponds to a penalization on the second-order differences between the coefficients and regularizes the wiggliness in the realizations of $f(x)$. More details on P-Splines are contained in the following chapter.

2.4 Scaling issue for IGMRFs

As mentioned, IGMRFs are often used in LGMs to account for spatial or temporal dependence. The scale of such effects is controlled by the variance parameter σ^2, which is usually considered random and assigned an hyperprior. The priors on the σ^2 parameters control the degree of smoothness of the trends, so that posterior results can be quite sensitive to prior specification.

Sørbye and Rue 2014 noted that the scale of an IGMRF is not directly controlled by the variance parameter σ^2, but rather that it changes for different models, as well as for different dimensions n. Consider for example that a process on the line in the $[0, 1]$ interval is modelled using a first-order random walk for regular locations as defined in the previous section. Consider now a design in which the number of observations' locations is $n = 11$, so that the process is defined on $x_1 = 0, x_2 = 0.1, ..., x_{10} = 0.9, x_{11} = 1$ as in Equation 2.6:

$$u_{i+1} - u_i \overset{iid}{\sim} N(0, \sigma^2) \quad i = 1, ..., 10.$$

Then, by definition we know that:

$$Var[u_{i+1} - u_i] = \sigma^2.$$

If we now consider a different design with $n = 21$, the random walk would be defined on $x'_1 = 0, x'_2 = 0.05, x'_3 = 0.1, ..., x'_{19} = 0.9, x'_{20} = 0.95, x'_{21} = 1$ where $x'_{2i-1} = x_i$:

$$u'_{i+1} - u'_i \overset{iid}{\sim} N(0, \sigma^2_{\text{new}}) \qquad i = 1, ..., 20.$$

If we now compute the variance between u_{i+1} and u_i, we obtain that $\sigma^2 = 2\sigma^2_{\text{new}}$ since:

$$Var[u_{i+1} - u_i] = Var[u'_{i+2} - u'_i] = 2\sigma^2_{\text{new}}.$$

We can note that different designs therefore imply different meanings of the variance parameter for a first-order random walk. Similar conclusions can be derived for second-order random walks and IGMRFs in general. Thus, we can understand how the scale of the effects is not only controlled by the variance parameter σ^2 but also by the design, i.e. the number of locations on the support. Imposing the same hyperprior on σ^2 and σ^2_{new} will induce two different assumptions a priori on the scale of the effect or the degree of smoothness of the process. This phenomenon can be called *scaling issue* and appears clearly when we consider the pattern of the marginal variance of each entry of the process u_i conditional on the scale parameter σ^2. This quantity is well-defined only after having imposed appropriate linear constraints that transform the IGMRF into a proper GMRF. Since a first-order random walk is an IGMRF of first-order, the constraint $\sum_{i=1}^{n} u_i = 0$ is necessary. Under this condition, we can find that the marginal variance of each u_i is equal to:

$$Var[u_i|\sigma^2] = \sigma^2[\boldsymbol{Q}^*]_{ii}$$

where $\boldsymbol{Q}^*$ is the generalized inverse of $\boldsymbol{Q}$ and $\boldsymbol{Q}_{ij}$ is the entry at row i and column j. Figure 2.2 reports these quantities for $n = 11$ and $n = 21$, conditional on the value $\sigma^2 = 1$.

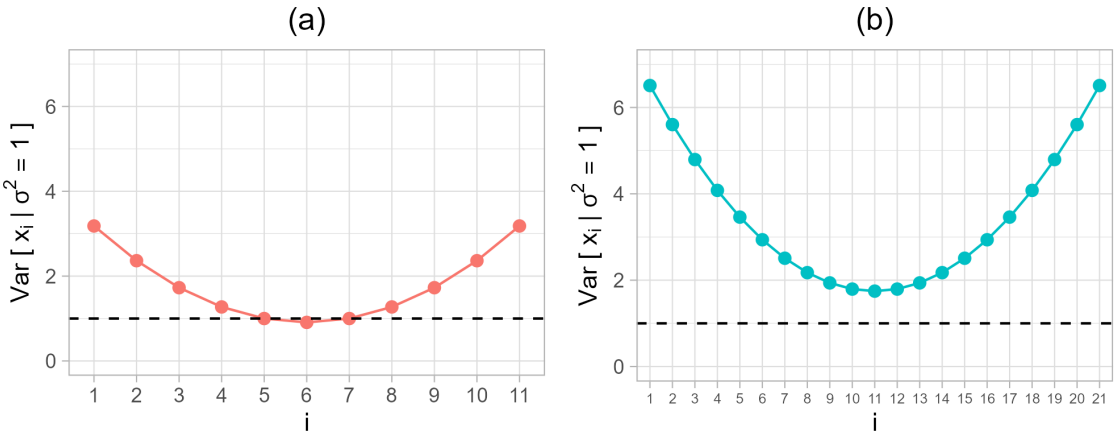

Figure 2.2: Marginal variance of u_i for a first-order random walk subject to $\sum_{i=1}^{n} u_i = 0$ and $\sigma^2 = 1$: (a) $n = 11$; (b) $n = 21$. The dashed black line is set at 1 to represent the value of σ^2.

First, we can note that the marginal variances do not have a constant value for all x_i in neither of the cases. Secondly, the two patterns are different for the two designs, and the value $\sigma^2 = 1$ is not a good summary of the patterns, especially for the case $n = 21$. This phenomenon happens for the IGMRFs in general, as the pattern changes for IGMRFs of different kinds and dimensions.

To solve the scaling issue, Sørbye and Rue 2014 first suggested that IGMRF effects should be appropriately scaled for a consistent interpretation of the σ^2 parameters over different models and, therefore, for a mindful specification of their hyperpriors. Specifically, Sørbye and Rue 2014 proposed to compute *reference standard deviations* σ_{ref}, defined as the geometric mean of the marginal variance given $\sigma^2 = 1$:

$$\begin{aligned}\sigma_{\text{ref}} &= \sqrt{\exp\left[\frac{1}{n}\sum_{i=1}^{n}\log(Var[u_i|\sigma^2 = 1])\right]} \\ &= \sqrt{\exp\left[\frac{1}{n}\sum_{i=1}^{n}\log([\boldsymbol{Q}^*]_{ii})\right]}.\end{aligned}$$

The value of σ_{ref} is different for each IGMRF model and reference standard deviations for univariate and bivariate IGMRFs of first- and second-order with varying dimension n have been reported in Spyropoulou and Bentham 2024. Sørbye and Rue 2014 then proposed to specify the hyperpriors on $\sigma^2 \cdot \sigma^2_{\text{ref}}$ to ensure that the interpretation of the hyperprior does not change from model to model. This procedure is equivalent to dividing the $\boldsymbol{u}$ effects by σ_{ref}. After scaling, the same hyperprior on σ^2 reflects the same assumption about the scale of the deviation of the effect from

its null space for all IGMRF models, regardless of the type and dimension.

Sørbye and Rue 2014 showed the practical impact of their proposal on posterior estimation through real case studies in disease mapping and ecology.

2.5 Penalized Complexity priors

Simpson et al. 2017 presented a novel approach for the derivation of prior distributions, which is referred to as Penalized Complexity (PC) priors. PC priors do not have a fixed functional form, but rather are built using a set of 4 principles or steps, which can be potentially applied to any parameter. Note that the principled nature of PC priors makes them invariant to reparametrization, which is a highly desirable quality that puts them in the same category as the revolutionary class of Jeffreys' priors.

Consider a model component with density $\pi(\boldsymbol{x}|\theta)$ where θ is the unknown parameter for which a prior must be specified.

1. **Occam's razor.** The principle that simpler models should be preferred until there is enough evidence for more complex models is adopted. Following this idea, a *base model* b, i.e. the value of θ that corresponds to the simplest model, is specified. Desirably, the base model shall be the preferred value in the prior distribution, while divergence from it should be penalized.

2. **Measure of complexity.** A measure of complexity intended as divergence from the base model is defined using the KLD between $\pi(\boldsymbol{x}|\theta)$ and $\pi(\boldsymbol{x}|\theta = b)$. The KLD as a function of θ is used to define the distance $d(\theta)$ from the base model as:

$$d(\theta) = \sqrt{2 \cdot \mathrm{KLD}[\pi(\boldsymbol{x}|\theta)||\pi(\boldsymbol{x}|\theta = b)]}.$$

   This particular transformation of the KLD is chosen so that $d(\theta)$ behaves more like a proper distance metric, since asymptotically it becomes equal to the true Fisher information distance (Simpson et al. 2017).

3. **Constant rate penalization.** The goal of penalizing model complexity is achieved specifying am Exponential distribution on $d(\theta)$:

$$d(\theta) \sim \mathrm{Exp}(\delta).$$

The implied prior on θ is found applying the change-of-variable formula:

$$\pi(\theta) = \delta \exp[-\delta \cdot d(\theta)] \cdot \Big|\frac{\mathrm{d}d(\theta)}{\mathrm{d}\theta}\Big|.$$

The Exponential distribution is chosen as it has a constant decay rate, which is the most reasonable assumption without additional knowledge. This choice has the advantage of ensuring the memoryless property. However, it is not the best option whenever sparse regression is desirable: in this case, heavier tails are necessary (Simpson et al. 2017).

4. **User-defined scaling.** The Exponential distribution hyperparameter δ is specified using a tail probability statement in terms of an upper bound U and tail probability α on an interpretable transformation $g(\theta)$ of the original parameter.

$$P(g(\theta) > U) = \alpha.$$

A PC prior distribution will be denoted here by $\mathrm{PC}_b(\delta)$, as in Hem, Fuglstad, and Riebler 2024.

We can now focus on the construction of a PC prior for the variance parameter σ^2 of a GMRF or IGMRF. In this context, the model simplicity principle suggests to set the base model to 0, which corresponds to the absence of the effect, i.e. the simplest possible model. Simpson et al. 2017 derived the PC_0 prior for σ^2 and found the following functional form:

$$\pi(\sigma^2) = \frac{\delta}{2\sqrt{\sigma^2}} \exp\left(-\delta\sqrt{\sigma^2}\right)$$

which corresponds to a Weibull distribution with scale hyperparameter $1/\delta^2$ and shape hyperparameter 1/2. Under different parametrizations, the distribution becomes an Exponential with hyperparameter δ on the standard deviation $\sqrt{\sigma^2}$ or a type-2 Gumbel distribution on the precision $1/\sigma^2$. Simpson et al. 2017 suggested to set the hyperparameter using a tail probability statement on the standard deviation:

$$P(\sqrt{\sigma^2} > U) = \alpha.$$

The Exponential hyperparameter is equal to $\delta = \frac{-\log(\alpha)}{U}$.

The use of the PC_0 prior for variance parameters of random effects of LGMs has become increasingly popular in recent years, due to its good theoretical properties, and superior performance in comparison to more traditional choices (e.g. Inverse-

Gamma), specifically in avoiding overfitting. Extensive simulation studies about the performance of PC priors have been performed in Klein and Kneib 2016.

Along with variances, PC priors have been applied to many other types of parameters, including: tail dependence (Kereszturi, Tawn, and Jonathan 2016), degrees of freedom for P-splines (Ventrucci and Rue 2016), Matérn parameters (Geir-Arne Fuglstad and Rue 2019), autoregressive correlation (Sørbye and Rue 2017), skewness (Ordonez et al. 2024).

2.6 Hierarchical Decomposition priors

Fuglstad et al. 2020 proposed a new method for the prior specification of the variance parameters for the random effects of LGMs. Consider again the linear predictor of an LGM as defined in Equation 2.3. In common practice, the Gaussian distribution of $\boldsymbol{\alpha}$ is specified as:

$$\begin{aligned} \mu &\sim N(0, \sigma_I^2) \\ \boldsymbol{\beta} &\sim N_P(0, \sigma_F^2 \boldsymbol{I}_P) \\ \boldsymbol{u}_r | \sigma_r^2 &\sim N_{K_r}(\boldsymbol{0}, \sigma_r^2 \boldsymbol{Q}_r^*) \quad r = 1, ..., R. \end{aligned}$$

Traditionally, σ_I^2 and σ_F^2 are fixed to large values and the prior specification focuses on the hyperparameters $\boldsymbol{\sigma} = [\sigma_1^2, ..., \sigma_R^2]$, along with potential likelihood parameters $\boldsymbol{\psi}$. If the precision matrices are not actually fixed but depends on some correlation parameters (e.g. spatial Matern processes or autoregressive temporal processes), such parameters are considered fixed to a reasonable value during the prior specification for $\boldsymbol{\sigma}$ and their priors are separately specified (Fuglstad et al. 2020).

The usual approach to define the prior on the variance parameters is to assume independence such that:

$$\pi(\sigma_1^2, ..., \sigma_R^2) = \prod_{r=1}^{R} \pi(\sigma_r^2).$$

The independence assumption ignores potential prior knowledge regarding the relative importance of different effects. This limitation can be addressed by specifying a joint prior distribution for the variance parameters. However, effectively capturing prior information about the relationships between the individual σ_j^2 through a joint distribution is challenging. Fuglstad et al. 2020 presented a framework in which an intuitive joint prior can be specified on these parameters, in a user-friendly way. This is achieved specifying the prior on a reparametrization of the original $\sigma_1^2, ..., \sigma_R^2$

parameters.

The reparametrization starts with the definition of the *total latent variance*:

$$W = \sum_{r=1}^{R} \sigma_r^2 \tag{2.9}$$

where W is the total variance in the linear predictor after having accounted for the fixed effects. If all random effects are homogeneous, i.e. $Var[\boldsymbol{D}_r^T(z_{ir})\boldsymbol{u}_r|\mu, \boldsymbol{\beta}, \boldsymbol{\sigma}] = \sigma_r^2$ for $i = 1, ..., n$, then $W = Var[\eta_i|\mu, \boldsymbol{\beta}, \boldsymbol{\sigma}]$ thanks to additivity. If some effects are heteregeneous, i.e. $Var[\boldsymbol{D}_r^T(z_{ir})\boldsymbol{u}_r|\mu, \boldsymbol{\beta}, \boldsymbol{\sigma}] \neq \sigma_r^2$, as in the case of IGMRFs, then $W \approx Var[\eta_i|\mu, \boldsymbol{\beta}, \boldsymbol{\sigma}]$ for $i = 1, ..., n$ as long as the effects have been appropriately scaled, as described in Section 2.4.

The remaining parameters are found building a *hierarchical decomposition* tree, which decomposes the total latent variance W through successive splits. It is desirable to design the splits such that the desired comparisons between effects (i.e. for which there is relevant information a priori) are reflected into the child nodes of a specific split. Fuglstad et al. 2020 formally described the design of a tree $\mathcal{T}$. The root node always contains W. The first split divides the root node in at least two child nodes, each containing one or more random effects' variance parameters σ_r^2. More splits are added until all the child nodes contain a single σ_r^2. We denote the total number of splits S, the number of branches at split s as K_s, the parent node at split s as a set P_s containing the indices of the effects in the node, and the child nodes as sets $C_{s,1}, ..., C_{s,K_s}$. The new parameters $\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_S$ are S vectors of proportions defined dividing the variance parameters in the child nodes by the sum of the variance parameters from the parent node at each split:

$$\boldsymbol{\phi}_s = \frac{1}{\sum_{r \in P_s} \sigma_r^2} \left(\sum_{r \in C_{s,1}} \sigma_r^2, ..., \sum_{r \in C_{s,K_s}} \sigma_r^2 \right) \quad s = 1, ..., S. \tag{2.10}$$

Note that $0 \leq \phi_{sk} \leq 1$, $\forall k = 1, ..., K_s$ and $\sum_{k=1}^{K_s} \phi_{sk} = 1$.

This hierarchical reparametrization offers a more intuitive way to specify a joint prior on the original parameters, as users now need to choose priors on interpretable proportion parameters that compare the relative importance of different groups of random effects.

In order to exemplify the decomposition tree design, we can consider the simple case of an LGM with only three random effects and respective variances $\sigma_1^2, \sigma_2^2, \sigma_3^2$. The total latent variance is then $W = \sigma_1^2 + \sigma_2^2 + \sigma_3^2$. One potential design of the decomposition tree is depicted in Figure 2.3: first, the first two effects are separated

by the third one (split $s = 1$); then, a second distinction ($s = 2$) is made between the first two effects.

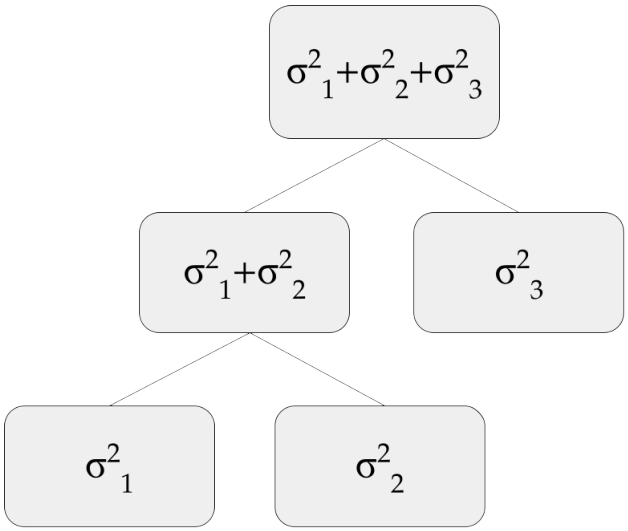

Figure 2.3: Example of Hierarchical Decomposition tree.

In this case, the tree has $S = 2$ total splits, both having two branches ($K_1 = K_2 = 2$), with parent nodes $P_1 = [1, 2, 3], P_2 = [1, 2]$, and child nodes $C_{1,1} = [1, 2], C_{1,2} = [3], C_{2,1} = [1], C_{2,2} = [2]$. In this example, the new vector parameters are actually single proportions that can be written as $\boldsymbol{\phi}_1 = [\phi_1, 1 - \phi_1]$ and $\boldsymbol{\phi}_2 = [\phi_2, 1 - \phi_2]$ where:

$$\phi_1 = \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}$$
$$\phi_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}.$$

Fuglstad et al. 2020 defined a class of priors on the reparametrization from Equations 2.9-2.10, called Hierarchical Decomposition (HD) priors, under the reasonable assumption that each parameter vector $\boldsymbol{\phi}_s$ will only depend on vector $\boldsymbol{\phi}_t$ if t is a descendant split of s (bottom-up approach).

Definition 2.5 (HD priors for LGMs)**.** *A hierarchical decomposition prior on the variance parameters $\sigma_1^2, ..., \sigma_R^2$ of a latent Gaussian model is defined as:*

$$\pi(\sigma_1^2, ..., \sigma_R^2) = \pi(W|\{\boldsymbol{\phi}_s\}_{s=1}^S) \prod_{s=1}^{S} \pi(\boldsymbol{\phi}_s|\{\boldsymbol{\phi}_t\}_{t \in D_s}) \tag{2.11}$$

where D_s denotes the set containing the indices of descendant splits of split s.

For the example presented above, Equation 2.11 becomes:

$$\pi(\sigma_1^2, ..., \sigma_3^2) = \pi(W|\phi_1, \phi_2)\pi(\phi_1|\phi_2)\pi(\phi_2).$$

While the conditioning on descendant splits' parameters is a sensible assumption, it is practically inconvenient in terms of computation. In practice, Fuglstad et al. 2020 therefore resorted to the use of *simplified conditioning*: the prior distributions on W and $\boldsymbol{\phi}_s$ are specified not as a function of the descendant parameters $\boldsymbol{\phi}_t$ but only considering a baseline value of them denoted by $\boldsymbol{\phi}_t^0$. An HD prior under simplified conditioning is technically made up by independent priors on the new parameters:

$$\pi(\sigma_1^2, ..., \sigma_R^2) = \pi(W) \prod_{s=1}^{S} \pi(\boldsymbol{\phi}_s).$$

For the example above, the prior simplifies to:

$$\pi(\sigma_1^2, ..., \sigma_3^2) = \pi(W)\pi(\phi_1)\pi(\phi_2).$$

The first step of the specification of an HD prior requires choosing a distribution for the total latent variance W. Following the work of Gelman, Simpson, and Betancourt 2017, Fuglstad et al. 2020 suggested to specify $\pi(W)$ taking into account the likelihood of the model. In the case of a Gaussian likelihood, it is recommended to add the additional Gaussian noise to the decomposition tree and assign a scale-invariant Jeffreys distribution to $W = \sum_{r=1}^{R} \sigma_r^2 + \sigma_\epsilon^2$. Under different likelihood models, the use of a PC prior with base model $W = 0$ is suggested, in order to shrink towards a model with no random effects unless there is evidence in the data for their importance.

The second step of an HD prior consists in the specification of prior distributions for the $\boldsymbol{\phi}_s$ vectors. The choice of the prior depends on the type of information the user has about the corresponding split. The user may or may not have prior beliefs about the relative importance of the child nodes of a given split. Fuglstad et al. 2020 identified two distinct categories of splits according to the type of information available to the users and proposed two corresponding approaches for the prior specification of the corresponding proportion vectors.

- **Indifference between branches**. If the user wants to express ignorance between the partition of the variance among the child nodes of a given split, an exchangeable prior can be chosen for $\boldsymbol{\phi}_s$, such as for example a symmetric Dirichlet distribution $\text{Dir}(q, .., q)$. If the user wants to reflect complete igno-

rance about the partition between the branches, the hyperparameter q can be set to 1 to obtain a Uniform distribution on the simplex space.

- **Preference towards a branch**. If instead the user wants to express a preference towards a branch, PC priors allow for a penalization of deviations from a preferred value of the partition. Fuglstad et al. 2020 derived the explicit form of the PC prior for proportion parameters for a relevant class of models. Franco-Villoria, Ventrucci, and Rue 2022 did the same for a subclass of models using Kronecker product IGMRFs.

Fuglstad et al. 2020 provided details for an intuitive specification of the hyperparameters of HD priors.

The simulation study performed by Fuglstad et al. 2020 showed that HD priors are competitive with respect to independent PC_0 priors on the σ_r^2 parameters, which represent a state-of-the-art alternative for LGMs and have been popularized by the R-INLA package. HD priors have been so far applied in various fields: demography (Fuglstad et al. 2020), genomics (Hem et al. 2021), disease mapping (Franco-Villoria, Ventrucci, and Rue 2022), forestry (Marques, Wiemann, and Kneib 2023).

The application of HD priors has been made easier by the work of Hem, Fuglstad, and Riebler 2024, who developed the user-friendly `makemyprior` package for the design of the decomposition tree and the prior specification on the consequent set of new parameters through the use of intuitive probability statements.

A more realistic tree design and HD prior specification are discussed in Chapter 4, for the treatment of species' distribution models from ecology.

2.7 R2D2 priors literature review

Fuglstad et al. 2020 cited as a reference the Dirichlet-Laplace prior introduced by Bhattacharya et al. 2015, which is a global-local shrinkage prior derived in the context of Bayesian variable selection. This is the first out of many works from a branch of literature that uses the same reparametrization idea of the HD prior with a different purpose, namely to induce a shrinkage prior on the linear coefficients and thus perform sparse regression. In addition to the purpose, the main difference between this separate class of priors and the HD class lies in the fact that the latter aims at fully exploiting prior knowledge through an application-specific decomposition tree, while the specification of this class of global-local shrinkage priors makes use of a single split with as many branches as the number of variance parameters.

Consider the case in which the linear predictor of Equation 2.3 only contains

linear effects:

$$\eta_i = \mu + \sum_{p=1}^{P} x_{ip}\beta_p.$$

Assume now that the Gaussian distribution of LGMs is replaced with a Double-Exponential (or Laplace) on the linear coefficients $\beta_p|\sigma_p^2 \sim \text{DE}(0, \sigma_p^2)$ where $Var[\beta_p] = \sigma_p^2$ for all $p = 1, ..., P$. Note that the Double-Exponential is a typical choice in a sparse regression context for its shrinkage properties.

The Dirichlet-Laplace (DL) prior uses the same reparametrization idea of HD priors to specify a joint prior distribution on the $\sigma_1^2, ..., \sigma_P^2$ parameters of this model. Let $W = \sum_{p=1}^{P} \sigma_p^2$ and $\boldsymbol{\phi} = [\phi_1, ..., \phi_P]$ where $\phi_p = \sigma_p^2/W$. Assuming that the covariates have all been standardized so that $Var_{i=1}^{n}[x_{ip}] = 1,\ p = 1, ..., P$, then:

$$Var[x_{ip}\beta_p|\mu, \sigma_1^2, ..., \sigma_p^2] \approx \sigma_p^2 \qquad i = 1, ..., n.$$

Hence, W can be interpreted as the total latent variance and ϕ_p as the individual contribution of the p linear effect. From the point of view of global-local shrinkage priors, W is the global scale parameter, while ϕ_p are the local scale parameters.

The DL prior refers to a prior induced on the linear coefficients β_p obtained through the following specification:

$$\begin{aligned} \beta_p|\boldsymbol{\phi}, W &\sim \text{DE}(0, W\phi_p) \quad p = 1, ..., P \\ \boldsymbol{\phi} &\sim \text{Dir}(q, ..., q) \\ W &\sim \text{Gamma}(P \cdot q, 1/2). \end{aligned} \tag{2.12}$$

The symmetric Dirichlet hyperparameter q regulates the sparsity level on the vector $\boldsymbol{\phi}$ and consequently the amount of shrinkage towards 0 on the implied distributions of $\boldsymbol{\beta}$ (the smaller q, the stronger the shrinkage).

The DL prior has been proven to have better tail and concentration properties on the marginal prior $\pi(\beta_p)$ than alternative global-local shrinkage priors. Moreover, it has an optimal posterior contraction rate and provides efficient estimation (Bhattacharya et al. 2015). The DL has been extensively studied for linear variable selection (Zhang and Bondell 2018), and also extended to the more complicated setting of non-linear regression (Wei et al. 2020).

The R2D2 prior builds upon the DL proposal. The idea behind this alternative is to implicitly induce a desirable prior on the linear coefficients by directly eliciting a prior on a more intuitive measure of goodness-of-fit of the model, namely the R^2 as defined in Zhang et al. 2022. A posteriori, the R^2 is an intuitive measure of model

fit, while a priori, it can be interpreted as the proportion of variance explained by the model for future data. To define the R2D2 prior, we first assume that the response has a Gaussian likelihood: $y_i|\eta_i, \sigma_\epsilon^2 \sim N(\eta_i, \sigma_\epsilon^2)$. Secondly, the Gamma distribution on W from Equation 2.12 is replaced by a Beta(a, b) distribution on the marginal version of the R^2 as defined by Zhang et al. 2022:

$$\begin{aligned} R^2 =& \frac{\frac{1}{n}\sum_{i=1}^{n} Var[\eta_i|\mu, \sigma_1^2, ..., \sigma_P^2]}{\frac{1}{n}\sum_{i=1}^{n} Var[y_i|\mu, \sigma_1^2, ..., \sigma_P^2, \sigma_\epsilon^2]} \\ =& \frac{W}{W + \sigma_\epsilon^2}. \end{aligned}$$

The R2D2 prior is defined as the induced prior on β_p when the Beta parameter a is set to $a = P \cdot q$. Note that the Beta distribution choice on R^2 corresponds to a generalized Beta prime distribution on $W|\sigma_\epsilon^2 \sim \text{GBP}(a, b, 1, \sigma_\epsilon^2)$, which has already been studied as a potential prior for variance parameters, for instance by Bai and Ghosh 2021. The marginal distribution $\pi(\beta_p)$ induced by the R2D2 prior has been found to have excellent shrinkage properties, since it simultaneously displays heavier tails and higher concentration at 0 than other popular global-local shrinkage priors, including the DL.

In more recent works, the use of R2D2 priors has been studied in the more popular context of LGMs, which offers both theoretical and computational advantages over the use of Double-Exponential distributions.

Aguilar and Bürkner 2023 extended the application of the R2D2 prior to linear multilevel models, i.e. both random intercepts and random slopes are included. The proposed R2D2M2 prior retains similar properties despite the extension to a multilevel structure and the shift from the Double-Exponential to the Gaussian distribution for the coefficients. Additionally, the authors highlighted the interpretability advantage offered by the R2D2 perspective, by acknowledging the intuitiveness of its hyperparameters and setting them according to interpretable quantities measuring model complexity. The R2D2M2 has recently been further extended to encompass more flexible choices than the symmetric Dirichlet for the prior of $\boldsymbol{\phi}$, namely a logistic Normal distribution (Aguilar and Bürkner 2024).

More recently, Yanchenko, Bondell, and Reich 2024b applied R2D2 priors to the broader class of generalized linear models under the LGM framework. This work mainly focuses on the derivation of exact or approximate induced prior on W for different likelihood choices but overlooks potential challenges due to the inclusion of complex effects or IGMRFs. Finally, the context of spatial models is discussed in Yanchenko, Bondell, and Reich 2024a, which concludes the existing literature on

R2D2 priors at the time of writing to the best of our knowledge.

Chapter 3

Standardization procedure for the use of Variance Partitioning priors in Latent Gaussian Models

3.1 Introduction

Bayesian hierarchical mixed models have become very popular in many fields of application, such as epidemiology, environmental studies, ecology, etc (Lawson 2018, Clark and Gelfand 2006, Ovaskainen et al. 2017). The traditional approach to prior specification of variance parameters in this class of models consists of choosing independent and often identical priors on the scale parameters. The problem of the selection of appropriate prior distributions for such parameters has long been studied in the Bayesian literature. The Inverse-Gamma, widely popular for its conjugacy property, has been found to perform poorly in practice as it leads to overfitting (Gelman 2006, Frühwirth-Schnatter and Wagner 2010, Lunn et al. 2009). Recently, new proposals have been presented to deal with these parameters. Among them, the prior derived according to the Penalized Complexity (PC) approach has been found to perform better than traditional competitors (Simpson et al. 2017).

However, the poor performance of traditional choices is not the only issue related to the usual approach to prior specification of variance parameters. Another limiting factor is the strong assumption of mutual independence between variance parameters, which is always adopted in practice even when prior information may suggest otherwise. In recent years, this has been challenged and the independence assumption has been relaxed.

Fuglstad et al. 2020 presented a new framework based on the design of a hierarchical decomposition of the variance in the linear predictor with the aim of deriving

more intuitive parameters. Specifically, a total variance is considered along with multiple sets of proportions defined on a case-by-case basis, according to a decomposition tree. Specifying independent priors on these new parameters is equivalent to assuming a joint prior distribution on the original variances. This "Hierarchical Decomposition (HD) reparametrization" offers two main benefits over the traditional independence assumption: first, the freedom in the tree design allows the definition of new parameters, coherent with the structure of the model and relevant to the available prior information; second, it is easier to carry out the prior specification on proportions due to their bounded nature. Intuitive ways to set priors on these proportions are suggested in Fuglstad et al. 2020, e.g. through the use of PC priors (Simpson et al. 2017). This approach is however so far limited to a subset of random effects, e.g. it has not been applied to effects for continuous covariates. One of the issues causing this limitation consists in the intuitive interpretation of the new parameters, which is only guaranteed if all the original variance parameters can be correctly interpreted as the contribution to the total variance due to their corresponding effects. The marginal variance of many popular effects is not constant and equal to the corresponding variance parameter, but is instead a function of the covariate that varies over its support, i.e. "non-stationary" variance: some examples include polynomial effects, ICAR models, P-Spline smoothing models, etc. Fuglstad et al. 2020 use the geometric mean scaling method presented in Sørbye and Rue 2014 to deal with this problem, limited to the context of discrete IGMRF cases. Moreover, linear effects are treated as fixed and not considered as part of the total variance.

Another stream of research considers joint priors on the scale parameters because of the desirable properties obtained on the implied marginal prior of linear effects. Bhattacharya et al. 2015 first exploited this idea for the definition of a novel global-local shrinkage prior. More recently, Zhang et al. 2022 introduced the $R2D2$ prior, again to obtain a desirable shrinkage prior: the model fit is controlled globally through a Beta distribution on the R^2 and a symmetric Dirichlet is imposed on the proportions of total variance. The $R2D2$ prior has shown to have better theoretical properties than its main competitors under the use of a Laplace kernel on the linear effects. Similar properties have also been found under the assumption of a Gaussian kernel by Aguilar and Bürkner 2023. Although this approach was originally developed for high-dimensional settings, this method can be applied more generally and has been recently extended to the GLMM class of models (Yanchenko, Bondell, and Reich 2024b,Yanchenko, Bondell, and Reich 2024a). Since this approach has been developed with a different objective than Fuglstad et al. 2020, there is no particular

attention to the interpretability of the proportion parameters. The case of effects with a non-constant variance over the support of the covariate only appears in the context of linear effects and is dealt with classical standardization.

Both lines of research make use of a reparametrization of the original scale parameters, which can be generally defined as *Variance Partitioning (VP) reparametrization* (Franco-Villoria, Ventrucci, and Rue 2022). However, none of the proposals considers all types of potential random effects that a user might wish to introduce in the linear predictor. In particular, the problem of "non-stationary" variance processes has not yet been addressed generally but only for specific cases. Hence, the VP reparametrization still has a limited scope of application in practice.

In this chapter, we argue that the two lines of research can be unified under a more general framework, whose scope of application can be much wider than the one so far delimited in the literature. To make this claim, it is necessary to extend the class of models to which the VP reparametrization can be correctly applied. This is achieved by the development of a general procedure to guarantee an intuitive interpretation of the variance parameters of non-stationary effects, extending the work of Sørbye and Rue 2014. Our proposal maintains the traditional difference between fixed and random effects, even when they are both assigned a random variance parameter, where the novel distinction is based on which quantities are of interest in terms of inference (Gelman et al. 2013,Hodges 2013). Finally, the paper discusses the case of effects with Intrinsic Gaussian Markov Random Fields (IGMRF) priors in this extended framework (Rue and Held 2005). We believe that the proposed approach will allow the exploitation of the advantages of variance partitioning priors in a wider scope of applications and fields. This will be particularly beneficial in applications where it is desirable to introduce expert knowledge in the prior specification, since this becomes much easier under the VP framework, following the guidelines of Fuglstad et al. 2020.

The remainder of the chapter is structured as follows. Section 3.2 defines in detail the class of models under consideration, namely Latent Gaussian Models, and carefully reviews the literature about the VP reparametrization and the issue regarding non-stationary effects; finally, IGMRFs are presented as a particular class of non-stationary effects that present additional challenges. A general standardization procedure that guarantees the correct implementation of variance partitioning priors is presented in Section 3.3. Section 3.4 presents a plethora of popular examples, often employed in applications. Section 3.5 reports simulation studies and an application to real data to assess the practical implications of the theoretical claims made in Section 3.3. Finally, Section 3.6 summarizes the main contributions of the

work and outlines potential future lines of research.

For a user-friendly practical implementation of standardization, the `scaleGMRF` R package has been developed and made publicly available at `https://github.com/LFerrariIt/scaleGMRF`.

3.2 Background

3.2.1 Traditional prior approach in LGMs

Consider the general framework of Latent Gaussian Models presented in Section 2.1, which covers a wide range of common model classes, such as GLMM, GAM, GAMMS, VCM models. The definition of the linear predictor of Equation 2.2 in terms of the basis/coefficients notation is helpful in showing the similarity between the fixed and random components, as both are expressed using a known design matrix and a set of unknown coefficients to be estimated.

In LGMs, all components of the latent additive model for the linear predictor are assumed to be Gaussian conditional on variance hyperparameters. In particular, here we assume that $\mu, \beta_1, ..., \beta_P, \boldsymbol{u}_1, ..., \boldsymbol{u}_R$ from Equation 2.2:

$$\begin{aligned} \mu &\sim N(0, \sigma_I^2) \\ \boldsymbol{\beta} &\sim N(0, \sigma_{F,p}^2) \quad p = 1, ..., P \\ \boldsymbol{u}_r &\sim N(0, \sigma_r^2 \boldsymbol{Q}_r^*) \quad r = 1, ..., R \end{aligned}$$

where $\boldsymbol{Q}_r$ is a known, symmetric, positive semi-definite matrix $\forall r = 1, ...R$ and $\boldsymbol{Q}_r^*$ denotes its the generalized inverse. When $\boldsymbol{Q}_r$ is positive definite, then $\boldsymbol{Q}_r^* = \boldsymbol{\Sigma}_r$ is the corresponding covariance matrix. If instead $\boldsymbol{Q}_r$ is rank-deficient (e.g. IGMRF effects), the covariance matrix does not formally exist.

The variance parameters σ_I^2 and $\sigma_{F,p}^2$ for $p = 1, ...P$ are fixed to large values to ensure flat priors on the fixed effects, while an actual prior is specified on σ_r^2 parameters of the random effects:

$$\sigma_I^2, \sigma_{F,1}^2, ..., \sigma_{F,P}^2 = 1000 \tag{3.1}$$

$$\sigma_1^2, ..., \sigma_R^2 \sim \pi(\sigma_1^2, ..., \sigma_R^2). \tag{3.2}$$

Moreover, the inferential focus is usually on the P fixed effects $\boldsymbol{\beta} = [\beta_1, ..., \beta_P]$ and the R variance parameters for each set of random effects $\boldsymbol{\sigma} = [\sigma_1^2, ..., \sigma_R^2]$, along with μ and $\boldsymbol{\psi}$. Traditionally, the elements of $\boldsymbol{\sigma}$ are assumed independent between each

other such that:

$$\pi(\boldsymbol{\beta}, \boldsymbol{\sigma}) = \left[\prod_{p=1}^{P} \pi(\beta_p)\right] \left[\prod_{r=1}^{R} \pi(\sigma_r^2)\right]. \tag{3.3}$$

The following section discusses an alternative approach for the prior specification of LGMs that challenges the set of assumptions summarized in Equation 3.3. The aim is to obtain an easier, more intuitive way to include prior beliefs into the model about the relative importance of the effects, intended here as their contributions to the linear predictor variability.

3.2.2 Variance partitioning priors

Consider the model presented in Equation 2.3 and following. The variance partitioning approach to prior specification consists of adopting a reparametrization of J different variance parameters of a model into a single variance V and a set of proportions $\boldsymbol{\omega}$.

Definition 3.1 (VP priors)**.** *Consider a set of variance parameters $\sigma_1^2, ..., \sigma_J^2$. We define the VP parameters as:*

$$\begin{aligned} V &= \sum_{j=1}^{J} \sigma_j^2, \\ \boldsymbol{\omega} &= \left[\omega_1 = \frac{\sigma_1^2}{V}, \ldots, \omega_{J-1} = \frac{\sigma_{J-1}^2}{V}, \omega_J = 1 - \sum_{j=1}^{J-1} \omega_j\right]. \end{aligned} \tag{3.4}$$

We call VP those priors implied on the original variance parameters by the specification of independent priors on the VP parameters, i.e.:

$$\pi(\sigma_1^2, \ldots, \sigma_J^2) = \pi(V)\pi(\boldsymbol{\omega})|\boldsymbol{J}|;$$

where $\boldsymbol{J}$ represents the Jacobian associated with the transformation $\boldsymbol{\sigma} \to V, \boldsymbol{\omega}$:

$$\boldsymbol{J} = \begin{bmatrix} \frac{\mathrm{d}V}{\mathrm{d}\sigma_1^2} & \ldots & \frac{\mathrm{d}V}{\mathrm{d}\sigma_J^2} \\ \ldots & \ldots & \ldots \\ \frac{\mathrm{d}\omega_{J-1}}{\mathrm{d}\sigma_1^2} & \ldots & \frac{\mathrm{d}\omega_{J-1}}{\mathrm{d}\sigma_J^2}. \end{bmatrix}$$

The VP reparametrization is the common denominator of all the works presented in Section 3.1, which have made use of this approach in different ways. While the

strategy to set the prior on the VP parameters varies with the goal and perspective adopted in each paper, the main benefit of this method is always an improvement in the user-friendliness of the prior specification procedure. This advantage is achieved through the definition of new parameters for which prior specification becomes more intuitive for the user, i.e. their beliefs can be more immediately translated into distributions for such parameters. This intuitive advantage is first achieved by simplifying the difficult problem of specifying a joint prior for a set of variance parameters through the separation between when the concept of magnitude of the overall variability (i.e. V) and that of relative importance of the different sources (i.e. $\boldsymbol{\omega}$) are divided and independently specified. However, this is usually only the first step and VP priors achieve their goal through further separate transformations of V and $\boldsymbol{\omega}$.

After obtaining the VP parameters, the prior for V can be chosen in order to include information about the scale of the variability in the response, for example when compared to the residual variability due to the likelihood. This is what is proposed by the R2D2 literature that considers the coefficient of determination R^2, defined as a function of V and $\boldsymbol{\psi}$, to include beliefs about the goodness-of-fit of the model. On the other hand, a prior for $\boldsymbol{\omega}$ can capture prior beliefs about the relative importance of each effect with respect to the others. In order to obtain the desirable comparisons between the effects, Fuglstad et al. 2020 suggests a further reparametrization of $\boldsymbol{\omega}$ based on the design of a case-specific decomposition tree to partition the variance contributions in branches. The design can often be guided not only by actual prior knowledge but also by the structure of the model and Occam's razor principle (e.g. splitting main and interaction effects and preferring the former). According to the chosen tree, a set of simplices is derived from $\boldsymbol{\omega}$ and independently specified. The shift from a variance scale to the simplex one greatly simplifies the task of specifying both marginal and joint priors since the parameters are now immediately interpretable for a user as proportional contributions to the variance in a simple 0-1 scale. See Fuglstad et al. 2020 for details about specific prior choices for different prior assumptions.

In summary, the appeal of VP prior methods lies in the possibility of obtaining parameters that are more easily interpretable for the user when compared to the original ones. However, it is crucial to understand that this interpetability is not guaranteed: V and $\boldsymbol{\omega}$ do not always equate to their *intuitive interpretation*, i.e. respectively as the total variance due to the J effects and the set of proportional contributions of the individual effects to this total variance. This holds only if each σ_j^2 actually matches its own *intuitive interpretation*, being the variance contribution

of its corresponding effect as intended by the user. However, this is not true in general.

The necessity for the interpretability advantage of VP priors has led so far the literature to define the concept of intuitive interpretation for σ_r^2 in different ways and restrict the field of applications only to those specific effects for which this intuitive interpretation requirement is respected, either by design or after adjustments. Here, we review the two main lines of research exploiting the VP reparametrization in general, specifically focusing on how they dealt with the interpretability issue (either implicitly or directly).

In the following, we shall refer for convenience to $Var_{\boldsymbol{u}}[f(Z)|\sigma^2, Z = z]$ as the *conditional variance* of a generic effect $f(Z)$, while to $Var_{\boldsymbol{u},Z}[f(Z)|\sigma^2]$ as the *marginal variance* of the effect.

Hierarchical Decomposition approach

Fuglstad et al. 2020 presented the *hierarchical decomposition* (HD) framework considering the model class of Equation 2.3 and applying the VP reparametrization to the variance parameters of the random effects $\sigma_1^2, ..., \sigma_R^2$. Dropping the indices for convenience, the intuitive interpretation for each σ^2 has been defined in this context as the conditional variance of effect $f(Z) = \boldsymbol{D}^T(Z)\boldsymbol{u}$, given the value of the covariate. Hence, the intuitive interpretation requirement can be defined as:

$$Var_{\boldsymbol{u}}[f(Z)|\sigma^2, Z = z] = \sigma^2 \quad \forall z \in \mathcal{Z}.$$

The requirement does not hold for all possible random effects, as the conditional variance for a generic effect is:

$$Var_{\boldsymbol{u}}[f(Z)|\sigma^2, Z = z] = \sigma^2 \cdot [\boldsymbol{D}^T(z)\boldsymbol{Q}^*\boldsymbol{D}(z)]. \tag{3.5}$$

Defining $g(z) = \boldsymbol{D}^T(z)\boldsymbol{Q}^*\boldsymbol{D}(z)$, it is clear that $g(z)$ is not always a constant function at 1: therefore, not all effects are going to respect the intuitive interpretation requirement. According to whether or not this is true, effects have been labelled as *homogeneous* or *heterogeneous* (lexicon used in Fuglstad et al. 2020).

- **Homogeneous effect**: $g(z) = 1, \ \forall z \in \mathcal{Z}$.
- **Heterogeneous effect**: $\exists z \in \mathcal{Z}$ such that $g(z) \neq 1$.

All stationary processes have a constant marginal variance by definition so that they are always homogeneous as long as $\boldsymbol{Q}^*$ is scaled to be a correlation matrix.

Examples of popular effects that fall in this category include i.i.d. group effects, Matern-based spatial effects, and stationary autoregressive models for temporal correlation.

In the presence of heterogeneous effects instead, V and $\boldsymbol{\omega}$ do not have their intuitive interpretation, as $Var(\eta|\mu, \beta_1, ..., \beta_P, \sigma_1^2, ..., \sigma_R^2, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}) \neq \sum_{j=1}^{R} \sigma_j^2$. This is a relevant problem as many common models fall in the heterogeneous category, including polynomial effects and non-stationary spatial and temporal models, e.g. random walks, ICAR model, etc.

Fuglstad et al. 2020 discussed the inclusion of a particularly important class of heterogeneous effects, i.e. IGMRFs (see Section 3.2.3), through the proposal of Sørbye and Rue 2014. First, a *reference variance* σ^2_{ref} is defined for each effect as being a location summary of the conditional variance over the support $\mathcal{Z}$ of Z having cardinality N_Z. In particular, Sørbye and Rue 2014 proposed the use of the geometric mean such that:

$$\sigma^2_{\text{ref}} = \exp\left(\frac{1}{N_Z}\sum_{z\in\mathcal{Z}} \log\{Var_{\boldsymbol{u}}[f(Z)|\sigma^2 = 1, Z = z]\}\right).$$

Then, the effect can be scaled according to this reference quantity such that σ^2 becomes, in fact, equal to the geometric mean of the marginal variance; this is achieved in practice either by dividing the covariance matrix by σ^2_{ref} (i.e. multiplying the precision matrix by σ^2_{ref}) or the basis matrix by σ_{ref}. This solution provides that the intuitive interpretation requirement does not hold exactly but it does approximately:

$$Var_{\boldsymbol{u}}[f(Z)|\sigma^2, Z = z] \approx \sigma^2 \quad \forall z \in \mathcal{Z}. \tag{3.6}$$

In terms of application, the HD approach has only considered cases in which the covariates Z_r are categorical or discrete, while the case of a continuous covariate is not discussed. The extension towards fixed effects is deemed as a desirable objective and the authors suggest considering the concept of *explained variance* for the derivation of a quantity comparable to the σ_r^2 parameters of the random effects.

R2D2 approach

The founding work in the R2D2 literature was developed as a new proposal in the field of global-local shrinkage priors and, thus, it focuses specifically on linear effects Zhang et al. 2022. Further works have extended the scope of application to the inclusion of i.i.d. group effects (Aguilar and Bürkner 2023), general i.i.d. random effects Yanchenko, Bondell, and Reich 2024b, and more recently to spatially

dependent data modelled using a Matern covariance function (Yanchenko, Bondell, and Reich 2024a). We shall consider here the work of Yanchenko, Bondell, and Reich 2024b as the main reference for the comparison to the HD approach of Fuglstad et al. 2020.

In line with the global-local shrinkage prior literature, the traditional assumption of a fixed $\sigma^2_{F,p}$ for all fixed effects (Equation 3.1) is discarded and the model parameters become $\boldsymbol{\sigma} = \sigma^2_{F,1}, ..., \sigma^2_{F,P}, \sigma^2_1, ..., \sigma^2_R$, i.e. a prior is also specified for $\sigma^2_{F,1}, ..., \sigma^2_{F,P}$. At this point, the R2D2 prior is a method based on a VP reparametrization on all the $P + R$ variance parameters. An alternative way to view this approach consists of considering a model with no fixed effects ($P = 0$) and viewing the linear effects as specific instances of the R random effects, where the basis matrix simplifies to a single (linear) function of the covariate.

For this subclass of models, the R2D2 literature adopts a different perspective than the one of Fuglstad et al. 2020 in the application of the VP reparametrization: $Z_1, .., Z_R$ are marginalized out such that the variance of each effect is no longer a function of the covariates' values. From this perspective, we can assume that the intuitive interpretation criterion for each σ^2 has been implicitly defined as the conditional variance, i.e.:

$$Var_{Z,\boldsymbol{u}}[f(Z)|\sigma^2] = \sigma^2. \tag{3.7}$$

However, again, this requirement does not hold for general random effects from the LGM class defined in Section 3.2.1, as the marginal variance for a generic effect can be found using the law of total variance:

$$\begin{aligned} Var_{Z,\boldsymbol{u}}[f(Z)|\sigma^2] &= E_Z\{Var_{\boldsymbol{u}}[f(Z)|\sigma^2, Z]\} + Var_Z\{E_{\boldsymbol{u}}[f(Z)|\sigma^2, Z]\} \\ &= E_Z\{Var_{\boldsymbol{u}}[f(Z)|\sigma^2, Z]\} \\ &= E_Z[\sigma^2 \boldsymbol{D}^T(Z)\boldsymbol{Q}^*\boldsymbol{D}(Z)] \\ &= \sigma^2 \cdot E_Z[\boldsymbol{D}^T(Z)\boldsymbol{Q}^*\boldsymbol{D}(Z)] \end{aligned} . \tag{3.8}$$

Based on whether or not the desired intuitive interpretation requirement is respected, the distinction between homogeneous and heterogeneous effects can be redefined here as follows:

- **Homogeneous effect**: $E_Z[\boldsymbol{D}^T(Z)\boldsymbol{Q}^*\boldsymbol{D}(Z)] = 1$.
- **Heterogeneous effect**: $E_Z[\boldsymbol{D}^T(Z)\boldsymbol{Q}^*\boldsymbol{D}(Z)] \neq 1$.

The problem of heterogeneous effects has not been discussed so far in the R2D2 literature, as the proposal of Yanchenko, Bondell, and Reich 2024b limits its scope of

application to effects with a certain structure in the basis and covariance matrices. In particular, $\boldsymbol{Q}_r^* = \boldsymbol{I}$ and $\boldsymbol{D}_r(Z_r) = [\mathbb{I}(Z_r = 1), ..., \mathbb{I}(Z_r = K_r)]^T$ for $r = 1, ..., R$. It is easy to prove that all effects are homogeneous under these restrictions, regardless of the distribution of the covariate (i.e. $E_{Z_r}[\boldsymbol{D}_r^T(Z_r)\boldsymbol{Q}_r^*\boldsymbol{D}_r(Z_r)] = 1,\ r = 1, ..., R$). Moreover, the problem of including IGMRF effects in the model has also not yet been addressed in this context. Finally, the historical distinction between fixed and random effects is ignored in the definition of the intuitive interpretation for the variance parameters, as effects are treated indiscriminately.

3.2.3 Intrinsic Gaussian Markov Random Fields (IGMRFs)

IGMRFs are a peculiar category of effects, which are very popular in modelling spatial (i.e. ICAR model), temporal (i.e. random walks), and non-linear effects of continuous covariates (i.e. P-Splines). Consider an IGMRF effect $\boldsymbol{u}|\sigma^2 \sim N_K(\boldsymbol{0}, \sigma^2\boldsymbol{Q}^*)$, where the precision matrix $\boldsymbol{Q}$ has rank-deficiency $d > 0$ and therefore a non-empty null space $\boldsymbol{S}$ with d columns $\boldsymbol{S}$, i.e. $\boldsymbol{QS} = \boldsymbol{0}$. IGMRFs are a particular type of improper GMRFs where the null space has a specific form. For example, IGMRFs of order d on the line are defined by a null space matrix $\boldsymbol{S}_{(d-1)}$:

$$\boldsymbol{S}_{(d-1)}(\boldsymbol{k}) = \begin{bmatrix} \boldsymbol{k}^0 & \boldsymbol{k}^1 & ... & \boldsymbol{k}^{(d-1)} \end{bmatrix} \tag{3.9}$$

where $\boldsymbol{k}$ is a column vector of locations on the line. Note that $\boldsymbol{S}_{(d-1)}$ corresponds to a Vandermonde matrix (Hoffman and Kunze 1971) of degree $d-1$ on locations $\boldsymbol{k}$.

Being improper Gaussian models, the parameters μ and $\boldsymbol{Q}$ of an IGMRF no longer represent the mean and precision as they do not formally exist. Moreover, σ^2 no longer controls the deviation from the mean of $\boldsymbol{u}$, and thus, it loses its intuitive interpretation necessary for a sensible prior specification. To understand the true meaning of σ^2, we make use of the decomposition of an IGMRF discussed in Rue and Held 2005, Section 3.4.1:

$$\boldsymbol{u} = \boldsymbol{H}_{(d-1)}\boldsymbol{u} + (\boldsymbol{I} - \boldsymbol{H}_{(d-1)})\boldsymbol{u}. \tag{3.10}$$

Equation 3.10 shows how $\boldsymbol{u}$ can be decomposed into a polynomial trend of degree $d-1$ and a residual term, using the *hat matrix* $\boldsymbol{H}_{(d-1)} = \boldsymbol{S}_{(d-1)}[\boldsymbol{S}_{(d-1)}^T\boldsymbol{S}_{(d-1)}]^{-1}\boldsymbol{S}_{(d-1)}^T$, which can project a generic $\boldsymbol{u}$ to the corresponding polynomial trend. Noting that $\boldsymbol{QH}_{(d-1)} = \boldsymbol{0}$, the probability density of $\boldsymbol{u}$ can be rewritten as:

$$\pi(\boldsymbol{u}) \propto \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{u} - \boldsymbol{H}_{(d-1)}\boldsymbol{u})^T\boldsymbol{Q}(\boldsymbol{u} - \boldsymbol{H}_{(d-1)}\boldsymbol{u})\right]. \tag{3.11}$$

Equation 3.11 clearly shows how σ^2 does not measure the dispersion around the mean of the process but rather the dispersion around the polynomial trend of order $d-1$ in $\boldsymbol{u}$, such that $\sigma^2 \to 0$ only shrinks the model to the polynomial trend rather than towards $\boldsymbol{u} = \mathbf{0}$. As a consequence, the σ^2 of a generic effect $\boldsymbol{D}^T(Z)\boldsymbol{u}$ cannot be interpreted in relation to the variability in η caused by covariate Z.

The only solution for an interpretation of the σ^2 parameter is introducing linear constraints on the coefficients in the form $\boldsymbol{S}^T_{(d-1)}\boldsymbol{u} = \mathbf{0}$, which can be called *null space constraints*. Under these constraints, the vector $\boldsymbol{u}$ is now a proper GMRF with a well-defined mean and covariance matrix:

$$E[\boldsymbol{u}|\sigma^2] = \mathbf{0} \text{ subject to } \boldsymbol{S}^T_{(d-1)}\boldsymbol{u} = \mathbf{0}$$
$$Cov[\boldsymbol{u}|\sigma^2] = \sigma^2\boldsymbol{Q}^* \text{ subject to } \boldsymbol{S}^T_{(d-1)}\boldsymbol{u} = \mathbf{0}.$$

3.3 Standardization procedure

We present a unifying framework that builds upon and fills the gap in the current joint scope of application of the HD and the R2D2 lines of research. The goal is achieved extending the scope of application of VP priors to the general class of LGMs and to both their fixed and random branches. To so so, we obtain a procedure that guarantees that the VP parameters from Equation 3.4 match the intuitive interpretation the user has about them (i.e. total variance and proportional contributions to this variance), regardless of the type of effects present in a model, including the popular IGMRFs. The importance of this goal lies in the possibility of extending the benefits of VP prior specification to a broader class of LGMs, hence to more applications and fields.

To fully exploit the advantages of VP priors, we start by considering as random the variance parameters σ^2_j of all effects in the model, including for those effects traditionally treated as fixed. We shall call this class of models as *fully random LGM*. This choice is inspired by R2D2 priors, and more generally global-local shrinkage priors, which use priors on the variance parameters of linear effects to easily introduce a specific type of prior information, namely sparsity. We generalize this principle with the fully random LGM as this specification allows to reflect any type of prior assumption about the contributions of the effects through the use of VP priors.

We then adopt the approach of recent works that have highlighted how the distinction between fixed and random effects should possibly be reframed in terms of user's inferential interest. Under this perspective, we are able to restore the

fixed/random categorization for the effects of fully random LGMs.

The most crucial point of the proposal consists in a formal definition of *intuitive interpretation* of the σ_j^2 parameters as variance contributions of the corresponding effects. The final definition combines aspects from both the HD and the R2D2 approaches and addresses the concern of Fuglstad et al. 2020 about the need for a different treatment of fixed effects.

We then present solutions to ensure that the σ_j^2 parameters match their intuitive interpretation for both fixed and random effects. Namely, we introduce a scaling procedure and a 0-mean constraint necessary to this purpose.

The issue of sensibly introducing the particular class of IGMRFs is separately discussed in Section 3.3.6. The main problem is that the variance parameter of an IGMRF does not control fully the variance of the effect, but only partially. As a solution, we propose to re-express IGMRFs through two separate effects, a polynomial and a residual one, so that the 2 corresponding variance parameters are able together to represent the overall variability of the process.

Once the intuitive interpretation conditions are satisfied by all effects, we are finally able to derive exact expressions for the interpretation of the VP parameters that we believe best reflect the intuition the user has about them. This result ensures that VP priors correctly reflect prior beliefs about the total variance in the linear predictor and the relative contributions of each effect to this variance.

The next section will present how the procedure can be implemented in practice to some of the most popular effects used in LGMs: random intercept effects, effects, linear effects, random slopes effects, discrete IGMRFs, P-Splines.

3.3.1 Fully random Latent Gaussian Models

We define a new class of Latent Gaussian Models in which the P covariates $X_1, .., X_P$ and the R covariates $Z_1, ..., Z_R$ from Section 3.2.1 are now collected in a single vector $\boldsymbol{X} = [X_1, ..., X_J]$ where $J = P + R$ and all their corresponding effects are assigned a random variance parameter.

Model 1 (Fully random Latent Gaussian model)**.**
Let $\boldsymbol{X} = [X_1, ..., X_J]$ be a set of covariates with $\boldsymbol{X} \sim \pi(\boldsymbol{x})$, and possible realizations $x \in \mathfrak{X}_j,\ \forall j = 1, ..., J$. Let a response $Y \sim Dist(\eta, \boldsymbol{\psi})$. The parameter η is defined as an additive model including an overall intercept μ and J random effects $f_j(X_j)$. Each effect $f_j(X_j)$ is defined through a known basis matrix $\boldsymbol{D}_j(X_j)$ with K_j basis functions and a set of Normally distributed coefficients $\boldsymbol{u}_j$, with null mean, known

precision matrix $\boldsymbol{Q}_j$*, and scalar parameter* σ_j^2*:*

$$\begin{aligned}\eta &= \mu + \sum_{j=1}^{J} f_j(X_j) \\ &= \mu + \sum_{j=1}^{J} \boldsymbol{D}_j^T(X_j)\boldsymbol{u}_j \\ \boldsymbol{u}_j|\sigma_j^2 &\sim N_{K_j}(\boldsymbol{0}, \sigma_j^2 \boldsymbol{Q}_j^*) \qquad\qquad j = 1, ..., J.\end{aligned}$$

If any $\boldsymbol{Q}_j$ *has a non-empty null space* $\boldsymbol{S}_j$*, then the null space constraints* $\boldsymbol{S}_j^T\boldsymbol{u} = \boldsymbol{0}$ *are imposed.*

The parameters of the model that need prior specification are $\boldsymbol{\sigma} = [\sigma_1^2, ...\sigma_J^2]$*, along with* μ *and* $\boldsymbol{\psi}$*.*

Note that the definition of Model 1 does not impose assumptions about $\boldsymbol{X}$, meaning that there can be more than one effect per covariate, e.g. a linear and a non-linear one. Care must however be taken in the design of the effects to avoid identifiability issues.

In Model 1, all the variance parameters $\sigma_1^2, ..., \sigma_J^2$ are treated as unknown quantities for which a prior must be specified. All effects in the model can therefore be considered random according to the criterion made explicit in Equations 3.1-3.2. In practice, the shift to a random variance for fixed effects changes the implied prior on the coefficients, which is no longer almost flat. This choice is necessary to extend the advantages of VP priors to all effects of the model, even to those traditionally treated as fixed. Under the VP framework, the novel prior induced on the fixed effects coefficients will not be flat and not independent from the rest of the model parameters. However, if it is desirable to preserve a flat prior on some fixed effects or it is not desirable to include some of them in the VP reparametrization, these effects can simply be viewed as part of the overall intercept μ from Model 1. In this context, we simply attempt to solve the problem of the inclusion of fixed effects in the VP framework whenever this is desirable, which may not always be the case for all fixed effects in a model.

3.3.2 Redefinition of fixed and random effects

We propose a redefinition of the concepts of fixed and random effects based on a more modern perspective. Different attempts have been made in the literature to identify an intrinsic binary categorization between effects that would go beyond prior specification. Particularly, we refer to the comment of by Gelman et al. 2013

in Section 15.6:

"*We believe that much of the statistical literature on fixed and random effects can be fruitfully re-expressed in terms of finite-population and super-population inferences [...] The difference between fixed and random effects is thus not a difference in inference or computation but in the ways that these inferences will be used.*"

On a similar note, Hodges 2013 introduced a categorization of (traditional) random effects into new-style/old-style ones:

- **old-style random effects**: *"The levels of a random effect [...] are draws from a population, and the draws are not of interest in themselves but only as samples from the larger population, which is of interest."*
- **new-style random effects**: *"[...] a random effect may have levels that are not draws from any population, or that are the entire population, or that may be a sample but a new draw from the random effect could not conceivably be drawn, and in all these cases the levels themselves are of interest".*

According to the finite-population/superpopulation inference distinction of Gelman et al. 2013, new-style effects could be considered fixed and old-style ones could be considered random, regardless of the fact that they would be traditionally classified as random.

Here, we propose a concise classification criterion between for the effects in Model 1 based on the concept of *parameter of interest*, intended as the quantity, between $\boldsymbol{u}_j$ and σ_j^2, that is the inferential target for a given effect.

Definition 3.2 (Fixed and random effects)**.**
Consider Model 1. Assume that the effects of Model 1 has been ordered such that the vector $\boldsymbol{\theta} = [\boldsymbol{u}_1, ..., \boldsymbol{u}_L, \sigma^2_{L+1}, ..., \sigma^2_J]$ *contains the "parameters of interest" for the user. The* L *effects for which the coefficients* $u_j \in \boldsymbol{\theta}$ *are called "fixed", while the* $J - L$ *effects for which the variance parameter* $\sigma_j^2 \in \boldsymbol{\theta}$ *are called "random".*

Definition 3.2 leaves the user free to classify each effect in an application-specific manner, according to their subjective assumptions and specific research questions. Nevertheless, general guidelines can be outlined to clarify which effects will typically fall in these categories. For example, traditional fixed effects, e.g. linear effects, polynomial effects of higher order) will still be treated as fixed in the vast majority of applications. On the other hand, cluster effects will usually fall in the random category, specifically when not all groups/levels present in the population have been sampled; however, cluster effects with a small number of distinct levels, all present in the sample, are more likely to be considered fixed (e.g. gender, ethnicity in

epidemiological studies). Spatio-temporal effects are often introduced to model the correlation in the residuals of a model and as such might be categorized as random; however, there are specific applications in which the actual trends are of interest (i.e. fixed), such as for example in disease mapping (Lawson 2018, Moraga 2019). There is more ambiguity about complex non-linear effects of continuous covariates modelled through splines, for which we suggest to make case-specific considerations. A thorough discussion on this categorization can be found in Hodges 2013 using the new-style/old-style taxonomy.

Given this novel distinction between fixed and random effects, adding a random variance parameter layer for the fixed effects to the purpose of using VP priors can be seen simply as an artificial step to conveniently induce a joint prior on $\boldsymbol{\theta}$ by specifying a joint prior on $\boldsymbol{\sigma}$.

3.3.3 On the intuitive interpretation of the variance parameters

The novel VP parameters V and $\boldsymbol{\omega}$ from Equation 3.4 can be interpreted as respectively the total variance in the linear predictor and the proportional contributions to this variance only if each σ_j^2 actually matches its own *intuitive interpretation*, being defined as the variance contribution of its corresponding effect as intended by the user. As reviewed in Section 3.2.2, the literature has proposed multiple definitions of variance contribution. In what follows, we shall discuss and consider the most sensible definition of this quantity.

On the one hand, the variance contribution of an effect could be simply defined as the variance of the effect conditional on the model parameters $\boldsymbol{\sigma}$. This would lead to the same requirement used in the R2D2 literature (Equation 3.7). However, we have also argued how the vector of model parameters $\boldsymbol{\sigma}$ is only artificially introduced a priori for the purpose of obtaining an easier prior specification procedure, through the use of VP priors. The quantities that are actually of interest for the user are contained in the vector of parameters of interest $\boldsymbol{\theta}$. In fact, we argue that the user has a different definition of variance contribution of an effect according to its classification as a fixed or random one, in other words according to whether they are interested in the finite-population inference or the super-population one. To see this, we can consider how variance contributions are estimated a posteriori.

For example, consider a linear effect $f_1(X_1) = X_1 \cdot u_1$ for covariate X_1. A linear effect is usually considered fixed, as the interest lies on the linear coefficient itself. The variance contribution of such an effect is usually estimated conditioning on the linear coefficient u_1, i.e. using $Var_{X_1}[X_1 \cdot u_1 | u_1]$. This estimate for the variance

contribution is used for example in the Bayesian R^2 metric proposed by (Gelman et al. 2019). On the other hand, consider an i.i.d. effect for a categorical covariate X_2 with K levels, i.e. $f_2(X_2) = \sum_{k=1}^{K} \mathbb{I}[X_2 = k] \cdot u_{2,k}, u_{2,k} \overset{iid}{\sim} N(0, \sigma_2^2)$. The estimates for the single levels $u_{2,k}$ are in the majority of applications not of interest (e.g. school effect on students' performance, batch effect in production measurements) so the effect can be categorized as random. The amount of variance imputable to X_2 is often estimated using $Var_{X_2, u_{2,1}, \dots, u_{2,K}}[f_2(X_2)|\sigma_2^2]$, which is simply equal to variance parameter σ_2^2. We argue that this different quantification of variance contributions between fixed and random effects holds in general. Specifically, we can use the terminology of Gelman et al. 2013 to state that variance contributions are estimated using the:

- *finite-population variance* for fixed effects (also called explained variance):

$$Var_{X_j}[f_j(X_j)|\boldsymbol{u}_j];$$

- *super-population variance* for random effects:

$$Var_{X_j, \boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2].$$

Note how the use of the marginal version of the variance with respect to the covariate X_j (i.e. the R2D2 approach) is a requirement for the definition of finite-population variances, as the conditional version would always be null. This result shows how the variability observed in the response for such effects is to be attributed to the to the spread of the corresponding covariate, since the trend is conditioned upon or "fixed".

The distinction in the definition of variance contribution for fixed and random effects can actually be summarized using the concept of *variance of interest*, i.e. the variance conditional on the parameters of interest:

$$Var_{X_j, \boldsymbol{u}_j}[f_j(X_j)|\boldsymbol{\theta}].$$

The concept of variance of interest allows to summarize in a single, neat expression how variance contributions are intended by the user on the basis of its inferential interest. Hence, we believe that the variance of interest represents the most sensible definition of intuitive interpretation for the σ_j^2 parameters. However, a problem arises in the case of fixed effects, since the finite-population variance is not a function of σ_j^2. Hence, we propose to use instead the *expected variance of interest*, defined as

the expectation of the variance of interest conditional on the model parameters $\boldsymbol{\sigma}$:

$$E_{\boldsymbol{\theta}}\{Var_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\boldsymbol{\theta}]|\boldsymbol{\sigma}\}. \tag{3.12}$$

Deriving the expected variance of interest separately for fixed and random effects, we can finally formally define the intuitive interpretation of the σ_j^2 parameters for all effects of Model 1.

Definition 3.3 (Intuitive interpretation of σ_j^2 parameters)**.**
Consider Model 1 with $\boldsymbol{\theta} = [\boldsymbol{u}_1, ..., \boldsymbol{u}_L, \sigma_{L+1}^2, ..., \sigma_J^2]$. *We say that* $\sigma_1^2, ..., \sigma_J^2$ *match their intuitive interpretation if:*

- *for fixed effects* $(j = 1...., L)$

$$\sigma_j^2 = E_{\boldsymbol{u}_j}\{Var_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\}; \tag{3.13}$$

- *for random effects* $(j = L + 1...., J)$

$$\sigma_j^2 = Var_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2]. \tag{3.14}$$

Definition 3.3 uses the super-population variance to describe the intuitive interpretation of the σ^2 parameters of random effects, while it uses the expected finite-population variance $E_{\boldsymbol{u}_j}[s_j^2|\sigma_j^2]$ (from now on, simply denoted as $E[s_j^2]$) for fixed ones. Section A.1 of the Appendix proves that Equations 3.13-3.14 are the two special cases of the expected variance of interest (Equation 3.12).

This proposal answers the concern raised in Fuglstad et al. 2020 about the introduction of fixed effects and is consistent with the authors' suggestion of considering the concept of *explained variance*, since the condition of Equation 3.13 requires that σ_j^2 must be equal to the expectation of the finite-population variance, i.e. the expectation of the explained variance. On the other hand, the definition for random effects is coherent with the R2D2 approach (Equation 3.7). In summary, the new definition can be seen as a generalization of previous approaches in which the inferential interest of the effects is taken into account.

If the conditions of Definition 3.3 are satisfied, we claim that the prior $\pi(\sigma_1^2, ..., \sigma_J^2)$ correctly reflects the assumptions of the user about the variance contributions of the effects. We can now discuss how all the effects of Model 1 can be adjusted so that Definition 3.3 is respected.

3.3.4 Scaling procedure

We propose a simple scaling procedure that ensures that the condition of Definition 3.3 on random effects is satisfied.

Proposition 1 (Scaling procedure)**.**
Consider Model 1. Let C_j be defined as the variance of the process $f_j(X_j)$ given $\sigma_j^2 = 1$:

$$\begin{aligned} C_j &= Var_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2 = 1] \\ &= E_{X_j}\left\{Var_{\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2 = 1, X_j]\right\} \\ &= E_{X_j}[\boldsymbol{D}_j^T(X_j)\boldsymbol{Q}_j^*\boldsymbol{D}_j(X_j)]. \end{aligned} \tag{3.15}$$

If a new effect is defined as $\widetilde{f}_j(X_j) = f_j(X_j)/\sqrt{C}_j$, then:

$$Var_{X_j,\boldsymbol{u}_j}[\widetilde{f}_j(X_j)|\boldsymbol{\sigma}] = \sigma_j^2$$

Proof. Using the derivation from Equation 3.8, it can be proved that:

$$\begin{aligned} Var_{X_j,\boldsymbol{u}_j}\left[\widetilde{f}_j(X_j)|\sigma_j^2\right] &= \frac{\sigma_j^2 E_{X_j}[\boldsymbol{D}_j^T(X_j)\boldsymbol{Q}_j^*\boldsymbol{D}_j(X_j)]}{C_j} \\ &= \frac{C_j}{C_j}\sigma_j^2 = \sigma_j^2. \end{aligned}$$

□

Proposition 1 proves that if each effect is scaled by the square root of the corresponding *scaling constant* C_j, then their super-population variance becomes equal to σ_j^2. As such, the scaling procedure is a necessary and sufficient step to ensure that the σ_j^2 parameters of random effects match their intuitive interpretation (Equation 3.14).

The scaling procedure can be implemented in practice either by dividing the basis matrix by $\sqrt{C_j}$ or by multiplying the precision matrix by C_j. In the computation of the scaling constants C_j, it is important to consider some technical aspects. First, all linear constraints imposed on the process must be considered before the application of the scaling procedure, so that the correct covariance matrix is used in the computation of the scaling constant C_j. Secondly, note that there is no guarantee that C_j will be non-null and finite for all potential models, as it is a function of $\pi(x_j), \boldsymbol{D}_j^T(X_j), \boldsymbol{Q}_j$: hence, the couples $f_j(X_j)$ and $\pi_j(x)$ should always be formed to ensure that $0 < C_j < \infty$, $\forall j$. Finally, it might be easier in practice to approximate the values of C_j using a Monte Carlo simulation, sampling N values $x_1, ..., x_N$ from

$X_j \sim \pi_j(x)$:

$$\widehat{C}_j = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{D}_j^T(x_i) \boldsymbol{Q}_j^* \boldsymbol{D}_j(x_i).$$

The scaling procedure from Proposition 1 has been derived to satisfy Equation 3.14. However, it is also coherent with the different approach taken by Fuglstad et al. 2020 and Sørbye and Rue 2014, which suggested the use of a location summary (i.e. the geometric mean) of the conditional variance as a scaling constant for the effects. With the notation of Model 1, we can redefine the reference variance as:

$$\sigma^2_{\text{ref}} = \text{GM}_{X_j} \left\{ Var_{\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2 = 1, X_j] \right\} \tag{3.16}$$

where $GM_X(\cdot) = \exp\{E_X[\log(\cdot)]\}$. Comparing Equation 3.15 and 3.16, Proposition 1 can be viewed as a variant of the proposal of Sørbye and Rue 2014, in which the geometric mean is replaced by the arithmetic mean. This result emphasizes how the marginal variance approach adopted in the R2D2 priors' literature automatically proposes a solution to the problem of choosing a location summary of the conditional variance, as envisioned by Sørbye and Rue 2014. In conclusion, note how the expectation-based scaling of Proposition 1 has advantages over the geometric mean approach (see also Section 3.3.7). For example, the geometric mean approach would return a null scaling constant for linear effects while the newly proposed scaling would not. This result holds more generally for all effects for which $\exists x \in \mathcal{X}$ such that $\pi(x) > 0$ and $Var_{\boldsymbol{u}}[f(x)|\sigma^2, X = x] = 0$.

3.3.5 0-mean constraint for fixed effects

To understand how the intuitive interpretation requirement (Definition 3.3) can also be satisfied for fixed effects, we rewrite Equation 3.13 as follows:

$$E_{\boldsymbol{u}_j}\{Var_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\} = Var_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2] - Var_{\boldsymbol{u}_j}\{E_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\}. \tag{3.17}$$

See the proof in Section A.2 of the Appendix.

Equation 3.17 shows how the intuitive interpretation for fixed effects (Equation 3.13) is equal to the one of random ones (Equation 3.14) minus a certain quantity. The last term of Equation 3.17 represents the variance with respect to all realizations of $\boldsymbol{u}$ of the process mean over the X_j support. This result shows how Definition 3.3 requires σ_j^2 to only measures deviation of the process from its mean and ignore the

additional variability in the process due to the uncertainty around its true mean. As such, applying Proposition 1 is insufficient for fixed effects. In order to remove the uncertainty around the mean, we can assume a fixed value of the mean, such as 0 for convenience. This assumption would remove the last term of Equation 3.17 and ensures that the intuitive interpretation for fixed effects becomes equal to the one for random ones. It is therefore clear that, whenever the mean of the process is fixed at 0, the intuitive interpretation requirement from Definition 3.3 will always hold regardless of the type of effect, as long as the effect has been correctly scaled as described in Proposition 1.

Proposition 2 (0-mean constraint)**.**
Consider Model 1 with $\boldsymbol{\theta} = [\boldsymbol{u}_1, ..., \boldsymbol{u}_L, \sigma^2_{L+1}, ..., \sigma^2_J]$*. Under a 0-mean constraint on all fixed effects, i.e.*

$$E_{X_j}[f_j(X_j)|\boldsymbol{u}_j] = 0 \quad j = 1, .., L,$$

we obtain that

$$E_{\boldsymbol{u}_j}\{Var_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma^2_j\} = Var_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\sigma^2_j] \quad j = 1, .., L.$$

Proof. If $E_{X_j}[f_j(X_j)|\boldsymbol{u}_j] = 0$, then $E_{\boldsymbol{u}_j}\{Var_{X_j}[f(X)|\boldsymbol{u}_j]|\sigma^2\} = Var_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\sigma^2_j]$ follows immediately from Equation 3.17. □

Thanks to Proposition 2, we can conclude that the σ^2_j parameter of a fixed effect matches its intuitive interpretation (Equation 3.13) if:

1. the effect respects the 0-mean constraint $E_{X_j}[f_j(X_j)|\boldsymbol{u}_j] = 0$ for every realization of the $\boldsymbol{u}_j$;
2. the effect has been scaled such that $Var_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\boldsymbol{\sigma}] = \sigma^2_j$ (Proposition 1)

In order to investigate when the 0-mean constraint holds, we further derive Equation 3.17 (proof in Section A.3) dropping the index j for convenience:

$$E[s^2] = \sigma^2 \cdot [1 - \text{tr}(\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Q}^*)] \tag{3.18}$$

$$\boldsymbol{a} = \begin{bmatrix} E_X[D_1(X)] \\ ... \\ E_X[D_K(X)]. \end{bmatrix} \tag{3.19}$$

Whenever $\text{tr}(\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Q}^*) = 0$, the 0-mean constraint is respected since $E[s^2] = \sigma^2$. On the one hand, a process can be designed such that this requirement is always met

(e.g. $\boldsymbol{a} = \mathbf{0}$). For instance, if the process has a single basis function $D(X)$, it is sufficient to subtract its expectation and redefine the basis as $D(X) - E_X[D(X)]$ (e.g. linear effect). Alternatively, a general process with more than one basis function can be constrained to respect the 0-mean constraint using linear constraints $\boldsymbol{a}^T\boldsymbol{u} = 0$, since $\boldsymbol{a}^T\boldsymbol{u} = E_X[f(X)|\boldsymbol{u}]$. The conditional distribution of a general Gaussian vector under linear constraints can be found according to the formulae in Rue and Held 2005 (Section 2.3.3). After conditioning, $\boldsymbol{a}$ will be in the null space of the new precision matrix, such that the argument of the trace operator in Equation 3.18 will be a null matrix.

Note that the scaling step should always be done after the imposition of the 0-mean constraint, as the correct scaling constant might be different for the new constrained process.

3.3.6 IGMRF effects

We finally discuss the case of IGMRF effects. In this section, we propose a procedure for the sensible inclusion of these popular effects, such that the intuitive interpretation of the variance parameters is guaranteed. The indices j are dropped for convenience.

Consider an effect $f(X)$ for unidimensional covariate X such that $f(X) = \boldsymbol{D}^T(X)\boldsymbol{u}$ where $\boldsymbol{u}$ of dimension K is an IGMRF of order d on regular locations $\boldsymbol{k} = [1, 2, ..., K-1, K]$:

$$\boldsymbol{u}|\sigma^2 \sim N(\mathbf{0}, \sigma^2\boldsymbol{Q}^*)$$
$$\boldsymbol{Q}\boldsymbol{S}_{(d-1)} = \mathbf{0}.$$

As pointed out in Section 3.2.3, IGMRFs cannot be straightforwardly introduced in the VP approach, as their corresponding variance parameters do not measure deviations of the process from its mean but rather only from its polynomial trend $\boldsymbol{H}_{(d-1)}\boldsymbol{u}$. In terms of $f(X)$, σ^2 then only measures deviations of the process from $\boldsymbol{D}^T(X)\boldsymbol{H}_{(d-1)}\boldsymbol{u}$ and therefore does not control the overall variance of $f(X)$ from its null mean. Imposing the null space constraints ensures that σ^2 rightly measures the deviation of $\boldsymbol{u}$ from their null mean. However, the constraints are insufficient to correctly introduce IGMRFs as the constrained process does not correspond to the original $f(X)$. The only way to control the overall variability and not only the scale of the deviation from the null space consists in considering an alternative representation of the process made up by two separate components. Let us define a

new process $f(X)$ as:

$$f(X) = f_t(X) + f_r(X)$$

- *residual term* $f_r(X)$

$$f_r(X) = \boldsymbol{D}^T(X)\boldsymbol{u}$$
$$\boldsymbol{u}|\sigma_r^2 \sim N(\boldsymbol{0}, \sigma_r^2 \boldsymbol{Q}^*) \text{ subject to } \boldsymbol{S}_{(d-1)}^T \boldsymbol{u} = \boldsymbol{0}$$

  $f_r(X)$ is set equal to the original IGMRF effect conditional on $\boldsymbol{S}_{(d-1)}^T \boldsymbol{u} = \boldsymbol{0}$. Conditioning on these constraints does not modify the precision matrix $\boldsymbol{Q}$ (Rue and Held 2005) but simplifies Equation 3.11, since now $\boldsymbol{H}_{(d-1)}\boldsymbol{u} = \boldsymbol{0}$:

$$\pi(\boldsymbol{u}) \propto \exp\left[-\frac{1}{2\sigma^2}\boldsymbol{u}^T \boldsymbol{Q}_d \boldsymbol{u}\right] \tag{3.20}$$

  The null space constraints ensure that the realizations of $\boldsymbol{u}$ will always have a null polynomial trend of degree $d-1$. As a direct consequence, the trend $\boldsymbol{D}^T(X)\boldsymbol{H}_{(d-1)}\boldsymbol{u}$ will also be null. Under the constraints, the process $f_r(X)$ can now be considered proper such that its associated scale parameter, called σ_r^2, will now properly control the deviation of the process from the null space.

- *trend term* $f_t(X)$

$$f_t(X)|\sigma_t^2 \sim N(\boldsymbol{0}, \sigma_t^2 \boldsymbol{Q}_t^*)$$

  $f_t(X)$ is introduced to account for the constraints imposed on $f_r(X)$ and ensures that $f(X)$ is equivalent to the original process, which is retrieved by setting $\sigma_t^2 \to \infty$. $f_t(X)$ must be specified such that it models the trend that the scale parameter σ^2 fails to consider. If so, then the two parts of variability of the original process are now respectively controlled by σ_t^2 and σ_r^2: the former controls the deviation from the null mean to the trend, while the latter measures the remaining variability around the trend.

In order to understand how $f_t(X)$ can be specified, we shall consider first the simpler case where $D_k(X) = I(X = k)\ \forall k$, i.e. where each u_k coefficient reflects one of the values of the support of $X \sim \text{Unif}([1, K])$. In this case, imposing $\boldsymbol{S}_{(d-1)}^T \boldsymbol{u} = \boldsymbol{0}$ ensures that the process $f_r(X)$ has a null polynomial trend, i.e.:

$$\int_{\mathbb{X}} x^m f_r(x) \cdot \pi(x)\, \mathrm{d}x = 0 \quad m = 0, ..., d-1. \tag{3.21}$$

See the proof in Section A.4 of Appendix. Hence, the original process is restored when $f_t(X)$ is designed as a polynomial effect of degree $d-1$ on X.

In the case of $d = 1$, $f_t(X)$ is redundant as the polynomial trend is simply a constant effect with respect to X, whose inclusion would clash with the intercept parameter μ already included in the linear predictor. In this case, $f(X) = f_r(X)$ and the variance contribution corresponds to σ_r^2, after appropriate scaling (Proposition 1).

For $d = 2$, the process $f_t(X)$ should simply be specified as a linear effect on X, after standardization of the covariate to respect the 0-mean constraint and the scaling requirement:

$$\begin{aligned} f(X) &= \frac{X - E[X]}{\sqrt{Var[X]}} \cdot \beta + \sum_{k=1}^{K} \mathbb{I}(X = k) u_k \\ \beta|\sigma_t^2 &\sim N(0, \sigma_t^2) \\ \boldsymbol{u}|\sigma_r^2 &\sim N(\mathbf{0}, \sigma_r^2 \boldsymbol{Q}^*) \text{ subject to } \boldsymbol{S}_{(1)}^T \boldsymbol{u} = \mathbf{0}. \end{aligned} \tag{3.22}$$

After applying the scaling procedure also to the second term, the variance of $f(X)$ becomes equal to $\sigma_t^2 + \sigma_r^2$. The role of the two parameters is clear, as the former controls the variance due to the linear effect, while the latter measures the additional non-linear contribution.

For $d > 2$, $f_t(X)$ must represent a polynomial trend of degree $d-1$. In order to interpret the associated scale parameters, we propose to introduce $d-1$ separate effects $f_{t_1}, ..., f_{t_m}$ with polynomial basis functions $h_1(X), ..., h_{d-1}(X)$, such that each of them respectively represents the linear, quadratic, cubic, etc. contribution to the effect of X. The new, overall, model for covariate X will then be equal to d terms:

$$f(X) = \sum_{m=1}^{d-1} h_m(X) \beta_m + \sum_{k=1}^{K} \mathbb{I}(X = k) u_k$$

whose coefficients are all regulated by a separate variance parameter:

$$\begin{gathered} \beta_m \sim N(0, \sigma_m^2), \quad m = 1, ..., d-1 \\ \boldsymbol{u}|\sigma^2 \sim N(\mathbf{0}, \sigma_r^2 \boldsymbol{Q}^*) \text{ subject to } \boldsymbol{S}_{(d-1)}^T \boldsymbol{u} = \mathbf{0} \end{gathered}$$

such that each σ_m^2 will represent (after appropriate scaling) the m^{th}-degree contribution, and σ_r^2 will instead control all the residual deviation from the polynomial trend of degree $d-1$.

The intuitive interpretation of each of these new effects as exclusively the con-

tribution of their corresponding polynomial degree is achieved by designing each $h_m(X)$ as a polynomial fucntion of degree m:

$$h_m(X) = \sum_{l=0}^{m} a_l \cdot x^l \tag{3.23}$$

whose coefficients must be constrained such that any polynomial trend of degree $m-1$ is removed, i.e.:

$$\int_{\mathcal{X}} x^l \cdot h_m(x) \cdot \pi(x)\,\mathrm{d}x = 0 \qquad \forall\, l = 0, ..., m-1. \tag{3.24}$$

Under these constraints, we can find for example $h_1(X) = X - E[X]$ and $h_2(X) = X^2 - E[X^2] - \dfrac{Cov[X, X^2](X - E[X])}{Var[X]}$.

For a generic choice of $\boldsymbol{D}(X)$ instead, the roles of $f_t(X)$ and $f_r(X)$ are less intuitive. The constraints do not control the polynomial trend on $f_r(X)$ and therefore setting up a basis for $f_t(X)$ is not an immediate task. Moreover, it is not guaranteed that $f_r(X)$ now respects the 0-mean constraint, and if this had to be imposed to satisfy Definition 3.3, this could possibly ruin the sparsity of the precision matrix on the coefficients. Finally, the sum of the parameters σ_t^2 and σ_r^2 would still be equal to the variance of the overall effect of X after scaling, but it would not be possible to assign an intuitive interpretation to them individually.

It would be more convenient if instead the linear constraints imposed on $f_r(X)$ equated to imposing a null polynomial trend of degree $d-1$, such that the impact of this conditioning would be clear and easily adjustable through the specification of $f_t(X)$, as detailed above. The polynomial trend of degree d can be removed from a generic process $\boldsymbol{D}^T(X)\boldsymbol{u}$ through linear constraints on $\boldsymbol{u}$ building a specific matrix $\widetilde{\boldsymbol{S}}$ of dimension $K \times d$ such that $\widetilde{\boldsymbol{S}}^T\boldsymbol{u} = \boldsymbol{0}$ where:

$$\widetilde{\boldsymbol{S}}^T\boldsymbol{u} = \begin{bmatrix} \int_{\mathcal{X}} x^0 \cdot \boldsymbol{D}^T(x)\boldsymbol{u} \cdot \pi(x)\,\mathrm{d}x \\ \\ \int_{\mathcal{X}} x^1 \cdot \boldsymbol{D}^T(x)\boldsymbol{u} \cdot \pi(x)\,\mathrm{d}x \\ \\ \cdots \\ \int_{\mathcal{X}} x^{d-1} \cdot \boldsymbol{D}^T(x)\boldsymbol{u} \cdot \pi(x)\,\mathrm{d}x \end{bmatrix}. \tag{3.25}$$

The appropriate constraint matrix $\widetilde{\boldsymbol{S}}$ can be defined as:

$$\widetilde{\boldsymbol{S}} = \int_{\mathcal{X}} \boldsymbol{D}(x) \cdot \boldsymbol{S}_{(d-1)}(x) \cdot \pi(x)\,\mathrm{d}x \in \mathbb{R}^{K \times d} \tag{3.26}$$

where $\boldsymbol{S}_{(d-1)}(x)$ is a function-valued row vector of dimension $1\times(d)$ obtained evaluating the Vandermonde matrix at a generic x, and the integral is applied element-wise to the matrix in its argument.

More explicitly, the same matrix can be defined as $\widetilde{\boldsymbol{S}} = [\widetilde{\boldsymbol{s}}_0, ..., \widetilde{\boldsymbol{s}}_{d-1}]$ where:

$$\widetilde{\boldsymbol{s}}_m = \begin{bmatrix} \int_{\mathcal{X}} x^m \cdot D_1(x) \cdot \pi(x)\,\mathrm{d}x \\ \int_{\mathcal{X}} x^m \cdot D_2(x) \cdot \pi(x)\,\mathrm{d}x \\ ... \\ \int_{\mathcal{X}} x^m \cdot D_K(x) \cdot \pi(x)\,\mathrm{d}x \end{bmatrix} \quad m = 0, ..., d-1 \tag{3.27}$$

Since in general the columns of $\widetilde{\boldsymbol{S}}$ will be not proportional to the ones of $\boldsymbol{S}_{(d-1)}$, $\boldsymbol{S}_{(d-1)}^T\boldsymbol{u} = \boldsymbol{0}$ does not imply $\widetilde{\boldsymbol{S}}^T\boldsymbol{u} = \boldsymbol{0}$. Hence, we propose to replace the precision matrix of $\boldsymbol{u}$ with a new one, denoted by $\widetilde{\boldsymbol{Q}}$, whose null space is in fact $\widetilde{\boldsymbol{S}}$. We shall refer to this procedure as *Q modification*.

Replacing the original precision matrix with $\widetilde{\boldsymbol{Q}}$ ensures that the σ^2 parameter measures the deviation of $f_r(X)$ from its polynomial trend of degree $d-1$. Thus, conditioning $f_r(X)$ on the constraints $\widetilde{\boldsymbol{S}}^T\boldsymbol{u} = \boldsymbol{0}$ would ensure that σ_r^2 has the intuitive interpretation of measuring the residual variability beyond the polynomial trend.

As it is clear from Equation 3.25, $\widetilde{\boldsymbol{S}}$ is a function of both $\boldsymbol{D}(X)$ and $\pi(x)$, so that the actual design of $\widetilde{\boldsymbol{Q}}$ will be specific to each model. Among the possible precision matrices having a null space $\widetilde{\boldsymbol{S}}$, we propose a solution that preserves the sparsity property of IGMRFs by designing the new precision matrix through the following decomposition:

$$\widetilde{\boldsymbol{Q}} = (\boldsymbol{\Lambda}\widetilde{\boldsymbol{R}}^*\boldsymbol{\Lambda})^* \tag{3.28}$$

where $\boldsymbol{\Lambda}$ is a positive diagonal matrix of entries $\boldsymbol{\lambda} = [\lambda_1, ...\lambda_K]$ and $\widetilde{\boldsymbol{R}}$ is a square matrix with the same sparsity and sign structure of the original $\boldsymbol{Q}$. If sparsity is not an issue, then $\widetilde{\boldsymbol{Q}}$ can be directly used in place of the original $\boldsymbol{Q}$. If instead it is desirable to preserve sparsity, the model can be equivalently represented as:

$$\begin{aligned} f_r(X) &= \boldsymbol{D}^T(X)\boldsymbol{\Lambda}\boldsymbol{u} \\ \boldsymbol{u}|\sigma_r^2 &\sim N(\boldsymbol{0}, \sigma_r^2\widetilde{\boldsymbol{R}}^*) \text{ subject to } \widetilde{\boldsymbol{S}}^T\boldsymbol{\Lambda}\boldsymbol{u} = \boldsymbol{0} \end{aligned}$$

where both the basis $\boldsymbol{\Lambda}\boldsymbol{D}^T(X)$ and the precision matrix $\widetilde{\boldsymbol{R}}$ have the same sparsity structure as the ones of the original model.

Finding a solution that guarantees $\widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{S}} = \mathbf{0}$ is now equivalent to finding $\widetilde{\boldsymbol{R}}$ such that $\widetilde{\boldsymbol{R}}\boldsymbol{\Lambda}\widetilde{\boldsymbol{S}} = \mathbf{0}$. The appropriate entries of $\widetilde{\boldsymbol{R}}$ returning the desired null space can be found as a function of the known elements of $\boldsymbol{Q}, \widetilde{\boldsymbol{S}}$, and the unknown $\boldsymbol{\lambda}$ (see Example 5 and 6). Since $\widetilde{\boldsymbol{R}}$ can be written as a function of $\boldsymbol{\lambda}$, the whole new precision matrix $\widetilde{\boldsymbol{Q}}$ is known up to the choice of these entries. In order to obtain a model similar to the original one, $\boldsymbol{\lambda}$ can be chosen such that the Kullback-Liebler divergence between the reference Gaussian distribution with $\boldsymbol{Q}$ and the Gaussian with $\widetilde{\boldsymbol{Q}}$ is minimized:

$$\widehat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}>\mathbf{0}}{\arg\min}\ D_{\mathrm{KL}}\left(\ \mathcal{N}_{\widetilde{\boldsymbol{Q}}(\boldsymbol{\lambda})}\ ||\ \mathcal{N}_{\boldsymbol{Q}}\ \right). \tag{3.29}$$

The use of the KLD minimization is inspired by Rue and Tjelmeland 2002 where it is used to approximate Gaussian fields with GMRFs. Other alternative minimization criteria could be used, such as the conditional-mean least-squares criterion proposed by Cressie and Verzelen 2008. Examples 5-6 cover some popular models used in application for which the construction of a valid $\widetilde{\boldsymbol{Q}}$ is thoroughly discussed.

Thanks to the *Q modification*, we are able to correctly represent a generic effect $f(X)$ with an IGMRF prior of order d through d separate effects:

$$f(X) = \sum_{m=1}^{d-1} f_{t_m}(X) + f_r(X) \tag{3.30}$$

where the $d-1$ terms $f_{t_m}(X)$ are polynomial trend effects of order $m = 1, ..., d-1$:

$$\begin{aligned} f_{t_m}(X) &= h_m(X)\beta_m \qquad m = 1, ..., d-1 \\ \beta_m|\sigma_m^2 &\sim N(0, \sigma_m^2) \qquad m = 1, ..., d-1 \end{aligned}$$

and $f_r(X)$ is the residual term:

$$\begin{aligned} f_r(X) &= \boldsymbol{D}^T(X)\boldsymbol{u} \\ \boldsymbol{u}|\sigma_r^2 &\sim N(\mathbf{0}, \sigma_r^2\widetilde{\boldsymbol{Q}}^*) \text{ subject to } \widetilde{\boldsymbol{S}}^T\boldsymbol{u} = \mathbf{0} \end{aligned}$$

Note that imposing $\widetilde{\boldsymbol{S}}^T\boldsymbol{u} = \mathbf{0}$ ensures that the 0-mean constraint is always guaranteed for $f_r(X)$. The design of each $h_m(X)$ also guarantees a 0-mean constraint for the $f_{t_m}(X)$ effects. With respect to scaling instead, it is worth highlighting that all of the terms in $f(X)$, i.e $f_{t_1}(X), ..., f_{t_{d-1}}(X), f_r(X)$, must be scaled separately.

After scaling, we obtain that:

$$Var_{X,\beta_1,...,\beta_{d-1},\boldsymbol{u}}[f(X)|\boldsymbol{\sigma}] = \sum_{m=1}^{d-1} \sigma_m^2 + \sigma_r^2. \tag{3.31}$$

See the proof in Section A.5 of the Appendix.

While here the proposal only considers unidimensional IGMRFs, an analogous strategy could be employed for settings where X is multivariate.

3.3.7 Interpretation of the VP parameters

Consider again Model 1 with $\boldsymbol{\theta} = [\boldsymbol{u}_1, ..., \boldsymbol{u}_L, \sigma^2_{L+1}, ..., \sigma^2_J]$. Using the results from the previous sections, we can conclude that Definition 3.3 is satisfied under the following conditions:

$$\begin{aligned} E_{X_j}[f_j(X_j)|\boldsymbol{u}_j] = 0 \quad j = 1, ..., L \\ Var_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2] = \sigma_j^2 \quad j = 1, ..., J. \end{aligned} \tag{3.32}$$

A generalization of the *standardization procedure* consisting of the application of Propositions 2 and 1 (in order) returns Equation 3.32 and, therefore, guarantees that Definition 3.3 is satisfied.

It must be noted that this result technically holds without any modifications for IGMRFs as long as the null space constraints have been imposed. However, we recommend following the steps detailed in Section 3.3.6 for the introduction of IGMRF effects such that the original process is restored and identifiability issues between the effects are avoided. For example, an effect with an IGMRF of second-order should be included in the model through the introduction of a separate linear effect and the IGMRF effect under null space constraints, after the modification its precision matrix (if necessary).

A second important point worth discussing is the impact of the distributional assumption on the covariates. The choice of $\pi(\boldsymbol{x})$ is a crucial step, as it affects the definition of the 0-mean constraints $\boldsymbol{a}$ for fixed effect, the C scaling constants, as well as the potential $\widetilde{\boldsymbol{S}}$ matrices for IGMRF effects. In summary, this choice determines the actual interpretation that is assigned to the various σ^2 parameters. Therefore, it is important to understand what the role of this choice is in this procedure: $\pi(\boldsymbol{x})$ should be specified with the goal of obtaining the optimal case-specific interpretation of each of the σ^2, rather than with the aim of correctly reproducing the actual distribution of $\boldsymbol{X}$. The empirical distribution observed in the data may be considered a sensible choice, as for example it is the traditional choice a posteriori for the estimation of variance contribution of fixed effects (Gelman et al.

2019). However, it represents just a special case among all the choices that can be made. For example, there might be applications in which the empirical distribution may not be representative and there is instead prior information about a more likely distribution assumption. If a covariate is not in fact considered random but rather fixed, a Uniform distribution (either for a categorical, discrete or finite continuous support) is arguably the most sensible choice, as it simply highlights the support of interest for the determination of the variance contribution.

Once it has been proven that the σ_j^2 parameters match their intuitive intepretation, this result can be used to derive expressions for the VP parameters that are interpretable for the user.

Remark 1 (Interpretation of the VP parameters)**.**
Consider Model 1 with $\boldsymbol{\theta} = [\boldsymbol{u}_1, ..., \boldsymbol{u}_L, \sigma^2_{L+1}, ..., \sigma^2_J]$*. If the conditions from Definition 3.3 holds, then the VP parameters defined as in Equation 3.4 are equal to:*

$$V = E_{\boldsymbol{\theta}}\{Var_{\boldsymbol{X},\boldsymbol{u}_1,...,\boldsymbol{u}_J}[\eta|\mu, \boldsymbol{\theta}]|\boldsymbol{\sigma}\}$$

$$\omega_j = \frac{E_{\boldsymbol{\theta}}\{Var_{X_j,\boldsymbol{u}_j}[f(X_j)|\boldsymbol{\theta}]|\boldsymbol{\sigma}\}}{V} \quad j = 1, ..., J.$$

Proof. See Section A.6 of the Appendix. □

Remark 1 proves how the parameter V can be correctly interpreted as the expected variance of interest of the linear predictor, while the entries of $\boldsymbol{\omega}$ correctly represent the proportional contributions of each effect to V. Note that the result from Remark 1 is achievable thanks to the use of an expectation-based scaling method (Proposition 1), which can exploit the linearity property of the expectation. The same result would not hold under the use of the geometric mean scaling method of Sørbye and Rue 2014.

Furthermore, note that Model 1 has been used to illustrate how the VP reparametrization can be correctly applied to all the latent components of a model, regardless of the type of effects. It is likely that there will be many applications that can benefit from this extension as now the prior knowledge about the relative importance of *all* effects can be easily introduced using VP priors. However, there might also be instances in which there are effects for which the comparison to other effects in terms of variance contribution does not make sense and for which traditional prior choices, independent from the other effects, are desired instead. Consider for example the case in which the first effect $f_1(X_1)$ falls in this category. The variance σ_1^2 can be set fixed to a large value, e.g. $\boldsymbol{u}_1 \sim N(\boldsymbol{0}, 1000 \cdot \boldsymbol{Q}_1^*)$, such that the model

parameters are now $\boldsymbol{u_1}$ and $\boldsymbol{\sigma} = [\mu, \sigma_2^2, ..., \sigma_J^2]$. The VP reparametrization can still be applied to the remaining variance parameters $\sigma_2^2, ..., \sigma_J^2$: The same steps must be taken to ensure the interpretability of each $\sigma_j^2,\ j = 2, ..., J$ but now the interpretation of the VP parameters does no longer refer to the whole linear predictor variability but rather to the residual variance, remaining after having accounted for the $f_1(X_1)$ effect.

Finally, it may happen that an LGM does not strictly respect the assumption of Model 1 about fixed precision matrices and that there are additional correlation parameters (e.g. spatial Matern processes and autoregressive temporal ones). In this case, VP priors can still be applied using the approach proposed by Fuglstad et al. 2020: *" [such models] can be integrated into the [VP] prior framework by first defining priors on the correlation parameters, and then constructing the joint prior for the variance parameters with the correlation parameters fixed to reasonable values".* Note that it must be checked that the standardization procedure does not depend on the correlation parameters.

3.4 Examples

In order to illustrate how the proposed method would work in practice on a variety of different effects, a range of popular models is reviewed and the application of the theoretical results from Section 3.3 is illustrated for each of them.

Example 1. Random intercepts

Consider a categorical covariate X with K levels represented by $1, ..., K$. The traditional model consists of using i.i.d. coefficients $u_1, ..., u_K$, where each of them is linked to a level of the covariate through basis functions $D_k(X) = \mathbb{I}[X = k]$. This effect is usually called a random intercept model:

$$f(X) = \sum_{k=1}^{K} \mathbb{I}(X = k) \cdot u_k \tag{3.33}$$

$$\boldsymbol{u}|\sigma^2 \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}). \tag{3.34}$$

This model does not need scaling as its C constant (Equation 3.15) is equal to 1. However, as mentioned in Section 3.3.5, there might be cases where group effects are directly of interest, i.e. the effect should be treated as a fixed one.

Given $\pi(X = k) = p_k$, the 0-mean constraint to be imposed if the effect is treated as fixed is $\boldsymbol{a}^T\boldsymbol{u} = 0$ where $\boldsymbol{a} = [p_1, ..., p_K]^T$. Under this constraint, the precision

matrix of $\boldsymbol{u}$ changes to:

$$\boldsymbol{Q} = \left[\boldsymbol{I} - \frac{\boldsymbol{a}\boldsymbol{a}^T}{\boldsymbol{a}^T\boldsymbol{a}}\right]^*.$$

As a consequence, the scaling constant for the constrained model is found to be:

$$C = 1 - \frac{\sum_{k=1}^{K} p_k^3}{\sum_{k=1}^{K} p_k^2}. \tag{3.35}$$

See the proof in Section A.7 of the Appendix.

If X is Uniformly distributed with $p_k = 1/K$, $\forall k$, the constant C then becomes $\frac{K-1}{K}$ and converges to 1 as the number of levels grows, i.e. as the variance in the mean of the process goes to 0 and the the 0-mean constraint becomes less and less relevant. Alternatively, the levels of X might have unequal probabilities of appearing in a certain population and it might be important to consider the actual distribution in the estimation of the variance contribution of this effect. In this second scenario, the distribution of X can be estimated using the empirical distribution in the dataset or set accordingly to some additional prior information about the distribution of X in the population of interest.

Example 2. Linear effects

Consider a continuous covariate X. A linear effect for such a covariate is an interesting special case with respect to the achievement of the correct interpretation for its corresponding σ^2 parameter:

$$f(X) = X \cdot u \tag{3.36}$$

$$u|\sigma^2 \sim N(0, \sigma^2). \tag{3.37}$$

Since the coefficient u is always the inferential focus, linear effects can be categorized always as fixed effects. As such, a 0-mean constraint should be imposed to guarantee that the condition in Definition 3.3 is met (Proposition 2). In the linear case, the combination of Propositions 1 and 2 results in the traditional standardization procedure:

$$\widetilde{f}(X) = \frac{X - E_X[X]}{\sqrt{Var_X[X]}} \cdot u \tag{3.38}$$

since $C = E_X[(X - E_X[X])^2]$.

This result is coherent with the choice of standardizing all the covariates from

the R2D2 literature and proves that standardization is a special case of the more general procedure presented here. Moreover, it can be noted that the linear effect case greatly reduces the requirement to define a distribution on X, as it is sufficient to specify a finite mean and variance $E[X] < \infty, Var[X] < \infty$ for the covariate. Contrary to the previous cases, several distributions are reasonable for a continuous covariate and there is no evident default choice apart from using the empirical distribution. Among potential sensible choices, a Uniform distribution could be used if the variability in a certain range is of interest (e.g. controlled factors in experiments).

Example 3. Random slopes

Consider now the interaction effect between a categorical covariate X_1 with K observed levels sampled from a larger population (random effect), and a continuous covariate X_2, for which a linear trend is a sensible model (fixed effect). This interaction corresponds to the random slope model:

$$f(X_1, X_2) = \boldsymbol{D}^T(X_1, X_2)\boldsymbol{u} \tag{3.39}$$

$$= \sum_{k=1}^{K} \mathbb{I}(X_1 = k) \cdot X_2 \cdot u_k \tag{3.40}$$

$$\boldsymbol{u}|\sigma^2 \sim N_K(\boldsymbol{0}, \sigma^2 \boldsymbol{I}). \tag{3.41}$$

It can be proved that $f(X_1, X_2)$ does not require the scaling procedure as long as both effects have already been scaled individually and independence is assumed between X_1 and X_2: in particular, it is only necessary to standardize X_2 (since it is treated as fixed), as discussed in Example 2. In fact, this result is not limited to the random slope model but it extends to all interaction terms designed using Kronecker products.

Proposition 3 (Interaction terms)**.** *Consider two independent random variables X_1 and X_2 and define the following two effects for them:*

$$\begin{aligned}
\widetilde{f}_1(X_1) &= \boldsymbol{D}_1^T(X_1)\boldsymbol{u}_1 \\
\boldsymbol{u}_1|\sigma_1^2 &\sim N(\boldsymbol{0}, \sigma_1^2 \boldsymbol{Q}_1^*) \\
\widetilde{f}_2(X_2) &= \boldsymbol{D}_2^T(X_2)\boldsymbol{u}_2 \\
\boldsymbol{u}_2|\sigma_2^2 &\sim N(\boldsymbol{0}, \sigma_2^2 \boldsymbol{Q}_2^*).
\end{aligned}$$

If these two effects have been scaled according to Proposition 1, the corresponding

interaction $f(X_1, X_2)$ term defined as:

$$\begin{aligned} f(X_1, X_2) &= \boldsymbol{D}^T(X_1, X_2)\boldsymbol{u} \\ \boldsymbol{D}(X_1, X_2) &= \boldsymbol{D}_1(X_1) \otimes \boldsymbol{D}_2(X_2) \\ \boldsymbol{u}|\sigma^2 &\sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{Q}_1^* \otimes \boldsymbol{Q}_2^*) \end{aligned}$$

does not need scaling as its variance is already equal to σ^2.

Proof. See Section A.8 of the Appendix.

These types of interactions based on Kronecker products are very popular for modelling for spatio-temporal effects (Knorr-Held 2000,Franco-Villoria, Ventrucci, and Rue 2022).

Example 4. IGMRFs for discrete spatial/temporal effects

Consider X to be on a discrete, finite support taking values $1, ..., K$. X can represent for example either a discrete time or spatial areal data. In this setting, the basis is usually simply defined to map from the X support to the $\boldsymbol{u}$ support, i.e. $D_k(X) = \mathbb{I}[X = k], \quad k = 1, ..., K$, so that the coefficients $u_1, ..., u_K$ then directly represents the process on the original support of the covariate X.

Various models can be used to describe the presence of temporal/spatial correlation, including stationary autoregressive ones. IGMRFs are a popular non-stationary alternative to model $\boldsymbol{u}$. In particular, first- and second-order random walks for regular locations on the line are often used for temporal correlation and corresponds respectively to IGMRF of order $d = 1$ and $d = 2$: as such, the null space of these models is equal to $\boldsymbol{S}_{(0)}$ and $\boldsymbol{S}_{(1)}$ where $\boldsymbol{k} = [1, ..., K]^T$. On the other hand, the popular ICAR model for areal data (Besag and Kooperberg 1995) corresponds to an IGMRF of order $d = 1$ on a lattice, whose null space is equal to $\boldsymbol{1}$, i.e $\boldsymbol{S}_0$. More details in Section 2.3 of Chapter 2). In all these three cases, the null space is evaluated on regular locations $\boldsymbol{k} = [1, ..., K]^T$ and they can be represented with the following general notation:

$$f(X) = \sum_{k=1}^{K} \mathbb{I}(X = k) \cdot u_k \tag{3.42}$$

$$\boldsymbol{u}|\sigma^2 \sim N_K(\boldsymbol{0}, \sigma^2 \boldsymbol{Q}^*) \text{ where } \boldsymbol{Q}\boldsymbol{S}_{(d-1)} = \boldsymbol{0}. \tag{3.43}$$

$\boldsymbol{Q}$ will change according to d and K for IGMRFs on the line, while its definition for an ICAR model on the lattice will also depend on the specific geometry of the areal data.

Although space and time are never considered random variables in application, Model 1 requires the specification of a distribution for X. In absence of contrasting reasons, a discrete Uniform represents the most sensible choice, as each location is given equal importance in the derivation of the variance contribution due to the spatial/temporal effect. Under this assumption, this class of effects from Equation 3.42 is the special case discussed in Section 3.3.6, as it can be proved that the columns of $\widetilde{\boldsymbol{S}}$ are proportional to the ones of $\boldsymbol{S}_{(d-1)}$ (see Section A.4 of Appendix) such that:

$$\boldsymbol{S}_{(d-1)}^T \boldsymbol{u} = \boldsymbol{0} \implies \widetilde{\boldsymbol{S}}^T \boldsymbol{u} = \boldsymbol{0}.$$

Hence, the modification of $\boldsymbol{Q}$ is not necessary in this case. These models can therefore be correctly introduced redefining $f(X) = f_t(X) + f_r(X)$ where:

$$\begin{aligned} f_r(X) &= \sum_{k=1}^{K} \mathbb{I}(X = k) \cdot u_k \\ \boldsymbol{u}|\sigma^2 &\sim N_K(\boldsymbol{0}, \sigma^2 \boldsymbol{Q}^*) \text{ subject to } \boldsymbol{S}_{(d-1)}^T \boldsymbol{u} = \boldsymbol{0}. \end{aligned}$$

$f_t(X)$ will be absent for IGMRFs of first-order (i.e. first-order random walks and ICAR models), while $f_t(X)$ shall contain a linear effect for X for second-order random walks (see Section 3.3.6).

At this point, it is necessary to scale the $f_r(X)$ (as well as the potential $f_t(X)$), according to Proposition 1. Figure 3.1 compares the square root of scaling constants C for a first-order random walk, second-order random walk, and a first-order IGMRF on a regular grid for different values of K.

Our scaling method is then compared to the original proposal of Sørbye and Rue 2014, which first dealt with the problem of heterogeneous effects in the context of IGMRFs through the concept of *reference variance* σ^2_{ref} (Equation 3.16). As discussed in Section 3.3.4, the expectation-based approach offers some nice properties shown in Remark 1. Figure 3.1 reports both $\sqrt{C}$ and σ_{ref} for a direct evaluation of their difference for all scenarios and different values of K. We can see how the trends of the two scaling constants diverge as K grows, suggesting that the impact of a choice over the other will be more significant for a larger K; however, this is likely to be counterbalanced by the fact that a larger K requires more data for the computation of the posterior distribution on the variance parameters, hence a consequent smaller role of the prior. The comparison is assessed in practice in Section 3.5.1.

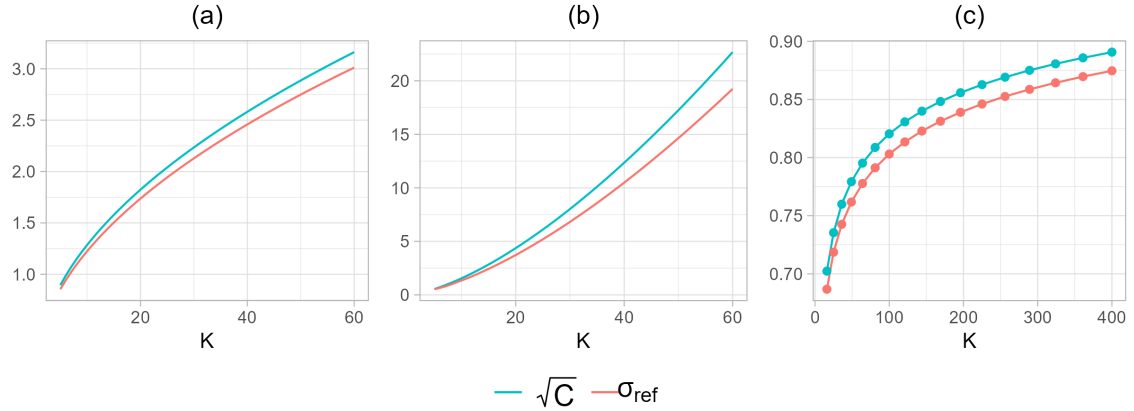

Figure 3.1: Comparison between the square root of the scaling constant $\sqrt{C}$ and the reference standard deviation σ_{ref} for: (a) a first-order random walk on the line for K regular locations; (b) a second-order random walk on the line for K regular locations; (c) an ICAR model on a regular $\sqrt{K} \times \sqrt{K}$ lattice

.

Example 5. P-Splines

Consider again a continuous variable X on a finite interval $[m, M]$, where $m, M \in \mathbb{R}$. P-Splines are a popular smoothing method (Fahrmeir, Kneib, and Lang 2004, Wood 2017) used to represent smooth non-linear effects of continuous covariates (Eilers and Marx 1996, Lang and Brezger 2004). P-Splines are usually defined through a cubic B-Spline basis on equidistant knots with a large number of basis functions (e.g. $K = 20, 30$) denoted by $\boldsymbol{B}(X) = [B_1(X), ..., B_K(X)]^T$; this high flexibility is then regularized through the choice of a penalty function on the coefficients, usually a second-order differences penalization which corresponds to a second-order random walk on the coefficients. This type of P-Spline model can be expressed as:

$$f(X) = \boldsymbol{B}_K^T(X)\boldsymbol{u}$$
$$\boldsymbol{u}|\sigma^2 \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{Q}^*_{\text{RW2}})$$

where the precision matrix $\boldsymbol{Q}_{\text{RW2}}$ is defined as Equation 2.8 of Chapter 2. As mentioned in Section 3.3.6, a second-order random walk is an IGMRF of order $d = 2$, such that $\boldsymbol{Q}_{\text{RW2}}\boldsymbol{S}_{(1)} = \boldsymbol{0}$. Hence, it is necessary again to verify whether the model requires a Q modification procedure. To do so, it is first necessary to derive $\widetilde{\boldsymbol{S}}$ as defined in Equation 3.26:

$$\underset{K\times 2}{\widetilde{\boldsymbol{S}}} = \left[\int_m^M \boldsymbol{B}(x) \cdot \pi(x)\, \mathrm{d}x \quad , \quad \int_m^M x \cdot \boldsymbol{B}(x) \cdot \pi(x)\, \mathrm{d}x\right].$$

We explicitly denote the elements of $\widetilde{\boldsymbol{S}}$ as:

$$\underset{K\times 2}{\widetilde{\boldsymbol{S}}} = \begin{bmatrix} \widetilde{S}_{1,0} & \widetilde{S}_{1,1} \\ \widetilde{S}_{2,0} & \widetilde{S}_{2,1} \\ \dots & \dots \\ \widetilde{S}_{K,0} & \widetilde{S}_{K,1} \end{bmatrix}.$$

Considering the simplest case of $X \sim \text{Unif}(m, M)$, $\widetilde{\boldsymbol{S}}$ is available in closed form (see Section A.9 in the Appendix for the exact derivation). Figure 3.2 compares $\widetilde{\boldsymbol{S}}$ for this case to the original polynomial design matrix $\boldsymbol{S}_{(1)}$ when $K = 20$ and the endpoints of the X support are $m = 0, M = 1$. While it is clear that the two sets of vectors are not pairwise proportional, the divergence from proportionality only occurs at the boundaries of the support, in particular in the first and last 3 locations, and this is true regardless of the value of K. $\widetilde{\boldsymbol{S}}$ will be different for other choices of $\pi(x)$ and might not be available in closed form.

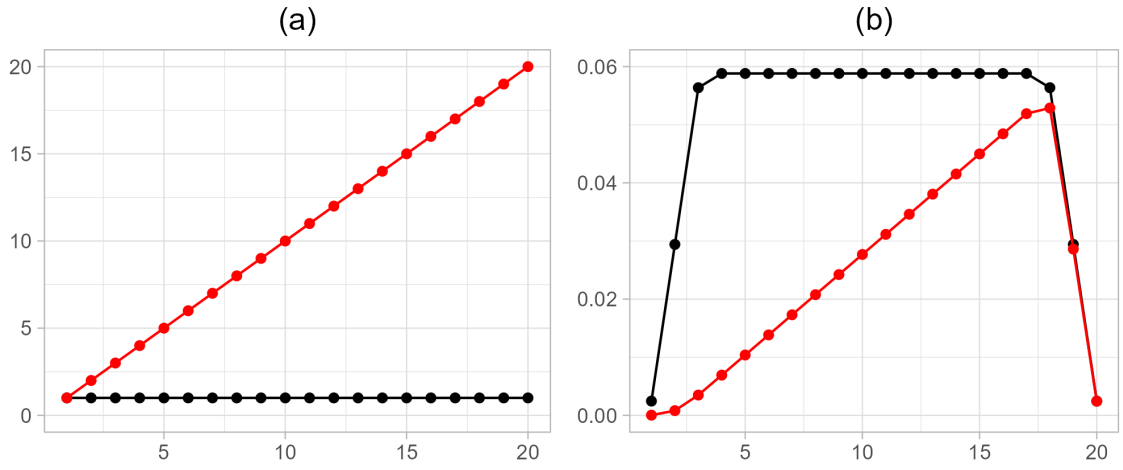

Figure 3.2: Comparison of (a) the columns of $\boldsymbol{S}_{(1)}$ and (b) the columns of $\widetilde{\boldsymbol{S}}$ for $K = 20$, assuming $X \sim \text{Unif}(0, 1)$

Contrary to the previous example, the matrix $\widetilde{\boldsymbol{S}}$ will not have in general columns proportional to the ones of $\boldsymbol{S}_{(1)}$ and therefore the modification of the precision matrix will be necessary. Once $\widetilde{\boldsymbol{S}}$ has been derived for the given $\pi(x)$ (either analytically or through numerical approximation), the next step consists in designing of a valid $\widetilde{\boldsymbol{Q}}$. In this case, the solution for matrix $\widetilde{\boldsymbol{R}}$ can be found using the entries of $\boldsymbol{W}$ defined as:

$$\boldsymbol{W} = \boldsymbol{G} - \boldsymbol{Q}_{\text{RW2}} \tag{3.44}$$

where $\boldsymbol{G}$ is a diagonal matrix with the same diagonal of $\boldsymbol{Q}_{\text{RW2}}$; thereby, $\boldsymbol{W}$ contains the negative of all the non-diagonal entries of $\boldsymbol{Q}_{\text{RW2}}$. It can be shown that $\widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{S}} = \boldsymbol{0}$, where $\widetilde{\boldsymbol{Q}} = (\boldsymbol{\Lambda}\widetilde{\boldsymbol{R}}^{*}\boldsymbol{\Lambda})^{*}$, if $\widetilde{\boldsymbol{R}}$ is defined as:

$$\widetilde{\boldsymbol{R}} = \widetilde{\boldsymbol{G}} - \widetilde{\boldsymbol{W}} \tag{3.45}$$

$$\widetilde{W}_{k,l} = \begin{cases} \dfrac{(l-k)\cdot W_{k,l}}{\lambda_k \widetilde{S}_{k,0}\cdot\lambda_l\widetilde{S}_{l,1} - \lambda_k\widetilde{S}_{k,1}\cdot\lambda_l\widetilde{S}_{l,0}} & k \neq l \\ 0 & k = l \end{cases} \tag{3.46}$$

$$\widetilde{G}_{k,l} = \begin{cases} 0 & k \neq l \\ \dfrac{1}{\lambda_k\widetilde{S}_{k,0}}\cdot\left[\sum_{j=1}^{K}\widetilde{W}_{k,j}\cdot\lambda_j\widetilde{S}_{j,0}\right] & k = l \end{cases} \tag{3.47}$$

See the proof in Section A.10 for the proof. This design of $\widetilde{\boldsymbol{R}}$ has the advantage of being easily implementable and general for a given $\widetilde{\boldsymbol{S}}$. Moreover, the precision matrix $\boldsymbol{Q}$ would not be modified (up to a scaling constant) if the columns of $\widetilde{\boldsymbol{S}}$ were proportional to the ones of $\boldsymbol{S}_{(1)}$.

Finally, the matrix $\widetilde{\boldsymbol{Q}}$ that best approximates the original $\boldsymbol{Q}$ is found optimizing $\boldsymbol{\Lambda}$ according to Equation 3.29. The optimal $\boldsymbol{\Lambda}$ is numerically found for the case of $X \sim \text{Unif}(0,1)$ using the R code reported in Section B.

Once $\widetilde{\boldsymbol{Q}}$ is also available, the P-Spline model can be redefined using Equation 3.30:

$$\begin{aligned} f(X) &= f_t(X) + f_r(X) \\ f_t(X) &= \frac{X - E[X]}{\sqrt{Var[X]}}\cdot\beta \\ f_r(X) &= \boldsymbol{B}_K^T(X)\boldsymbol{u} \\ \beta|\sigma_t^2 &\sim N(0,\sigma_t^2) \\ \boldsymbol{u}|\sigma_r^2 &\sim N(\boldsymbol{0},\sigma_r^2\widetilde{\boldsymbol{Q}}) \text{ subject to } \widetilde{\boldsymbol{S}}^T\boldsymbol{u} = \boldsymbol{0} \end{aligned} \tag{3.48}$$

As usual, the last step consists in scaling the effects such that the variance parameters σ_t^2 and σ_r^2 match their intuitive interpretation. While the linear effect only requires to be standardized as in Equation 3.48 (see Section 3.4), we define the scaled version of $f_r(X)$ explicitly:

$$\widetilde{f}_r(X) = \frac{f_r(X)}{C}.$$

The scaling constant C for the residual effect varies with K. Figure 3.3 displays the conditional variance for different values of K, before scaling in panel (a), and after scaling (c). Panel (b) compares the scaling constants C to the σ^2_{ref} in smaller

dots, under the assumption of $X \sim \text{Unif}(0, 1)$: the difference between the summary metrics diverges as K grows, suggesting the use of C specifically for those cases in which the number of nodes is large. The values of C for various K are reported in Table 3.1 and have been numerically approximated, as discussed in Section 3.3.4.

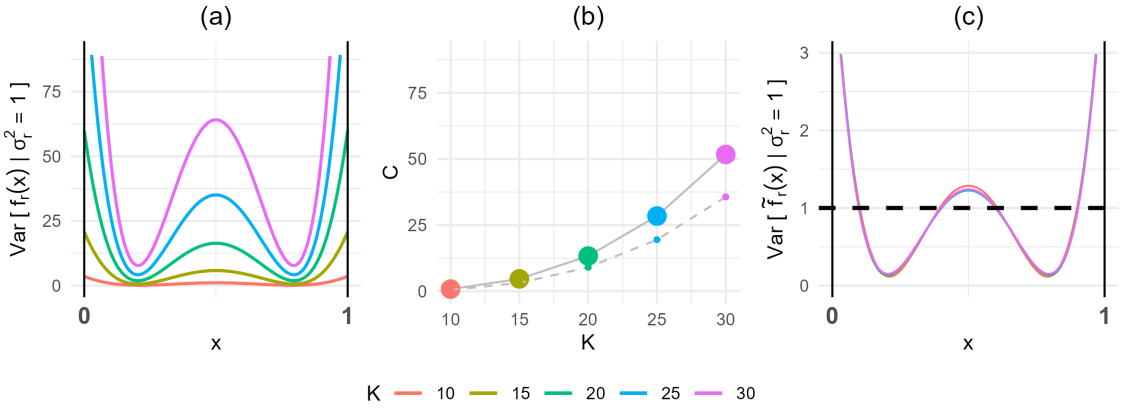

Figure 3.3: (a) Conditional variance of $f_r(X)$ given $X = x$ and $\sigma_r^2 = 1$ for different numbers of basis functions $K = 10, 15, 20, 25, 30$; (b) corresponding scaling constants C (bigger dots and solid line) and σ^2_{ref} (smaller dots and dashed line) assuming $X \sim \text{Unif}(0, 1)$; (c) conditional variance of $\widetilde{f}_r(X)$ after scaling with a dashed line indicating the value of $\sigma_r^2 = 1$.

K	C	K	C	K	C
6	0.045	11	1.308	20	13.328
7	0.123	12	1.924	25	28.438
8	0.266	13	2.686	30	51.693
9	0.496	14	3.603	40	129.476
10	0.835	15	4.695	50	259.966

Table 3.1: Scaling constants for $f_r(X)$ from Equation 3.48 for different values of K and $X \sim \text{Unif}(0, 1)$.

In order to appreciate how $\widetilde{f}_r(X)$ changes after the Q modification, its behaviour is compared to the original model that simply uses $\boldsymbol{Q}_{\text{RW2}}$ in Figure 3.4, after having appropriately scaled the precision matrix.

First, it can be noted that the generalized inverse of $\widetilde{\boldsymbol{Q}}$ displays a similar pattern to the original $\boldsymbol{Q}^*$. The conditional variance of the process as a function of x also has a similar W-shape (or quartic) in both models. The big difference motivating the use of the modified version of the model is illustrated in the bottom panel of Figure 3.4 where realizations of $f_r(x)$ under the constraints have been drawn. Considering the linear trends of these realizations, the original model has non-null trends despite the constraints, while the new model by design removes the linear trends.

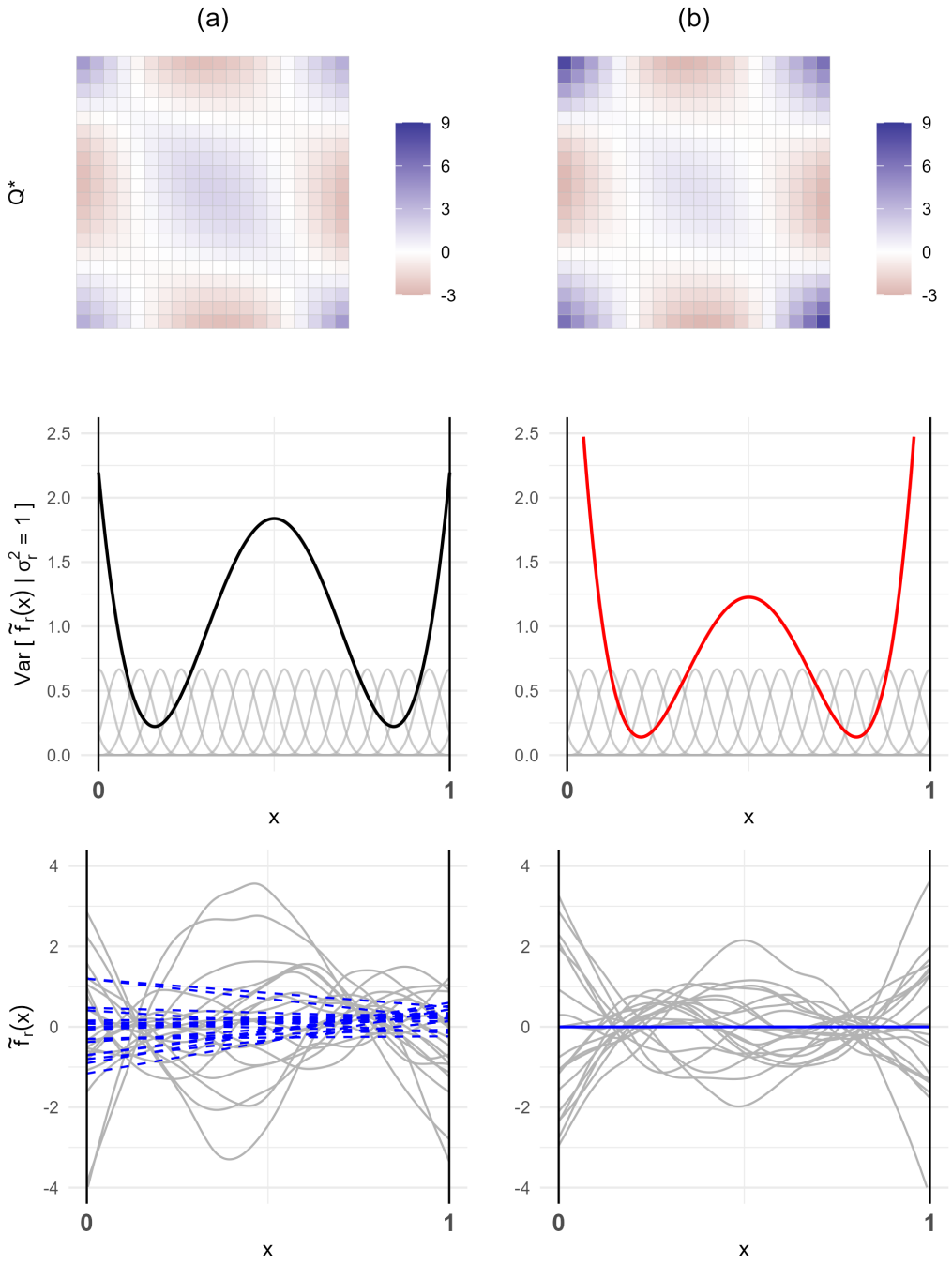

Figure 3.4: Properties of $\widetilde{f}_r(X)$ using (a) $\boldsymbol{Q}_{\text{RW2}}$ and (b) $\widetilde{\boldsymbol{Q}}$ for $K = 20$ and $X \sim \text{Unif}(0, 1)$. Top panel: generalized inverse of the precision matrix on the coefficients. Middle panel: conditional variance of $\widetilde{f}_r(X)$ given $X = x, \sigma_r^2 = 1$ and illustration of the basis of B-Splines (grey). Bottom panel: realizations of $\widetilde{f}_r(x)$ (grey) with corresponding linear trends (blue) when $\sigma_r^2 = 1$.

As K grows, the difference between $\boldsymbol{S}_{(1)}$ and $\widetilde{\boldsymbol{S}}$ becomes less and less relevant and the Q modification has less impact. In fact, the conditional variance of the

process after the Q modification tends to approximate the conditional variance of the original model when $K \to \infty$. Figure 3.5 shows how the conditional variances of $\widetilde{f}_r(X)$ for different values of K are extremely similar after the Q modification (Panel (b)), which is not true when the Q modification procedure is not applied (Panel (a)). Therefore, the Q modification has the additional advantage of neutralizing the effect of K on the conditional variance function, which regulates the overall, global shape of the realizations (K still controls the local flexibility of the model). This result is desirable, as it further reduces the difference in the meaning of variance contributions between different P-Spline designs.

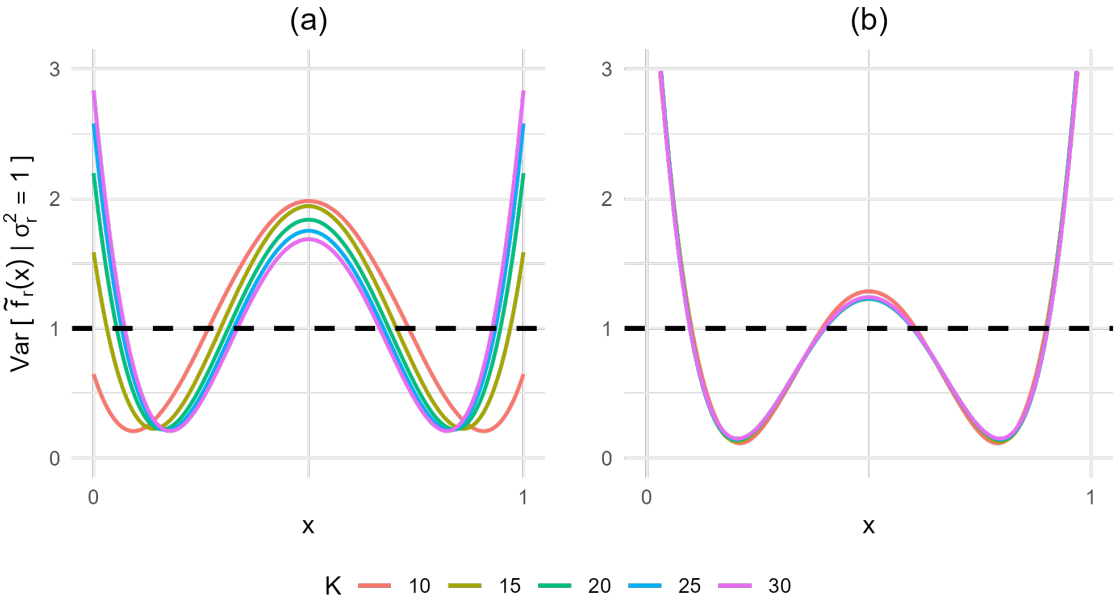

Figure 3.5: Conditional variance of $\widetilde{f}_r(X)$ given $\sigma_r^2 = 1$ for different values of K: (a) before the Q modification, (b) after the Q modification.

Under different assumptions on $\pi(x)$, the results will be different as both the scaling and the Q modification procedures are affected by this choice. Alternatively, the covariate X could be transformed using the probability integral transform to guarantee that the Uniform distribution assumption is always met (Wei et al. 2020). While the flexibility of a P-Spline model allows to well approximate the relationship between X and Y even after the transformation, this approach has some drawbacks such as different flexibility levels to different covariates, and loss of the interpretatibility of σ_t^2 and σ_r^2 as the linear and non-linear contribution of the effect.

Example 6. P-Splines with IGMRF of order 1

Let X_1 and X_2 be two spatial coordinates, delimited on the respective supports $[m_1, M_1]$ and $[m_2, M_2]$. In the case of a spatial effect, it is not reasonable to assume that X_1 and X_2 are random, so that the only case of interest is the assumption of a Uniform distribution over a closed surface of interest. We focus here on the simplest case in which the support is the rectangle $[m_1, M_1] \times [m_2, M_2]$.

Geostatistical data is often modelled using a two-dimensional effect with a Matern autocorrelation structure. However, this approach is computationally expensive as the corresponding precision matrix is dense. Moreover, the range parameter is often considered a random quantity to be estimated: in this case, this model would not fall in the class described in Section 2. Bivariate smoothing can offer a non-parametric alternative which can overcome this issue. In particular, two-dimensional P-Splines have been used in Fahrmeir, Kneib, and Lang 2004 to model spatial heterogeneity: this choice offers sparsity and does not rely on additional nuisance parameters as the Matern does.

The basis for this bivariate P-Spline model $\boldsymbol{B}(X_1, X_2)$ of dimension $K_1 \cdot K_2$ is built as the Kronecker product between 2 one-dimensional cubic B-Splines $\boldsymbol{B}_1(X_1)$ and $\boldsymbol{B}_2(X_2)$, respectively with support $[m_1, M_1]$ and $[m_2, M_2]$ and dimension K_1 and K_2:

$$\boldsymbol{B}(X_1, X_2) = \boldsymbol{B}_1(X_1) \otimes \boldsymbol{B}_2(X_2). \tag{3.49}$$

As in the case of univariate P-Splines, choosing a non-i.i.d. structure on the coefficients can prevent overfitting by regularizing the wiggliness of the curve. In this setting, the elements of $\boldsymbol{u}$ represent nodes in a regular grid of dimension $K_1 \times K_2$. As such, it makes sense to assume an ICAR model on them (i.e. a two-dimensional first-order IGMRF), which is the standard choice for areal spatial data. Under this choice, we can define the effect as:

$$f(X_1, X_2) = \boldsymbol{B}^T(X_1, X_2)\boldsymbol{u} \tag{3.50}$$

$$\boldsymbol{u}|\sigma^2 \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{Q}^*_{\text{ICAR}}) \tag{3.51}$$

The precision matrix $\boldsymbol{Q}_{\text{ICAR}}$ is defined as:

$$\boldsymbol{Q}_{\text{ICAR}} = \boldsymbol{G} - \boldsymbol{W} \tag{3.52}$$

where $\boldsymbol{W}$ is the adjacency matrix of first-order neighbours on the regular grid $K_1 \times K_2$ and $\boldsymbol{G}$ is a diagonal matrix with entries equal to the total number of neighbours

per cell:

$$G_{k,l} = \mathbb{I}[k = l] \cdot \sum_{j=1}^{K_1 \times K_2} W_{k,j}. \tag{3.53}$$

In order to correctly include this model in the VP framework, it is first necessary to compute $\widetilde{\boldsymbol{S}}$. In this example:

$$\underset{K\times 1}{\widetilde{\boldsymbol{S}}} = \left[\int_{m_1}^{M_1}\int_{m_2}^{M_2} \boldsymbol{B}(x_1, x_2) \cdot \pi(x_1, x_2)\,\mathrm{d}x_2\,\mathrm{d}x_1\right]$$

where its elements are denoted as:

$$\underset{K\times 1}{\widetilde{\boldsymbol{S}}} = \begin{bmatrix} \widetilde{S}_1 \\ \widetilde{S}_2 \\ \dots \\ \widetilde{S}_{K_1 \cdot K_2 \cdot} \end{bmatrix}$$

The result for a bivariate Uniform distribution on the covariates are proved to be independent of the values of m_1, M_1, m_2, M_2 (see Section A.11 of the Appendix). Again, this matrix is not proportional to $\boldsymbol{S}_{(0)} = \mathbf{1}_{K_1 \cdot K_2 \times 1}$, so that the precision matrix must again be modified. For a general $\widetilde{\boldsymbol{S}}$, a valid $\widetilde{\boldsymbol{Q}}$ can be found using a simplification of the procedure presented in Example 5, making use again of the diagonal/non-diagonal entries decomposition from Equation 3.52. It can be shown that $\widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{S}} = \mathbf{0}$ if $\widetilde{\boldsymbol{Q}} = (\boldsymbol{\Lambda}\widetilde{\boldsymbol{R}}^*\boldsymbol{\Lambda})^*$ and:

$$\widetilde{\boldsymbol{R}} = \widetilde{\boldsymbol{G}} - \widetilde{\boldsymbol{W}} \tag{3.54}$$

$$\widetilde{W}_{k,l} = \frac{W_{k,l}}{\lambda_k \cdot \widetilde{S}_k \cdot \lambda_l \widetilde{S}_l} \tag{3.55}$$

$$\widetilde{G}_{k,l} = \frac{G_{k,l}}{\lambda_k^2 \cdot \widetilde{S}_k^2} \tag{3.56}$$

See the proof in Section A.12 of the Appendix.

Given $\widetilde{\boldsymbol{Q}}$ the final model can be redefined. No trend term is to be added in the linear predictor since the IGMRF is of order 1:

$$\begin{aligned} f(X_1, X_2) &= f_r(X_1, X_2) \\ f_r(X_1, X_2) &= \boldsymbol{B}^T(X_1, X_2)\boldsymbol{u} \\ \boldsymbol{u}|\sigma^2 &\sim N(\mathbf{0}, \sigma_r^2 \widetilde{\boldsymbol{Q}}^*) \text{ subject to } \widetilde{\boldsymbol{S}}^T\boldsymbol{u} = \mathbf{0}. \end{aligned} \tag{3.57}$$

Table 3.2 reports a numerical approximation of the scaling constants C for different

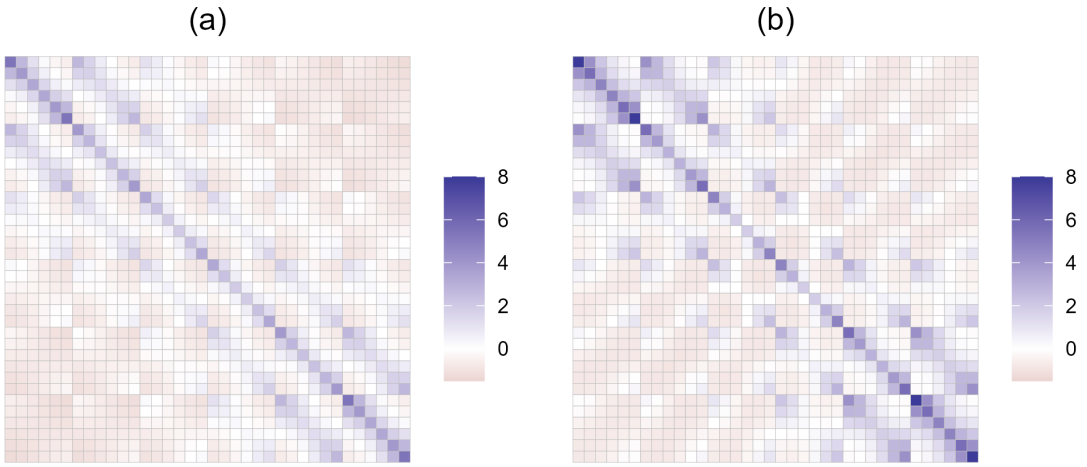

Figure 3.6: Representation of the entries of (a) $\boldsymbol{Q}^*_{\text{ICAR}}$ and (b) $\widetilde{\boldsymbol{Q}}^*$

values of $K_1 = K_2$ (see Section 3.3.4). Since $\widetilde{\boldsymbol{S}}$ does not depend on the boundaries of the rectangular support of $[X_1, X_2]$, both $\widetilde{\boldsymbol{Q}}$ and C are also independent of m_1, M_1, m_2, M_2:

$K_1 = K_2$	6	8	10	12	14	16
C	0.136	0.222	0.286	0.337	0.377	0.412

Table 3.2: Scaling constants for $f_r(X)$ from Equation 3.57 for different values of $K_1 = K_2$.

Focusing on the Uniform distribution, Figure 3.6 shows the generalized inverses of $\boldsymbol{Q}_{\text{ICAR}}$ and $\widetilde{\boldsymbol{Q}}$ for $K_1 = K_2 = 6$ after scaling: again, the original covariance pattern is maintained.

An interesting special case of this model is when $K_2 = 1$ and $\boldsymbol{B}_2(X_2)$ is replaced in Equation 3.49 with $\boldsymbol{D}(X_2) = 1$: the model becomes a univariate P-Spline model with an IGMRF of order 1 structure on the coefficients, which can be then modified following Equation 3.54. This model could be useful for smoothing in applications where the separation between the linear and non-linear contributions are not of interest or inconvenient. For a range of values of K_2, Figure 3.7 shows the conditional variance of $f_r(X_1)$ and $\widetilde{f}_r(X_1) = f_r(X_1)/\sqrt{C}$ given $\sigma^2 = 1$, as well as the corresponding scaling constants C and σ^2_{ref}. First, the conditional variance pattern is different from the one of Figure 3.3. Secondly, it can also be noted that the two scaling constants tend to diverge as K_1 grows, which means that the impact of using σ^2_{ref} in place of C becomes less negligible if K_1 is large. The values of C for different values of K_1 are also reported in Table 3.3.

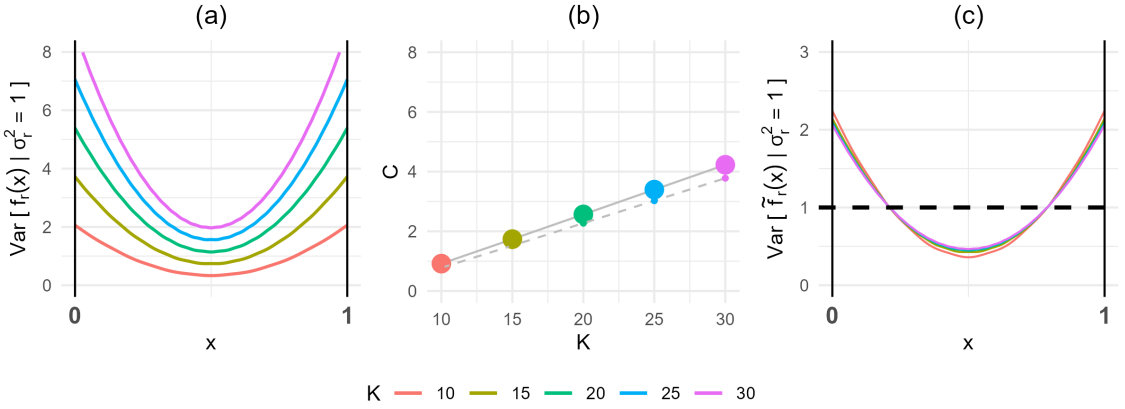

Figure 3.7: (a) Conditional variance of $f_r(X_1)$ given $X_1 = x$ and $\sigma_r^2 = 1$ for different numbers of basis functions $K_1 = 10, 15, 20, 25, 30$; (*b*) corresponding scaling constants C (bigger dots and solid line) and σ^2_{ref} (smaller dots and dashed line) assuming $X_1 \sim \text{Unif}(0, 1)$; (c) conditional variance of $\widetilde{f}_r(X_1)$ after scaling with a dashed line indicating the value of σ_r^2.

K_1	C	K_1	C	K_1	C
6	0.289	11	1.080	20	2.566
7	0.440	12	1.244	25	3.397
8	0.596	13	1.408	30	4.229
9	0.756	14	1.573	40	5.895
10	0.917	15	1.738	50	7.562

Table 3.3: Scaling constants for $f_r(X)$ from Equation 3.57 for different values of K_1 and $K_2 = 1$.

3.5 Empirical results

The R-INLA software has been used to fit the models used in the following sections (Rue, Martino, and Chopin 2009).

3.5.1 Impact of 0-mean constraint

The 0-mean constraint proposed in Section 3.3.5 has clear theoretical implications, as it guarantees that the variance parameters of fixed effects can correctly be interpreted as the expected values of the finite-population variances. However, imposing the constraint might have in practice a negligible impact on posterior inference. Under a 0-mean constraint, specifying a prior on a given σ^2 parameter implies the same prior on the expected finite-population variance $E[s^2]$ since $\sigma^2 = E[s^2]$. Without the constraint instead, $E[s^2] = \sigma^2 \cdot [1 - \text{tr}(\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Q}^*)]$ (Equation 3.18 proved in Section

A.3). and the implied prior on $E[s^2]$ becomes:

$$\pi(E[s^2]) = \pi_{\sigma^2}\left(\frac{E[s^2]}{1-\text{tr}[\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Q}^*]}\right)\cdot\frac{1}{1-\text{tr}[\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Q}^*]}.$$

Hence, if the user represents its prior beliefs about the variance contribution specifying a prior on σ^2, the actual prior on the variance contribution as they intend it (i.e. $E[s^2]$) will be different from the prior on $\pi(\sigma^2)$. We shall call *distortion* this difference between the desired and the actual prior of a certain quantity for which the user is interested in introducing prior beliefs.

The practical implications of this distortion (which is caused by not using a 0-mean constraint) are here investigated on the simple Gaussian random intercept model for a categorical covariate X, with a varying number of groups K and equally likely outcomes, i.e. $X \sim \text{Uniform}([1, K])$ (see Example 1). Under this model, $\boldsymbol{a} = [K^{-1}, ..., K^{-1}]^T$ such that $E[s^2] = \sigma^2 \cdot \frac{K-1}{K}$. The impact of the 0-mean constraint will be more relevant when the distortion in the implied prior is bigger, which is in this case obtained with a small K. Even so, the impact of this "distorted" prior may still become negligible if the data are informative enough: thus, it is obvious to expect a larger impact of the 0-mean constraint imposition when the number of observations is small. For this reason, the practical impact of the constraint on posterior inference is checked through a simulation study on a small sample size scenario. 200 datasets are simulated with a number of groups $K = 4$ and equal number of observations per group $N_g = 5$ from the following model:

$$Y_{ik}|u_k, \sigma^2_\epsilon \sim N(u_k, \sigma^2_\epsilon) \qquad \forall i = 1, ..., N_g, \quad k = 1, ..., K$$

where σ^2_ϵ is set to 1 and $\boldsymbol{u} = [u_1, u_2, u_3, u_4]$ are i.i.d. samples from a Standard Normal, which have been standardized to have exactly mean 0 and $s^2_1 = \frac{1}{K}\sum_{k=1}^K u_k^2 = 1$. The datasets are then fitted using the random intercept model in the following two ways:

- *No constraint*: $\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2_1\boldsymbol{I})$.
- *Constraint*: $\boldsymbol{u} \sim N\left(\boldsymbol{0}, \sigma^2_1\frac{K}{K-1}\left[\boldsymbol{I} - \frac{1}{K}\right]\right)$ subject to $K^{-1}\sum_{k=1}^K u_k = 0$ (see Example 1).

Note how the covariance matrix is multiplied by scaling constant in the *No constraint*.

Finally, the two models are fitted with three different priors:

(a) Default Inverse-Gamma priors for variance parameters in the INLA software, i.e. with shape hyperparameter set to 1 and rate hyperparameter set to $5e-5$ (**IG priors**):

$$\sigma_1^2, \sigma_\epsilon^2 \overset{iid}{\sim} \text{IG}(1, 5e-5).$$

(b) Penalized complexity priors with null base model (**PC priors**):

$$\sigma_1^2, \sigma_\epsilon^2 \overset{iid}{\sim} \text{PC}_0\left(\frac{-\log(0.05)}{3}\right)$$

where the hyperparameter is set as in Fuglstad et al. 2020, such that $P(\sigma > 3) = 0.05$.

(c) VP prior as used in Fuglstad et al. 2020 (**VP prior**):

$$V = \sigma_1^2 + \sigma_\epsilon^2 \sim \text{Jeffreys}$$
$$\omega = \frac{\sigma_1^2}{V} \sim \text{Unif}(0, 1).$$

The priors implied on s_1^2 under the absence of a 0-mean constraint are analytically derived in Section A.13 of Appendix.

By simulation, the value of s_1^2 is fixed to 1. Results about the estimation of s_1^2 are reported in Figure 3.8 in terms of posterior medians, along with the difference between the two methods. Even in this small sample size context, the difference in the estimates is very small for all three prior choices. In terms of performance, the IG priors perform poorly regardless the constraint being applied or not. The other two specifications return very similar results. However, despite the similarity in the posterior results, the distribution of the differences between the *No constraint* and *Constraint* is closer to 0 for prior (b) than for prior (c). This result suggests that the application of the 0-mean constraint is more relevant when using VP priors than independent PC ones, which seem to be more robust to the distortion caused by the *No constraint* case.

Since for larger values of K and N_g the differences are expected to be even less severe, we conclude that the 0-mean constraint is practically relevant for the random intercept model only if both K and N_g are extremely small. Otherwise, we expect that not imposing the constraint will return a negligible bias in the s_1^2.

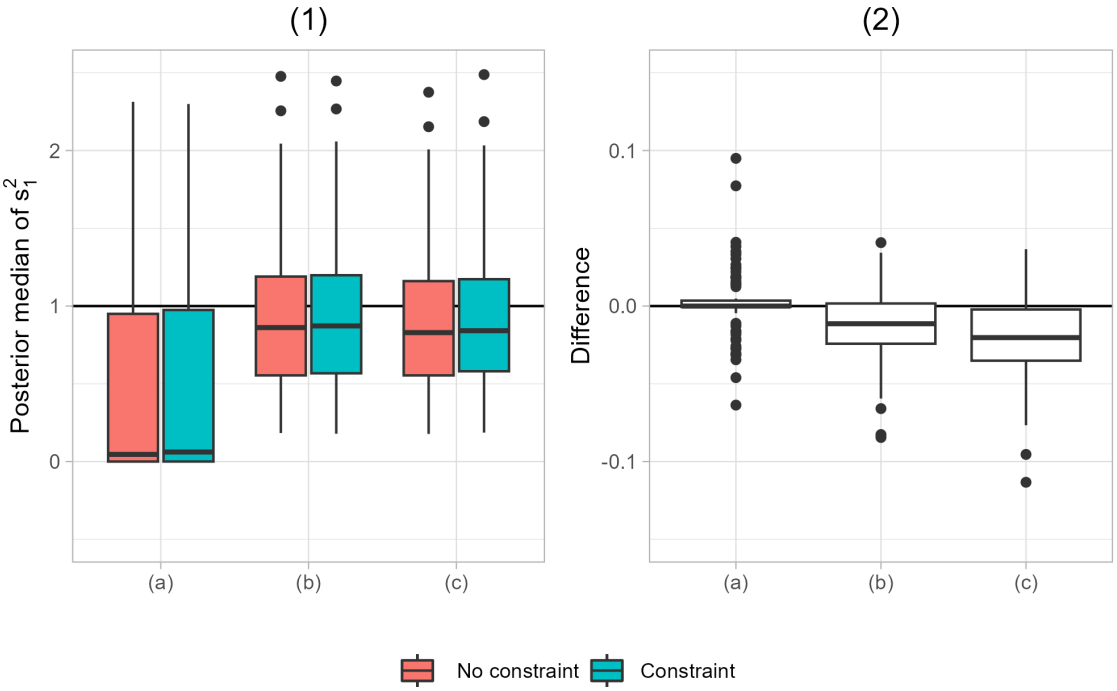

Figure 3.8: (1) Posterior medians of s_1^2 for 200 datasets using the *No constraint* model (red) and the *Constraint* (blue) for three prior specifications: (a) IG priors, (b) PC priors, (c) VP prior. (2) Difference in the posterior median estimates of s_1^2 between the *No constraint* and the *Constraint* models for the same three prior choices.

3.5.2 Impact of scaling

Section 3.3.4 illustrates how scaling theoretically allows the variance parameters to match their intuitive interpretation. The importance of this procedure resides in the fact that it ensures that specifying a prior on each σ_j^2 implies the same prior on its intuitive interpretation as in Definition 3.3. Otherwise, this is not the case and the implied prior on the variance could be greatly distorted from what was the intended prior. The same is obviously true if a prior specification is chosen on V and $\boldsymbol{\omega}$, which implies the same prior on their interpretations as defined in Remark 1 only if the scaling procedure has been applied.

Although the benefits of expectation-based scaling are evident in theory, it might be that the impact of not scaling (or scaling by another location summary) has no practical implications in some scenarios. This could happen for the following two reasons: the distortion on the implied prior on the variance contributions for a particular prior specification choice is negligible; the information in the data is overwhelming so that the impact of the prior on posterior inference is negligible, despite the distortion in the prior.

With regard to the former point, the distortion can be evaluated deriving the implied priors on the quantities of interest. This analysis can be useful in comparing the severity of the distortion caused by different prior choices and identifying those that appear to be less robust to the scaling issue. Here, we study the phenomenon using a simple model with only two effects:

- an effect for a discrete, Uniformly distributed, covariate X on $[1, K]$, with basis functions $D_k(X_k) = \mathbb{I}(X = x_k)$ and an IGMRF of first-order on the coefficients;
- an i.i.d effect ϵ for any observation.

The model can be written as:

$$\begin{array}{rcl} \eta_i & = & \mu + \sum_{k=1}^{K} \boldsymbol{I}(X_i = k) u_k + \epsilon_i \quad i = 1, ..., n \\ \boldsymbol{u}|\sigma_1^2 & \sim & N_K(\boldsymbol{0}, \sigma_1^2 \boldsymbol{Q}^*) \text{ subject to } \boldsymbol{S}_{(0)}^T \boldsymbol{u} = \boldsymbol{0} \\ \boldsymbol{\epsilon}|\sigma_\epsilon^2 & \sim & N(0, \sigma_\epsilon^2 \boldsymbol{I}) \end{array} \tag{3.58}$$

where the vector of parameters of interest is defined as $\boldsymbol{\theta} = [\mu, \sigma_1^2, \sigma_\epsilon^2]$.

From the original parameters, we can derive the VP parameters V and ω (Equation 3.4), along with the total variance T and the proportion of variance φ due to $f(X)$ using the expressions from Remark 1:

$$\begin{array}{ll} V = \sigma_1^2 + \sigma_\epsilon^2 & \omega = \dfrac{\sigma_1^2}{\sigma_1^2 + \sigma_\epsilon^2} \\ T = Var[Y|\boldsymbol{\theta}] & \varphi = \dfrac{Var[f(X)|\boldsymbol{\theta}]}{Var[\eta|\boldsymbol{\theta}]}. \end{array}$$

To understand the distortion caused by not scaling, we shall focus on the prior implied on φ for a given prior specification either on $\sigma_1^2, \sigma_\epsilon^2$ or on V, ω.

If the effect $f(X)$ is correctly scaled (i.e. the scaling constant from Proposition 1 is equal to $C = 1$, then we obtain that$V = T$ and $\omega = \varphi$: the prior specified on ω will therefore be equal to the one implied on φ. However, if the effect is not scaled, then $C = \text{tr}[\boldsymbol{Q}^*]$ and we obtain that:

$$\omega = \frac{\frac{\varphi T}{C}}{\frac{\varphi T}{C} + (1 - \varphi)T}.$$

As a consequence, the implied prior on φ is:

$$\pi(\varphi) = \pi_\omega \left(\frac{\varphi}{\varphi + C - \varphi C} \right) \cdot \frac{C}{[\varphi + C - \varphi C]^2}. \tag{3.59}$$

Equation 3.59 highlights how the lack of scaling when it is necessary (i.e. $C \neq 1$) causes a distortion between the desired prior distribution on φ, i.e. $\pi(\omega)$, and the actual prior on φ. While it is clear that the difference grows as C is further from 1, the actual impact of $C \neq 1$ depends on the choice $\pi_\omega(\cdot)$.

Along with the case of scaling according to the procedure of Section 3.3.4 and non-scaling, we shall investigate also what happens when the geometric mean method by Sørbye and Rue 2014 is applied, i.e. the effect is scaled by σ^2_{ref} as defined in Equation 3.16. Scaling by the constant σ^2_{ref} would return a new value of C equal to C/σ^2_{ref}. Thus, the prior implied on φ would become:

$$\pi(\varphi) = \pi_\omega \left(\frac{\varphi}{\varphi + \frac{C}{\sigma^2_{\text{ref}}} - \varphi \frac{C}{\sigma^2_{\text{ref}}}} \right) \cdot \frac{\frac{C}{\sigma^2_{\text{ref}}}}{\left[\varphi + \frac{C}{\sigma^2_{\text{ref}}} - \varphi \frac{C}{\sigma^2_{\text{ref}}} \right]^2}.$$

Since the impact of scaling on posterior inference is greater when the role of the prior is larger, the priors implied on φ are studied here for two models usually employed in contexts of a single observation for any value of the X support (i.e. $n = K$).

- **Local level model for time series** (Durbin and Koopman 2012): $\boldsymbol{Q}$ is th e precision matrix of a first-order random walk on $K = 25$ regular locations ($C = 4.16$, $\sigma^2_{\text{ref}} = 3.77$).
- **BYM model for areal data** (Besag, York, and Mollié 1991): $\boldsymbol{Q}$ is the precision matrix for an ICAR model on the graph of the 366 Sardinia districts used in Riebler et al. 2016 ($C = 0.514$, $\sigma^2_{\text{ref}} = 0.486$).

The analytical form of $\pi(\varphi)$ is derived for the 3 different prior choices (a)-(b)-(c) detailed in the previous section, which all implies a symmetric prior on ω (see Section A.14 of Appendix for exact derivation). Figures 3.9 and 3.10 report the resulting priors, respectively for the local level model and the BYM model. The black lines in the 2 figures represent the prior elicited on ω, i.e. the desired prior on φ, as well as its actual prior in the case of expectation-based scaling; on the other hand, the blue lines represent the implied priors under geometric mean scaling and red ones under no scaling at all. Under no scaling, $C > 1$ favours large values of φ over smaller ones more than the desired prior (Figure 3.9), and vice versa for $C < 1$ (Figure 3.10). Additionally, a very small difference appears between the expectation-based and the geometric mean scaling methods for all priors in both examples. However, by design, it can be said that the geometric mean method will always favour larger values more than the desired prior, as the geometric mean is by construction always equal or smaller than the arithmetic mean (i.e. $C/\sigma^2_{\text{ref}} > 1$). Thirdly, some prior

specifications appear to be more robust than others, namely the PC prior approach implies very similar priors regardless of the chosen scaling strategy in both examples.

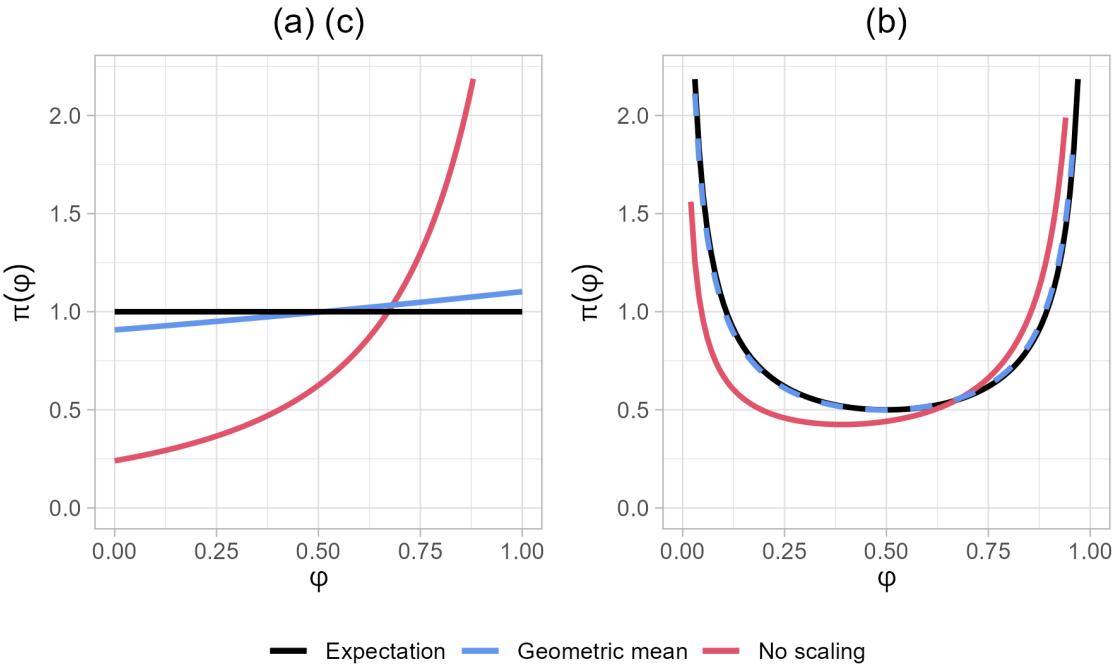

Figure 3.9: Implied prior on φ for the local level model for the different prior choices from Section 3.5.1: (a) IG priors; (b) PC priors; (c) VP prior. The results for the Inverse-Gamma (a) and the VP (c) prior choices are identical.

To understand how this prior distortion affects posterior inference, 200 datasets are simulated and fitted for both models with $T = 1$ and 3 different values for φ: 0.2, 0.5, 0.8.

Local level model Y_i is simply set equal to η_i from Equation 3.58 and a single observation is simulated for each of the $K = 25$ locations on X.

Figure 3.11 reports the bias of the the posterior mean of φ. In this context, we are interested in assessing the differences between the different scaling methods rather than assessing the goodness of estimation. The results show evidence for a non-negligible difference between the posterior estimates of φ of a scaled model (either with geometric mean or expectation-based constants) and an unscaled one (Figure 3.11). As expected, this difference is non-negligible in the case of the VP prior (c), while the PC priors choice appears more robust, and even more so the IG prior. In most scenarios, there is no relevant difference between the estimates obtained using the geometric mean or the expectation-based scaling strategy.

If we then consider performance, the IG prior appears to be the worst choice as it consistently underestimates φ, specifically when the true value is large. Better

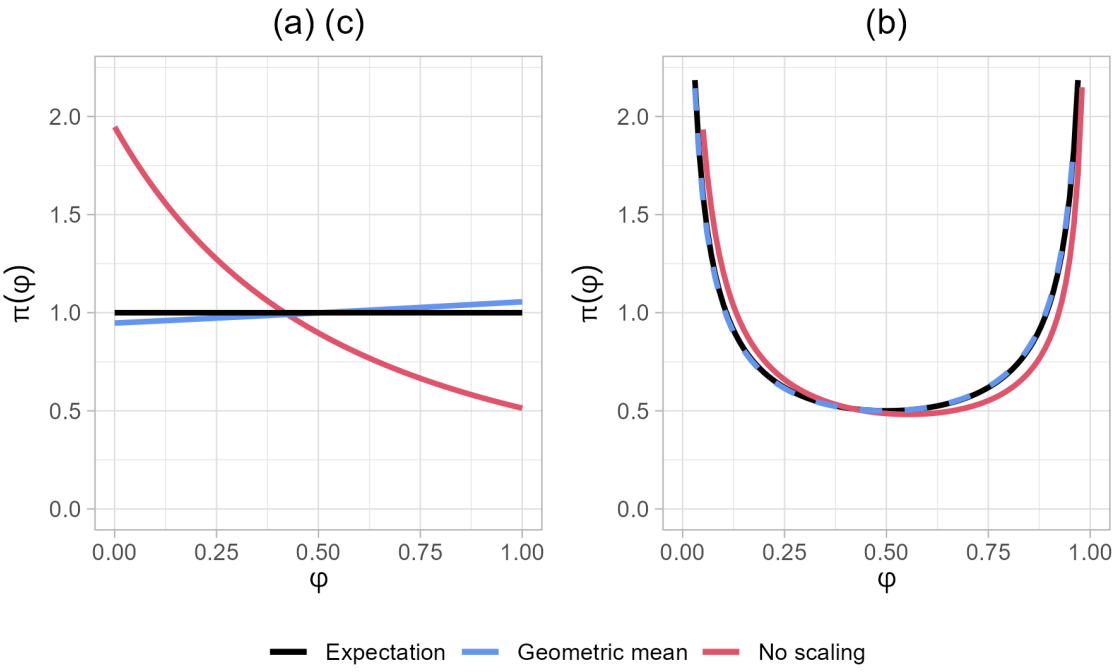

Figure 3.10: Implied prior on φ for the BYM model for the different prior choices from Section 3.5.1: (a) IG priors; (b) PC priors; (c) VP prior. The results for the Inverse-Gamma (a) and the VP (c) prior choices are identical.

results are obtained with PC priors, whose bias' distributions are almost centered at 0 but quite dispersed. Arguably the best among the three, the VP prior returns instead estimates with less variance but a small systematic bias (positive when $\varphi < 0.5$ and negative for $\varphi > 0.5$): this is due to the Uniform distribution on the φ, which pushes the posterior estimates towards the center when the data is not informative enough.

Similar conclusions are drawn looking at the posterior estimates for the total variance parameter T (Figure 3.12). Not scaling returns smaller estimates for T under the PC priors and the VP prior choices. Again, the VP prior after correct scaling performs overall better than the competitors.

BYM model A Poisson likelihood is instead chosen for the BYM simulation since this model is popularly used for epidemiological count data: $Y_i \sim \text{Poisson}(E_i \cdot \exp(\eta_i))$. $E_i = 15$ is fixed for all $N = 366$ locations. The prior for the total variance is changed from Section 3.5.1 to $V \sim \text{PC}_0\left(\frac{-\log(0.05)}{3}\right)$, as proposed by Fuglstad et al. 2020 for non-Gaussian likelihood models and also used in Riebler et al. 2016.

Posterior estimates are more precise in this second example because of the larger sample size (Figure 3.13). Almost no differences appear in the results under the first two choices, while the no scaling method consistently estimates smaller values

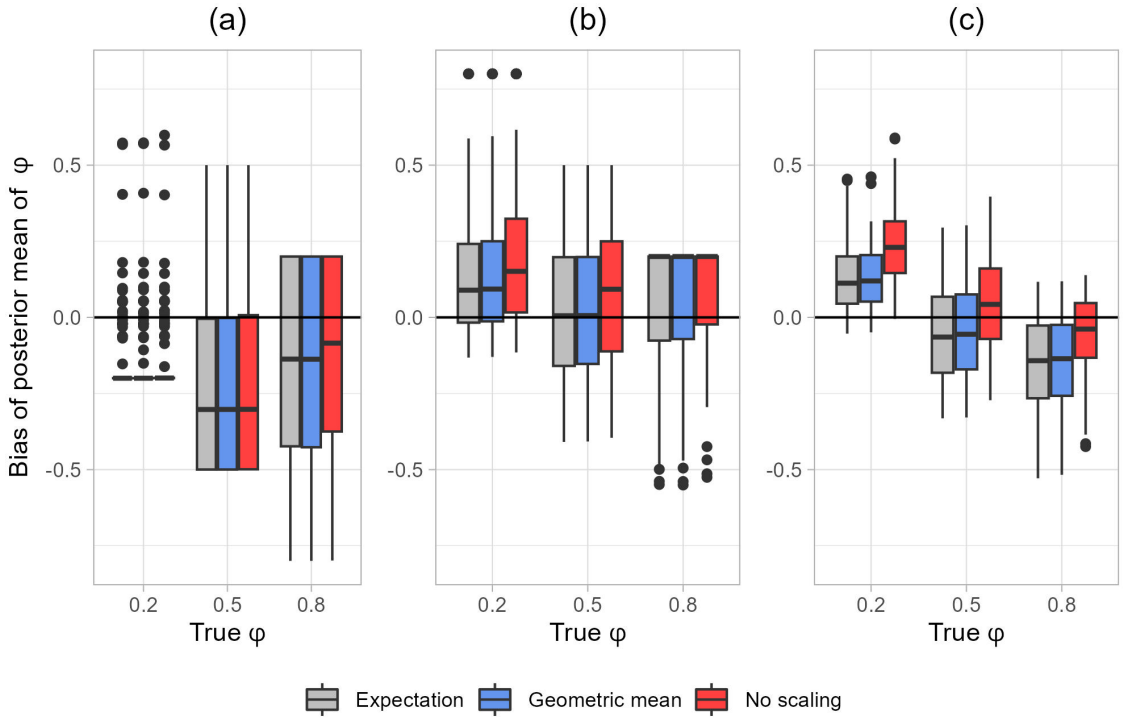

Figure 3.11: Bias of the posterior mean of φ for the local level model with the following prior choices: (a) IG priors; (b) PC priors; (c) VP prior.

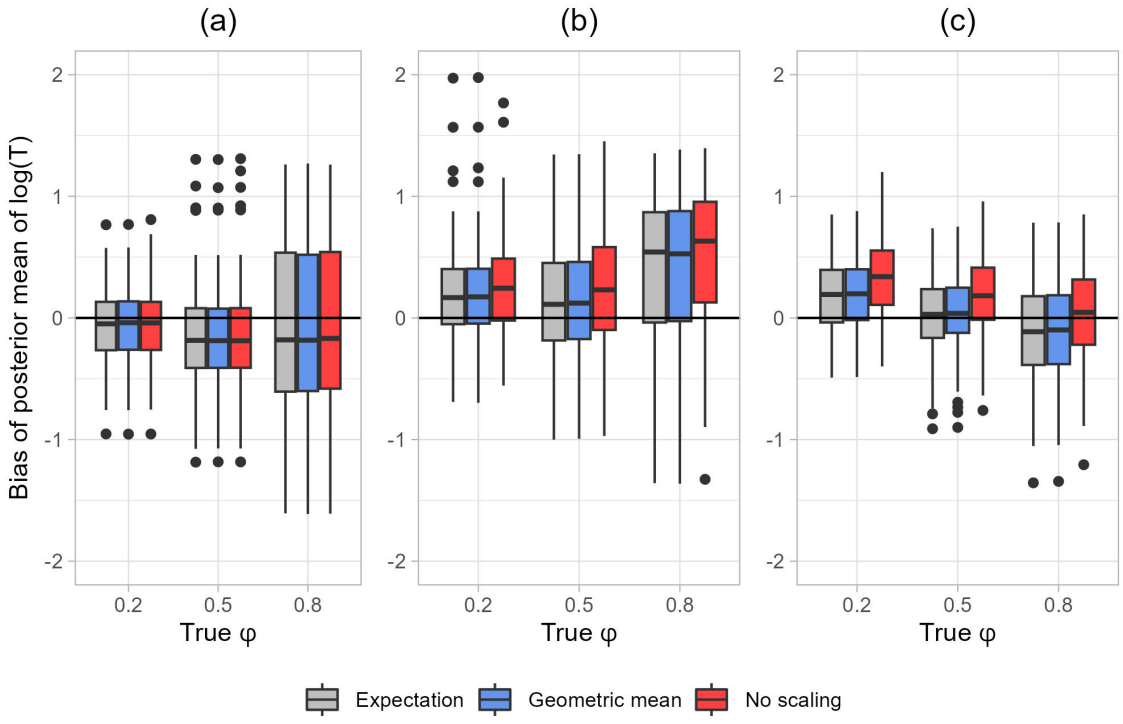

Figure 3.12: Bias of the posterior mean of T in log scale for the local level model with the following prior choices: (a) IG priors; (b) PC priors; (c) VP prior.

of φ when the VP prior is used . Again, no relevant differences appear between the geometric mean and the expectation-based scaling methods.

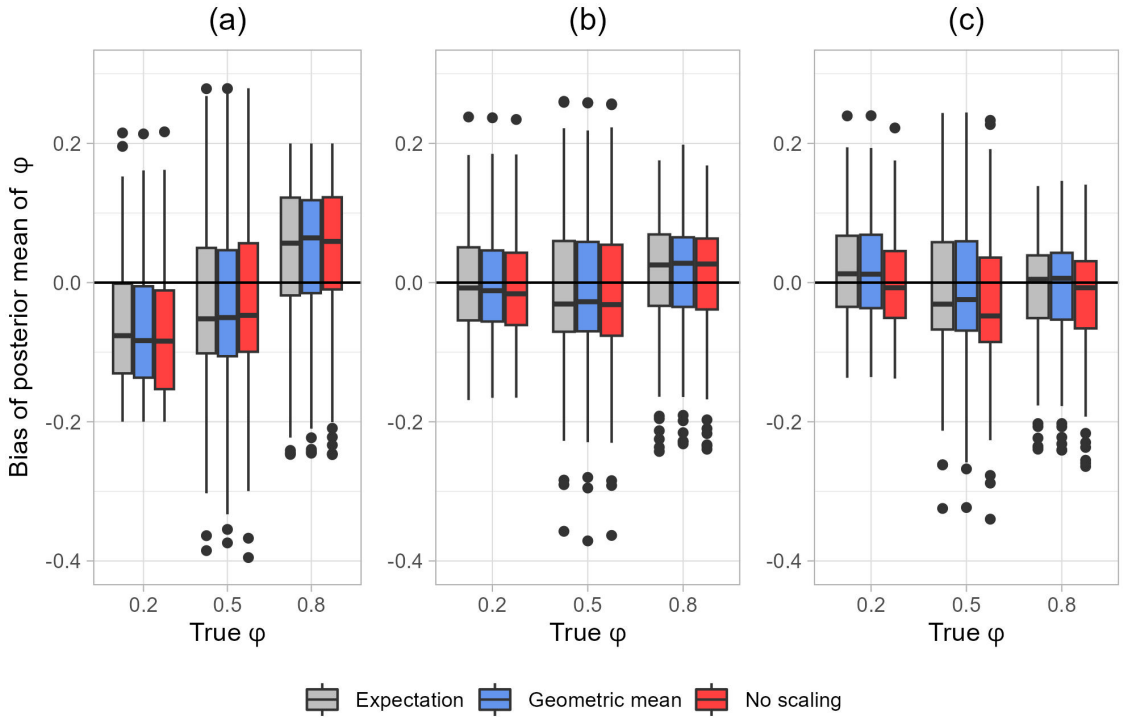

Figure 3.13: Bias of the posterior mean of φ for the BYM model with the following prior choices: (a) IG priors; (b) PC priors; (c) VP prior with $V \sim \text{PC}_0(U = 3, \alpha = 0.05)$.

In terms of performance, the posterior estimates are less impacted by the prior choice in this scenario, but the VP prior has an advantage in the $\varphi = 0.8$ scenario, both for φ itself as well as for T (see Figure 3.14).

In conclusion, we particularly recommend implementing the scaling procedure when working with spatio-temporal data with no repeated measurements for location, as the bias caused by non-scaling is likely to be less negligible because of the larger role played by the prior in these contexts. Moreover, the importance of scaling has proven higher when a VP prior is used, while the state-of-the-art method for a traditional i.i.d. prior specification (i.e. PC) is found to be more robust to the presence of heterogeneous effects. This confirms that, although scaling is always necessary in theory, VP priors suffer more severely from misinterpretation due to non-scaling, since not only the individual marginal priors are distorted but also the joint dependence structure. As such, we advise against the use of VP prior specifications without proper scaling of the effects. Finally, the VP priors provide the best posterior estimates for both φ and T among the three prior choices.

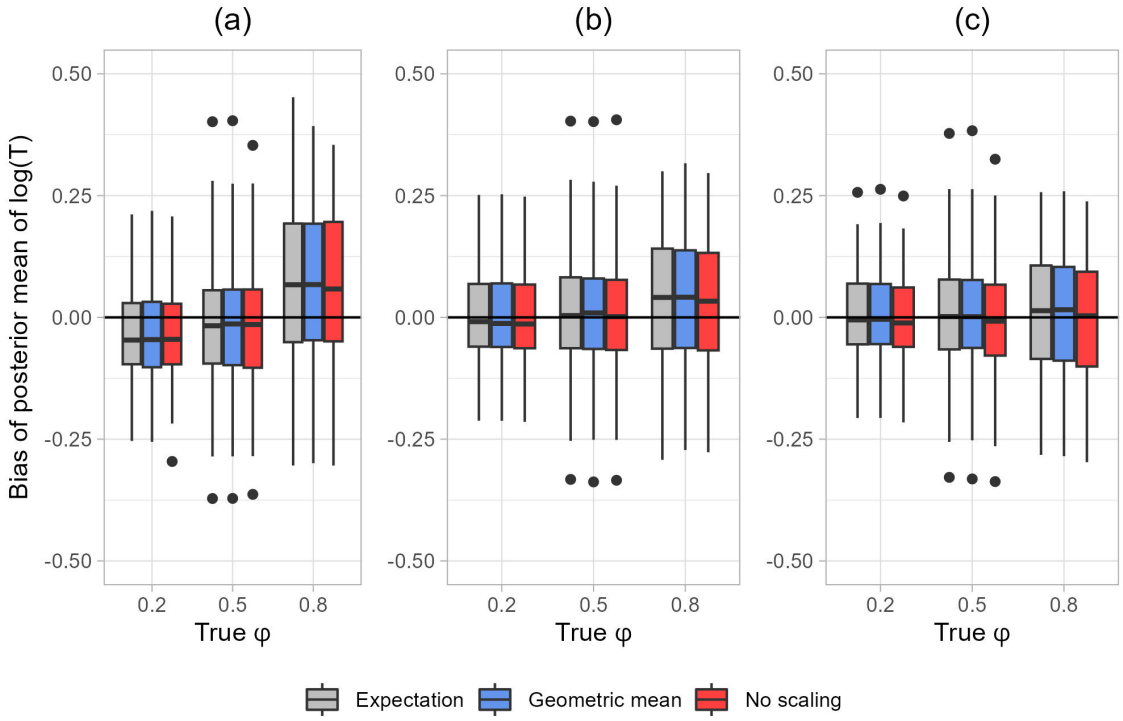

Figure 3.14: Bias of the posterior mean of T in log scale for the BYM model with the following prior choices: (a) IG priors; (b) PC priors; (c) VP prior with $V \sim \text{PC}_0(U = 3, \alpha = 0.05)$.

3.5.3 Impact of Q modification

As discussed in Section 3.3.6, if the scale parameter of an IGMRF effect is to be interpreted as its variance contribution (as in the VP framework), it is necessary to separate into a polynomial and a residual component. Note that this procedure is only necessary when the interpretation of the scale parameter is relevant, such as for example for a more intuitive prior specification.

In order to correctly separate into a polynomial trend and a residual part, we have proposed the Q modification procedure for generic basis choice. If the procedure is not applied, this creates potentially an identifiability issue between the effect with an IGMRF prior and the separate polynomial trend. This problem arises for example in the case of a P-Spline (Equation 3.48).

Here, the importance of the Q modification in practice is tested through a simple simulation study. 200 datasets are generated for a Gaussian response, whose linear predictor includes a linear effect $f_t(X)$ and a non-linear effect $f_r(X)$ (with null linear

trend). For $i = 1, ..., 300$, we generate $X_i \sim \text{Unif}(0, 1)$ and:

$$\begin{aligned} Y_i &\sim N(\eta_i, \sigma_\epsilon^2) \\ \eta_i &= f_t(X_i) + f_r(X_i) \end{aligned}$$

where:

$$\begin{aligned} f_t(X_i) &= (X_i - 0.5)\sqrt{12} \cdot \beta \\ f_r(X_i) &= \cos(2\pi X_i). \end{aligned}$$

The proportional contribution of the non-linear effect to the variance in the linear predictor is denoted by φ and defined as:

$$\varphi = \frac{Var_X[f_r(X)]}{Var_X[f_t(X)|\beta] + Var_X[f_r(X)]}$$

The parameters are set to $\beta = \sqrt{0.5}$ and $\sigma_\epsilon^2 = 1$, where the value of β is chosen so that $\varphi = 0.5$.

The response is then fitted using a model containing a P-Spline effect from Equation 3.48 with $K = 10$:

$$\begin{aligned} Y &\sim N(\eta, \sigma_\epsilon^2) \\ \eta &= \mu + f_t(X) + f_r(X) \\ f_t(X) &= \frac{X - E[X]}{\sqrt{Var[X]}} \cdot \beta \\ f_r(X) &= \boldsymbol{B}_K^T(X)\boldsymbol{u} \\ \beta|\sigma_t^2 &\sim N(0, \sigma_t^2). \end{aligned}$$

Two different priors are compared on the coefficients $\boldsymbol{u}$, i.e. the traditional second-order random walk precision matrix versus its modified version:

- $\boldsymbol{u} \sim N\left(\mathbf{0}, \frac{\sigma_r^2}{C}\boldsymbol{Q}_{\text{RW2}}^*\right)$ subject to $\boldsymbol{S}_{(1)}^T\boldsymbol{u} = \mathbf{0}$.
- $\boldsymbol{u} \sim N\left(\mathbf{0}, \frac{\sigma_r^2}{C}\widetilde{\boldsymbol{Q}}^*\right)$ subject to $\widetilde{\boldsymbol{S}}^T\boldsymbol{u} = \mathbf{0}$ with $\widetilde{\boldsymbol{Q}}$ and $\widetilde{\boldsymbol{S}}$ as in Example 5.

Note that the scaling procedure is applied to each model according to its corresponding constant ($C = 1.432$ for $\boldsymbol{Q}_{\text{RW2}}$ and $C = 0.835$ for $\widetilde{\boldsymbol{Q}}$).

Finally, three different prior specifications are used to fit the models:

(a) IG priors: $\sigma_t^2, \sigma_r^2, \sigma_\epsilon^2 \sim \text{IG}(1, 5e-5)$;

(b) PC priors: $\sigma_t^2, \sigma_r^2, \sigma_\epsilon^2 \sim \text{PC}_0(U = 3, \alpha = 0.05)$;

(c) VP prior: $V = \sigma_t^2 + \sigma_r^2 \sim \text{Jeffreys}, \omega = \sigma_t^2/V \sim \text{Unif}(0,1), \sigma_\epsilon^2 \sim \text{IG}(1, 5e-5)$.

The comparison is evaluated using the posterior mean of the linear coefficients $\widehat{\beta} = E[\beta|\boldsymbol{y}]$ and one of the possible estimates for φ defined as:

$$\widehat{\varphi} = \frac{Var_X\left[\boldsymbol{B}^T(X) \cdot \widehat{\boldsymbol{u}}\right]}{Var_X[(X - 0.5)\sqrt{12} \cdot \widehat{\beta}] + Var_X\left[\boldsymbol{B}^T(X)\widehat{\boldsymbol{u}}\right]}$$

where $\widehat{\boldsymbol{u}} = E[\boldsymbol{u}|\boldsymbol{y}]$. From the simulation set-up, we expect $\widehat{\beta} = \sqrt{0.5}$ and $\widehat{\varphi} = 0.5$.

Results are reported in Figure 3.15. First, we can note how the results from prior choice (a) highlight the identifiability issue created by the fact that the P-Spline effect is not constrained to a null linear trend: without the Q modification, the linear contribution is partially or totally absorbed by the P-Spline effect and the quantities displayed in the plots are unable to inform about the role of linear and non-linear components. For the prior choices (b) and (c), the impact of the Q modification is reduced, but an improvement in the estimation of both $\widehat{\beta}$ and $\widehat{\varphi}$ is still visible: the estimates are less biased and less dispersed after the Q modification.

Overall predictive performance is evaluated using a summary of the Conditional Predictive Ordinates (Pettit 1990), which are readily reported by INLA:

$$\text{CPO} = -\sum_{i=1}^{300} \pi(y_i|\boldsymbol{y}_{-i}).$$

CPO is reported in the bottom panels of Figure 3.15: the use of the modified precision matrix does not cause any loss in performance as it returns equivalent goodness-of-fit.

Further simulations with different linear/non-linear ratios in the trend show similar results in terms of the improvements achieved by the Q modification. In the scenario explored above, K has been set to a relative small value to highlight the impact of the Q modification. However, as K grows, the issue that the Q modification addresses becomes less and less relevant and the difference in estimation performance may be considered negligible, e.g. $K > 25$.

3.5.4 Case study: leukaemia in North West England

We consider the dataset analysed by Henderson, Shimakura, and Gorst 2002, already studied in Kneib and Fahrmeir 2007 and Sørbye and Rue 2014. The dataset contains survival times of $N = 1043$ patients, diagnosed with adult acute myeloid leukaemia

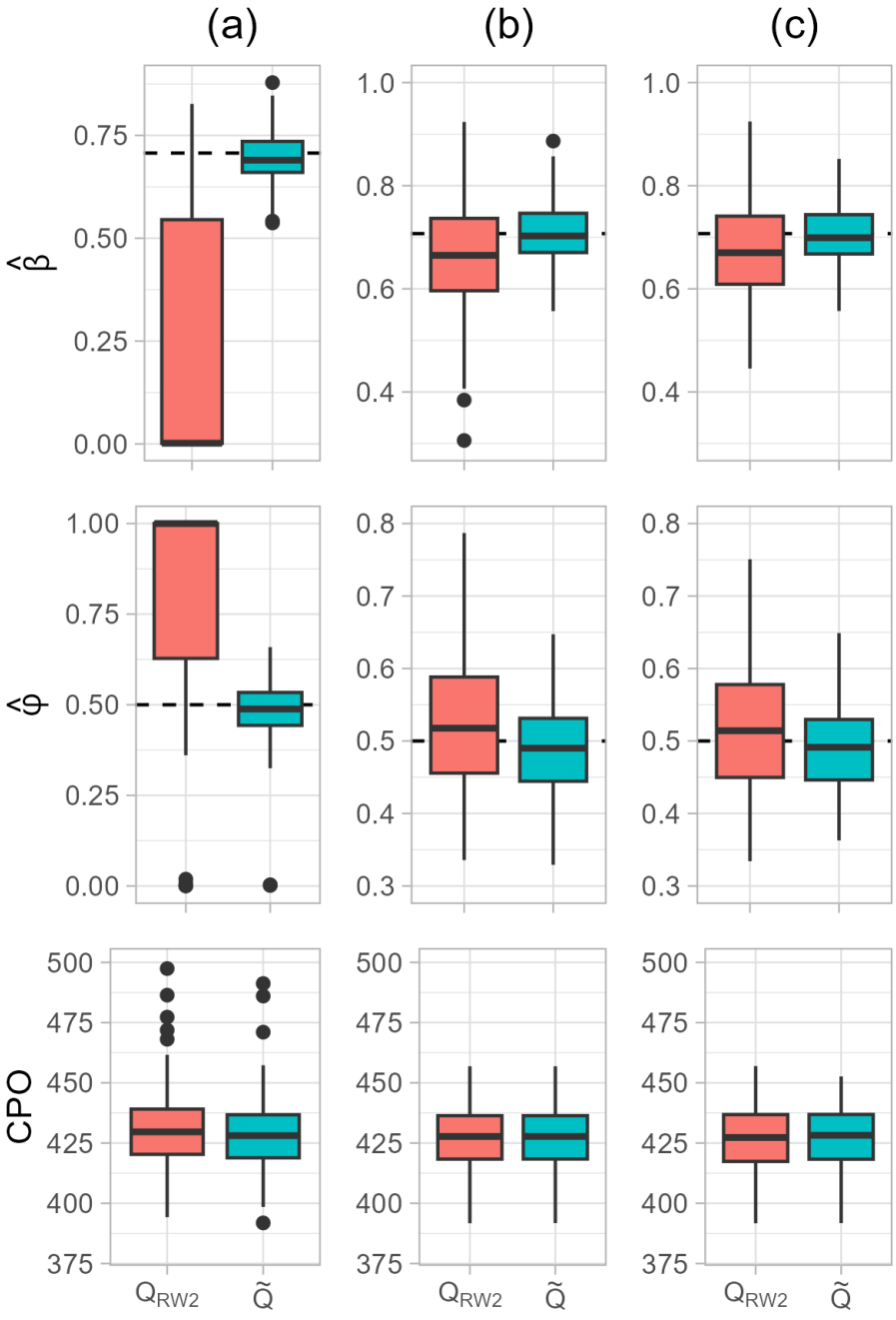

Figure 3.15: Comparison between the use of the original precision matrix $\boldsymbol{Q}_{RW2}$ and the modified version $\widetilde{\boldsymbol{Q}}$. Distribution of $\widehat{\beta}$, $\widehat{\varphi}$ along with their true values (dashed lines), and CPO. Each column represents a different prior choice: (a) IG priors; (b) PC priors; (c) VP prior.

between 1982 and 1998 in the North West England (UK). The following covariates are reported for each patient: Age, Wbc (white blood cells count at diagnosis), Tpi (Townsend social deprivation index), Sex, $District$ (district of residence).Martino, Akerkar, and Rue 2011 illustrated how a survival analysis can be carried out in INLA (i.e. via an LGM) under the assumption of a piecewise log-constant proportional hazard model (Breslow 1972). In this case, the linear predictor of the model (i.e. the log-hazard function) is specified as:

$$\eta = \mu + f_1(Age) + f_2(Wbc) + f_3(Tpi) + f_4(Sex) + f_T(Time) + f_S(District)$$

where $Time$ is a discretization of the survival time in $K_T = 27$ intervals.

The effects $f_1(Age)$, $f_2(Wbc)$, $f_3(Tpi)$ are modelled as P-spline effects with $K = 50$ basis functions (Equations 3.48). We denote by $\sigma^2_{t1}, \sigma^2_{t2}, \sigma^2_{t3}$ the variance parameters of the trend terms, and by $\sigma^2_{r1}, \sigma^2_{r2}, \sigma^2_{r3}$ the parameters of the residual terms. $f_4(Sex)$ is set to a group effect (Example 3.4), while a Besag model (Besag and Kooperberg 1995) based on the adjacency matrix of the districts is used for $f_S(District)$, and a first-order random walk is chosen for $f_T(T)$: these effects are respectively associated with variance parameters $\sigma^2_4, \sigma^2_S, \sigma^2_T$. All effects are treated as fixed, a discrete Uniform distribution is assumed for $Sex, Time, District$, and a continuous one for Age, Wbc, Tpi, on their respective empirical ranges. On the basis of this distributional choice, the necessary 0-mean and null space constraints are imposed on the effects.

In terms of prior, the VP reparameterization from Definition 3.1 is applied to all the 9 variance parameters and a simple HD prior is assumed for convenience: $V \sim$ Jeffreys and $\boldsymbol{\omega} \sim \text{Dir}(1, \dots, 1)$. A more thoughtful prior design could entail, for instance, the use of PC_0 priors on the proportions $\sigma^2_{rp}(\sigma^2_{tp} + \sigma^2_{rp})^{-1}$ for $p = 1, 2, 3$ to penalize non-linearity.

Figure 3.16 reports the posterior mean of the effects for Wbc, Tpi before and after the application of the scaling step of the standardization procedure: lack of appropriate scaling significantly affects the smoothness of the estimated functions, making them more wiggly than necessary. The remaining effects are not significantly affected by scaling. Overall, the scaled solution reports results that are coherent with past analyses of the dataset, suggesting for example a linear trend for both Age and Wbc (Kneib and Fahrmeir 2007). Very similar results are obtained if only the linear effects are standardized (i.e. standardizing the covariates), which can be considered the default approach usually adopted in practice. The geometric mean scaling could not be applied to all the effects in the model due to the presence of the linear ones; however, the results are also very similar when expectation scaling is applied to

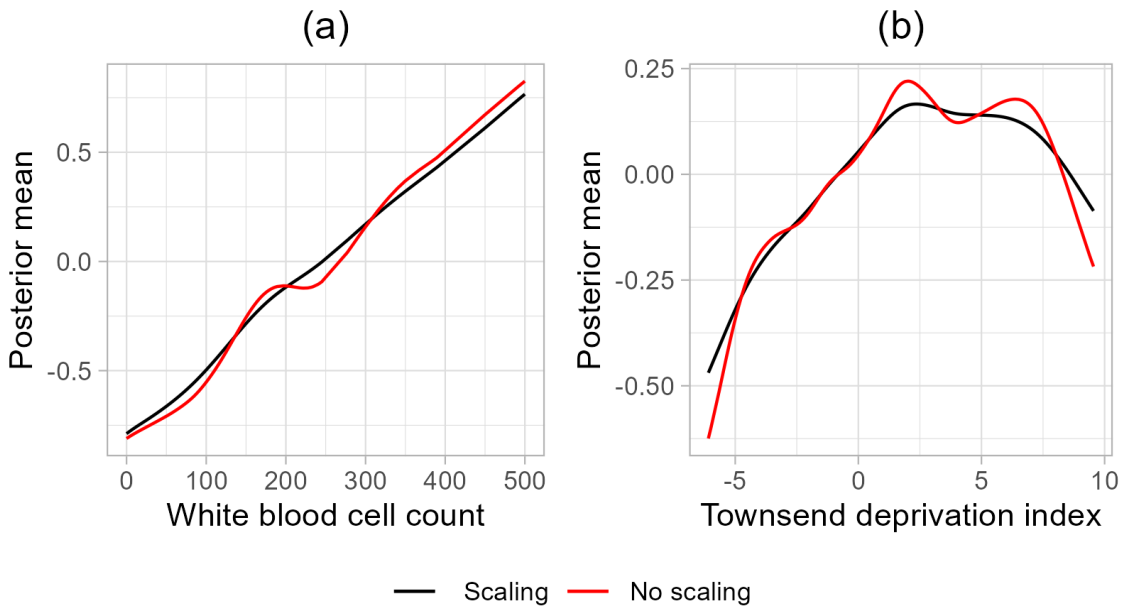

Figure 3.16: Posterior means of (a) $f_2(Wbc)$ and (b) $f_3(Tpi)$ before (red) and after expectation scaling (black).

linear effects and the geometric mean one is used on the remaining effects. Finally, the impact of scaling is greatly reduced if independent PC priors are used instead of the VP prior, proving once more the robustness of the PC framework.

Additional results and R code scripts to replicate the analysis are available at `https://github.com/LFerrariIt/scaleGMRF/tree/master/Leuk_application`.

3.6 Discussion

The motivation behind the work presented in this chapter was the unification into a single framework of two prior specification approaches recently introduced in the literature, namely HD priors (Fuglstad et al. 2020) and R2D2 priors (Zhang et al. 2022). The desire for a single framework comes from the fact that both approaches find their foundation on a reparametrization of the variance parameters of the model, which we have called *variance partitioning* (VP) reparametrization, but that neither of them considers the generic class of LGMs. The VP reparametrization returns a more intuitive set of parameters in a proportion scale (and a single total variance), for which it is easier to specify prior distributions in a way that reflects prior information. We have called VP priors all those priors that can be generated using the VP reparametrization. With the aim of extending the advantages of VP priors to the whole class of LGMs, we have investigated how we can obtain an intuitive interpretation for the variance parameters of a model, such that the user can actually

easily interpret the VP parameters as proportions of variance contributions of the different effects.

In this chapter, a proposal has been made about a formal definition of intuitive interpretation of the variance parameters of LGMs that considers both the individual inferential focus of each effect and the impact of the distribution assumptions made on the effects' covariates. The proposal merges the different point of views present in the VP priors' literature and combines them into a unified, consistent framework that works for all effects of an LGM, both fixed and random ones. Although this contribution was developed in the context of VP priors, we argue that correctly interpreting the variance parameters is fundamental to correctly specify any prior, regardless of the chosen reparametrization.

In order to achieve the goal of obtaining a match between the variance parameters and the intuitive interpretation users have about them, we have proposed a *standardization procedure* that requires a 0-mean constraint for fixed effects and a scaling step for both fixed and random ones. Simulation results proved the practical importance of this standardization procedure, especially of its scaling step. To the purpose of illustrating the procedure and future reference, scaling constants have been either tabulated or analytically derived for a plethora of popular effects, commonly introduced in LGMs. Note how the procedure has been referred to as a generalization of standardization, as it simplifies to the traditional standardization procedure when it is applied to linear effects (i.e. removal of the mean and scaling).

The scaling procedure has been inspired by the work of Sørbye and Rue 2014 but it has been differently derived. In particular, the expectation-based scaling presented in the chapter derives straightforwardly from the formal definition of intuitive interpretation, while the use of the geometric mean in Sørbye and Rue 2014 is justified as being an appropriate location summary for a variance. Moreover, in its original formulation, the geometric mean scaling was used to scale effects of covariates with discrete and finite support, in particular IGMRFs. Although this method can be easily extended to a continuous approach, it would not return a finite scaling constant if the conditional variance viewed as a function of $X = x$ is unbounded over its support, as well as if it is 0 in some points. Thus, the introduction of a probability distribution on the covariate and the shift from geometric mean to expectation are necessary steps (not sufficient, though) to derive a finite scaling constant for many popular continuous processes (e.g. polynomial trends including linear ones). Additionally, thanks to the linearity property of expectation, we are able to obtain clear interpretations for the VP parameters. In practice, however, the difference between the two scaling methods is found to be negligible as in the

instances studied through simulations. Further investigation might be required to assess whether this is always the case.

Finally, it is worth mentioning that our simulations confirm the competitiveness of VP priors against common alternatives, such as independent PC priors on the variance parameters. However, PC priors have been found to be less sensitive to the application of the standardization procedure than VP priors. This result further highlights the benefits of using the increasingly popular PC priors for variance parameters.

With respect to the topic of IGMRFS, we have proposed an alternative representation of these effects through a separation between the polynomial and the residual components, with corresponding σ_t^2 and σ_r^2. The separation is necessary to quantify fully the variance contribution of an IGMRF, which can only be represented using both σ_t^2 and σ_r^2. Additionally, we have also introduced the idea of applying a *Q modification* to the precision matrix of an IGMRF when it is used in combination with complex basis choices, e.g. P-Splines. The Q modification guarantees an equivalent but more convenient way to specify such effects and the relevance of this procedure has been confirmed in practice through simulations. In the specific context of P-Splines, the use of the Q modification returns a useful neat separation between the linear and non-linear contributions. In this context, the motivation for this representation was intuitive prior specification, but other works have made similar proposal in the literature for different reasons (Currie and Durban 2002, Bach and Klein 2024). Hence, we argue that this innovative representation of P-Spline processes has the potential to be relevant in smoothing theory, behind the context of prior specification where it has been developed.

In practical terms, the work in this chapter made possible to correctly deploy the VP approach beyond the original scope of HD and R2D2 priors. This is the main novelty point of this work, which we believe can open the door to the application of VP methods in yet unexplored contexts. In terms of future research, widening the scope of this framework opens the possibility to further study the performance of VP priors, which is particularly interesting in those fields where there is in fact available expert knowledge to be exploited. Secondly, the treatment of the covariates as random poses a new problem by itself and we aim to investigate the impact of the distribution choice $\pi(\boldsymbol{X})$ and the potential bias caused by its misspecification on posterior inference. Finally, the introduction of smooth non-linear effects for continuous regressors (e.g. P-Splines) can help extend the research on the use of VP priors to tackle typical regression problems, such as variable selection, collinearity, confounding, in a beyond-linear framework (see for instance the work of Wei et al.

2020).

Chapter 4

Variance Partitioning priors for Species Distribution Models

4.1 Introduction

Bayesian hierarchical modelling has gained prominence in various scientific domains due to its flexibility in modelling complex relationships, as well as the comprehensive insights derived from posterior inference, which enable robust parameter estimation and uncertainty quantification. These advantages come at the cost of having to specify a prior distribution on the model parameters. This is particularly challenging for variance parameters (Lambert et al. 2005,Gelman 2006).

The VP prior approach presented in Chapter 3 represents an alternative to the usual i.i.d. vague prior set up used for the specification of variance parameters. VP priors allow for the inclusion of prior expert knowledge about the relative importance of different effects in an intuitive manner. This outcome is achieved in two steps using the Hierarchical Decomposition (HD) framework proposed by Fuglstad et al. 2020. First, a reparametrization into a set of more intuitive parameters must be designed using an appropriate decomposition tree, coherent with the information that is actually available to the user. Subsequently, prior distributions and their hyperparameters are set in a way that reflects prior beliefs on these new quantities. The outcome is a joint prior distribution on the original variance parameters that effectively integrates expert knowledge.

The use of VP priors has been found to be competitive when compared to other state-of-the-art alternatives in Hem et al. 2021 and Marques, Wiemann, and Kneib 2023. VP priors can be designed using the HD approach and the `makemyprior` R package developed by Hem, Fuglstad, and Riebler 2024, which provides a graphical user interface for the visualization of the decomposition tree, as well as user-friendly

commands for the imputation of prior beliefs.

The innovative approach of VP priors is anticipated to be particularly valuable in fields with specific characteristics. On the one hand, numerous research areas involve data collected via relatively standard designs, such as multi-level or spatio-temporal structures. In such contexts, VP priors can be readily constructed by developing decomposition trees that acknowledge the model structure (see Section 4.4.1) and adopting a principle of either parsimony or ignorance in prior specification, in lack of more pertinent information. Such applications of the HD approach are found in disease mapping (Franco-Villoria, Ventrucci, and Rue 2022, Riebler et al. 2016), demography (neonatal mortality case study in Fuglstad et al. 2020), forestry (Marques, Wiemann, and Kneib 2023). Other fields with similar characteristics in which the HD approach has not yet been trialled include environmental sciences, such as agriculture applications (latin square design simulation in Fuglstad et al. 2020).

Furthermore, VP priors can be particularly useful in those contexts in which expert knowledge is directly available in the scale of proportions of variance, e.g. disease mapping (Wakefield 2007), genomics (Holand et al. 2013), or ecology (Peres-Neto et al. 2006). Using the direct contribution of expert knowledge, VP priors have so far only been applied to plant breeding data by Hem et al. 2021.

Although the R2D2 literature (Yanchenko, Bondell, and Reich 2024b, Aguilar and Bürkner 2023) has used VP priors in the presence of linear effects, no application has so far exploited the advantages of the HD framework to jointly specify the prior of both the fixed and random effects' variance parameters. Building upon the foundation laid in Chapter 3, we extend the Hierarchical Decomposition framework to encompass the overall variance in the linear predictor, including the contribution of fixed effects. This expansion offers substantial benefits. Primarily, it enables a comprehensive utilization of prior knowledge regarding the relative importance of all effects. Moreover, this approach facilitates the integration of prior beliefs and modeling assumptions through a unified, homogeneous framework.

This chapter demonstrates the advantages of our extended Hierarchical Decomposition approach in the specific context of species distribution models (SDMs), a foundational tool of ecological research used to map species' occurrence (Ovaskainen et al. 2017, Sofaer et al. 2019). The ecological domain is an ideal testing ground for our proposal due to the well-defined model structure of SDMs (Ovaskainen et al. 2017), and the availability of expert knowledge on variance partitioning (Pettit 1990). Additionally, the typical structure of SDMs always entails both a fixed component, accounting for environmental factors, as well as random effects to account

for additional variability due to spatio-temporal correlation or other sampling conditions. This setting allows us to discuss the three main stages necessary for the specification of an HD prior and to come up with default solutions for each step specifically tailored for the context of SDMs:

1. standardization procedure to obtain the correct interpretation of the variance parameters;

2. design of the Hierarchical Decomposition tree;

3. prior specification on the HD reparametrization.

The remainder of the chapter is organized as follows. Section 4.2 is devoted to a formal definition of single species distribution models as LGMs. Section 4.3 then focuses on the application of the standardization procedure to ensure that the variance parameters of an SDM accurately reflect their intuitive interpretations. We then extend the Hierarchical Decomposition (HD) approach to SDMs in Section 4.4, providing recommendations for tree design and prior specification based on a review of relevant applications of VP priors in the literature. Finally, Section 4.5 presents a method to perform variance partitioning estimation a posteriori that aligns with the prior framework established in the previous sections and highlights the advantages of this proposal over a more traditional approach. Section 4.6 concludes the chapter with a discussion.

The proposal is exemplified through a real-world case study on the dataset provided by Hui et al. 2023, which reports presence/absence data for 39 fish species in the North Atlantic from the NOAA-NEFSC survey (*NEFSC Fall Bottom Trawl Survey* 2024). The complex SDM required by the nature of this dataset gives us the opportunity to highlight and discuss some of the critical challenges that may arise in practice during the specification of a VP prior. The theoretical concepts presented in each section are immediately followed by their practical application to the case study.

4.2 Species Distribution Models (SDMs)

Ecology can be described as the scientific field that studies the abundance and the distribution of species (Begon and Townsend 2020). Species are often not studied individually, but rather within their community. A community is made up by the living organisms that reside in a certain area (Ovaskainen et al. 2017). The main research questions in the field of community ecology include the following:

- *"are species associated to habitat characteristics?"*
- *"does species' occurrence show spatial or temporal pattern unexplained by changes in the habitat?"*
- *"what are the drivers of the variability observed in the occurrence of a species?"*

These questions show the interest of community ecology in estimating variance contributions.

The type of data available to ecologists to answer these and other questions usually consists of a response Y, capturing occurrence of a species. Occurrence can be recorded in terms of presence-absence, count of individuals or percentage estimate of biomass (Ovaskainen et al. 2017). Along with the response, the location and time of observation $(\boldsymbol{Z}, T)$ are usually available, along with a set of environmental covariates $(\boldsymbol{X})$, collected because assumed to be associated with occurrence. Additional information about species' traits and phylogeny can be combined to the observational data to answer related research questions but this scenario is not discussed here.

In order to neatly answer the research questions of interest, occurrence is often formally modelled using a so-called *species distribution model* (SDM). This type of models assumes that the distribution of species' occurrence, over an area of interest and a given period of time, is the result of the combination of 3 different factors:

- **Abiotic factors.** Environmetal covariates, such as habitat characteristics, play an essential role in the determination of species' occurrence (Elith and Leathwick 2009).
- **Biotic factors.** This concept includes all within and between-species interactions (e.g. predation, competition, and mutualism) that may cause additional variability in the observed occurrence. These interactions are often taken into account by modelling the residual spatio-temporal correlation, still present in the data after having controlled for environmental covariates (Dormann et al. 2012, Araújo and Luoto 2007).
- **Stochastic processes.** Additional processes (e.g. such as ecological drift or environmental stochasticity) can affect occurrence in a way that cannot be explained by any environmental covariate or spatio-temporal correlation. These processes introduce variability that is often challenging to quantify but is critical for understanding the full range of factors influencing species distributions (Elith and Leathwick 2009).

SDMs can be used individually to map the distribution of single species. In terms of fitting, the Bayesian implementation is often favoured in ecology due to the immediate uncertainty quantification, which is particularly useful to accompany the estimation of quantities such as contributions to the variation of different factors.

4.2.1 SDMs as Latent Gaussian Models

Consider Y being a random variable measuring occurrence of a single species, distributed according to an Exponential family distribution with location parameter η and additional parameters $\boldsymbol{\psi}$: the stochasticity component enters through the choice of this distribution.

$$Y \sim Dist(\eta, \boldsymbol{\psi}).$$

Let $\boldsymbol{X} = [X_1, ..., X_P]$ be a set of environmental covariates assumed to be associated with Y and $\boldsymbol{Z} = [Z_1, Z_2]$, and T represent the spatio-temporal location of Y. The supports of each of the variables is denoted by calligraphic letters, such that $X_p \in \mathcal{X}_p,\ p = 1, ..., P$, $T \in \mathcal{T}$,$\boldsymbol{Z} \in \mathcal{Z}$. Each support is assumed to be bounded.

In an SDM, the location parameter η is specified as a function of $\boldsymbol{X}, \boldsymbol{Z}, T$. A generic SDM can be written as:

$$\eta = \mu + \sum_{j=1}^{J} f_j(\boldsymbol{X}) + \sum_{l=1}^{L} f_l(\boldsymbol{Z}, T). \tag{4.1}$$

This model class is able to accommodate main effects for each of the covariates $X_1, ..., X_P$ (linear or smooth non-linear effects), as well as a spatial and temporal effect, but also interactions between the different environmental variables, a spatio-temporal interaction, additional unstructured noise, etc. For simplicity, we further assume that each $f_j(\boldsymbol{X})$ can be at most a function of two of the P environmental covariates, e.g. $f(X_1, X_2)$. This formulation does not account for potential interactions between covariates and spatio-temporal locations, which significantly complicates the model and are anyway rarely included in SDMs.

The $f_j(\cdot)$ and the $f_l(\cdot)$ functions can be conveniently approximated choosing to adopt the LGM framework. We adopt here the same notation of Model 1 from

Chapter 3:

$$\begin{aligned} \eta &= \mu + \sum_{j=1}^{J} \boldsymbol{D}_j^T(\boldsymbol{X})\boldsymbol{u}_j + \sum_{l=1}^{L} \boldsymbol{G}_l^T(\boldsymbol{Z}, T)\boldsymbol{v}_l \\ \boldsymbol{u}_j|\sigma_{A,j}^2 &\sim N(\boldsymbol{0}, \sigma_{A,j}^2 \boldsymbol{Q}_{A,j}^*) \qquad j = 1, ..., J \\ \boldsymbol{v}_l|\sigma_{B,l}^2 &\sim N(\boldsymbol{0}, \sigma_{B,l}^2 \boldsymbol{Q}_{B,l}^*) \qquad l = 1, ..., L. \end{aligned} \tag{4.2}$$

A stands for *Abiotic* factors and B for *Biotic* ones. Note that in the case of rank-deficient precision matrices, the null space constraints must be imposed on the corresponding random coefficients (see Section 3.2.3 of Chapter 3).

Under the choice to adopt the LGM framework, the VP approach can then be used to introduce prior knowledge or assumptions about the relative importance of all these effects, through a joint prior on the scale parameters.

4.2.2 Case study: NOAA-NEFSC fall bottom trawl survey data

In what follows, we use a community ecology dataset, which consists of a subset of the NOAA-NEFSC fall bottom trawl survey (*NEFSC Fall Bottom Trawl Survey* 2024), processed and studied in Hui et al. 2023 and made publicly available online at `https://github.com/fhui28/CBFM` (last access: October 2024). This particular case study was selected because of the relatively large number of species, i.e. $N_{\text{species}} = 39$. The main research question consists in the quantification of the variance contributions of the different model components, i.e. environmental factors and residual spatial and temporal contributions.

The dataset contains $N = 5892$ observations collected over the U.S. Northeast continental shelf, collected during the fall season of each year from 2000 to 2019; the spatial locations are unique. The occurrence of $N_{\text{species}} = 39$ different demersal fish species is collected in the form of a presence/absence response Y. Figure 4.1 reports some exploratory plots for an overview of the data.

The following environmental covariates are also part of the dataset: *Surface temperature* (X_1), *Bottom temperature* (X_2), *Surface salinity* (X_3), *Bottom salinity* (X_4), *Depth* (X_5). In order to control for sampling conditions, a binary covariate is also reported to represent the type of *Survey vessel* (X_6) that collected the data: this variable is treated as an additional environmental condition, for a total of $P = 6$ covariates. The spatial location is reported using the UTM coordinate system ($\boldsymbol{Z} = [Z_1, Z_2]$). The time of the observation is precisely reported , but here we choose to use the year as the T covariate, to acknowledge that the data is only collected during the fall period of each year and not continuously.

The dataset has been originally used to illustrate a new approach to Joint Species Distribution Models (JSDMs) (Warton et al. 2015), i.e. all the data is modelled jointly using a multivariate structure. In this setting, instead, each of the species' occurrence will be treated marginally and fitted using an individual, although identical, SDM (*stacked* SDM approach, see Guisan and Rahbek 2011). The design of the common SDM has been inspired by the joint model proposed in Hui et al. 2023: separability between the spatial and temporal effect and non-linear effects of the continuous environmental covariates are assumed, based on exploratory analysis from in Hui et al. 2023. Avoiding an index for species for convenience, each response Y is assumed to follow a Bernoulli distribution with parameter $g^{-1}(\eta)$ where $g(\cdot)$ is the logit link and the linear predictor η is defined as:

$$\eta = \mu + \sum_{p=1}^{6} f_p(X_p) + f_S(\boldsymbol{Z}) + f_T(T). \tag{4.3}$$

The model is then specified as an LGM. In particular, $f_p(X_p)$ will be specified as a P-Spline effect for the 5 continuous environmental covariates, i.e. $p = 1, ..., 5$ (see Example 5 of Chapter 3); $f_6(X_6)$ will be specified as a simple i.i.d. effect; $f_S(\boldsymbol{Z})$ will be specified as a two-dimensional P-Spline effect with an IGMRF of order 1 prior on the coefficients (see Example 6 of Chapter 3); $f_T(T)$ will be specified as a first-order random walk. The model thus implicitly contains the following variance parameters:

$$\sigma_1^2, ..., \sigma_6^2, \sigma_S^2, \sigma_T^2.$$

Note that P-Splines were chosen over the use of random walks for irregular locations, which represents a popular alternative commonly implemented in INLA (Lindgren and Rue 2008). The reason for this choice lies in the fact that the design of the latter model depends on the observed values of the covariates in the data: hence, the standardization procedure will be different in each application. On the contrary, low-rank smoothers such as P-Splines are more convenient, as the specification of the basis and the precision matrix does not depend on the data (at most, only on its range): as a consequence, we can precompute the appropriate modified precision matrix and the scaling constants for all applications using these effects without the need for the case-specific data.

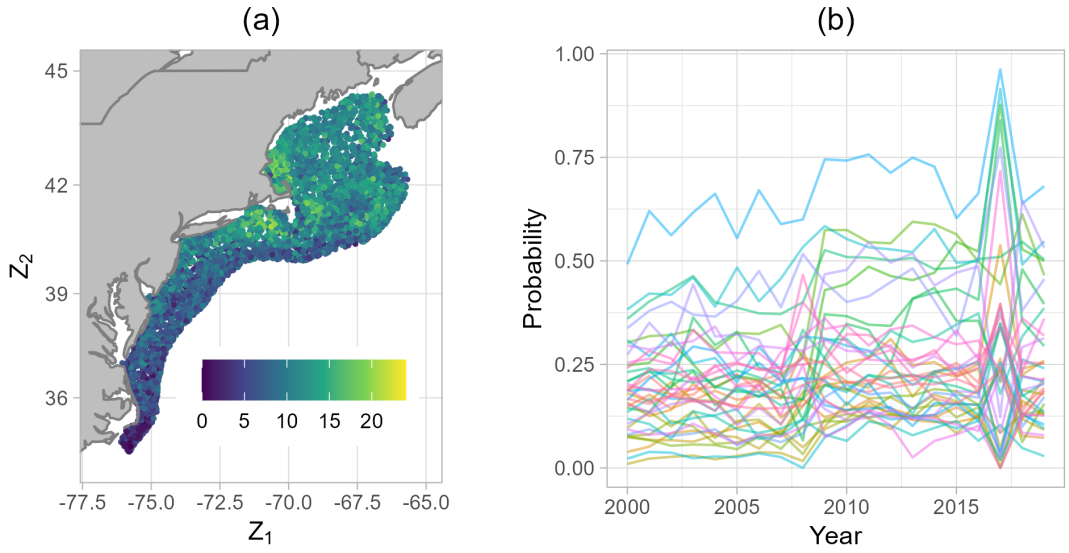

Figure 4.1: Overview of the case study dataset: (a) number of different species detected in each location; (b) proportions of presence observations for each of the 39 species over time.

4.3 Interpretation of the σ^2 parameters in SDMs

The correct application of a VP prior is guaranteed by the application of the standardization procedure presented in Chapter 3, which ensures that each of the scale parameters matches its intuitive interpretation, defined as the variance contribution of the corresponding effect as intended by the user. Definition 3.3 provides a different definition of intuitive interpretation for fixed and random effects, as described in Section 3.3.2.

In the context of SDMs, the first J effects from Equation 4.2 can be generally categorized as fixed, since environmental variables are usually expected to be directly responsible for affecting occurrence and the estimated trends are interesting to experts, to study the correlation between species' characteristics and their habitat. The model then includes L effects included to capture the residual spatio-temporal correlation, still present after having accounted for environmental conditions. In this case, it is not immediate to categorize these effects as either fixed or random: the categorization should be made based on the specific application at hand and research questions. For convenience, we consider here the case in which the spatio-temporal effects can be categorized as random ones, which is appropriate when the actual realizations of such effects are not of direct interest. Thus, we can define the vector of parameters of interest for SDMs as $\boldsymbol{\theta} = [\boldsymbol{u}_1, ..., \boldsymbol{u}_J, \sigma^2_{B,1}, ..., \sigma^2_{B,L}]$.

According to Definition 3.3, the intuitive interpretation associated to the $\sigma^2_{A,j}$

parameters is therefore the *expected finite-population variance*, i.e.:

$$\sigma^2_{A,j} = E_{\boldsymbol{u}_j}\{Var_{\boldsymbol{X}}[f_j(\boldsymbol{X})|\boldsymbol{u}_j]|\sigma^2_{A,j}\} \quad j = 1, ..., J$$

which is guaranteed under a 0-mean constraint on the corresponding effects, followed by appropriate scaling. A 0-mean constraint can be guaranteed by imposing the appropriate linear constraint $E_{\boldsymbol{X}}[\boldsymbol{D}_j^T(\boldsymbol{X})\boldsymbol{u}_j|\boldsymbol{u}_j] = 0$ (Proposition 2). The scaling step is then satisfied computing the scaling constant $C_{A,j} = Var_{\boldsymbol{X},\boldsymbol{u}_j}[f_j(\boldsymbol{X})|\sigma^2_{A,j} = 1]$ (Proposition 1) and multiplying the precision matrix of the random coefficients by this constant (alternatively, dividing the basis matrix by $\sqrt{C_{A,j}}$):

$$\boldsymbol{u}_j|\sigma^2_{A,j} \sim N\left(\boldsymbol{0}, \sigma^2_{A,j}\frac{\boldsymbol{Q}^*_{A,j}}{C_{A,j}}\right).$$

On the other hand, the intuitive interpretation for the $\sigma^2_{B,l}$ parameters is the *super-population variance*. To satisfy the intuitive interpretation requirement for these random effects, scaling is sufficient to obtain:

$$\sigma^2_{B,l} = Var_{\boldsymbol{Z},T,\boldsymbol{v}_l}[f_l(\boldsymbol{Z}, T)|\sigma^2_{B,l}] \quad l = 1, ..., L.$$

The appropriate scaling constants are $C_{B,l} = Var_{\boldsymbol{Z},T,\boldsymbol{v}_l}[f_l(\boldsymbol{Z}, T)|\sigma^2_{B,l} = 1]$, $l = 1, ..., L$.

If the spatio-temporal effects are introduced with IGMRF priors on the coefficients (at it happens in the case study), the null space constraints that must be imposed on the process imply a 0-mean constraint. Hence, the effects respect the 0-mean requirement even though it is not necessary for the interpretation of their σ^2 parameters if they are considered random.

4.3.1 Distribution assumption on the covariates

Since both the scaling procedure and the 0-mean constraints rely on the distributional choice on the covariates, the joint distribution $\pi(\boldsymbol{x}, \boldsymbol{z}, t)$ plays a crucial role in determining the actual interpretation of the scale parameters. A priori, this choice is important to correctly define the meaning of *variance contribution* of an effect, according to the experts' interest. A posteriori, this distribution determines the estimation of the variance contributions, and thus it influences subsequent conclusions.

For the standardization procedure from Chapter 3, it is sufficient to specify only the joint distribution of the covariates that make up a single effect, while the dependence structure between covariates that do not share an effect is irrelevant

and independence can be assumed for simplicity. Thus, the choice of $\pi(\boldsymbol{x}, \boldsymbol{z}, t)$ can actually simplify to $\pi(\boldsymbol{x})\pi(\boldsymbol{z}, t)$ for the class of models from Equation 4.2, which can be further simplified to $\prod_{p=1}^{P} \pi(x_p)\pi(\boldsymbol{z})\pi(t)$ when interaction terms are absent from the model.

The simplest distributional assumption that can be made consists in using the empirical distribution as it is, i.e. $\pi(\boldsymbol{x}_i, \boldsymbol{z}_i, t_i) = 1/N$. This choice can be reasonable when the data can come from survey studies, in which the sampling design has been chosen by researchers (Hayward et al. 2015, Burgazzi et al. 2020). Under this choice, the experts must be reminded at the prior specification stage to state their beliefs only considering the sampling locations, or at most a broader population they think is well represented by their sampling design. However, the use of the empirical distribution is discouraged whenever the dataset is not generated from a predefined sampling design (e.g. fish and wildfile national registries (*U.S. Fish and Wildlife Service Open Data* n.d.) or citizen science projects (Sullivan et al. 2009,*iNaturalist* 2024): in this scenario, it is unlikely that the empirical distribution is representative of the distribution the experts are interested in, because of sampling bias.

A safer choice might consist in the selection of a Uniform distribution over a spatio-temporal support containing all the data points, and a Uniform distribution for each $\pi(x_p)$ over reasonable ranges of values for each environmental covariate. This choice is convenient for different reasons. First, a Uniform distribution provides a simple intuition to the user about the meaning of the concept of variance contribution, as it simplifies this quantity to the variance of the trend within the support of interest, and the user does not have to take into account the probability distribution. Secondly, using a Uniform distribution also simplifies the computation of the necessary 0-mean constraints, scaling constants, and of the Q modification procedure, useful for example in the case of P-Splines effects.

Despite the advantages of choosing Uniform distributions, it is still necessary to estimate their parameters, namely the range extremes $[m, M]$. The choice of these ranges is a challenging point, since they greatly affect the final quantification of the variance contributions. In theory, experts can be questioned about what range is most of interest for them, or simply what values can be reasonably selected as extrema based on previous information. In practice, statistical tools can facilitate this choice through a direct estimation from the available data. While the sample range is the simplest and most intuitive estimator, there are more robust alternatives that are less susceptible to extreme values, based for example on quantile ranges or variance estimation. We argue in favour of the sample range, unless high leverage points can be detected, as they can be particularly detrimental to the estimation of

the variance contribution.

To understand the potential impact of high-leverage points, Panel (a) of Figure 4.2 shows a toy example in which $N = 25$ points have been generated for $X \sim \text{Uniform}(0, 1)$, but the last one has for some reason been reported as $x_N = 3$. The N^{th} observation has of course a high leverage value ($h_N = 0.729$), which is much larger than the mean leverage at ($\bar{h} = 0.04$). While in this example, the value is known not to come from the same distribution as the rest of the sample, in practice we can choose to either remove or keep high-leverage points for the estimation of the sample range. Panel (b) of Figure 4.2 reports the density of the Uniform distribution over the two possible ranges: the blue area is obtained removing the high-leverage point, while the red one results from keeping it. Not removing high-leverage points can be detrimental in the definition of variance contributions a priori, and of course for their posterior estimation.

This is true regardless of whether or not the point is also highly influential. Figure 4.3 explores two scenarios in which the effect of x on y is modelled through a linear effect, either with or without the high-leverage point. In the first scenario (a), the point x_N is also highly influential in the estimation of the linear regression, as we can see that the estimation of the linear trend greatly changes when the point is considered. As the estimate for the linear coefficient changes, this directly impacts the estimation of the variance contribution of x. The second scenario (b) instead displays the case of a point which is not highly influential, i.e. the estimation of the linear coefficient and trend are almost invariant before and after the inclusion of the point. However, Panel (b) represents how the estimation of the variance contribution is nevertheless greatly affected by the inclusion of the point, as is the linear trend: when the point is included, the linear trend goes on much higher and thus the variance due to x is estimated at a much higher value.

This simple toy example illustrates how high-leverage points are dangerous when the interest is in the quantification of variance contributions, regardless of their influence. As such, we suggest removing them before the estimation of the Uniform distribution extremes through the sample range. After the specification, points falling outside the range must either be removed or treated as missing and their value imputed with the closest possible values within the range, i.e. the minimum or the maximum. This removal/imputation step is necessary to ensure that the data are coherent with the chosen distributional assumption $\pi(\boldsymbol{x}, \boldsymbol{s}, t)$, but more importantly to avoid that these high leverage points influence the estimation of the coefficients (which should only depend on the data in the range of interest), and consequently affect the estimation of the corresponding variance contributions.

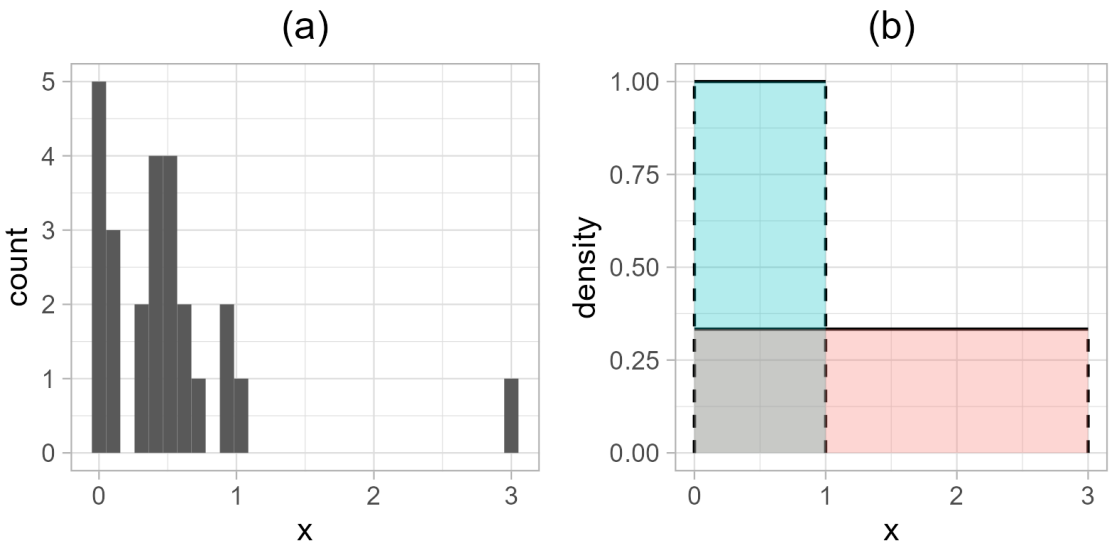

Figure 4.2: Distribution of $\boldsymbol{x}$ points from the toy example: (a) histogram; (b) representation of the density of the estimated Uniform distribution before (red) and after (blue) having removed the high-leverage point $x_N = 3$.

In conclusion, we recommend combining expert knowledge, external data sources, and statistical tools (e.g. leverage), to come up with reasonable ranges for each of the covariates individually.

4.3.2 Case study: interpretation of the σ^2 parameters

Consider again Equation 4.3:

$$\eta = \mu + \sum_{p=1}^{6} f_p(X_p) + f_S(\boldsymbol{Z}) + f_T(T).$$

This model implicitly contains the following variance parameters: $\sigma^2_1, ..., \sigma^2_6, \sigma^2_S, \sigma^2_T$. In order to correctly specify their prior, regardless of the chosen reparametrization, it is necessary to first obtain that each σ^2 actually represents the variance contribution of the corresponding effect, whose definition varies according to whether an effect is treated as fixed or as random (see Chapter 3). In this case, we assume that the effects of the environmental covariates are fixed, since the interest lies in the trends, while the spatial and temporal effects can be treated as random, since the focus is on the estimation of the variance rather than the realized trends. Thus, we want

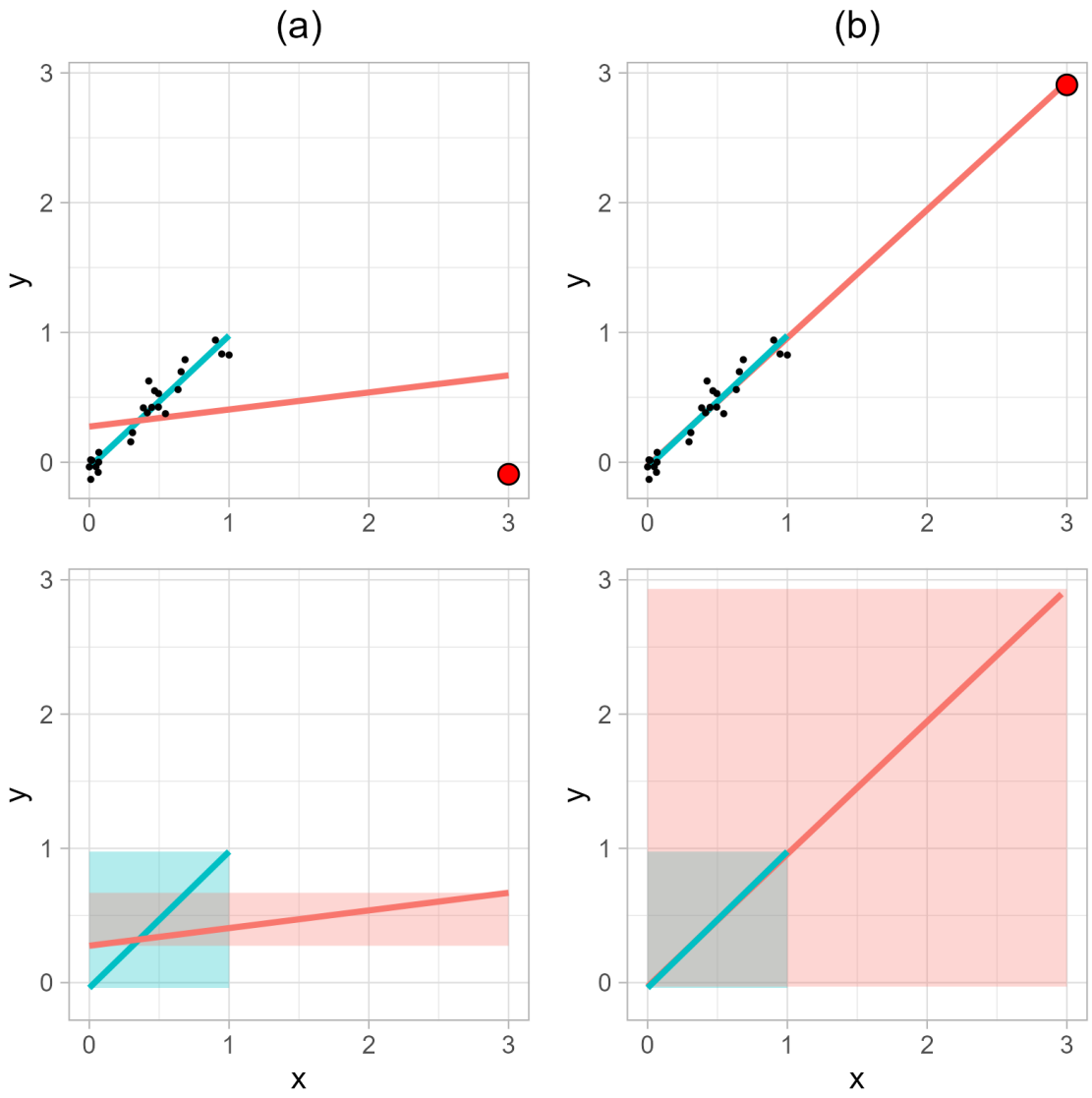

Figure 4.3: Linear regression with (red) and without (blue) the high-leverage point at $x_N = 3$, represented by the red dot. The bottom plots informally represent via shaded areas the estimates of the variance contributions over the chosen $\pi(x)$. (a) The high-leverage point is highly influential. (b) The high-leverage point is not highly influential.

.

that the following equations are true before moving on to prior specification:

$$\sigma_p^2 = E_{\boldsymbol{u}_p}\{Var_{X_p}[f_p(X_p)|\boldsymbol{u}_p]|\sigma_p^2\} \qquad p = 1, ..., 6 \tag{4.4}$$

$$\sigma_S^2 = Var_{\boldsymbol{Z}}[f_S(\boldsymbol{Z})|\sigma_S^2] \tag{4.5}$$

$$\sigma_T^2 = Var_T[f_T(T)|\sigma_T^2]. \tag{4.6}$$

In the following subsections, we detail and justify the specification of each effect, as well as the choice of the marginal probability distributions of the associated covariates, which influence the procedure necessary to guarantee Equations 4.4-4.6.

Environmental covariates' effects

The effects of the 5 continuous environmental covariates are modelled using P-Splines (Lang and Brezger 2004, Eilers and Marx 1996), in order to accommodate for possible smooth non-linear effects. In particular, we choose to use $K_X = 20$ basis functions and an IGMRF prior of order 2 on the coefficients. From the discussion of Example 5 in Chapter 3, we know that the interpretability of the variance parameters for these effects require multiple steps, all relying on the chosen distribution for the corresponding covariate. In this case, we choose to assume a Uniform distribution for all the covariates. This choice is motivated by both convenience and interpretability. First, using the same distribution is an advantage, as the procedure to correctly introduce the P-Spline effects (i.e. Q modification and scaling) becomes the same for all covariates. Additionally, Uniformity guarantees an interpretational advantage for the user: when a Uniform is chosen, the variance with respect to the covariate is simply the variance of the trend measured in all location over the support, which is more intuitive for a user than to having to take into account different probability densities at each point.

Thus, considering $X_p \sim \text{Unif}(m_p, M_p)$, each $f_p(X_p)$ is defined as:

$$\begin{aligned} f_p(X_p) &= f_{Lp}(X_p) + f_{Np}(X_p) \\ f_{Lp}(X_p) &= \frac{X_p - E[X_p]}{SD[X_p]}\beta_p \\ f_{Lp}(X_p) &= \boldsymbol{B}^T(X_p)\boldsymbol{u}_p \end{aligned}$$

where:

$$\begin{aligned} &\beta_p|\sigma^2_{Lp} \sim N(0, \sigma^2_{Lp}) \\ &\boldsymbol{u}_p|\sigma^2_{Np} \sim N\left(\mathbf{0}, \sigma^2_{Np}\frac{\widetilde{\boldsymbol{Q}}^*_X}{C_X}\right) \text{ subject to } \widetilde{\boldsymbol{S}}^T_X\boldsymbol{u}_p = \mathbf{0} \end{aligned}$$

and $\boldsymbol{B}(X_p)$ is the B-Spline basis defined on equidistant knots on the interval $[m_p, M_p]$, $\widetilde{\boldsymbol{Q}}_X$ is the modified precision matrix from Example 5 of Chapter 3, and $\widetilde{\boldsymbol{S}}_X$ is its null space of dimension $K_X \times 2$.

Alternatively, the effects $f_p(\cdot)$ can be expressed as:

$$\begin{aligned} f_p(X_p) &= \sqrt{\sigma_p^2}[\sqrt{1-\omega_{Np}}\widetilde{f}_{Lp}(X_p) + \sqrt{\omega_{Np}}\widetilde{f}_{Np}(X_p)] \\ &= \sqrt{\sigma_p^2}[\sqrt{1-\omega_{Np}} \cdot \frac{X_p - E[X_p]}{SD[X_p]}\widetilde{\beta}_p + \sqrt{\omega_{Np}} \cdot \boldsymbol{B}^T(X_p)\widetilde{\boldsymbol{u}}_p] \end{aligned}$$

where:

$$\begin{aligned} &\widetilde{\beta}_p \sim N(0,1) \\ &\widetilde{\boldsymbol{u}}_p \sim N\left(\mathbf{0}, \frac{\widetilde{\boldsymbol{Q}}_X^*}{C_X}\right) \text{ subject to } \widetilde{\boldsymbol{S}}_X^T \boldsymbol{u}_p = \mathbf{0}. \end{aligned}$$

The effect is guaranteed to respect the 0-mean constraint since $\widetilde{\boldsymbol{S}}_X^T \boldsymbol{u}_p = \mathbf{0} \implies E_{X_p}[f_p(X_p)|\boldsymbol{u}_p] = 0$. Finally, $C_X \approx 13.33$ guarantees that:

$$\sigma_{Lp}^2 + \sigma_{Np}^2 = E_{\boldsymbol{u}_p}\{Var_{X_p}[f_p(X_p)|\boldsymbol{u}_p]|\sigma_p^2\} \quad p = 1, ..., 5.$$

See the proof in Section A.5 in the Appendix.

Hence, Equation 4.4 is satisfied if we define the overall contribution of the $f_p(X_p)$ effect by $\sigma_p^2 = \sigma_{Lp}^2 + \sigma_{Np}^2$.

The final point that must be discussed is the choice of $[m_p, M_p]$. Figure 4.4 shows how the empirical distribution of the covariates compares to a Uniformity assumption over the chosen ranges. The empirical ranges are used for all the covariates, after having excluded points deemed as of high leverage (3 in total): these extreme values are replaced with the closest non-excluded point, i.e. either the maximum or minimum of the corresponding range. The chosen ranges are coherent with the knowledge available about the behaviour of these environmental covariates over the U.S. Northeast continental shelf (*NEFSC Fall Bottom Trawl Survey* 2024).

Although it is not necessary to specify the joint distribution for the environmental covariates, we report here some considerations about their dependence structure. Figure 4.5 shows that the covariates are correlated with each other, as to be expected: in particular, the surface/bottom couples are correlated, and the bottom covariates are in turn correlated with the water depth. These correlations might affect posterior estimates, so we shall consider this issue in the analysis of the results.

Survey vessel type effect

The covariate X_6 that indicates the type of vessel used to collect the data is coded such that its support is $[-1, 1]$, with each of two outcomes assigned equal proba-

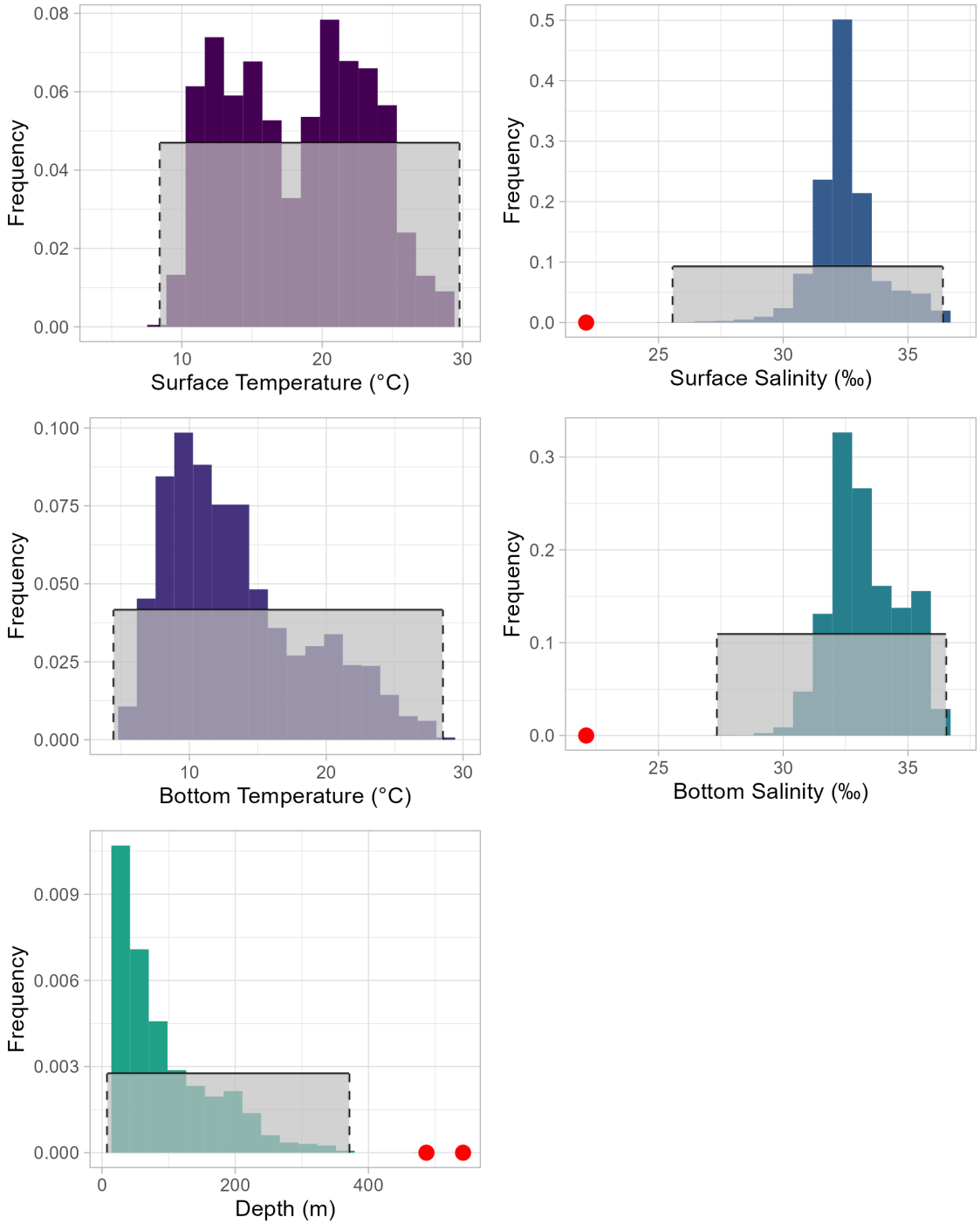

Figure 4.4: Histogram of the 5 continuous environmental covariates, along with the corresponding Uniform distribution (grey shaded area) used as $\pi(x_p)$. The red dots represent 3 observations found to have extreme leverage: 2 points present extremely high values of *Depth*, while a single observation presents abnormally low values of both *Surface salinity* and *Bottom salinity*.

bility 0.5. As such, $E[X_6] = 0, Var[X_6] = 1$ so that the covariate is automatically

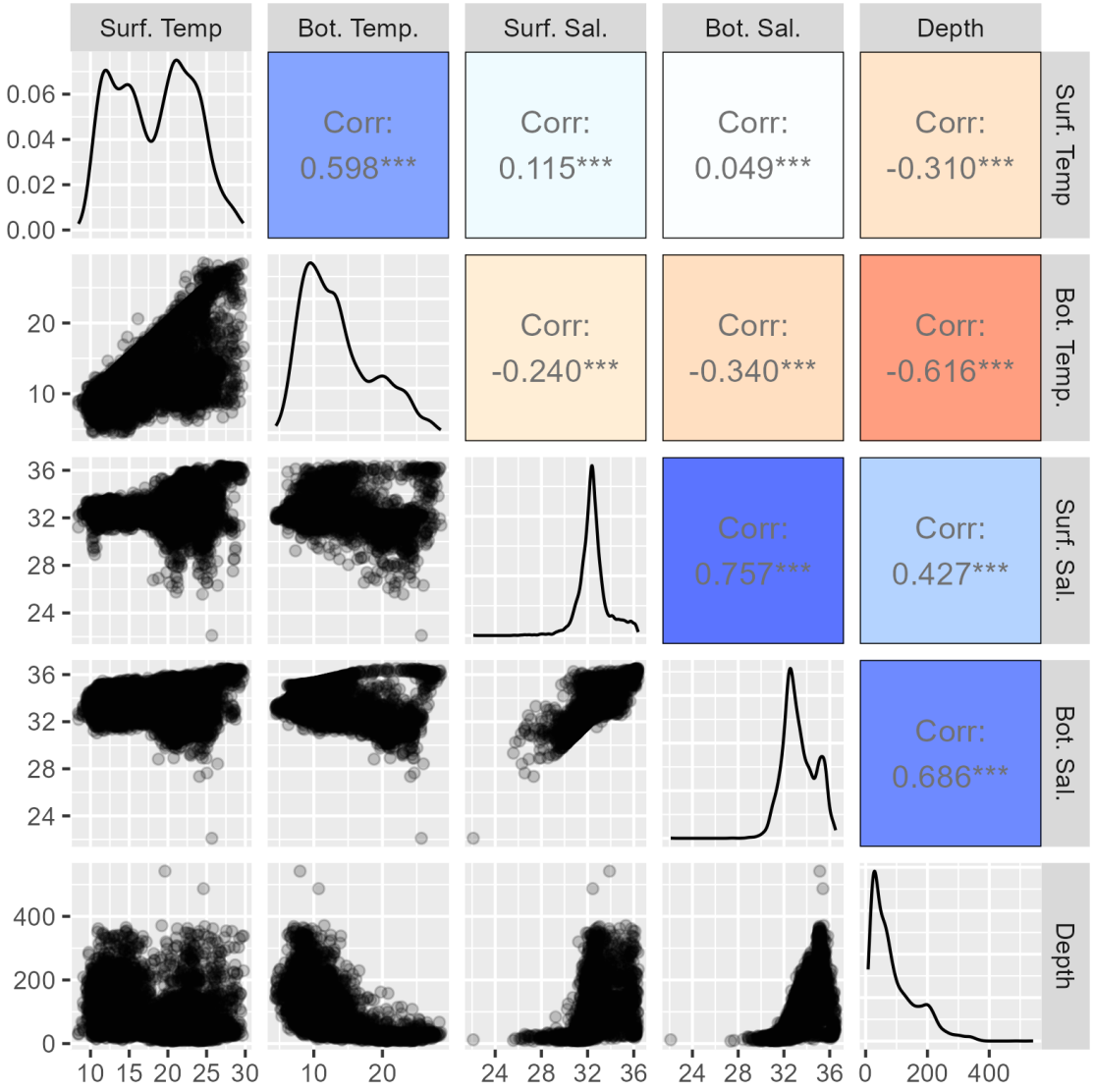

Figure 4.5: Scatterplot and Pearson's correlations between the 5 environmental covariates.

standardized. We then assume a linear effect on X_6:

$$f_6(X_6) = X_6\beta_6$$
$$\beta_6|\sigma_6^2 \sim N(0, \sigma_6^2)$$

which guarantees Equation 4.4 for $p = 6$.

Spatial effect

For the spatial effect, we start by considering what could be a reasonable distributional assumption for the $\boldsymbol{Z}$ coordinates. The locations in the sample all belong to

a specific area of the North Atlantic ocean called U.S. Northeast continental shelf. As it can be seen from Figure 4.6 (a), the area has not been uniformly sampled, with some regions more thoroughly sampled than others. As discussed in Section 4.3.1, it is convenient to replace the empirical distribution with a Uniform one, which equally distributes importance all over an area of interest, and thus allows for a more intuitive definition of variance contribution attributable to the spatial component. In this application, we opt for this alternative and define the *area of interest* as a polygon that includes all locations found in the sample. The polygon is designed to cover all the observed spatial locations and found using a concave hull algorithm (Gombin, Vaidyanathan, and Agafonkin 2020): the resulting shape is represented in Figure 4.6 (a) as the shaded green area. To quickly approximate desired quantities with respect to this distribution, a large sample of points equally distributed over the polygon area is then generated.

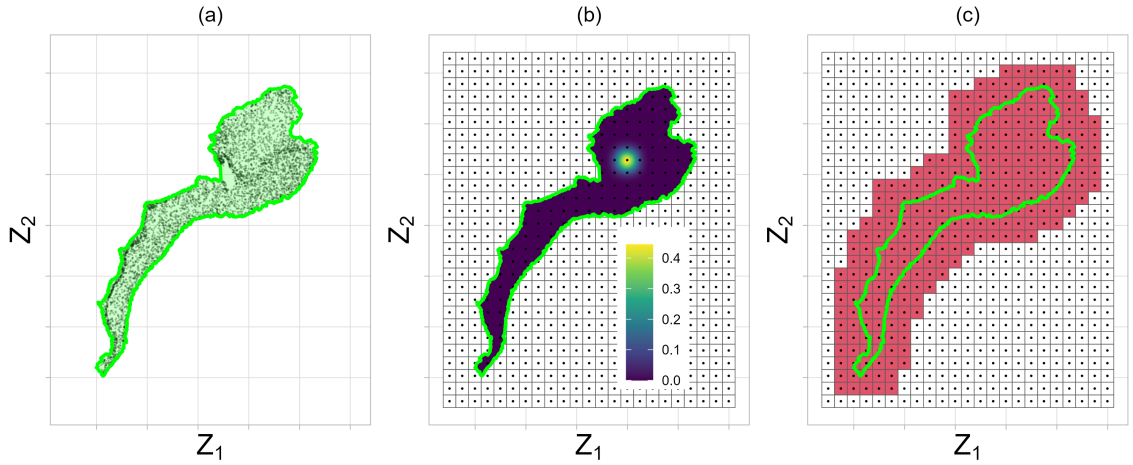

Figure 4.6: (a) Representation of the sampling locations over the Z_1, Z_2 spatial coordinates system with black dots, along with the chosen concave polygon containing all data points, outlined and shaded in green. (b) 50x50 km grid with cell centroids (black dots), along with the value of one B-Spline basis function centered in a grid cell over the area of interest. (c) The red grid cells indicate that the corresponding basis functions (centered in their centroids) have non-null values over the polygon represented by the solid green line.

.

Once the distribution for $\boldsymbol{Z}$ has been well defined, the spatial effect can be then specified and scaled accordingly. Here, we choose two-dimensional P-splines (Lang and Brezger 2004, Fahrmeir, Kneib, and Lang 2004) to model the variability over the continuous spatial coordinates. In the dataset, the distance between locations highly varies so that multiple resolutions could be used to model the spatial trend. Here we

choose to capture a fairly large-scale spatial pattern creating a grid of 50x50 km cells and centering a 2D B-Spline in each cell. A two-dimensional B-Spline basis can be simply created using the Kronecker product between two univariate B-Spline basis (in this case both with 50 km distance between basis functions). Figure 4.6 (b) shows the grid and the values of one of the basis functions over the polygon of interest. With respect to the precision matrix, we could simply use an IGMRF of first-order derived considering the overall rectangular grid. However, this method would be appropriate only if we actually want to model the whole rectangular surface, as in Example 6 of Chapter 3. In this case, however, the interesting area is only the irregular polygon. Thus, we decide to remove all basis functions that have null values all over the polygon area. The grid cells whose corresponding basis functions have been retained ($K_S = 267$) are colored in red in Figure 4.6 (c). From this red grid or lattice, we can retrieve an adjacency matrix $\boldsymbol{W}$ that reports the neighbours of each remaining basis function. $\boldsymbol{W}$ is used to create a precision matrix $\boldsymbol{Q}_S$, which corresponds to an IGMRF of first order over the irregular lattice highlighted in Figure 4.6 (c):

$$\boldsymbol{Q}_S = \text{diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}.$$

The model can therefore be defined as:

$$\begin{aligned} f_S(\boldsymbol{Z}) &= \boldsymbol{B}^T(\boldsymbol{Z})\boldsymbol{v}_S \\ \boldsymbol{v}_S|\sigma_S^2 &\sim N\left(\mathbf{0}, \sigma_S^2 \frac{\boldsymbol{Q}_S^*}{C_S}\right) \text{ subject to } \boldsymbol{S}_S^T\boldsymbol{v}_S = \mathbf{0} \end{aligned}$$

where $\boldsymbol{B}(\cdot,\cdot)$ is the bivariate B-Spline created as detailed above, and $\boldsymbol{S}_S$ is the null space of $\boldsymbol{Q}_S$.

Similar to what happens in Example 6 of Chapter 3, however, the constraint $\boldsymbol{S}_S^T\boldsymbol{v}_S = 0$ does not imply $E_{\boldsymbol{Z}}[f_S(\boldsymbol{Z})|\boldsymbol{v}_S] = 0$ but only $\sum_{k=1}^{K_S} v_{S,k} = 0$. This means that the effect is subject to a constraint that is not directly interpretable in the scale of the spatial coordinates, which is inconvenient and creates an identifiability issue with the intercept parameter μ. Following the discussion of Section 3.3.6 of Chapter 3, we would rather replace the precision matrix with a modified version $\widetilde{\boldsymbol{Q}}_S$ with null space $\widetilde{\boldsymbol{S}}_S$ such that $\widetilde{\boldsymbol{S}}_S^T\boldsymbol{v}_S = 0 \implies E_{\boldsymbol{Z}}[f_S(\boldsymbol{Z})] = 0$, which is more convenient even though not strictly necessary to guarantee Equation 4.5.

The design of an appropriate $\widetilde{\boldsymbol{Q}}_S$ starts from finding explicitly the desired null space, which in this case is $\widetilde{\boldsymbol{S}}_S = E_{\boldsymbol{Z}}[\boldsymbol{B}(\boldsymbol{Z})]$. From Equation 3.28 of Section 3.3.6,

we know that $\widetilde{\boldsymbol{Q}}_S$ can be found as:

$$\widetilde{\boldsymbol{Q}}_S = (\boldsymbol{\Lambda}\widetilde{\boldsymbol{R}}^*\boldsymbol{\Lambda})^*$$

where $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda}) > 0$ and $\widetilde{\boldsymbol{R}}$ must be defined as in Equation 3.54 to be a function of the entries of $\boldsymbol{Q}_S, \widetilde{\boldsymbol{S}}_S, \boldsymbol{\Lambda}$ (see the proof in Section A.12 of the Appendix). In Chapter 3, it is recommended to set the vector $\boldsymbol{\lambda}$ equal to $\widehat{\boldsymbol{\lambda}}$ such that:

$$\widehat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}>\mathbf{0}}{\arg\min}\ D_{\text{KL}}\ (\ \mathcal{N}_{\widetilde{\boldsymbol{Q}}_S(\boldsymbol{\lambda})}\ ||\ \mathcal{N}_{\boldsymbol{Q}_S}\ ).$$

This optimization step is more costly in real-world applications than in the examples of Chapter 3. For instance, this case study requires the optimization of a vector of relative large dimension, i.e. 267 parameters. Additionally, the distribution on $\boldsymbol{Z}$ is not Uniform on a rectangular surface. This entails an additional challenge in comparison to Example 6, where reasonable constraints can be imposed during the optimization of $\boldsymbol{\lambda}$ such that they reflect the symmetrical structure of the desired null space (whose entries can be derived analytically): such constraints can greatly reduce the effective number of parameters to be optimized and thus cut the optimization time. This is not possible in this case study, where the entries of $\widetilde{\boldsymbol{S}}_S$ have been numerically approximated and do not display symmetrical patterns due to the irregular support of $\pi(\boldsymbol{z})$. As a consequence, the optimization remains more involved and consuming, and it might be argued whether implementing the optimization is actually useful or can be avoided. To highlight the importance of the optimization step, we compare the solution under the optimal $\widehat{\boldsymbol{\lambda}}$, which we call $\widetilde{\boldsymbol{Q}}_{\hat{\boldsymbol{\lambda}}}$, with a solution $\widetilde{\boldsymbol{Q}}_1$ where the vector is arbitrarily set to $\boldsymbol{\lambda} = \mathbf{1}$ and numerical optimization is not deployed. Figure 4.7 compares the diagonal entries, i.e. the marginal variances, of three different covariance matrices: (a) the original $\boldsymbol{Q}_S^*$; (b) the modified version without optimization $\widetilde{\boldsymbol{Q}}_{\mathbf{1}}^*$; (c) the modified version after the optimization $\widetilde{\boldsymbol{Q}}_{\hat{\boldsymbol{\lambda}}}^*$. We can conclude that the cost of the optimization is worthwhile, since the non-optimized version (b) is a completely inadequate approximation of the original model (a), while the numerically optimized solution (c) effectively reconstructs the original variance pattern.

Finally, the matrix $\widetilde{\boldsymbol{Q}}_S = \widetilde{\boldsymbol{Q}}_{\hat{\boldsymbol{\lambda}}}$ can be used to redefine the spatial effect:

$$\begin{aligned} &f_S(\boldsymbol{Z}) = \boldsymbol{B}^T(\boldsymbol{Z})\boldsymbol{v}_S \\ &\boldsymbol{v}_S|\sigma_S^2 \sim N\left(\mathbf{0}, \sigma_S^2\frac{\widetilde{\boldsymbol{Q}}_S^*}{C_S}\right) \text{ subject to } \widetilde{\boldsymbol{S}}_S^T\boldsymbol{v}_S = \mathbf{0} \end{aligned} \tag{4.7}$$

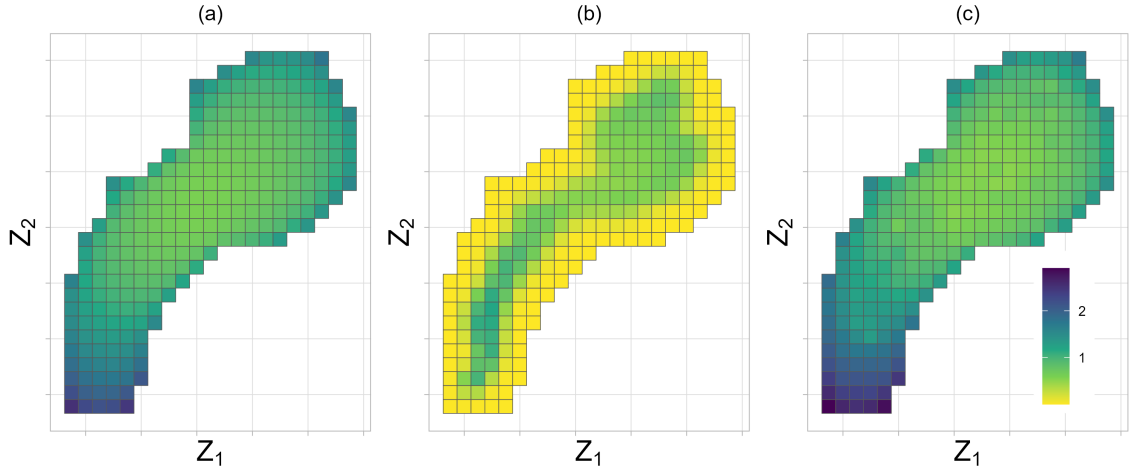

Figure 4.7: Each of the grid cells containing a basis function is colored by the value of the marginal variance of its corresponding coefficient, found as the diagonal values of the generalized inverse of the precision matrix: (a) $\boldsymbol{Q}_S^*$; (b) $\widetilde{\boldsymbol{Q}}_{\mathbf{1}}^*$; (c)$\widetilde{\boldsymbol{Q}}_{\widehat{\boldsymbol{\lambda}}}^*$.

where the scaling constant necessary to guarantee Equation 4.5 is found to be $C_S \approx 0.582$.

In the practical implementation, the basis matrix is redefined as $\boldsymbol{\Lambda B}(\cdot,\cdot)$ and the precision matrix as $\widetilde{\boldsymbol{R}}$, so that all the matrices involved preserve their sparsity structure.

Temporal effect

A random walk effect is introduced for the T covariate, which takes $K_T = 20$ different values representing distinct years (2000-2019). Thus, the effect can be written as:

$$f_T(T) = \sum_{m=1}^{20} \mathbb{I}(T = m) v_{Tm}$$
$$\boldsymbol{v}_T | \sigma_T^2 \sim N\left(\mathbf{0}, \sigma_T^2 \frac{\boldsymbol{Q}_T^*}{C_T}\right) \text{ subject to } \boldsymbol{S}_T^T \boldsymbol{v}_T = \mathbf{0}$$

where $\boldsymbol{Q}_T$ is the precision matrix of a first-order random walk on 20 equidistant locations (Equation 2.7) and its null space $\boldsymbol{S}_T = \mathbf{1}$. A reasonable distribution for T gives equal probability to each value, such that the variance contribution assigns equal importance to each of the years in the sample. Under this assumption, the null space constraint $\boldsymbol{S}_T^T \boldsymbol{v}_T = \mathbf{0}$, which is necessary for the interpretability of variance parameters of IGMRF effects, corresponds to a 0-mean constraint $E_T[f_T(T)] = 0$. Hence, a Q modification is not necessary in this case. Finally, specifying the scaling

constant to $C_T = 3.325$ guarantees Equation 4.6.

4.4 Hierarchical Decomposition approach for SDMs

Once each of the variance parameters match their intuitive interpretation, VP priors can be used on LGM species distribution models to include a wide range of prior assumptions in an intuitive manner.

In the VP approach, the prior distribution is specified on the following reparametrization of the original $J + L$ parameters:

$$\begin{aligned} V &= \sum_{j=1}^{J} \sigma^2_{A,j} + \sum_{l=1}^{L} \sigma^2_{B,l} \\ \boldsymbol{\omega} &= \left[\frac{\sigma^2_{A,1}}{V}, ..., \frac{\sigma^2_{A,J}}{V}, \frac{\sigma^2_{B,1}}{V}, ..., \frac{\sigma^2_{B,L}}{V} \right]. \end{aligned} \tag{4.8}$$

Further reparametrizations of the random vector $\boldsymbol{\omega}$ (although, often conveniently written also involving V) can be used to introduce specific knowledge about the relative importance of different effects. This is achieved using the Hierarchical Decomposition (HD) approach by Fuglstad et al. 2020, which consists of two steps: design of the reparametrization through a variance decomposition tree; prior specification on the new parameters such that it reflects prior beliefs. Ideally, both steps should be considered on a case-by-case basis such that prior information is optimally exploited. However, it is often the case that expert knowledge is not as precise or strictly pertinent to the case study as expected. It is more reasonable to assume that many assumptions about the relative importance of effects come from modelling principles, such as parsimony, or from general ecology theory.

In what follows, we discuss the two steps of the HD approach for a generic SDM as defined in Equation 4.2. We propose a default decomposition tree that exploits the relatively fixed structure of this class of models (as suggested by Fuglstad et al. 2020) and makes use of the tree design principles that can be derived from the HD literature. We believe that the reparametrization provided by this default tree will already represent an undoubtful improvement in terms of intuitiveness for ecologists called to specify a prior for SDMs. Subsequently, we also discuss potential prior choices for the parameters coming from each split of tree, again on the basis of choices made in applications from other fields.

4.4.1 Default HD tree

The reparametrization of $\boldsymbol{\omega}$ from Equation 4.8 into proportions that have a direct intuition for the experts is designed with the help of a *tree* in the HD approach. The tree decomposes V (its root node) through successive splits, each with two or more branches, until all leaf nodes contain a single σ^2. The new parameters are found dividing the elements in the child nodes of a split by the sum of the elements in its parent node. By design, the parameters are therefore simplices and sole reparametrizations of $\boldsymbol{\omega}$.

According to Fuglstad et al. 2020, the *"tree structure must be selected so that the desired comparisons can be made"*: this means that the resulting proportion parameters should measure the relative importance between groups of effects for which the user has a direct intuition. The tree design should therefore be in theory application-specific and expert-driven.

However, it is often the case that the available prior information is not directly pertinent to the case study at hand, but rather comes from the combination of general field knowledge and adoption of modelling principles (e.g. simplicity). For this reason, we propose a default decomposition tree for the generic SDM of Equation 4.2, which can represent a baseline for users that wish to implement the HD approach with minimum effort. The aim of the design is to derive new parameters for which ecologists have a broad intuition, regardless of the application at hand (e.g. proportional contribution to the variance of environmental factors). The best tree design for this goal is here created combining the theory behind SDMs, modelling principles, and considerations from the HD literature so far. Such tree can then be tweaked at will, whenever the user possesses more sophisticated information.

From the applications of the HD framework presented so far in the literature, we retrieve a few principles about tree design that can guide us in building a reasonable default choice for SDMs:

(a) ***The initial splits of a tree (i.e. the ones closer to the root node) usually separate groups of similar effects, either in terms of covariates or structure, into separate branches***. This is for example found in the tree design for the latin square design application found in Fuglstad et al. 2020, where the first split distinguishes between residual effect and all the other effects related to a covariate (also found in the neonatal mortality application), while the second split distinguishes between the effect of the treatment on the response versus the plot design effect. A further example is found in Hem et al. 2021, where phenotypic values are modelled using genomic information and the tree design starts with a split between genetic

and environmental factors. Since the branches of the split have a clear distinct interpretation, there might be more precise prior information about this type of split than for others.

(b) ***Secondary splits are often used to divide a group of similar effects (e.g. all linear effects) using multiple branches.*** This type of split is found in the R2D2 literature, where all linear effects are divided into singletons at the first split. Additional examples include the work of Marques, Wiemann, and Kneib 2023 where the tree design for multiple spatio-temporal effects only comprises of a single multi-branch split that treats all random effects equally. Because of the similarity, it might be the case that exchangeable priors are sufficient to expresse the user's prior beliefs about these splits.

(c) ***Binary splits are often used to divide two effects, functions of the same covariates, but with different levels of flexibility***. Some obvious cases include separating an additive interaction versus a non-additive one (Franco-Villoria, Ventrucci, and Rue 2022,Hem et al. 2021), an unstructured versus a structured effect (Riebler et al. 2016, treatment effect split in the latin square design case study of Fuglstad et al. 2020), higher-level cluster effect versus a nested-level one (county versus cluster effect of the neonatal mortality case study in Fuglstad et al. 2020). In this type of splits, the principle of model simplicity or parsimony would recommend a preference towards the branch containing the effect providing less flexibility. If there are more than two effects with different levels of flexibility, subsequent binary splits can be used in place of the multiple-branch split. (Fuglstad et al. 2020)

On the basis of principles (a)-(c), we build a default decomposition tree for a general SDM represented in Figure 4.8. The graph shall be read such that each node corresponds to the sum of all the variance parameters σ^2 corresponding to the effects described in the label of the node.

Level 1

The first split is inspired by principle (a). The total variance in the linear predictor can be partitioned into the two main sources of variability recognized by the SDM theory: the contribution of abiotic/environmental effects; the contribution of the additional variability, often still spatio-temporally structured. The first child node contains all the $\sigma^2_{A,j}$, while the second one contains all the $\sigma^2_{B,l}$. The new parameter coming from this binary split can be denoted by $\omega_A = \sum_{j=1}^{J} \omega_j$, which represents the proportion of total variance explained by the effects related to the $\boldsymbol{X}$

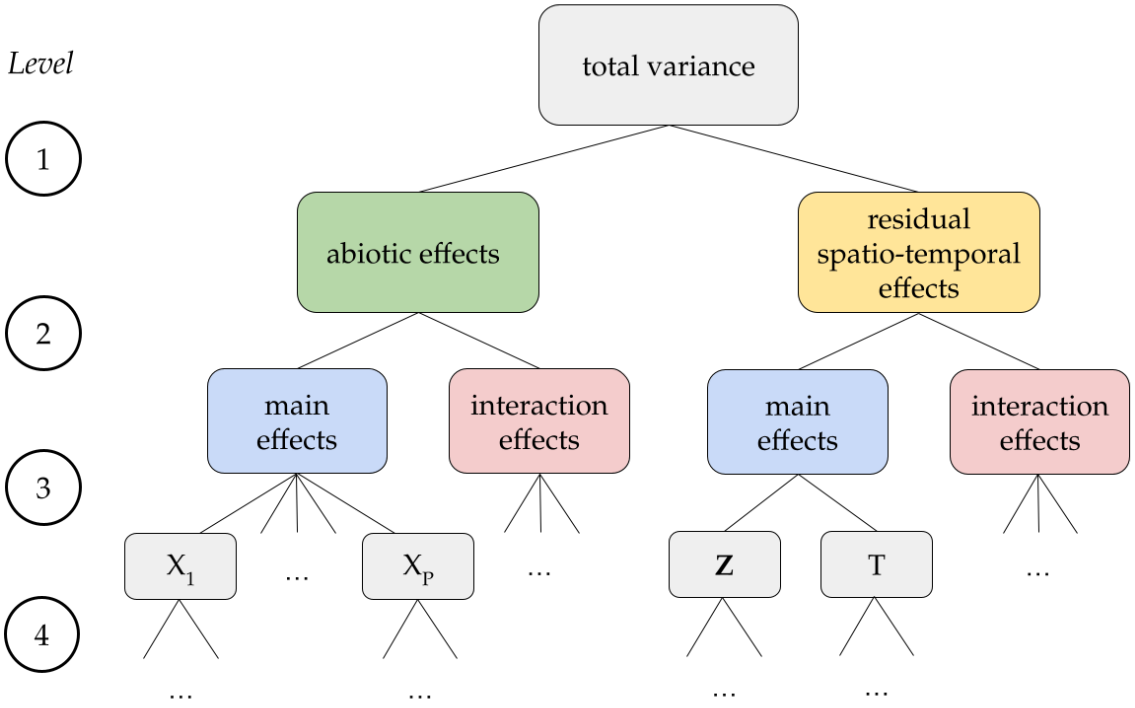

Figure 4.8: Default decomposition tree for SDMs. Each node represents the sum of all the variance parameters σ^2 from the corresponding effects.

covariates. It is often the case that the environmental covariates present a spatio-temporal structure, which may lead to confounding problems. This split allows to include prior assumptions about this phenomenon or check potential confounding problems a posteriori by looking at the posterior of ω_A. In the future, we aim at further investigating how the HD approach may help identify potential confounding phenomena.

Level 2

At the second level, we envision two splits, one for each of the two child nodes from the previous step. In both cases, the main effects (functions of individual covariates) are separated from interaction terms (functions of two covariates); note that $\boldsymbol{Z}$ is considered a single covariate. These splits are inspired by principle (c), as the presence of interactions entails a more flexible model. If no interactions are included in the model, this level can simply be pruned from the tree.

Level 3

Following again principle (a), the third level involves potential multi-branch splits, dividing effects on the basis of the covariates they are functions of. A first split is used to divide the effects of the different environmental covariates into multiple branches, while a second split separates the spatial effects from the temporal one. Interaction terms from either branch of Level 2 might require additional splits at Level 3: again, the effects should be organized into sub-branches based on the specific covariates involved.

Level 4

The previous three split levels make up the structure of a default decomposition tree that well adapts to most SDMs applications. If all the grey nodes in Figure 4.8 (including the omitted ones) are singletons, i.e. they contain a single σ^2 parameter, then the tree is complete as it is. However, the contribution of some covariates might be introduced in the model through multiple effects: if this is the case, the scheme of Figure 4.8 needs to be integrated with additional splits, until all the child nodes are actually singletons. In doing so, we suggest to use a fourth-level split strategy based on principle (c), in which effects with different degrees of flexibility are separated with successive binary branches. An example of split at level 4 is depicted in Figure 4.9, where the effect of an environmental covariate is partitioned into two branches, one for its linear component and the other for the non-linear one.

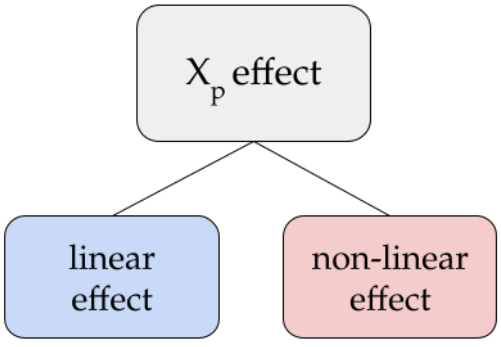

Figure 4.9: Example of split at level 4. The effect of covariate X_p is divided into its linear and non-linear component.

If there are more than two different levels of flexibility, we suggest the use of subsequent binary splits, rather than the use of multi-branch ones, since this choice simplifies the prior specification step (see Section 4.4.3). Fuglstad et al. 2020 found

that the use of successive binary splits in place of multi-brach ones has negligible impact on posterior inference.

4.4.2 Case study: HD tree

The case study model has a total of 13 variance parameters that need prior specification: $\sigma^2_{L1}, \sigma^2_{N1}, ..., \sigma^2_{L5}, \sigma^2_{N5}, \sigma^2_6, \sigma^2_S, \sigma^2_T$. The tree design for this application is represented in Figure 4.10. Level 2 of the default tree is pruned, while, at Level 4, a binary split between the linear and non-linear contribution is added for all the 5 environmental covariates that are continuous.

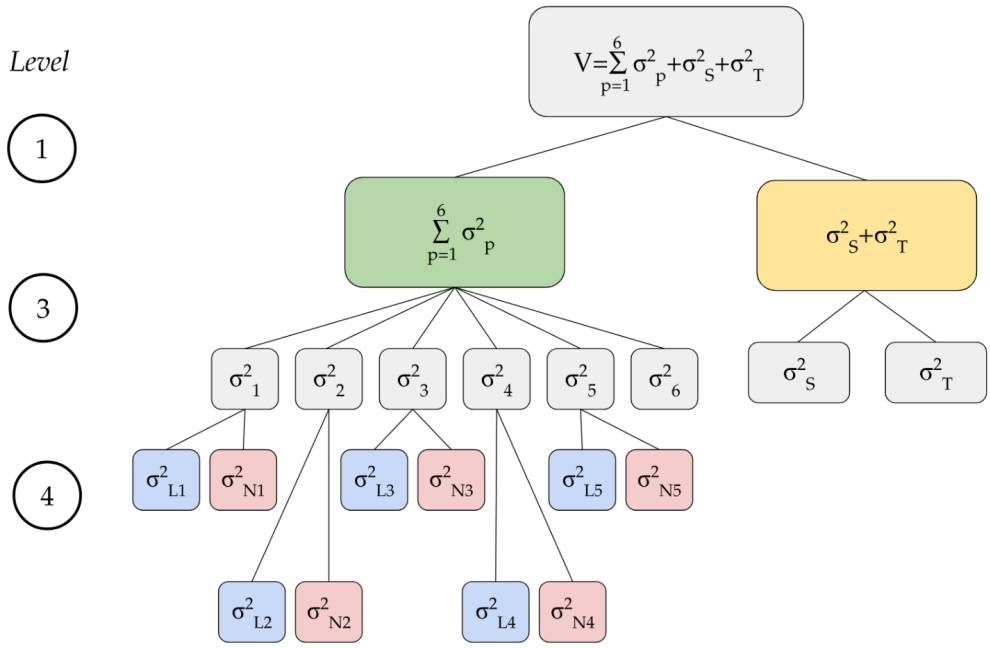

Figure 4.10: Decomposition tree for the case study based on the default proposal of Section 4.4.1.

Along with V, we obtain the following new parameters dividing the child nodes

by their parent node:

$$
\begin{aligned}
V &= \left(\sum_{p=1}^{5} \sigma_{Lp}^2 + \sigma_{Np}^2\right) + \sigma_6^2 + \sigma_S^2 + \sigma_T^2 \\
\omega_A &= \frac{(\sum_{p=1}^{5} \sigma_{Lp}^2 + \sigma_{Np}^2) + \sigma_6^2}{V} \\
\boldsymbol{\omega}_X &= \left[\frac{\sigma_{L1}^2 + \sigma_{N1}^2}{\omega_A \cdot V}, ..., \frac{\sigma_{L5}^2 + \sigma_{N5}^2}{\omega_A \cdot V}, \frac{\sigma_6^2}{\omega_A \cdot V}\right] \\
\omega_S &= \frac{\sigma_S^2}{\sigma_S^2 + \sigma_T^2} \\
\omega_{Np} &= \frac{\sigma_{Np}^2}{\sigma_{Lp}^2 + \sigma_{Np}^2} \qquad p = 1, ..., 5.
\end{aligned}
\tag{4.9}
$$

4.4.3 Guidelines for prior specification

Stsrting from the root node of the tree, we examine suitable prior distributions for the new HD parameters that are tailored for the SDM context. The original HD approach suggests building the joint prior in a bottom-up manner, where priors for higher-level parameters (closer to the root) depend on lower-level parameters. However, most applications of HD priors actually display in practice independent In what follows, we recommend independent priors on the new parameters.

Total variance V

The V parameter from Equation 4.8 is the total variance in the linear predictor: all priors proposed in the literature for variance parameters of Gaussian distributions can be used (e.g. Inverse-Gamma, Half-Cauchy, Jeffreys, etc.).

Recently, the PC prior for variance parameters has been proven particularly popular (Simpson et al. 2017, Klein and Kneib 2016). The $\text{PC}_0(\delta)$ prior for the parameter V is recommended by Fuglstad et al. 2020 when the chosen likelihood for Y is not Gaussian. The hyperparameter δ regulates the level of shrinkage towards the base model, in this case $V = 0$, and its tuning is done through tail probability statements, i.e. $P(\sqrt{V} > U) = \alpha$. When the likelihood is instead Gaussian, Fuglstad et al. 2020 suggests the use of the scale-invariant Jeffreys prior. In the context of SDMs, both choices have advantages, as the PC_0 prior can help regularize very complex models, while the popular Jeffreys prior does not require hyperparameter tuning.

Level 1

The split at this level generates a proportion parameter called ω_A. Indifference can be easily expressed through a Uniform distribution. If prior information is available, this can be introduced using an appropriate Beta distribution, as done in the R2D2 literature to reflect beliefs about the R^2 proportion (Yanchenko, Bondell, and Reich 2024b), or a PC prior with the desired base model and concentration parameter reflecting the user's uncertainty (Hem, Fuglstad, and Riebler 2024).

Level 3

Level 3 splits divide the effects into different branches according to the covariates they represent. If the user has no information about the partition among the branches, a Uniform on the simplex (i.e. $\text{Dir}(1, ..., 1)$), can be used to express complete ignorance. On the other hand, if the split is binary and the user has specific information about the relative importance of the covariates, an informative prior such as a Beta distribution can be used on the corresponding proportion parameters. If the split has instead multiple branches, it becomes more difficult to appropriately reflect available information about the partition through an appropriate prior choice. Additionally, experts might be unable to fully express this information in the first place. Nevertheless, they might still have an intuition about the number of important covariates, i.e. about the sparsity of the partition.

Following Fuglstad et al. 2020, we suggest that priors for multi-branch splits should be exchangeable, i.e. all branches are treated equally. Exchangeable priors can be used to introduce prior information about the sparsity level in the partition through the regulation of their hyperparameter. A symmetric Dirichlet $\text{Dir}(q, ..., q)$ is an example of an exchangeable prior on a simplex vector. If a vector $\boldsymbol{\omega}$ is distributed as a symmetric Dirichlet with hyperparameter q, its density is:

$$\pi(\boldsymbol{\omega}) = \frac{\Gamma(qP)}{\Gamma(q)^P} \prod_{p=1}^{P} \omega_p^{q-1}$$

The hyperparameter q controls the sparsity level and can be regulated to reflect assumptions about the partition of the variance among the branches. A value of $q = 1$ represents complete ignorance about the partition, i.e. a Uniform on all possible values of the simplex. Setting $q > 1$ reflects the belief that all covariates are equally important, while a value of $q < 1$ is set if only a few of them are assumed important. Hem, Fuglstad, and Riebler 2024 recommended setting q using the marginal prior on a single proportion ω_p such that $P(\text{logit}(1/4) < \text{logit}(\omega_p) - \text{logit}(1/P) <$

logit(3/4)) = 0.5. On the other hand, the work of Zhang et al. 2022 derived good theoretical properties on the induced prior on the coefficients of the R2D2 prior under the choice of $q < 1/2$. The derivation of such properties is extremely useful for variable selection contexts but it remains problematic to reflect specific assumptions about the sparsity level in other contexts. The concept of *effective number of non-zero coefficients* presented by Piironen and Vehtari 2017 has been used by Aguilar and Bürkner 2023 as an intuitive way to evaluate specific hyperparameters choices. However, this quantity depends on all the hyperparameters of the prior and not exclusively on q.

Here, we suggest to focus on a particular statistic based exclusively on the Dirichlet distribution that can help the user set up the hyperparameters in a way that reflects prior assumptions about the number of important covariates. Specifically, we consider the *quantity of the proportion of variance explained by the top k components*. For a given realization $\boldsymbol{\omega}_i$ of a random vector distributed according to a symmetric Dirichlet with parameter q, we define this quantity as the sum of the largest k entries of the vector. Figure 4.11 shows the distribution of 5000 realizations of this statistic for $P = 10,\ q = 0.5$ and values of $k = 1, 2, 5$.

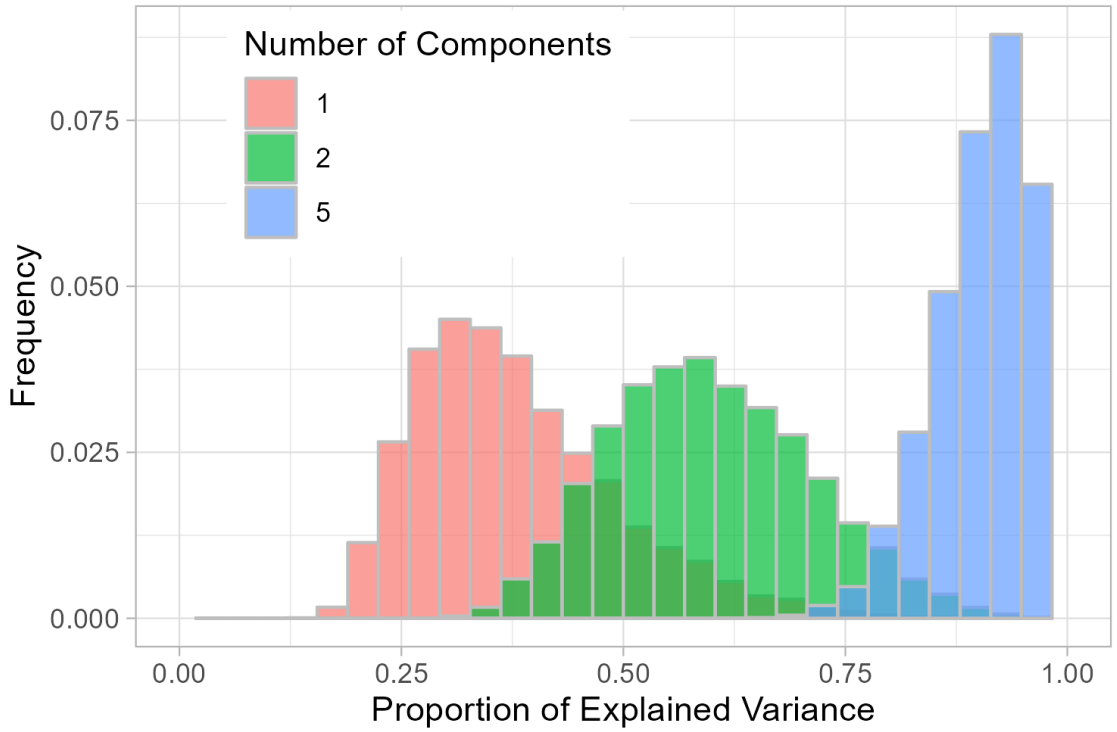

Figure 4.11: Distribution of the proportion of variance explained by the top k components: (red) $k = 1$; (green) $k = 2$; (blue) $k = 5$.

Figure 4.11 can be interpreted as follows. The largest proportion from a realiza-

tion of a symmetric Dirichlet distribution with $P = 10, q = 0.5$, usually has values between 25% and 50% with an average around 40%; the sum of the largest two proportions usually sum up to more than 50% with an average around 60%; finally, the largest 5 proportions sum up almost always to more than 75% and on average to around 90%.

In order to intuitively specify the hyperparameter q, the distribution of proportion of variance explained by the top k components can be compared for different values of q. Figure 4.12 reports on the x-axis the value of k, i.e. the number of components, while the y-axis represents the proportion of explained variance. The same distributions used in Figure 4.11 are here reported as vertical boxplots of different colors, where each color represents a different value of the q hyperparameter. Obviously, the statistic increases for all values $k < P$ for a smaller value of q. Figure 4.12 (b) reports more concisely only the mean of each of the distributions, along with dashed lines representing a proportion of 90%, 95%,99%.

Plots like the one in Figure 4.12 can be used to find the value of q that best reflects prior assumptions in the form: "the k most important covariates explain around α% of the total variance". For example, if we believe that the 3 most important covariates out of 10 will explain around 90% of the total variance, we may choose a value of $q = 0.2$.

To summarize some guidelines about prior specification for Level 3 parameters, Beta distributions are recommended for the binary splits, since they can be both vague (Uniform distribution is a special case) but also flexible enough to exploit potentially available prior information. For multi-branch splits, we suggest instead the use of symmetric Dirichlet priors, whose hyperparameter can be intuitively calibrated following the intuitive procedure outlined above.

Levels 2 & 4

Levels 2 and 4 create similar binary splits that separate effects with different degrees of flexibility. A reasonable choice is to define the proportion parameter as the proportional contribution of the more flexible branch and specify a PC_0 prior on it, in accordance to the principle of model simplicity.

The derivation of the PC_0 prior is however not as simple as for the V parameter. For each split, the prior must be derived separately, as the resulting function often depends on the choices of basis and precision matrices: this represents an inconvenience for the user. More importantly, the PC_0 can also depend on lower-level proportion parameters, which is computationally inefficient and not coherent with the recommendation of using only independent priors.

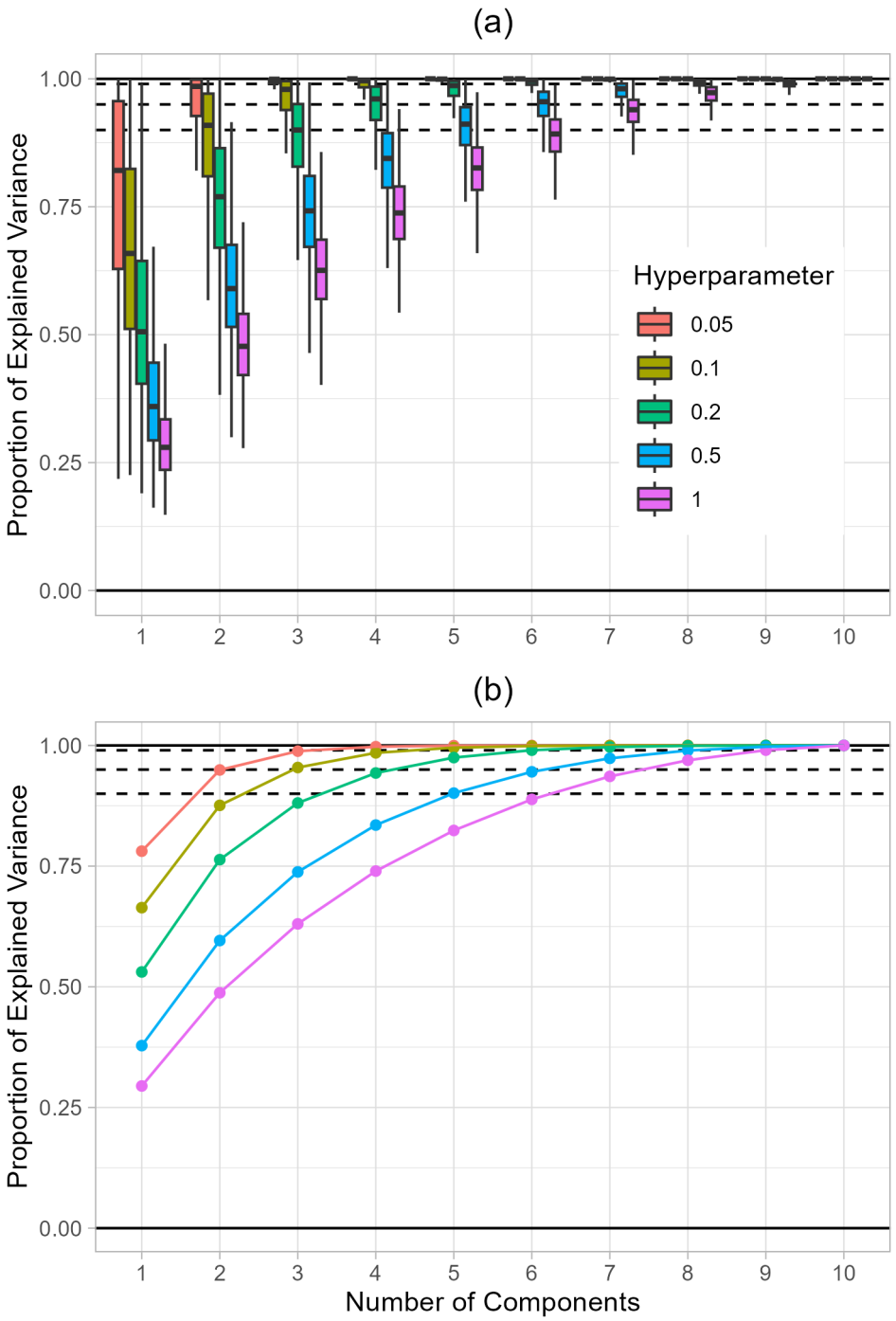

Figure 4.12: (a) Distribution of the proportion of variance explained by the top k components, grouped by the corresponding value of the hyperparameter q. (b) Mean of the samples for different values of the hyperparameter. The three dashed black lines represent the values of 90%, 95%, and 99%.

Nevertheless, the PC_0 often ends up having a simple functional form, completely independent of both the effects' matrices and potential lower-level parameters, namely a Truncated Exponential on the square root of the proportion parame-

ter. This result has been found by Fuglstad et al. 2020 and later by Franco-Villoria, Ventrucci, and Rue 2022, under two different conditions. In the following section, we prove that a simple condition, easy to check, can guarantee that the PC_0 prior actually reduces to this simplified version. We also show how this condition is respected in many relevant cases in the context of SDMs. Therefore, we argue that the the PC_0 prior can be recommended as the standard choice for the splits at Level 2 and 4. The hyperparameter of the PC prior can then be regulated using median or tail event statements according to the prior beliefs of the user. If in some cases the PC_0 does not simplify to a model-independent, we argue in favour of still using the same convenient functional form, because of the computational advantages, despite it no longer being the correct PC prior. The recommendation of Fuglstad et al. 2020 about using successive binary splits, in place of multi-branch ones, comes from the fact that the derivation of the PC prior greatly complicates in the latter case. We suggest to set up the tree such that the use of PC_0 prior at each split is reasonable.

4.4.4 PC_0 prior for proportions

In specific instances, the PC_0 for proportions of variance ω for a Gaussian model is found to have the following *simplified form* (Fuglstad et al. 2020, Franco-Villoria, Ventrucci, and Rue 2022):

$$\pi(\omega) = \frac{\delta \exp(-\delta\sqrt{\omega})}{2\sqrt{\omega}[1 - \exp(-\delta)]} \qquad 0 < \omega < 1 \tag{4.10}$$

which corresponds to an Exponential distribution on $\sqrt{\omega}$, truncated at $\sqrt{\omega} = 1$. This solution is called "simplified" as it does not depend on the choice of basis or precision matrices, nor on the parameters at lower splits of the decomposition tree.

Fuglstad et al. 2020 found this solution for the PC_0 prior under non-singularity conditions on the basis and precision matrices involved. Franco-Villoria, Ventrucci, and Rue 2022 developed an alternative proof for a specific spatio-temporal application using IGMRF effects, which returns the same result of Fuglstad et al. 2020, but relies on different assumptions.

This section intends to extend the proof of Franco-Villoria, Ventrucci, and Rue 2022 to a general case, in order to understand how common the simplified form of Equation 4.10 is in practice. First, notation for a general pair of effects is set up. Secondly, the KLD-based distance needed for the PC prior is derived. Thirdly, a simple condition that guarantees the simplified form, called *sum-of-rank condition*, is discussed in details. Finally, it is explained how the condition is respected in many important cases in which the use of a PC_0 prior might be desirable (see Section

4.4.3).

General set up

Let X_0 and X_1 be two generic covariates, which can also be multivariate. Note that X_0 and X_1 can also represent spatio-temporal covariates. Let $f(X_0, X_1; \omega)$ be the weighted sum of two effects with weight $\omega \in [0, 1]$:

$$f(X_0, X_1; \omega) = \sqrt{1-\omega} f_0(X_0) + \sqrt{\omega} f_1(X_1)$$

where:

$$\begin{aligned} f_0(X_0) &= \boldsymbol{D}_0^T(X_0)\boldsymbol{u}_0 \\ f_1(X_1) &= \boldsymbol{D}_1^T(X_1)\boldsymbol{u}_1 \\ \underset{K_0 \times 1}{\boldsymbol{u}_0} &\sim N(\boldsymbol{0}, \boldsymbol{Q}_0^*) \\ \underset{K_1 \times 1}{\boldsymbol{u}_1} &\sim N(\boldsymbol{0}, \boldsymbol{Q}_1^*). \end{aligned}$$

$\boldsymbol{D}_0(\cdot)$ is a basis matrix made up by K_0 functions and $\boldsymbol{D}_1(\cdot)$ is a basis matrix made up by K_1 functions. Let N_0 be the cardinality of the set of values $x \in \mathcal{X}_0$ such that $\pi_{X_0}(x) > 0$ and let N_1 be the same for X_1. If N_0 or N_1 are infinite, the probability distributions are discretized such that these quantities are still large but finite: this is convenient to avoid working with Gaussian vectors of infinite dimensions. Let $\boldsymbol{x}_0, \boldsymbol{x}_1$ be row vectors of dimension $N \leq N_0 \times N_1$ such that their i^{th} elements form a unique pair $x_{i0} \in \mathcal{X}_0, x_{i1} \in \mathcal{X}_1, \pi_{X_0,X_1}(x_{i0}, x_{i1}) > 0$: the pairs $x_{i0}, x_{i1},\ i = 1, ..., N$ are by design all possible realizations of the covariates that can be observed in the data. The basis $\boldsymbol{D}_0(\boldsymbol{x}_0)$ becomes a matrix of dimension $K_0 \times N$ and $\boldsymbol{D}_1(\boldsymbol{x}_1)$ matrix of dimension $K_1 \times N$. We always assume that $K_0 \leq N_0$ and $K_1 \leq N_1$.

The vector $f(\boldsymbol{x}_0, \boldsymbol{x}_1; \omega)$ of dimension N is distributed as a multivariate Gaussian, whose covariance matrix $\boldsymbol{\Sigma}(\omega)$ is equal to:

$$\begin{aligned} \boldsymbol{\Sigma}(\omega) &= (1-\omega)\boldsymbol{\Sigma}_0 + \omega\boldsymbol{\Sigma}_1 \\ \boldsymbol{\Sigma}_0 &= \boldsymbol{D}_0^T(\boldsymbol{x}_0)\boldsymbol{Q}_0^*\boldsymbol{D}_0(\boldsymbol{x}_0) \\ \boldsymbol{\Sigma}_1 &= \boldsymbol{D}_1^T(\boldsymbol{x}_1)\boldsymbol{Q}_1^*\boldsymbol{D}_1(\boldsymbol{x}_1). \end{aligned}$$

$f(\boldsymbol{x}_0, \boldsymbol{x}_1; \omega)$ will have a proper Gaussian distribution only if its covariance matrix is full rank. Thanks to the properties of matrix rank, an upper bound can be found for the rank of $\boldsymbol{\Sigma}(\omega)$ (called $R(\omega)$ from now on) as $\text{rank}(\boldsymbol{\Sigma}_0) + \text{rank}(\boldsymbol{\Sigma}_1)$. Noting that $R(0) = \text{rank}(\boldsymbol{\Sigma}_0)$ and $R(1) = \text{rank}(\boldsymbol{\Sigma}_1)$ and that the upper bounds for these

quantities are:

$$R(0) \leq \min[N_0, K_0] \qquad R(1) \leq \min[N_1, K_1] \tag{4.11}$$

we can find that an upper bound for $R(\omega)$ is:

$$R(\omega) \leq \min[N_0, K_0] + \min[N_1, K_1]. \tag{4.12}$$

Since it might be the case that this upper bound is less than N, the probability density $\pi(\boldsymbol{y}; \omega)$ of $f(\boldsymbol{x}_0, \boldsymbol{x}_1; \omega)$ can be written using the improper version (Rue and Held 2005), which simplifies to the classical one when $R(\omega) = N$:

$$\pi(\boldsymbol{y}; \omega) = \frac{1}{\sqrt{(2\pi)^{R(\omega)} \cdot |\boldsymbol{\Sigma}(\omega)|^*}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T \boldsymbol{\Sigma}^*(\omega)\boldsymbol{y}\right). \tag{4.13}$$

Note that $|\cdot|^*$ represents the generalized determinant (i.e. the product of non-null eigenvalues).

PC prior derivation

The first step in deriving the PC prior for ω is computing the KLD between $f(\boldsymbol{x}_0, \boldsymbol{x}_1; \omega)$ and $f(\boldsymbol{x}_0, \boldsymbol{x}_1; \omega_0)$, where ω_0 is the chosen base model. In this context, we only consider the case $\omega_0 = 0$; note that the case $\omega_0 = 1$ is equivalent. Once the KLD has been computed as a function of ω, the functional form of the PC prior is found assuming a (truncated) Exponential distribution on the distance $d(\omega; \omega_0) = \sqrt{2 \cdot KLD[\pi(\boldsymbol{y}; \omega)||\pi(\boldsymbol{y}; \omega_0)]}$ and solving for ω.

Using the density function of Equation 4.13 for $f(\boldsymbol{x}_0, \boldsymbol{x}_1; \omega)$, the KLD-based distance $d(\omega; \omega_0)$ simplifies to:

$$d(\omega; \omega_0) = \sqrt{\text{tr}\left[\boldsymbol{\Sigma}^*(\omega_0)\boldsymbol{\Sigma}(\omega)\right] - R(\omega) - \log\frac{|\boldsymbol{\Sigma}(\omega)|}{|\boldsymbol{\Sigma}(\omega_0)|} + [R(\omega_0) - R(\omega)]\log(2\pi)}. \tag{4.14}$$

See the proof in Section C.1 of Appendix.

This is still a complicated formula that does not give a simple functional form for the prior of ω, since it directly depends on the chosen basis and precision matrices of the effects involved.

Simplification under sum-of-ranks condition

The sum-of-ranks condition consists in checking whether the sum of the number of non-null eigenvalues of $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$, or equivalently the sum of their ranks, is less or equal to the dimension N.

$$R(0) + R(1) \leq N. \tag{4.15}$$

This inequality has many consequences. First, it implies that both $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ are singular, assuming that neither of them can be a zero matrix. Most importantly, it guarantees that $\mathbf{\Sigma}(\omega)$ can be rewritten in a more convenient form.

Let $\boldsymbol{e}_{0,1}, ..., \boldsymbol{e}_{0,R(0)}$be the $R(0)$ eigenvectors of $\mathbf{\Sigma}_0$ associated with non-null eigenvalues $\lambda_{0,1}, ..., \lambda_{0,R(0)}$. Let $\boldsymbol{e}_{1,1}, ..., \boldsymbol{e}_{1,R(1)}$ be the $R(1)$ eigenvectors of $\mathbf{\Sigma}_1$ associated with non-null eigenvalues $\lambda_{1,1}, ..., \lambda_{1,R(1)}$. If the sum-of-ranks condition (4.15) holds, then $\mathbf{\Sigma}(\omega)$ can be rewritten in terms of $\mathbf{\Lambda}_0 = \text{diag}(\lambda_{0,1}, ..., \lambda_{0,R(0)}, 0...., 0)$, $\mathbf{\Lambda}_1 = \text{diag}(0, ..., 0, \lambda_{1,1}, ..., \lambda_{1,R(1)})$ and a common eigenbasis $\boldsymbol{V} = [\boldsymbol{e}_{0,1}, ..., \boldsymbol{e}_{0,R(0)}, \mathbf{0}, ..., \mathbf{0}, \boldsymbol{e}_1, ..., \boldsymbol{e}_{R(1)}]$:

$$\mathbf{\Sigma}(\omega) = \boldsymbol{V}[(1-\omega)\mathbf{\Lambda}_0 + \omega\mathbf{\Lambda}_1]\boldsymbol{V}^T.$$

Rewriting $\mathbf{\Sigma}(\omega)$ ensures that the distance function (4.14) further simplifies (Section C.2 of the Appendix). However, this distance is not finite for $\omega_0 = 0$ and must be computed instead as a limit for $\omega_0 = 0$ (Franco-Villoria, Ventrucci, and Rue 2022). It can be proven that:

$$\lim_{\omega_0 \to 0} d(\omega; \omega_0) = R(1)\sqrt{\omega} \tag{4.16}$$

where $R(1)$ is a constant with respect to ω. Specifying an Exponential distribution on the distance (truncated at $d(1)$ as the upper bound for ω is 1), the final form of the PC prior is found using the change-of-variable formula:

$$\pi(\omega) = \frac{\widetilde{\delta} R(1) \exp(-\widetilde{\delta} R(1)\sqrt{\omega})}{2\sqrt{\omega}[1 - \exp(-\widetilde{\delta} R(1))]}.$$

The dependence on the constant $R(1)$, which contains parameters of the model is removed, as it becomes not identifiable with the hyperparameter of the Exponential distribution $\widetilde{\delta}$. Denoting $\delta = \widetilde{\delta} R(1)$, we obtain Equation 4.10, which proves that, under the sum-of-ranks condition, the simplified form of the PC_0 prior is guaranteed.

The δ hyperparameter can be chosen using probability statements such as $P(\omega < U) = \alpha$. However, in order to obtain a valid probability distribution (i.e. $\delta > 0$), Franco-Villoria, Ventrucci, and Rue 2022 noted that the following condition must

be respected:

$$\alpha \geq \sqrt{U}. \tag{4.17}$$

As a consequence, the median can be at most 0.25, which can be obtained with $\delta \to 0$.

Because the sum-of-ranks condition guarantees a simpler solution for the PC prior, it is convenient to start the derivation of the prior by checking this condition first. If the condition is not found to be satisfied using the upper bounds of Equation 4.11, the next step should be to derive the actual $R(0)$ and $R(1)$ and check the condition again.

Important cases

The sum-of-ranks assumption can be verified when $R(0)$ and $R(1)$ are available analytically, but this is rarely the case in practice. Therefore, the assumption can be checked instead using their upper bounds (Equation 4.11), which are usually explicitly available to the user. We discuss here three cases in which PC_0 priors on ω are a sensible choice, according to the discussion of Section 4.4.3. These cases are selected as they may result from the decomposition tree designed for SDMs (Figure 4.8).

- **Linear effect versus non-linear effect.**

  Consider the case in which X is a univariate variable and $X = X_0 = X_1$. In this case, $N_0 = N_1 = N$ since the two covariates are completely dependent. Consider now that $f_0(X)$ is a linear effect (after standardization of X) and $f_1(X)$ is a non-linear function of X with a finite number of K_1 coefficients, constrained by design to have a null intercept and linear trend: $f_1(X)$ can be for example the residual term of a P-Spline of second order (see Example 5 of Chapter 3). With regard to ranks, upper bound for $R(0)$ and $R(1)$ are found to be:

  $$R(0) \leq 1$$
  $$R(1) \leq K_1.$$

  As a consequence, their sum $R(0) + R(1) \leq K_1 + 1$ respects the condition of Equation 4.15 as long as $K_1 < N$. This is always true when X is a continuous variable, since K_1 is finite such that $K_1 << N$.

  This scenario can emerge from splits at Level 4 of the tree from Figure 4.8.

- **Linear main effects versus interaction effect.**

  Consider now the case in which $X_0 = [X_A, X_B]$ where X_A, X_B are two independent univariate covariates, and $X_1 = X_0$. Again, $N_0 = N_1 = N$ and $N = N_A \cdot N_B$, where N_A is the cardinality of the set of values $x \in \mathcal{X}_A$ such that $\pi_{X_A}(x) > 0$ and N_B is the same for X_B.

  Let $f_0(X_A, X_B)$ and $f_1(X_A, X_B)$ be defined as:

$$\begin{aligned} f_0(X_A, X_B) &= \sqrt{1-\phi}\widetilde{X_A}u_A + \sqrt{\phi}\widetilde{X_B}u_B \\ f_1(X_A, X_B) &= \widetilde{X_A}\widetilde{X_B}u_1 \end{aligned}$$

  where $\widetilde{X_A}, \widetilde{X_B}$ are the standardized versions of X_A, X_B.

  In this scenario, it is easy to work out that $R(1) \leq 1$. However, $R(0)$ depends on the value of ϕ. Nevertheless, an upper bound can still be found using the same formula of Equation 4.12:

$$\begin{aligned} R(0) &\leq \min[N_A, 1] + \min[N_B, 1] \\ &\leq 2. \end{aligned}$$

  Hence, the sum-of-ranks condition is respected as long as $3 \leq N_A \cdot N_B$. This is always true if X_A, X_B are both continuous covariates.

  This case covers the splits at Level 3 in the left branch of the tree from Figure 4.8, where the main effects of the environmental covariates are separated by the interaction terms.

- **Finite-dimensional main effects versus Kronecker product interaction effect.**

  Consider the scenario above where now $f_0(X_A, X_B)$ is defined as the weighted sum of two finite-dimensional effects for X_A and X_B with respectively K_A and K_B coefficients with $K_A \leq N_A, K_B \leq N_B$:

$$\begin{aligned} f_0(X_A, X_B) &= \sqrt{1-\phi}\boldsymbol{D}_A^T(X_A)\boldsymbol{u}_A + \sqrt{\phi}\boldsymbol{D}_B^T(X_B)\boldsymbol{u}_B \\ \underset{K_A \times 1}{\boldsymbol{u}_A} &\sim N(\boldsymbol{0}, \boldsymbol{Q}_A^*) \\ \underset{K_B \times 1}{\boldsymbol{u}_B} &\sim N(\boldsymbol{0}, \boldsymbol{Q}_B^*). \end{aligned}$$

  Let $f_1(X_A, X_B) = \boldsymbol{D}^T(X_A, X_B)\boldsymbol{u}$ be an interaction effect built using the Kronecker product between the basis and precision matrices of the main effecwrit-

ingts in $f_0(X_A, X_B)$, i.e. the IV type of interaction effect defined by Knorr-Held 2000 for inseparable spatio-temporal effects.

$$\boldsymbol{D}(X_A, X_B) = \boldsymbol{D}_A(X_A) \otimes \boldsymbol{D}_B(X_B)$$
$$\underset{K_A \cdot K_B \times 1}{\boldsymbol{u}} \sim N(\mathbf{0}, [\underset{K_A \times K_A}{\boldsymbol{Q}_A} \otimes \underset{K_B \times K_B}{\boldsymbol{Q}_B}]^*).$$

The upper bounds for $R(0)$ and $R(1)$ can be easily found:

$$R(0) \leq K_A + K_B$$
$$R(1) \leq K_A \cdot K_B.$$

Hence, the sum-of-ranks condition is satisfied as long as:

$$K_A + K_B + K_A \cdot K_B \leq N_A \cdot N_B.$$

This is always true if X_A and X_B are continuous covariates. The condition also holds in many other cases in which X_A and X_B are discrete, such as the particular example derived in Franco-Villoria, Ventrucci, and Rue 2022.

This last case covers the possible split between the main spatial and temporal effects and their non-additive interaction, which appears at Level 3 in the right branch of the tree from Figure 4.8.

4.4.5 Case study: VP prior

The VP prior approach requires the specification of priors on the new parameters $V, \omega_A, \boldsymbol{\omega}_X \omega_S, \omega_{N1}, ..., \omega_{N5}$ defined in Equation 4.9.

From Level 1 of the HD tree (Figure 4.10), we obtain ω_A, which represents the proportional contribution to the variance of the environmental effects. ω_A must be assigned a Uniform distribution, as no expert knowledge about the variance partitioning was available to us.

At Level 3, there are two splits. The first split distinguishes between the contributions of each of the 6 environmental covariates. The resulting simplex $\boldsymbol{\omega}_X$ is assigned a symmetric Dirichlet. In order to specify a reasonable value for the hyperparameter, we need to state our prior assumptions about the partition between the different covariates. A priori, we believe that the *Depth* covariate will play an important role, but that the correlation between *Surface temperature* and *Bottom temperature*, as well as between *Surface salinity* and *Bottom salinity* may cause only one of the two covariates in each pair to be selected; finally, we do not have prior

assumptions about the role of the *Survey Vessel*. We may reflect this assumption by stating, for example, that half of the covariates (i.e. 3) might be sufficient to explain at least 90% of the overall variability due to environmental factors, without the need to specify which ones are believed to be important. According to Figure 4.13, this assumption is best reflected by selecting a value of $q = 0.5$.

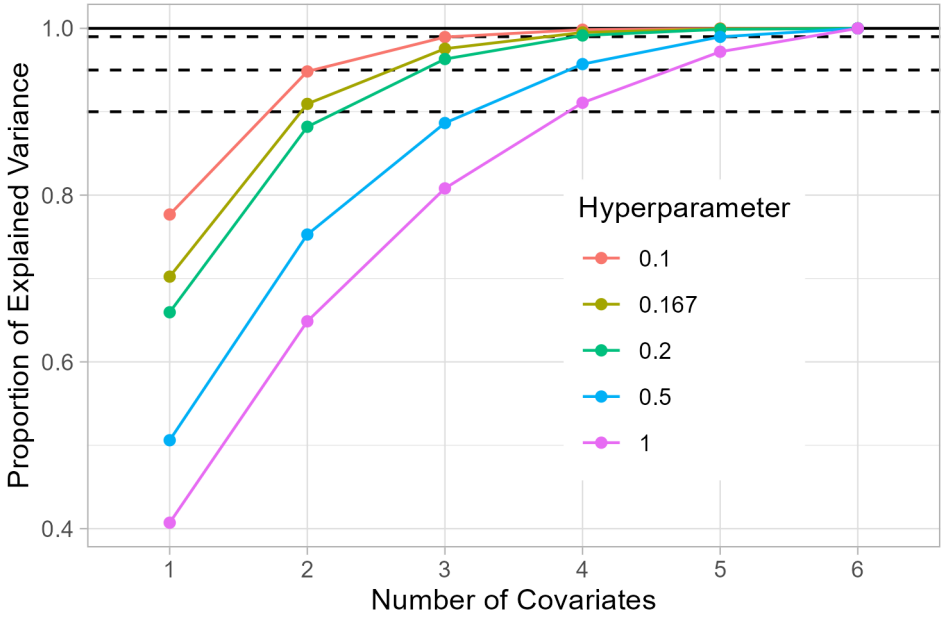

Figure 4.13: Mean of the samples of the proportion of variance explained by the top k components (k indicated in the x axis) using a symmetric Dirichlet of dimension $P = 6$, obtained with different hyperparameter values. The three dashed black lines represent the values of 90%, 95%, and 99%.

The second split from Level 3 separates the spatial contribution from the temporal one. The absence of information mandates the use of a Uniform distribution on ω_S.

The 5 splits at Level 4 are identical and separates the linear and non-linear contribution for all the 5 continuous environmental covariates. For each of the $p = 1, ..., 5$ splits, we get a proportion parameter ω_{Np}. For all the 5 parameters, we choose a highly flat PC_0 prior to induce little shrinkage. From the discussion of Section 4.4.4, we know that the PC_0 prior has its simplified functional form in this case, i.e. Equation 4.10. Due to the bound of Equation 4.17, the highest the median can be set is at 0.25 with $\delta \to 0$: we choose $\delta = 0.1$, which guarantees the median to be 0.238. Figure 4.14 shows that there is little difference between this choice and an extremely small value such as $\delta = 0.001$.

The prior specification is completed assuming a distribution for V. We compare the choice of a Jeffreys (VP1) and of a relatively flat PC_0 prior (VP2), i.e. $\delta =$

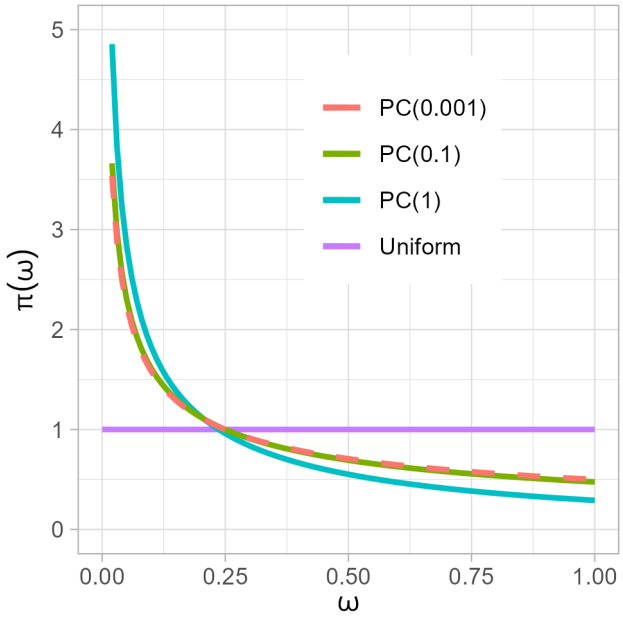

Figure 4.14: Comparison of multiple distributions for a generic proportion parameter ω. Along with the Uniform distribution, three PC_0 priors with different δ hyperparameter values are plotted.

0.1. Figure 4.15 compares the two chosen priors, after the choice of a convenient normalizing constant for the improper Jeffreys distribution.

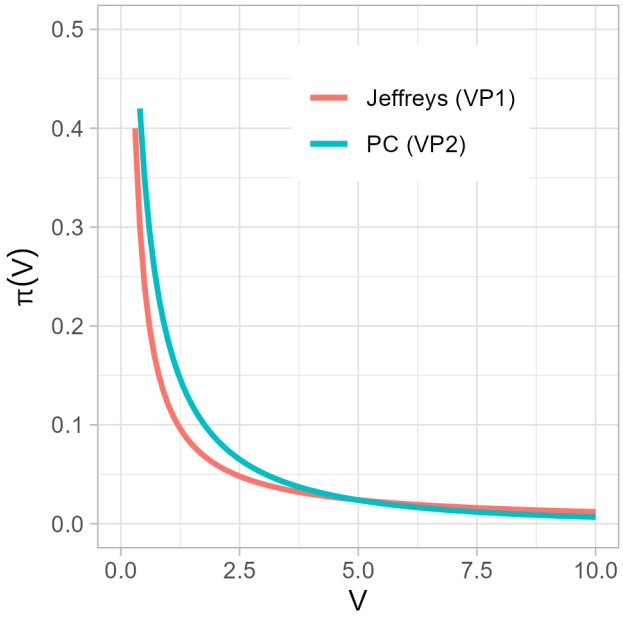

Figure 4.15: Comparison of the two different prior choices for the V parameter. The hyperparameter for the PC_0 prior is set to $\delta = 0.1$.

The joint prior on the original parameters can be written applying the change-

of-variable formula:

$$\pi(\sigma^2_{L1}, \sigma^2_{N1}, ..., \sigma^2_{L1}, \sigma^2_{N1}, \sigma^2_6, \sigma^2_S, \sigma^2_T) = \pi(V)\pi(\omega_A)\pi(\boldsymbol{\omega}_X)\pi(\omega_S)\prod_{p=1}^{5}\pi(\omega_{Np}) \cdot |\boldsymbol{J}|$$

where $V, \omega_A, \boldsymbol{\omega}_X\omega_S, \omega_{N1}, ..., \omega_{N5}$ must be rewritten in terms of $\sigma^2_{L1}, \sigma^2_{N1}, ..., \sigma^2_{L5}, \sigma^2_{N5}, \sigma^2_6, \sigma^2_S, \sigma^2_T$, and $\boldsymbol{J}$ is the Jacobian of the transformation from the original parameters to the new ones.

Performance evaluation. The performance of the proposed VP priors is compared to more traditional choices. In total, we consider 4 different prior specifications:

- **IG prior**: i.i.d. Inverse-Gamma(1,5e-5) on all the σ^2 parameters, which is the default specification in INLA;
- **PC prior**: i.i.d. $\text{PC}_0(\delta)$ prior on all the σ^2 parameters with hyperparameter δ such that $P(\sigma > 3) = 0.05$ (Fuglstad et al. 2020);
- **VP1 prior**: $V \sim$ Jeffreys, $\omega_A \sim \text{Unif}(0,1)$, $\boldsymbol{\omega}_X \sim \text{Dir}(0.5)$, $\omega_S \sim \text{Unif}(0,1)$, $\omega_{Np} \sim \text{PC}_0(0.1)$ for $p = 1, ..., 5$;
- **VP2 prior**: same as the VP1 prior but $V \sim \text{PC}_0(0.1)$.

The model is fitted using the INLA software (Rue, Martino, and Chopin 2009), which offers a joint posterior sample of all the model parameters, along with many other useful outputs. The dataset is divided into $\boldsymbol{y} = [\boldsymbol{y}^T_{\text{train}}, \boldsymbol{y}^T_{\text{test}}]^T$ where the training set $\boldsymbol{y}_{\text{train}}$ contains all observations up to 2015, while the test set $\boldsymbol{y}_{\text{test}}$ the remaining ones up to 2020, for a total of 1028 observations or $\approx 17\%$. The models are fitted on the training sets and the performance in terms of prediction over the test set is evaluated using the same metrics used in Hui et al. 2023 (log likelihood, Brier score, Tjur R^2), along with the more interpretable accuracy metric. All metrics are based on the point estimates $\hat{p}_i = \text{logistic}(E[\eta_i|\boldsymbol{y}_{\text{train}}, \boldsymbol{x}_i, \boldsymbol{z}_i, t_i,])$ for all instances i in the test set. Accuracy is defined as the percentage of instances in the test set accurately predicted using $\mathbb{I}[\hat{p}_i > 0.5]$. The log-likelihood metric is measured on the test set observations under a Bernoulli distribution with parameter $\hat{p}_i$. The Brier score corresponds to the mean squared error between y_i and $\hat{p}_i$. The Tjur R^2 is the difference between the mean of $\hat{p}_i$ for all i such that $y_i = 1$ and the mean of $\hat{p}_i$ for all i such that $y_i = 0$. Figure 4.16 reports the values of these 4 metrics for each of the 39 fish species in the dataset.

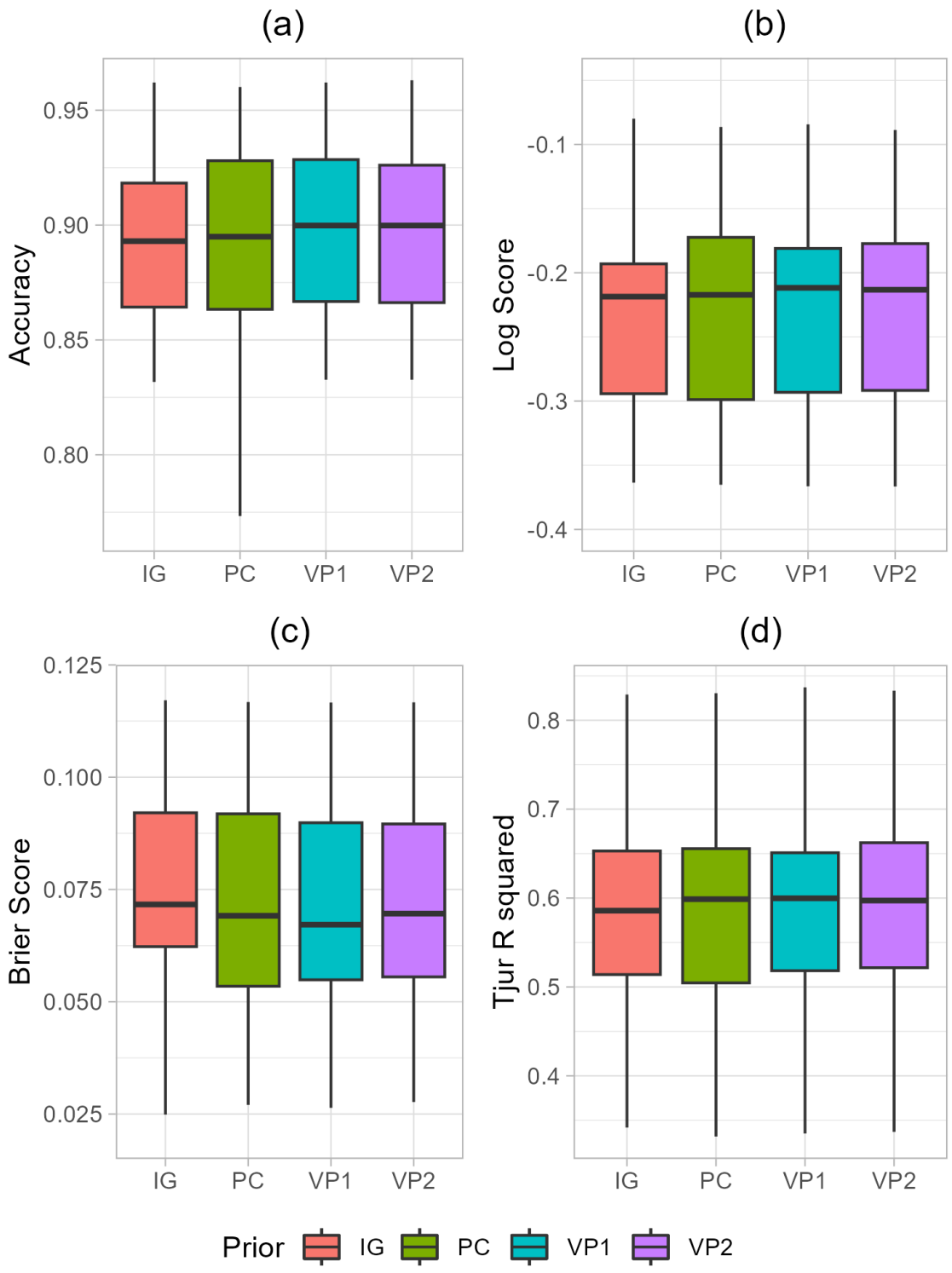

Figure 4.16: Comparison of the performance of the 4 different prior specifications on prediction on the test set: (a) accuracy; (b) log-likelihood; (c) Brier score; (d) Tjur R^2.

At first, we can note how the predictive performance is not largely impacted by the prior specification, thanks to the large number of observations in the dataset.

Nevertheless, there are still some differences in the results: both VP priors are competitive with respect to the popular PC prior choice, and represent an improvement with respect to the Inverse-Gamma choice, particularly noticeable in terms of accuracy and Brier score. Between the two VP options, the impact of the prior on V appears to be negligible and no clear preference emerges based on the four metrics examined.

Looking at the posterior means of the variance parameters for all the species (Figure 4.17), it can be noted how the IG prior option tends to shrink many more of the parameters to 0 than the other three alternatives, which might be the cause for the slightly worse prediction ability. While the estimations for the other priors exhibit greater consistency, the PC prior generally yields larger estimates for most of the parameters, when compared to the VP priors.

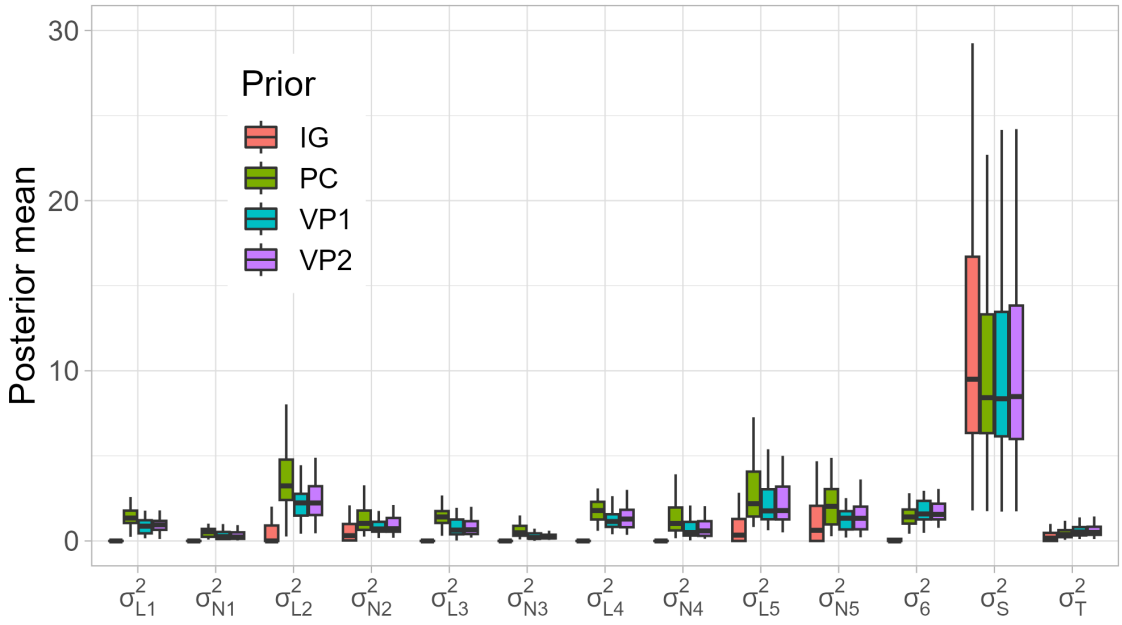

Figure 4.17: Posterior mean of the σ^2 parameters for the 39 species in the dataset under different prior specifications.

4.5 Posterior variance partitioning

This section focuses on proposing an alternative to the traditional estimation of variance partition in SDMs, through a method that is coherent with the prior framework presented so far.

Assume that the joint posterior distribution of the model parameters has been derived, either analytically or through simulation. In quantifying the variance con-

tributions of each effects, we shall not use functions of location estimators of the model parameters, such as posterior means or medians. Instead, we adopt a fully Bayesian approach as in Gelman et al. 2019 and define such quantities directly as functions of the model parameters. The posterior distribution of such metrics can then be approximated through simulation. First, a sample of joint realizations must be simulated from the parameters' posterior distribution:

$$\pi(\mu, \boldsymbol{u}_1, ..., \boldsymbol{u}_J, \boldsymbol{v}_1, ..., \boldsymbol{v}_L, \sigma^2_{A,1}, ..., \sigma^2_{A,J}, \sigma^2_{B,1}, ..., \sigma^2_{B,L}|\boldsymbol{y}).$$

Then, the function of interest must be applied to each realization to obtain a new sample that is distributed as the target posterior. This method is more computationally expensive but offers full information about the estimators. Now that we have selected an inferential approach, we can discuss the possible choices for the estimation of the variance contributions.

Traditionally, variance contributions of the effects of an SDM are estimated using sample variances of the trends, conditional on the coefficients (Ovaskainen et al. 2017, Hui et al. 2023):

$$W_{A,j} = Var^N_{i=1}[\boldsymbol{D}^T_j(\boldsymbol{x}_i)\boldsymbol{u}_j] \qquad j = 1, ..., J \tag{4.18}$$

$$W_{B,l} = Var^N_{i=1}[\boldsymbol{G}^T_l(\boldsymbol{z}_i, t_i)\boldsymbol{v}_l] \qquad l = 1, ..., L \tag{4.19}$$

where $\boldsymbol{x}_i, \boldsymbol{z}_i, t_i$ represent the N realizations of $\boldsymbol{X}, \boldsymbol{Z}, T$ from the data. This choice corresponds to the use of *finite-population variances* (as defined in Section 3.3.3 of Chapter 3).

The variance partition itself can then be estimated dividing each $W_{A,j}, W_{B,l}$ by their overall sum to obtain a proportions' vector $\boldsymbol{\omega}_{\text{trad}}$:

$$\boldsymbol{\omega}_{\text{trad}} = \frac{1}{\sum_{j=1}^{J} W_{A,j} + \sum_{l=1}^{L} W_{B,l}} [W_{A,1}, ..., W_{A,J}, W_{B,1}, ..., W_{B,L}]$$

where $0 \leq \omega_{\text{trad},m} \leq 1,\ m = 1, ..., J + L$ and $\sum_{m=1}^{J+L} \omega_{\text{trad},m} = 1$. The posterior means of the entries of the $\boldsymbol{\omega}_{\text{trad}}$ sum up to 1, as each of its realizations: due to this property, the posterior mean of $\boldsymbol{\omega}_{\text{trad}}$ represents the optimal location summary for the variance partition.

In accordance to the interprability conditions on the σ^2 parameters of an SDM from Section 4.3, we propose an alternative approach based on the concept of the *variance of interest*, i.e. the variance of the effects conditional on the parameters of interest, which have been defined for an SDM as $\boldsymbol{\theta} = [\boldsymbol{u}_1, ..., \boldsymbol{u}_J, \sigma^2_{B,1}, ..., \sigma^2_{B,L}]$. From Chapter 3, we know that the variances of interest is the finite-population variance

for fixed effects and the super-population for random ones. Following this logic, we could estimate the variance contributions for the SDM effects as:

$$s^2_{A,j} = Var_{\boldsymbol{X}}[\boldsymbol{D}_j^T(\boldsymbol{X})\boldsymbol{u}_j|\boldsymbol{u}_j] \qquad j = 1, ..., J \tag{4.20}$$

$$\sigma^2_{B,l} = Var_{\boldsymbol{Z},T}[\boldsymbol{G}_l^T(\boldsymbol{Z}, T)\boldsymbol{v}_l|\sigma^2_{B,l}] \qquad l = 1, ..., L. \tag{4.21}$$

Note that the symbol chosen for the quantities of Equation 4.21 is not ambiguous, as these equalities are actually true if the random effects have been appropriately scaled. In contrast to $W_{A,j}, W_{B,l}$, the estimators from Equation 4.20-4.21 actually respect the inferential focus of the effects and respect the assumptions made about the phenomenon, i.e. that it is actually modelled by the parameters $\boldsymbol{\theta}$, while the $\sigma^2_{A,j},\ j = 1, ..., J$ are simply introduced as "hyperparameters" for a more convenient prior specification.

It is important to note that, in Section 4.3, we have made the assumption that spatio-temporal effects are random ones for illustrative purposes. However, this is not always the case and if some or all spatio-temporal effects are to be considered fixed, then the finite-population variances $s^2_{B,l}$ should be used for these effects as well.

In addition to a change in the estimator, note how Equations 4.20-4.21 differ from the traditional approach as they do not make use of sample variance but rather assume that $\boldsymbol{X}, \boldsymbol{Z}, T$ are random variables. In particular, the same probability distribution $\pi(\boldsymbol{x}, \boldsymbol{z}, t)$ assumed a priori to obtain the interpretability of the σ^2 parameters should be used. This distinction brings additional advantages to the proposed method. First, the explicit specification of $\pi(\boldsymbol{x}, \boldsymbol{z}, t)$ raises the user's awareness about the actual meaning of the variance contribution definition and the impact that the covariate distributional choice has on it. Secondly, $\pi(\boldsymbol{x}, \boldsymbol{z}, t)$ can be defined in such a way that is portable between similar case studies so that the variance partitioning results for different datasets become directly and immediately compared. Moreover, in a context of sequential learning, the posterior distributions of the quantities from Equation 4.20-4.21 can be used for an informed prior specification of the σ^2 parameters for future studies. Furthermore, we reiterate how a suitable selection of $\pi(\boldsymbol{x}, \boldsymbol{z}, t)$, in contrast to the naive use of the empirical distribution, can significantly enhance the interpretability of posterior results for domain experts (e.g., straightforward Uniform distributions).

In order to compute the variance partition, we finally define $\boldsymbol{\omega}$ as:

$$\boldsymbol{\omega} = \frac{1}{\sum_{j=1}^{J} s^2_{A,j} + \sum_{l=1}^{L} \sigma^2_{B,l}} \left[s^2_{A,1}, ..., s^2_{A,J}, \sigma^2_{B,1}, ..., \sigma^2_{B,L}\right] \tag{4.22}$$

where $0 \leq \omega_m \leq 1,\ m = 1, ..., J + L$ and $\sum_{m=1}^{J+L} \omega_m = 1$. Again, the posterior means can be used as an estimate for the variance partition.

4.5.1 Case study: results

In what follows, the N_{species} models are fitted again on the whole dataset using the Jeffreys prior (VP1), chosen because of its scale-invariant property and the advantage of not having to regulate hyperparameters. A posterior sample on the model parameters is obtained using the `inla.posterior.sample()` function of the R-INLA package (Rue, Martino, and Chopin 2009).

We first compute the variance contributions of each of the covariates according to Equations 4.20-4.21:

$$\begin{aligned} s_p^2 &= Var_{X_p}[f_p(X_p)|\beta_p, \boldsymbol{u}_p] \qquad p = 1, ..., 6 \\ \sigma_S^2 &= Var_{\boldsymbol{Z},\boldsymbol{v}_S}[f_S(\boldsymbol{Z})|\sigma_S^2] \\ \sigma_T^2 &= Var_{T,\boldsymbol{v}_T}[f_T(T)|\sigma_T^2]. \end{aligned}$$

Note that $s_p^2 = s_{Lp}^2 + s_{Np}^2\ p = 1, ..., 5$ where:

$$\begin{aligned} s_{Lp}^2 &= Var_{X_p}[f_{Lp}(X_p)|\beta_p] \qquad p = 1, ..., 5 \\ s_{Np}^2 &= Var_{X_p}[f_{Np}(X_p)|\boldsymbol{u}_p] \qquad p = 1, ..., 5. \end{aligned}$$

See the proof in Section A.5 of the Appendix.

We then compute the variance partition for all the species using the entries of $\boldsymbol{\omega}$ defined as:

$$\boldsymbol{\omega} = \frac{1}{\sum_{p=1}^{6} s_p^2 + \sigma_S^2 + \sigma_T^2} \left[s_1^2, ..., s_6^2, \sigma_S^2, \sigma_T^2\right].$$

Figure 4.18 (a) reports the posterior mean of the entries of $\boldsymbol{\omega}$ for all the 39 species.

Our analysis indicates that the primary factors influencing occurrence variability are the spatial effect, the *Depth* effect, and the *Bottom temperature* one. This aligns with our expectations, as the selected species are all demersal and depth is a known driver of habitat suitability. The temporal effect has a relative small contribution, expect for a few species.

The type of plot as Figure 4.18 is often used in ecology to assess the contribution of different covariates. However, it does not provide any information about the uncertainty on the estimates for the variance partition. In the next subsection, we focus on the analysis of the results for a single species so that we can showcase more

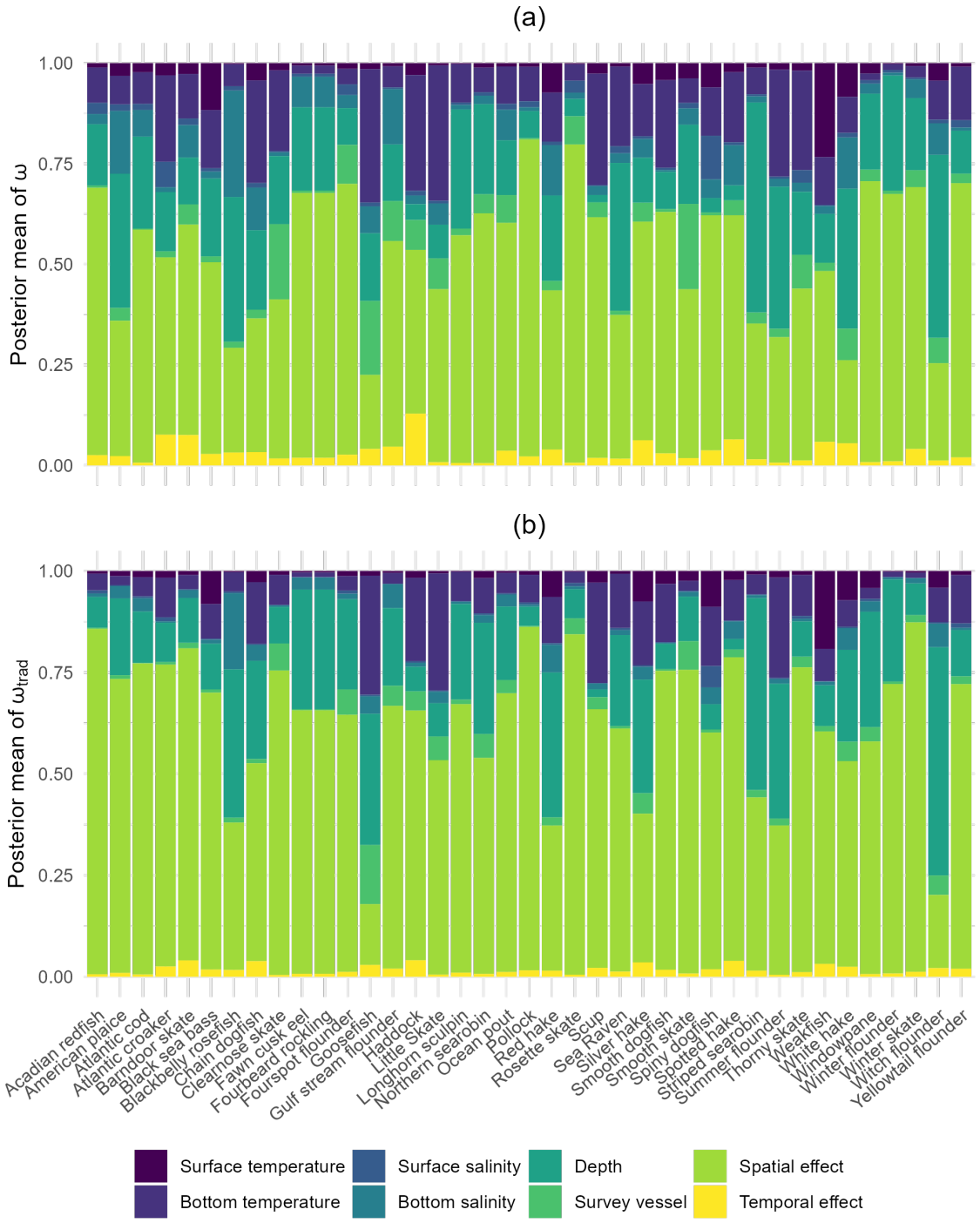

Figure 4.18: Variance partition estimates for the 39 different species: (a) Posterior means of the entries of $\boldsymbol{\omega}$; (b) posterior means of the entries of $\boldsymbol{\omega}_{\text{trad}}$.

outputs and summaries about the quantification of the variance contributions of the effects.

The proposed method of variance partitioning estimation is compared to the

more traditional approach, comparing $\boldsymbol{\omega}$ to $\boldsymbol{\omega}_{\text{trad}}$:

$$\boldsymbol{\omega}_{\text{trad}} = \frac{1}{\sum_{p=1}^{6} W_p + W_S + W_T} [W_1, ..., W_6, W_S, W_T]$$

where:

$$\begin{aligned} W_p &= Var_{i=1}^{N}[f_p(x_{ip})] \qquad p = 1, ..., 6 \\ W_S &= Var_{i=1}^{N}[f_S(\boldsymbol{z}_i)] \\ W_T &= Var_{i=1}^{N}[f_T(t_i)]. \end{aligned}$$

Figure 4.18 (b) reports the posterior means of $\boldsymbol{\omega}_{\text{trad}}$. The comparison shows that the discrepancy between the two methods is large in some species and modest but noticeable in others. This result suggests that the transition to the proposed approach has appreciable practical implications, along with the stronger theoretical foundations detailed above.

Analysis of the results of a single species distribution model

We now analyse the results for a single species, namely the *Summer flounder*, which has been chosen among the species with the largest overall occurrence level (28.3%).

From Figure 4.18 (a), it can be noted that the most importnant factors can be identified (in decreasing order) as the *Depth* effect, the spatial effect, and the *Bottom temperature* effect, similar to what happens for the majority of the species. To assess the uncertainty around this conclusion, Figure 4.19 reports the marginal distributions of both $s_1^2, ..., s_6^2, \sigma_S^2, \sigma_T^2$ and of the entries of $\boldsymbol{\omega}$. While there is quite a large uncertainty around the variance contribution of the top 3 factors mentioned above, the small role played by the remaining effects appears quite surely. Table 4.1 reports the posterior summaries of the entries of $\boldsymbol{\omega}$.

The *Depth* effect contribution has a larger mean estimate but also the widest credible interval; conversely, the *Bottom temperature* effect and the spatial effect contributions have both a slightly smaller mean and a smaller standard deviation. From these results, we may conclude that we are fairly certain about the most relevant drivers of occurrence variability for the *Summer flounder* (i.e. *Depth*, Bottom temperature, spatial effect) but there is quite a large uncertainty about the partition among these three factors.

Despite the use of smooth non-linear trends for the environmental covariates, the novel specification of P-Splines allows for the conservation of the immediate typical interpretation of the linear regression coefficients. Table 4.2 reports posterior

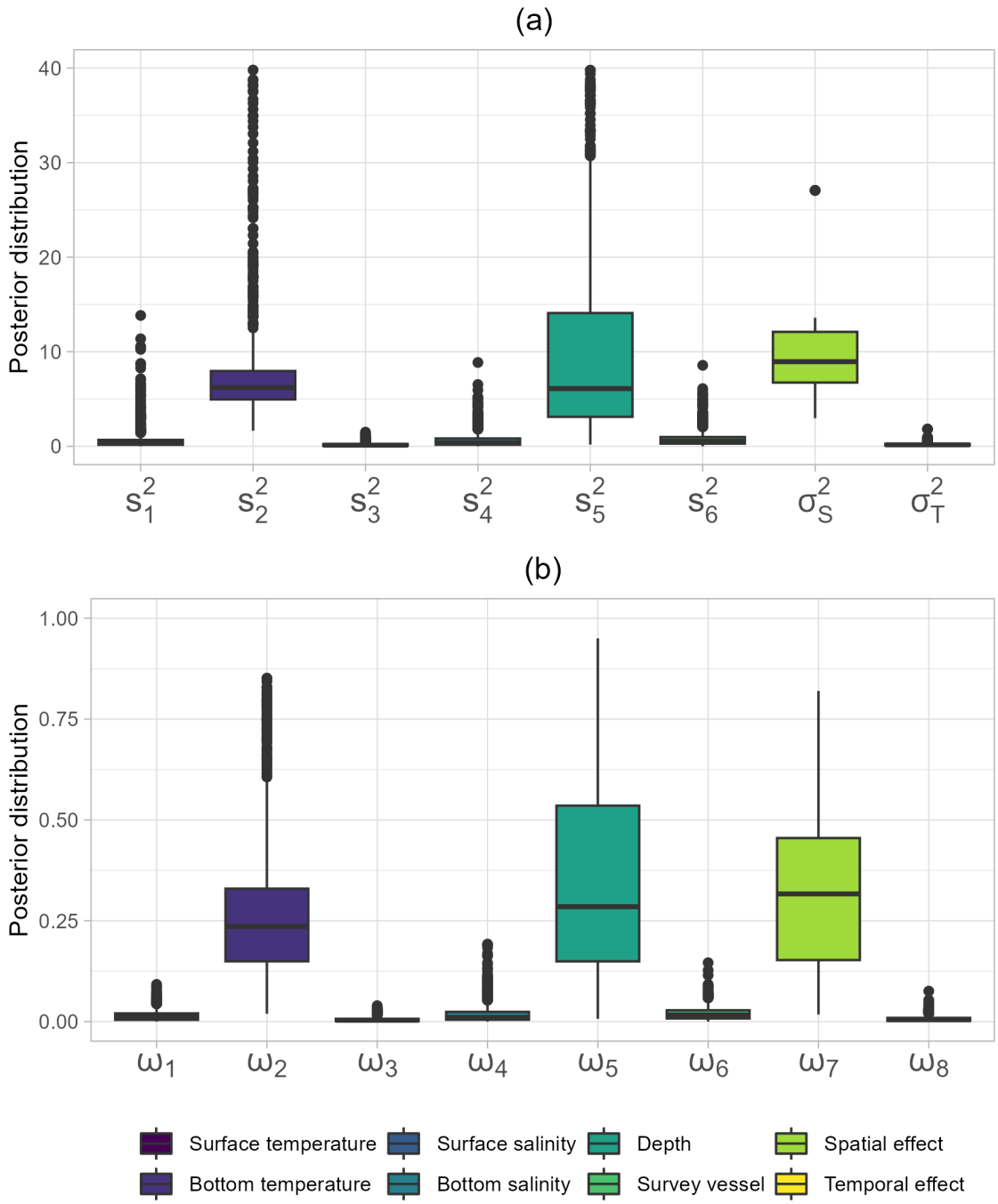

Figure 4.19: (a) Posterior marginal distribution of $s_1^2, ..., s_6^2, \sigma_S^2, \sigma_T^2$; (b) posterior marginal distribution of $\omega_1, ..., \omega_8$ for the *Summer flounder* species.

summaries for $\beta_1, ..., \beta_5$, along with β_6 *Survey vessel* dummy covariate.

The largest values in posterior mean belong to β_5 and β_2, which align with the result from the variance partitioning analysis. Moreover, the signs of the linear effects for *Depth* and *Bottom temperature* are consistent with the known habitat preferences of this species, which favours warmer, shallower waters (Packer et al.

Covariate	ω	Mean (%)	SD (%)	95% C.I. (%)
Surface temperature	ω_1	1.55	1.45	[0.05, 5.15]
Bottom temperature	ω_2	26.17	16.71	[4.29, 74.3]
Surface salinity	ω_3	0.49	0.59	[0.01, 2.17]
Bottom salinity	ω_4	1.85	2.69	[0.08, 8.56]
Depth	ω_5	36.41	25.65	[3.60,89.94]
Survey vessel	ω_6	2.04	1.76	[0.06, 6.63]
Spatial effect	ω_7	30.86	17.43	[3.43,61.05]
Temporal effect	ω_8	0.63	0.70	[0.02, 2.67]

Table 4.1: Posterior summaries of the entries of $\boldsymbol{\omega}$ expressed as percentage for the *Summer flounder*. The 95% credible interval is computed using the 2.5 and the 97.5 percentiles.

Covariate	β	Mean	SD	95% C.I.
Surface temperature	β_1	-0.59	0.47	[-2.07, 0.03]
Bottom temperature	β_2	2.69	1.53	[1.46, 7.77]
Surface salinity	β_3	-0.18	0.24	[-0.69, 0.24]
Bottom salinity	β_4	-0.34	0.42	[-1.29, 0.36]
Depth	β_5	-3.47	3.45	[-12.96, 0.50]
Survey vessel	β_6	0.57	0.29	[0.11, 1.32]

Table 4.2: Posterior summaries of the linear coefficients of the environmental covariates and the *Survey vessel* covariate for the *Summer flounder*. The 95% credible interval is computed using the 2.5 and the 97.5 percentiles.

1999). However, only β_2 and β_6 are significant if we consider the 95% credible intervals. Looking at the linear coefficients gives only a partial representation of the results, as the model allows for non-linearity in the effects of the environmental covariates. The large standard deviations reported for some of the covariates might be caused by the correlation between the environmental factors (Figure 4.5). The impact of a potential collinearity issue should be further studied in this case study, for example comparing the results with the ones from a simpler model with only the identified relevant covariates. Since the aim in this context is to illustrate the applicability of the HD priors, we do not consider further analysis on the subject here.

Posterior summaries of the overall trends $f_p(X_p)$ of the top 2 relevant covariates (i.e. *Depth* and *Bottom temperature*) are represented in Figure 4.20. Some non-linearity appears in the trend for the *Bottom temperature* effect, whose credible interval is quite tight in most of the support. Contrarily, the posterior mean of the trend for *Depth* appears quasi-linear, while the credible interval is much wider

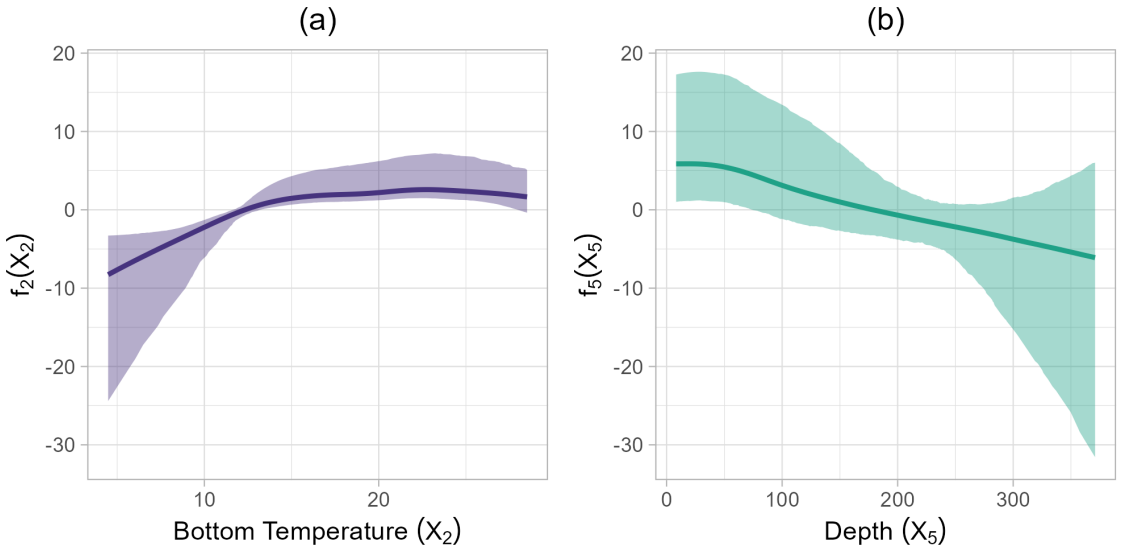

Figure 4.20: Posterior means (solid line) and 95% credible interval for the effects of the environmental covariates $f_p(X_p)$ for the *Summer flounder*: (a) *Bottom temperature*; (b) *Depth*.

and does not exclude fully linear effects. The difference in uncertainty level for the covariates is coherent with the mentioned results from Figure 4.19.

We further investigate the decomposition in linear and non-linear contribution to assess which conclusions can be made a posteriori about the necessity for smooth non-linear trends on the covariates. Figure 4.21 (a) reports separately the posterior summaries of the linear and non-linear components of the two effects. Additionally, we consider the posterior distribution of the non-linear proportion of these effects defined as:

$$w_{Np} = \frac{s^2_{Np}}{s^2_{Lp} + s^2_{Np}}.$$

Figure 4.21 (a) shows relatively strong evidence for non-linearity in the trend of the *Bottom temperature* covariate, while there is more uncertainty around the role played by the non-linear component in the *Depth* trend, whose mode is however close to 0. These results might suggest that a linear effect for *Depth* might be sufficiently flexible in this example.

Finally, the irregularity in the width of the credible bands in Figures 4.20-4.21 is likely due to the fact that the binary observations for the response are not uniformly present over the support of the covariates or that the presence and absence observations are not well separated in some areas: as a consequence, some of the P-Spline coefficients seem to better estimated than others due to their location on

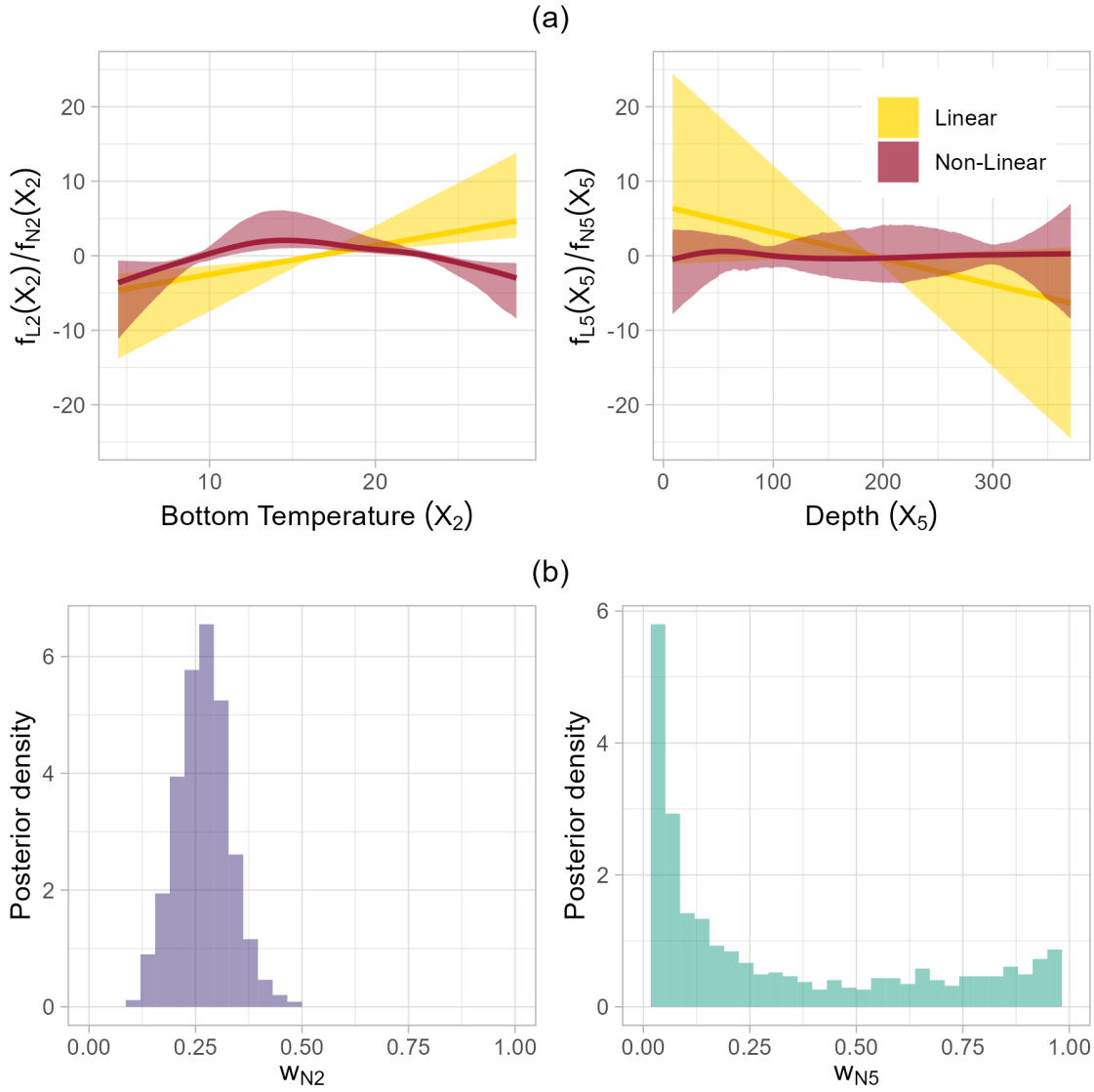

Figure 4.21: (a) Posterior means and 95% credible intervals of $f_{Lp}(X_p)$ and $f_{Np}(X_p)$ for $p = 2, 5$ for the *Summer flounder*; (b) posterior distribution of w_{N2} and w_{N5}.

the covariate support.

4.6 Discussion

The goal of this chapter was to showcase the benefits of the extension of VP priors proposed in Chapter 3. To do so, we have chosen the context of species distribution models and we have detailed a default way to set up a sensible VP prior on SDMs using the HD approach. First, we have presented a default tree design that groups effects according to their nature and can be used as a starting point for users. Secondly, we have outlined guidelines for the prior specification on the new parameters

based on a literature review of previous applications of VP priors. In the future, we aim at extending this default approach to fields with models that have a similar structure, such as disease mapping and environmental quality.

On the topic of prior specification, we can mention two minor contributions. First, we have presented a simple strategy for a more intuitive regulation of the hyperparameter of a symmetric Dirichlet. Secondly, we have extended the proof of Franco-Villoria, Ventrucci, and Rue 2022 on PC_0 prior for proportions, and proposed a simple condition to check whether the prior simplifies.

The application to the NOAA-NEFSC dataset gave us the opportunity to study the application of VP priors in the context of a complex model. First, we illustrated how to apply the theory from Chapter 3 on each effect of the model, so as to obtain the correct interpretation of the variance parameters. Secondly, we showcased the versatility of the default tree design explaining how users can prune and expand it, according to the case study at hand, following sensible modelling principles. Finally, in the prior specification step, we have proposed the use of PC_0 priors for the proportion of variance explained by the non-linear component of P-Spline effects. This prior proposal could be further investigated in the future to assess whether it can become a viable method to promote model simplicity whilst simultaneously still allowing for flexibility.

It is important to mention that in the analysis of the 39 available species, we chose to use a *stacked* models' approach, which treats species as mutually independent. This assumption is quite strict and it does not respect contemporary community ecology theory that recognizes the large role played by biotic interactions on occurrence levels. For this reason, the modelling standard of community ecology consists today in the use of *joint species distribution models* (JSDM) (Ovaskainen et al. 2017), which acknowledge potential correlations between species' occurrences. Future research could focus on exploring whether VP priors can be useful in the context of JSDMs, which do not fall in the category of LGMs as the precision matrices are not fixed but need to be estimated. A possible implementation could consist of a multi-step estimation procedure in which the correlation matrices and the other model parameters are separately estimated, similar to the method proposed by Hui et al. 2023 in a frequentist setting.

Finally, a new method to estimate variance partitioning a posteriori has been presented, along with its multiple advantages. The application to the case study suggested that the difference between the novel approach and the traditional one is non-negligible. We intend to further investigate this method and its properties in future applications.

Chapter 5

Conclusions

This thesis aimed at extending the applicability of VP priors to both the fixed and random branches of latent Gaussian models (Chapter 3), and at illustrating the advantages of this extension in a field requiring complex models with many different types of effects (Chapter 4).

The standardization procedure from Chapter 3 represents the main contribution of the thesis. It guarantees an interpretation for the VP parameters that is intuitive for the user and it is able to account for the different nature of fixed effects, as it was hoped for in the seminal work of Fuglstad et al. 2020. The simulation study confirmed the benefit of applying the standardization procedure, in particular of the scaling step, when VP priors are adopted. Negligible difference was noted in the simulated scenarios between our proposed scaling procedure and the one of Sørbye and Rue 2014. Finally, the VP priors are found once more to be competitive with the popular Penalized Complexity priors (PC, Simpson et al. 2017), although the latter showed to be more robust to the misapplication of the standardization procedure.

In studying the concept of variance contribution of effects, we have thoroughly discussed the class of IGMRF effects: we presented a novel representation of such effects made up two separate components, namely a polynomial term and a residual one. Moreover, we have proposed the idea of modifying the precision matrices of IGMRF effects (Q modification), when necessary to remove identifiability issues and guarantee a neat separation between the polynomial and residual contributions. The modification is necessary in all cases where a basis of spline functions is required, such as smooth (non linear) effects of covariates using penalized splines.

Chapter 4 focused on highlighting the practical benefits of the theory from Chapter 3, through the specification of VP priors for species distribution models. In doing so, we have developed a default strategy for the specification of VP priors for SDMs using the HD approach, consisting of guidelines for the design of the decomposition

tree and the prior specification step. The chosen case study also illustrates how the special case of a spatial continuous effect modelled using P-Splines can be correctly treated to accomodate the requirements of VP priors.

An additional contribution of this chapter is the derivation of a sufficient condition under which the PC_0 prior for proportions takes a simplified form, which extends the work of Franco-Villoria, Ventrucci, and Rue 2022. Finally, we have also proposed a new, more intuitive, method to perform variance partitioning estimation, which takes into account the inferential interest of the user, as well as the distributional assumption made on the covariates. The application to the case study showed that the new method can lead to substantially different conclusions in comparison to the traditional approach.

In terms of future work, we aim to apply VP priors to more ecology case studies and investigate whether their benefits can be extended to the context of Joint Species Distribution models, more commonly employed nowadays in community ecology. Additionally, the workflow discussed in Chapter 4 could also be applied to applications from other fields, such as disease mapping and environmental quality assessment.

With regard to more theoretical aspects, various research lines have emerged from the thesis. In particular, we argue that the P-Spline alternative representation and the use of a PC prior for the penalization of its non-linear component is in itself an interesting prior choice that might have the potential to be useful in different contexts. The quality of this proposal could be possibly assessed in the future via extensive simulation. In general, we believe that the Q modification is a relevant contribution of our work. In the future, it would be interesting to study the implication of Q modification in other types of models, such as space-time smoothing via penalized splines.

Other interesting points that may deserve further consideration include the impact of misspecification of the distributional assumption on the covariates, which might greatly affect conclusions about variance contributions, and more generally the proposed variance partitioning estimation method.

Bibliography

Aguilar, Javier Enrique and Paul-Christian Bürkner (2023). "Intuitive joint priors for Bayesian linear multilevel models: The R2D2M2 prior". *Electronic Journal of Statistics* 17.1, pp. 1711–1767.

Aguilar, Javier Enrique and Paul-Christian Bürkner (2024). "Generalized Decomposition Priors on R2". *arXiv preprint arXiv:2401.10180*.

Araújo, Miguel B and Miska Luoto (2007). "The importance of biotic interactions for modelling species distributions under climate change". *Global Ecology and Biogeography* 16.6, pp. 743–753.

Bach, Paul and Nadja Klein (2024). "Bayesian Effect Selection in Additive Models with an Application to Time-to-Event Data". arXiv: `2401.00840 [stat.ME]`. URL: `https://arxiv.org/abs/2401.00840`.

Bai, Ray and Malay Ghosh (2021). "On the beta prime prior for scale parameters in high-dimensional Bayesian regression models". *Statistica Sinica* 31.2, pp. 843–865.

Banerjee, Sudipto, Bradley P Carlin, and Alan E Gelfand (2003). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Begon, Michael and Colin R Townsend (2020). *Ecology: from individuals to ecosystems*. John Wiley & Sons.

Besag, Julian and Charles Kooperberg (1995). "On conditional and intrinsic autoregressions". *Biometrika* 82.4, pp. 733–746.

Besag, Julian, Jeremy York, and Annie Mollié (1991). "Bayesian image restoration, with two applications in spatial statistics". *Annals of the institute of statistical mathematics* 43, pp. 1–20.

Bhattacharya, Anirban et al. (2015). "Dirichlet–Laplace priors for optimal shrinkage". *Journal of the American Statistical Association* 110.512, pp. 1479–1490.

Breslow, N (1972). "Disussion of regression models and life-tables by cox, dr". *J. Roy. Statist. Assoc., B* 34, pp. 216–217.

Burgazzi, Gemma et al. (2020). "Communities in high definition: Spatial and environmental factors shape the micro-distribution of aquatic invertebrates". *Freshwater Biology* 65.12, pp. 2053–2065.

Clark, James Samuel and Alan E Gelfand (2006). *Hierarchical modelling for the environmental sciences: statistical methods and applications*. OUP Oxford.

Cressie, Noel and Nicolas Verzelen (2008). "Conditional-mean least-squares fitting of Gaussian Markov random fields to Gaussian fields". *Computational Statistics & Data Analysis* 52.5, pp. 2794–2807.

Currie, Iain D and Maria Durban (2002). "Flexible smoothing with P-splines: a unified approach". *Statistical Modelling* 2.4, pp. 333–349.

Dormann, Carsten F et al. (2012). "Correlation and process in species distribution models: bridging a dichotomy". *Journal of Biogeography* 39.12, pp. 2119–2131.

Durbin, James and Siem Jan Koopman (2012). *Time series analysis by state space methods*. Vol. 38. OUP Oxford.

Eilers, Paul HC and Brian D Marx (1996). "Flexible smoothing with B-splines and penalties". *Statistical science* 11.2, pp. 89–121.

Elith, Jane and John R Leathwick (2009). "Species distribution models: ecological explanation and prediction across space and time". *Annual review of ecology, evolution, and systematics* 40.1, pp. 677–697.

Fahrmeir, Ludwig, Thomas Kneib, and Stefan Lang (2004). "Penalized structured additive regression for space-time data: a Bayesian perspective". *Statistica Sinica*, pp. 731–761.

Franco-Villoria, Maria, Massimo Ventrucci, and Håvard Rue (2022). "Variance partitioning in spatio-temporal disease mapping models". *Statistical Methods in Medical Research* 31.8, pp. 1566–1578.

Frühwirth-Schnatter, Sylvia and Helga Wagner (2010). "Stochastic model specification search for Gaussian and partial non-Gaussian state space models". *Journal of Econometrics* 154.1, pp. 85–100.

Fuglstad, Geir-Arne et al. (2020). "Intuitive Joint Priors for Variance Parameters". *Bayesian Analysis* 15.4, pp. 1109 –1137.

Geir-Arne Fuglstad Daniel Simpson, Finn Lindgren and Håvard Rue (2019). "Constructing Priors that Penalize the Complexity of Gaussian Random Fields". *Journal of the American Statistical Association* 114.525, pp. 445–452.

Gelman, Andrew (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)". *Bayesian Analysis* 1.3, pp. 515 –534.

Gelman, Andrew and Jennifer Hill (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, Andrew, Daniel Simpson, and Michael Betancourt (2017). "The prior can often only be understood in the context of the likelihood". *Entropy* 19.10, p. 555.

Gelman, Andrew et al. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, Andrew et al. (2019). "R-squared for Bayesian regression models". *The American Statistician*.

Gombin, Joal, Ramnath Vaidyanathan, and Vladimir Agafonkin (2020). "concaveman: A Very Fast 2D Concave Hull Algorithm". URL: `https://CRAN.R-project.org/package=concaveman`.

Guisan, Antoine and Carsten Rahbek (2011). "SESAM–a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages". *Journal of Biogeography* 38.8, pp. 1433–1444.

Hayward, Matt W et al. (2015). "Ecologists need robust survey designs, sampling and analytical methods". *Journal of Applied Ecology* 52.2, pp. 286–290.

Hem, Ingeborg, Geir-Arne Fuglstad, and Andrea Riebler (2024). "makemyprior: Intuitive Construction of Joint Priors for Variance Parameters in R". *Journal of Statistical Software* 110.3, 1–39.

Hem, Ingeborg Gullikstad et al. (2021). "Robust modeling of additive and nonadditive variation with intuitive inclusion of expert knowledge". *Genetics* 217.3.

Henderson, Robin, Silvia Shimakura, and David Gorst (2002). "Modeling spatial variation in leukemia survival data". *Journal of the American Statistical Association* 97.460, pp. 965–972.

Hodges, James S (2013). *Richly parameterized linear models: additive, time series, and spatial models using random effects*. CRC Press.

Hoffman, Kenneth and Ray Kunze (1971). *Linear Algebra*. Prentice Hall.

Holand, Anna Marie et al. (2013). "Animal models and integrated nested Laplace approximations". *G3: Genes, Genomes, Genetics* 3.8, pp. 1241–1251.

Hrafnkelsson, Birgir (2023). *Statistical Modeling Using Bayesian Latent Gaussian Models: With Applications in Geophysics and Environmental Sciences*. Springer Cham.

Hui, Francis KC et al. (2023). "Spatiotemporal joint species distribution modelling: A basis function approach". *Methods in Ecology and Evolution* 14.8, pp. 2150–2164.

iNaturalist (2024). Accessed 2024-09-06. URL: `https://www.inaturalist.org/`.

Kendrick, David A (1981). *Stochastic control for economic models*. McGraw-Hill Inc., US.

Kereszturi, Monika, Jonathan Tawn, and Philip Jonathan (2016). "Assessing extremal dependence of North Sea storm severity". *Ocean Engineering* 118, pp. 242–259.

Klein, Nadja and Thomas Kneib (2016). "Scale-Dependent Priors for Variance Parameters in Structured Additive Distributional Regression". *Bayesian Analysis* 11.4, pp. 1071 –1106.

Kneib, Thomas and Ludwig Fahrmeir (2007). "A mixed model approach for geoadditive hazard regression". *Scandinavian Journal of Statistics* 34.1, pp. 207–228.

Knorr-Held, Leonhard (2000). "Bayesian modelling of inseparable space-time variation in disease risk". *Statistics in medicine* 19.17-18, pp. 2555–2567.

Lambert, Paul C et al. (2005). "How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS". *Statistics in medicine* 24.15, pp. 2401–2428.

Lang, Stefan and Andreas Brezger (2004). "Bayesian P-splines". *Journal of computational and graphical statistics* 13.1, pp. 183–212.

Lawson, Andrew B (2018). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. Chapman and Hall/CRC.

Lindgren, Finn and Håvard Rue (2008). "On the second-order random walk model for irregular locations". *Scandinavian journal of statistics* 35.4, pp. 691–700.

Lunn, David et al. (2009). "Rejoinder to commentaries on: "The BUGS project: Evolution, critique and future directions"". *Statistics in Medicine* 28.25, pp. 3081–3082.

Marques, Isa, Paul FV Wiemann, and Thomas Kneib (2023). "A Variance Partitioning Multi-level Model for Forest Inventory Data with a Fixed Plot Design". *Journal of Agricultural, Biological and Environmental Statistics* 28.4, pp. 706–725.

Martino, Sara, Rupali Akerkar, and Håvard Rue (2011). "Approximate Bayesian inference for survival models". *Scandinavian Journal of Statistics* 38.3, pp. 514–528.

Moraga, Paula (2019). *Geospatial health data: Modeling and visualization with R-INLA and shiny*. Chapman and Hall/CRC.

NEFSC Fall Bottom Trawl Survey (2024). Accessed: August 30, 2024. URL: `https://www.fisheries.noaa.gov/inport/item/22560`.

Ordonez, Jose A et al. (2024). "Penalized complexity priors for the skewness parameter of power links". *Canadian Journal of Statistics* 52.1, pp. 98–117.

Ovaskainen, Otso et al. (2017). "How to make more out of community data? A conceptual framework and its implementation as models and software". *Ecology letters* 20.5, pp. 561–576.

Packer, DB et al. (1999). "Summer Flounder, Paralichthys dentatus, life history and habitat characteristics". *NOAA Technical Memorandum NMFS-NE* 151.

Peres-Neto, Pedro R et al. (2006). "Variation partitioning of species data matrices: estimation and comparison of fractions". *Ecology* 87.10, pp. 2614–2625.

Pettit, LI (1990). "The conditional predictive ordinate for the normal distribution". *Journal of the Royal Statistical Society: Series B (Methodological)* 52.1, pp. 175–184.

Piironen, Juho and Aki Vehtari (2017). "Sparsity information and regularization in the horseshoe and other shrinkage priors". *Electronic Journal of Statistics* 11.2, pp. 5018 –5051.

Riebler, Andrea et al. (2016). "An intuitive Bayesian spatial model for disease mapping that accounts for scaling". *Statistical methods in medical research* 25.4, pp. 1145–1165.

Rue, Håvard and Leonhard Held (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

Rue, Håvard, Sara Martino, and Nicolas Chopin (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 71.2, pp. 319–392.

Rue, Håvard and Hååkon Tjelmeland (2002). "Fitting Gaussian Markov random fields to Gaussian fields". *Scandinavian journal of Statistics* 29.1, pp. 31–49.

Simpson, Daniel et al. (2017). "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors". *Statistical Science* 32.1, pp. 1 –28.

Sofaer, Helen R et al. (2019). "Development and delivery of species distribution models to inform decision-making". *BioScience* 69.7, pp. 544–557.

Sørbye, Sigrunn Holbek and Håvard Rue (2014). "Scaling intrinsic Gaussian Markov random field priors in spatial modelling". *Spatial Statistics* 8, pp. 39–51.

Sørbye, Sigrunn Holbek and Håvard Rue (2017). "Penalised complexity priors for stationary autoregressive processes". *Journal of Time Series Analysis* 38.6, pp. 923–935.

Spyropoulou, Maria-Zafeiria and James Bentham (2024). "Scaling priors for intrinsic Gaussian Markov random fields applied to blood pressure data". *Statistica Neerlandica* 78.3, pp. 491–504.

Sullivan, Brian L et al. (2009). "eBird: A citizen-based bird observation network in the biological sciences". *Biological conservation* 142.10, pp. 2282–2292.

U.S. Fish and Wildlife Service Open Data (n.d.). Accessed on 2024-09-06. URL: `https://gis-fws.opendata.arcgis.com`.

Ventrucci, Massimo and Håvard Rue (2016). "Penalized complexity priors for degrees of freedom in Bayesian P-splines". *Statistical Modelling* 16.6, pp. 429–453.

Wakefield, Jon (2007). "Disease mapping and spatial regression with count data". *Biostatistics* 8.2, pp. 158–183.

Wakefield, Jon et al. (2013). *Bayesian and frequentist regression methods*. Vol. 23. Springer.

Warton, David I et al. (2015). "So many variables: joint modeling in community ecology". *Trends in ecology & evolution* 30.12, pp. 766–779.

Wei, Ran et al. (2020). "Sparse Bayesian additive nonparametric regression with application to health effects of pesticides mixtures". *Statistica Sinica* 30.1, pp. 55–79.

Wood, Simon N (2017). *Generalized additive models: an introduction with R*. CRC press.

Yanchenko, Eric, Howard D. Bondell, and Brian J Reich (2024a). "Spatial regression modeling via the R2D2 framework". *Environmetrics* 35.2.

Yanchenko, Eric, Howard D Bondell, and Brian J Reich (2024b). "The R2D2 prior for generalized linear mixed models". *The American Statistician*, pp. 1–10.

Zhang, Yan and Howard D Bondell (2018). "Variable selection via penalized credible regions with Dirichlet–Laplace global-local shrinkage priors". *Bayesian Analysis* 13.3, pp. 823–844.

Zhang, Yan Dora et al. (2022). "Bayesian regression using a prior on the model fit: The r2-d2 shrinkage prior". *Journal of the American Statistical Association* 117.538, pp. 862–874.

Appendix

A Proofs of Chapter 3

A.1 Equations 3.13-3.14 from Section 3.3.3

Consider Equation 3.12. For the fixed effects, i.e. $j = 1, ..., L$, we have that $\boldsymbol{u}_j \in \boldsymbol{\theta}$. Hence, the expected variance of interest simplifies as follows:

$$E_{\boldsymbol{\theta}}\{Var_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\boldsymbol{\theta}]|\boldsymbol{\sigma}\} = E_{\boldsymbol{u}_j}\{Var_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\}$$

which is equal to the expected finite-population variance and equivalent to Equation 3.13.

On the other hand, $\sigma_j^2 \in \boldsymbol{\theta}$ for random effects $j = L + 1, ..., J$. Hence, the expected variance of interest is in this case equal to:

$$\begin{aligned} E_{\boldsymbol{\theta}}\{Var_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\boldsymbol{\theta}]|\boldsymbol{\sigma}\} &= E_{\sigma_j^2}\{Var_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2]|\sigma_j^2\} \\ &= Var_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2] \end{aligned}$$

The expectation $E_{\sigma_j^2}[\cdot|\sigma_j^2]$ disappears as the expectation with respect to a random variable conditional on the same random variable is just equal to its argument. Therefore, we find that the expected variance of interest for random effects is simply equal to the super-population variance (Equation 3.14).

A.2 Equation 3.17 from Section 3.3.5

Consider again the intuitive interpretation for the variance of a fixed effect and drop the index j for convenience. This quantity can be expressed as a difference between two terms if the variance is written in terms of difference of expectations:

$$\begin{aligned} E_{\boldsymbol{u}}\{Var_X[f(X)|\boldsymbol{u}]|\sigma^2\} &= E_{\boldsymbol{u}}\{E_X[f^2(X)|\boldsymbol{u}] - E_X^2[f(X)|\boldsymbol{u}]|\sigma^2\} \\ &= E_{\boldsymbol{u}}\{E_X[f^2(X)|\boldsymbol{u}]|\sigma^2\} - E_{\boldsymbol{u}}\{E_X^2[f(X)|\boldsymbol{u}]|\sigma^2\} \end{aligned}$$

At this stage, the order of integration in the first term can be changed as long as $E_{\boldsymbol{u}}\{E_X[f^2(X)|\boldsymbol{u}]|\sigma^2\}$ is finite (Fubini-Tonelli theorem). Inverting the expectations, the first term becomes equal to the marginal variance given σ^2. The second term can also be rewritten as a variance, noting that $E_{\boldsymbol{u}}\{E_X[f(X)|\boldsymbol{u}]|\sigma^2\} = 0$:

$$\begin{aligned} E_{\boldsymbol{u}}[Var_X[f(X)|\boldsymbol{u}]|\sigma^2] &= E_X\{E_{\boldsymbol{u}}[f^2(X)|X]|\sigma^2\} - E_{\boldsymbol{u}}\{E_X^2[f(X)|\boldsymbol{u}]|\sigma^2\} \\ &= Var_{X,\boldsymbol{u}}[f(X)|\sigma^2] - Var_{\boldsymbol{u}}\{E_X[f(X)|\boldsymbol{u}]|\sigma^2\} \end{aligned}$$

A.3 Equation 3.18 from Section 3.3.5

The second term of Equation 3.17 can be simplified as a function of a vector containing the expectations of the basis functions $\boldsymbol{a} = [E_X[D_1(X)], ..., E_X[D_K(X)]]^T$ with respect to X:

$$\begin{aligned} E_{\boldsymbol{u}}\{E_X^2[f(X)|\boldsymbol{u}]|\sigma^2\} &= E_{\boldsymbol{u}}\{E_X^2[\boldsymbol{D}(X)^T\boldsymbol{u}|\boldsymbol{u}]|\sigma^2\} \\ &= E_{\boldsymbol{u}}\left\{\left[\sum_{k=1}^{K} E_X[D_k(X)]\cdot u_k\right]^2 |\sigma^2\right\} \\ &= E_{\boldsymbol{u}}[(\boldsymbol{a}^T\boldsymbol{u})^2|\sigma^2] \end{aligned}$$

The argument of the final expectation is now simply a quadratic form, which can be neatly expressed in terms of $\boldsymbol{a}$ and $\boldsymbol{Q}$:

$$\begin{aligned} E_{\boldsymbol{u}}\{E_X^2[f(X)|\boldsymbol{u}]|\sigma^2\} &= E_{\boldsymbol{u}}[\boldsymbol{u}^T\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{u}|\sigma^2] \\ &= \sigma^2\mathrm{tr}[\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Q}^*] \end{aligned}$$

As a consequence, Equation 3.18 simplifies if effects have been scaled according to Proposition 1:

$$E_{\boldsymbol{u}}\{Var_X[f(X)|\boldsymbol{u}]|\sigma^2\} = \sigma^2 - \sigma^2\mathrm{tr}[\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Q}^*]$$

A.4 Equation 3.21 from Section 3.3.6

Equation 3.21 can be proven verifying that:

$$\begin{gathered}\int_{\mathcal{X}} x^0 \cdot \boldsymbol{D}^T(x)\boldsymbol{u} \cdot \pi(x)\,\mathrm{d}x = 0 \\ \int_{\mathcal{X}} x^1 \cdot \boldsymbol{D}^T(x)\boldsymbol{u} \cdot \pi(x)\,\mathrm{d}x = 0 \\ \ldots \\ \int_{\mathcal{X}} x^{d-1} \cdot \cdot \boldsymbol{D}^T(x)\boldsymbol{u} \cdot \pi(x)\,\mathrm{d}x = 0\end{gathered}$$

In the case of $X \sim \text{Unif}([1, K])$ and $D_k(X) = \mathbb{I}(X = k)$, these constraints simplify to:

$$\begin{gathered}\sum_{k=1}^{K} k^0 \cdot u_k = 0 \\ \sum_{k=1}^{K} k \cdot u_k = 0 \\ \ldots \\ \sum_{k=1}^{K} k^{d-1} \cdot u_k = 0\end{gathered}$$

Imposing $\boldsymbol{S}_{(d-1)}\boldsymbol{u} = \boldsymbol{0}$ guarantees each of these constraints (see Equation 3.9) and thus ensures that $f_r(X)$ has a null polynomial trend of degree $d-1$.

$$\begin{aligned}\widetilde{\boldsymbol{S}}^T\boldsymbol{u} &= \int_{\mathcal{X}} \boldsymbol{S}_{(d-1)}^T(x)\boldsymbol{D}^T(x)\boldsymbol{u} \cdot \pi(x)\,\mathrm{d}x \\ &= \sum_{k=1}^{K} \boldsymbol{S}_{(d-1)}^T(k) \left[\sum_{j=1}^{K} \mathbb{I}(k = j) \cdot u_j\right] \frac{1}{K} \\ &= \sum_{k=1}^{K} \boldsymbol{S}_{(d-1)}^T(k) \cdot u_k \cdot \frac{1}{K} \\ &= \frac{1}{K}\boldsymbol{S}_{(d-1)}^T\boldsymbol{u}\end{aligned}$$

Hence, $\widetilde{\boldsymbol{S}} \propto \boldsymbol{S}_{(d-1)}$.

A.5 Equation 3.31 from Section 3.3.6

Consider $f(X) = f_t(X) + f_r(X)$ as defined in Equation 3.30. First, we note that the variance of $f(X)$ with respect to both X and the coefficients $\beta_1, .., \beta_{d-1}, \boldsymbol{u}$ is equal to:

$$\begin{aligned}
Var_{X,\beta_1,...,\beta_{d-1},\boldsymbol{u}}[f(X)] &= Var_{X,\beta_1,...,\beta_{d-1},\boldsymbol{u}}\left[\sum_{m=1}^{d-1} f_{t_m}(X) + f_r(X)\right] \\
&= E_X\left\{Var_{\beta_1,...,\beta_{d-1},\boldsymbol{u}}\left[\sum_{m=1}^{d-1} f_{t_m}(X) + f_r(X)\right]\right\} \\
&= \sum_{m=1}^{d-1} E_X\{Var_{\beta_m}[f_{t_m}(X)]\} \\
&+ E_X\{Var_{\boldsymbol{u}}[f_r(X)]\} \\
&+ \sum_{l=1}^{d-2}\sum_{m>l}^{d-1} E_X\{Cov_{\beta_l,\beta_m}[f_{t_l}(X), f_{t_m}(X)]\} \\
&+ \sum_{m=1}^{d-1} E_X\{Cov_{\beta_m,\boldsymbol{u}}[f_{t_m}(X), f_r(X)]\}
\end{aligned}$$

If all the effects in $f(X)$ have been appropriately scaled as suggested in Proposition 1, then the variance simplifies to:

$$\begin{aligned}
Var_{X,\beta_1,...,\beta_{d-1},\boldsymbol{u}}[f(X)] &= \sum_{m=l}^{d-1} \sigma_m^2 + \sigma_r^2 \\
&+ \sum_{l=1}^{d-2}\sum_{m>l}^{d-1} E_X\{Cov_{\beta_l,\beta_m}[f_{t_l}(X), f_{t_m}(X)]\} \\
&+ \sum_{m=1}^{d-1} E_X\{Cov_{\beta_m,\boldsymbol{u}}[f_{t_m}(X), f_r(X)]\}
\end{aligned}$$

To show that all the covariance terms are null, we can use the properties of the distributions chosen on the parameters, specifically null mean and mutual independence

assumption between $\beta_1, ..., \beta_{d-1}, \boldsymbol{u}$. For $l = 1, ..., d-2$ and $m > l,\ m = 2, ..., d-1$:

$$\begin{aligned} E_X\{Cov_{\beta_l,\beta_m}[f_{t_l}(X), f_{t_m}(X)]\} &= E_X\{E_{\beta_l,\beta_m}[f_{t_l}(X) \cdot f_{t_m}(X)]\} \\ &- E_X\{E_{\beta_l}[f_{t_l}(X)] \cdot E_{\beta_m}[f_{t_m}(X)]\} \\ &= E_X\{E_{\beta_l,\beta_m}[h_l(X)\beta_l \cdot h_m(X)\beta_m]\} \\ &- E_X\{E_{\beta_l}[h_l(X)\beta_l] \cdot E_{\beta_m}[h_m(X)\beta_m]\}\} \\ &= 0 \end{aligned}$$

Additionally, for $m = 1, ..., d-1$:

$$\begin{aligned} E_X\{Cov_{\beta_m,\boldsymbol{u}}[f_{t_m}(X), f_r(X)]\} &= E_X\{E_{\beta_m,\boldsymbol{u}}[f_{t_l}(X) \cdot f_r(X)]\} \\ &- E_X\{E_{\beta_m}[f_{t_m}(X)] \cdot E_{\boldsymbol{u}}[f_r(X)]\} \\ &= E_X\{E_{\beta_m,\boldsymbol{u}}[h_m(X)\beta_m \cdot \mathbf{D}^T(X)\mathbf{u}]\} \\ &- E_X\{E_{\beta_m}[h_m(X)\beta_m] \cdot E_{\boldsymbol{u}}[\mathbf{D}^T(X)\mathbf{u}]\} \\ &= 0 \end{aligned}$$

Hence, we can finally write the variance of $f(X)$ as:

$$Var_{X,\beta_1,...,\beta_{d-1},\boldsymbol{u}}[f(X)] = \sum_{m=l}^{d-1} \sigma_m^2 + \sigma_r^2$$

which completes the proof.

However, we have an alternative method to prove that all the covariance terms are null even in the absence of the mutual independence assumption, thanks to the design constraints imposed on $f_{t_m}(X)$ and $f_r(X)$. This method is particularly useful a posteriori when the prior assumptions on the parameters do not hold anymore. First, we change the order of expectation so that for $l = 1, ..., d-2,\ m > l\ m = 2, ..., d-1$:

$$\begin{aligned} E_X\{Cov_{\beta_l,\beta_m}[f_{t_l}(X), f_{t_m}(X)]\} &= E_X\{E_{\beta_l,\beta_m}[h_l(X)\beta_l \cdot h_m(X)\beta_m]\} \\ &- E_X\{E_{\beta_l}[h_l(X)\beta_l] \cdot E_{\beta_m}[h_m(X)\beta_m]\}\} \\ &= E_{\beta_l,\beta_m}\{E_X[h_l(X)\beta_l \cdot h_m(X)\beta_m]\} \\ &- E_{\beta_l,\beta_m}\{E_X[h_l(X)\beta_l] \cdot E_X[h_m(X)\beta_m]\}\} \end{aligned}$$

Then, we note from Equation 3.24 that $E_X[h_m(X)\beta_m] = 0,\ m = 1, ..., d-1$ so that

we can simplify to:

$$\begin{aligned} E_X\{Cov_{\beta_l,\beta_m}[f_{t_l}(X), f_{t_m}(X)]\} &= E_{\beta_l,\beta_m}\{E_X[h_l(X)\beta_l \cdot h_m(X)\beta_m]\} \\ &= E_{\beta_l,\beta_m}\{\beta_l\beta_m \cdot E_X[h_l(X) \cdot h_m(X)]\} \end{aligned}$$

From the definition of $h_m(X)$ functions from Equation 3.23, we can then note that each $h_l(X)$ is a polynomial function of degree l. Knowing that by design $h_m(X)$ must respect the constraints of Equation 3.24, we can find that

$$\begin{aligned} E_X\{Cov_{\beta_l,\beta_m}[f_{t_l}(X), f_{t_m}(X)]\} &= E_{\beta_l,\beta_m}\left\{\beta_l\beta_m \cdot E_X\left[\sum_{j=1}^{j} a_j \cdot x^j \cdot h_m(X)\right]\right\} \\ &= E_{\beta_l,\beta_m}\left\{\beta_l\beta_m \cdot \sum_{j=1}^{j} a_j E_X\left[x^j \cdot h_m(X)\right]\right\} \\ &= 0 \end{aligned}$$

and this is true for all $l = 1, ..., d-2,\ m > l,\ m = 2, ..., d-1$. A similar logic can be used for the other covariance terms $E_X\{Cov_{\beta_m,\boldsymbol{u}}[f_{t_m}(X), f_r(X)]\},\ m = 1, ..., d-1$, using the constraints $\widetilde{\boldsymbol{S}}^T\boldsymbol{u} = \boldsymbol{0}$.

A.6 Remark 1

Remark 1 can be proven showing that $E_{\boldsymbol{\theta}}\{Var_{\boldsymbol{X},\boldsymbol{u}_1,...,\boldsymbol{u}_J}[\eta|\mu,\boldsymbol{\theta}]|\boldsymbol{\sigma}\} = \sum_{j=1}^{J}\sigma_j^2$.

First, we can note that:

$$E_{\boldsymbol{\theta}}\{Var_{\boldsymbol{X},\boldsymbol{u}_1,...,\boldsymbol{u}_J}[\eta|\mu,\boldsymbol{\theta}]|\boldsymbol{\sigma}\} = E_{\boldsymbol{\theta}}\{Var_{\boldsymbol{X},\boldsymbol{u}_1,...,\boldsymbol{u}_J}[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{\theta}]|\boldsymbol{\sigma}\}$$

Secondly, we rewrite $E_{\boldsymbol{\theta}}\{Var_{\boldsymbol{X},\boldsymbol{u}_1,...,\boldsymbol{u}_J}[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{\theta}]|\boldsymbol{\sigma}\}$ as:

$$E_{\boldsymbol{u}_1,...,\boldsymbol{u}_L}\left\{Var_{\boldsymbol{X},\boldsymbol{u}_{L+1},...,\boldsymbol{u}_J}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{u}_1, ..., \boldsymbol{u}_L, \sigma_{L+1}^2, ..., \sigma_J^2\right]|\sigma_1^2, ..., \sigma_L^2\right\}$$

This expression can be rewritten more concisely using the notation $\boldsymbol{U}_F = [\boldsymbol{u}_1, ..., \boldsymbol{u}_L]$, $\boldsymbol{U}_R = [\boldsymbol{u}_{L+1}, ..., \boldsymbol{u}_J]$,$\boldsymbol{\sigma}_F = [\sigma_1^2, ..., \sigma_L^2]$, $\boldsymbol{\sigma}_R = [\sigma_{L+1}^2, ..., \sigma_J^2]$ as:

$$E_{\boldsymbol{U}_F}\left\{Var_{\boldsymbol{X},\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{U}_F, \boldsymbol{\sigma}_R\right]|\boldsymbol{\sigma}_F\right\}$$

Then, we can write the argument of the expectation using the law of total variance

as:

$$
\begin{aligned}
Var_{\boldsymbol{X},\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{U}_F,\boldsymbol{\sigma}_R\right] &= E_{\boldsymbol{X}}\left\{Var_{\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{X},\boldsymbol{U}_F,\boldsymbol{\sigma}_R\right]|\boldsymbol{U}_F,\boldsymbol{\sigma}_R\right\} \\
&+ Var_{\boldsymbol{X}}\left\{E_{\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{X},\boldsymbol{U}_F,\boldsymbol{\sigma}_R\right]|\boldsymbol{U}_F,\boldsymbol{\sigma}_R\right\} \\
&= E_{\boldsymbol{X}}\left\{Var_{\boldsymbol{U}_R}\left[\sum_{j=L+1}^{J} f_j(X_j)|\boldsymbol{X},\boldsymbol{\sigma}_R\right]|\boldsymbol{\sigma}_R\right\} \\
&+ Var_{\boldsymbol{X}}\left[\sum_{j=1}^{L} f_j(X_j)|\boldsymbol{U}_F\right] \\
&= \sum_{j=L+1}^{J} E_{\boldsymbol{X}}\left\{Var_{\boldsymbol{U}_R}\left[f_j(X_j)|\boldsymbol{X},\boldsymbol{\sigma}_R\right]|\boldsymbol{\sigma}_R\right\} \\
&+ Var_{\boldsymbol{X}}\left[\sum_{j=1}^{L} f_j(X_j)|\boldsymbol{U}_F\right] \\
&= \sum_{j=L+1}^{J} E_{\boldsymbol{X}}\left\{E_{\boldsymbol{U}_R}\left[f_j^2(X_j)|\boldsymbol{X},\boldsymbol{\sigma}_R\right]\right\} \\
&+ Var_{\boldsymbol{X}}\left[\sum_{j=1}^{L} f_j(X_j)|\boldsymbol{U}_F\right]
\end{aligned}
$$

If a 0-mean constraint is imposed on the $j = 1, ..., L$ effects, then:

$$
Var_{\boldsymbol{X}}\left[\sum_{j=1}^{L} f_j(X_j)|\boldsymbol{U}_F\right] = E_{\boldsymbol{X}}\left\{\left[\sum_{j=1}^{L} f_j(X_j)\right]^2|\boldsymbol{U}_F\right\}
$$

so that:

$$
\begin{aligned}
Var_{\boldsymbol{X},\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{U}_F,\boldsymbol{\sigma}_R\right] &= \sum_{j=L+1}^{J} E_{\boldsymbol{X}}\left\{E_{\boldsymbol{U}_R}\left[f_j^2(X_j)|\boldsymbol{X},\boldsymbol{\sigma}_R\right]\right\} \\
&+ E_{\boldsymbol{X}}\left\{\left[\sum_{j=1}^{L} f_j(X_j)\right]^2|\boldsymbol{U}_F\right\}
\end{aligned}
$$

If we consider again $E_{\boldsymbol{U}_F}\left\{Var_{\boldsymbol{X},\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{U}_F,\boldsymbol{\sigma}_R\right]|\boldsymbol{\sigma}_F\right\}$, it can be written

as:

$$E_{\boldsymbol{U}_F}\left\{Var_{\boldsymbol{X},\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{U}_F,\boldsymbol{\sigma}_R\right]|\boldsymbol{\sigma}_F\right\} = \sum_{j=L+1}^{J} E_{\boldsymbol{X}}\left\{E_{\boldsymbol{U}_R}\left[f_j^2(X_j)|\boldsymbol{X},\boldsymbol{\sigma}_R\right]\right\} + E_{\boldsymbol{U}_F}\left\{E_{\boldsymbol{X}}\left\{\left[\sum_{j=1}^{L} f_j(X_j)\right]^2|\boldsymbol{U}_F\right\}|\boldsymbol{\sigma}_F\right\}.$$

Inverting the order of expectation, we get:

$$\begin{aligned}
E_{\boldsymbol{U}_F}\left\{Var_{\boldsymbol{X},\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{U}_F,\boldsymbol{\sigma}_R\right]|\boldsymbol{\sigma}_F\right\} &= \sum_{j=L+1}^{J} E_{\boldsymbol{X}}\left\{E_{\boldsymbol{U}_R}\left[f_j^2(X_j)|\boldsymbol{X},\boldsymbol{\sigma}_R\right]\right\} \\
&\quad + E_{\boldsymbol{X}}\left\{E_{\boldsymbol{U}_F}\left\{\left[\sum_{j=1}^{L} f_j(X_j)\right]^2|\boldsymbol{\sigma}_F\right\}\right\} \\
&= \sum_{j=L+1}^{J} E_{\boldsymbol{X}}\left\{E_{\boldsymbol{U}_R}\left[f_j^2(X_j)|\boldsymbol{X},\boldsymbol{\sigma}_R\right]\right\} \\
&\quad + \sum_{j=1}^{L} E_{\boldsymbol{X}}\left\{E_{\boldsymbol{U}_F}\left[\sum_{j=1}^{L} f_j^2(X_j)\right]|\boldsymbol{\sigma}_F\right\} \\
&= \sum_{j=1}^{J} E_{\boldsymbol{X}}\left\{E_{\boldsymbol{U}_F,\boldsymbol{U}_R}\left[f_j^2(X_j)|\boldsymbol{\sigma}_F,\boldsymbol{\sigma}_R\right]\right\} \\
&= \sum_{j=1}^{J} \sigma_j^2 \cdot E_{X_j}\left[\boldsymbol{D}_j^T(X_j)\boldsymbol{Q}_j^*\boldsymbol{D}_j(X_j)\right]
\end{aligned}$$

If scaling has been applied as in Proposition 1, then we know that:

$$\sum_{j=1}^{J} \sigma_j^2 E_{X_j}\left[\boldsymbol{D}_j^T(X_j)\boldsymbol{Q}_j^*\boldsymbol{D}_j(X_j)\right] = \sum_{j=1}^{J} \sigma_j^2$$

which completes the proof.

A.7 Equation 3.35 from Example 1

Equation 2.29 of Rue and Held 2005 can be used to find the covariance matrix for an originally i.i.d. effect under constraint $\boldsymbol{a}^T\boldsymbol{u} = \mathbf{0}$:

$$\begin{aligned}\boldsymbol{Q}^* &= \boldsymbol{I} - \boldsymbol{I}\boldsymbol{a}(\boldsymbol{a}^T\boldsymbol{I}\boldsymbol{a})^{-1}\boldsymbol{a}^T\boldsymbol{I} \\ &= \boldsymbol{I} - \boldsymbol{a}(\boldsymbol{a}^T\boldsymbol{a})^{-1}\boldsymbol{a}^T \\ &= \boldsymbol{I} - \frac{\boldsymbol{a}\boldsymbol{a}^T}{\boldsymbol{a}^T\boldsymbol{a}}\end{aligned}$$

$\boldsymbol{a}$ must be equal to $[p_1, ..., p_K]^T$ to obtain that $E_X[f(X)]$:

$$\begin{aligned}\boldsymbol{a}^T\boldsymbol{u} = \mathbf{0} = E_X[f(X)] &= \sum_{k=1}^{K} p_k f(k) \\ &= \sum_{k=1}^{K} p_k \left[\sum_{j=1}^{K} \mathbb{I}(k=j)\cdot u_j\right] \\ &= \sum_{k=1}^{K} p_k \cdot u_k \\ &= \begin{bmatrix} p_1 & p_2 & ... & p_K \end{bmatrix} \boldsymbol{u}\end{aligned}$$

Finally, C can be found applying Proposition 1, knowing $\boldsymbol{Q}^*$ and $\boldsymbol{a}$:

$$\begin{aligned}C &= \sum_{k=1}^{K} p_k \boldsymbol{D}(k)\left[\boldsymbol{I} - \frac{\boldsymbol{a}\boldsymbol{a}^T}{\boldsymbol{a}^T\boldsymbol{a}}\right]\boldsymbol{D}^T(k) \\ &= 1 - \sum_{k=1}^{K} p_k \boldsymbol{D}^T(k)\left[\frac{\boldsymbol{a}\boldsymbol{a}^T}{\boldsymbol{a}^T\boldsymbol{a}}\right]\boldsymbol{D}(k) \\ &= 1 - \frac{1}{\boldsymbol{a}^T\boldsymbol{a}}\sum_{k=1}^{K} p_k \boldsymbol{D}^T(k)\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{D}(k) \\ &= 1 - \frac{1}{\sum_{k=1}^{K} p_k^2}\sum_{k=1}^{K} p_k \left[\sum_{j=1}^{K} \mathbb{I}(j=k)p_j\right]^2 \\ &= 1 - \frac{1}{\sum_{k=1}^{K} p_k^2}\sum_{k=1}^{K} p_k^3 \\ &= 1 - \frac{\sum_{k=1}^{K} p_k^3}{\sum_{k=1}^{K} p_k^2}\end{aligned}$$

A.8 Proposition 3

Let $X_1 \perp\!\!\!\perp X_2$ so that $E_{X_1,X_2}[g(X_1)\cdot h(X_2)] = E_{X_1}[g(X_1)]\cdot E_{X_2}[h(X_2)]$ for measurable $g(X_1)$ and $h(X_2)$. Then, we can use the properties of the Kronecker product to show that:

$$\begin{aligned}
Var_{X_1,X_2,\boldsymbol{u}_1,\boldsymbol{u}_2}[f(X_1,X_2)|\sigma^2] &= \sigma^2 E_{X_1,X_2}[\boldsymbol{D}^T(X_1,X_2)\boldsymbol{Q}^*\boldsymbol{D}(X_1,X_2)] \\
&= \sigma^2 E_{X_1,X_2}[\boldsymbol{D}_1^T(X_1)\boldsymbol{Q}_1^*\boldsymbol{D}_1(X_1)\cdot\boldsymbol{D}_2^T(X_2)\boldsymbol{Q}_2^*\boldsymbol{D}_2(X_2)] \\
&= \sigma^2 E_{X_1,X_2}[\boldsymbol{D}_1^T(X_1)\boldsymbol{Q}_1^*\boldsymbol{D}_1(X_1)\cdot\boldsymbol{D}_2^T(X_2)\boldsymbol{Q}_2^*\boldsymbol{D}_2(X_2)] \\
&= \sigma^2 \cdot Var_{X_1,\boldsymbol{u}_1}[\widetilde{f}_1(X_1)|\sigma_1^2=1]\cdot Var_{X_2,\boldsymbol{u}_2}[\widetilde{f}_2(X_2)|\sigma_2^2=1].
\end{aligned}$$

Since $\widetilde{f}_1(X_1)$ and $\widetilde{f}_2(X_2)$ have been scaled, their marginal variances are both equal to 1 and the variance of the interaction effect simplifies, proving that the effect is already appropriately scaled by design:

$$Var_{X_1,X_2,\boldsymbol{u}_1,\boldsymbol{u}_2}[f(X_1,X_2)|\sigma^2] = \sigma^2$$

A.9 Derivarion of $\widetilde{\mathbf{S}}$ in Example 5

The columns of matrix $\widetilde{\boldsymbol{S}} = [\widetilde{\boldsymbol{s}}_0 \quad \widetilde{\boldsymbol{s}}_1]$ are defined as:

$$\begin{aligned}
\widetilde{\boldsymbol{s}}_0 &= \int_m^M \boldsymbol{B}(x)\cdot\pi(x)\,\mathrm{d}x \\
\widetilde{\boldsymbol{s}}_1 &= \int_m^M x\cdot\boldsymbol{B}(x)\cdot\pi(x)\,\mathrm{d}x
\end{aligned}$$

In order to explicitly derive $\widetilde{\boldsymbol{S}}$ is necessary to first derive explicit expressions for $\boldsymbol{B}(x)$. In order to derive analytically the B-Spline basis, it is first necessary to make explicit the degree d and the number of basis functions K so that a B-Spline can be denoted by $\boldsymbol{B}_K^{(d)}(X)$. Hence, the cubic B-Spline is actually $\boldsymbol{B}(X) = \boldsymbol{B}_K^{(3)}(X)$. Eilers and Marx 1996 defines analytically B-Splines using the recursion formula from

Equation 45:

$$B_k^{(0)}(X) = \mathbb{I}\left[\frac{k-1}{K-D} < \frac{X-m}{M-m} < \frac{k}{K-D}\right]$$

$$B_k^{(d)}(X) = \begin{cases} \frac{1}{d}\left[d + \frac{X-m}{M-m}(K-D) - k + 1\right] B_{k-1}^{(d-1)}(X) + \\ \frac{1}{d}\left[k - \frac{X-m}{M-m}(K-D)\right] B_k^{(d-1)}(X) & 1 \le k \le K-D+d \\ \\ 0 & \text{otherwise} \end{cases} \tag{0.1}$$

From the recursive formula of Equation 0.1 and following, it can be found that the elements of $\boldsymbol{B}(x) = [B_1(x), ..., B_K(x)]$ are:

$$\begin{aligned} B_k(x) = \Big[& \mathbb{I}(k-1 < \hat{x}(K-3) < k) \cdot g_1\left(\hat{x}(K-3) - (k-1)\right) + \\ & \mathbb{I}(k-2 < \hat{x}(K-3) < k-1) \cdot g_2\left(\hat{x}(K-3) - (k-2)\right) + \\ & \mathbb{I}(k-3 < \hat{x}(K-3) < k-2) \cdot g_3\left(\hat{x}(K-3) - (k-3)\right) + \\ & \mathbb{I}(k-4 < \hat{x}(K-3) < k-3) \cdot g_4\left(\hat{x}(K-3) - (k-4)\right) \Big] \mathbb{I}(0 < \hat{x} < 1) \end{aligned}$$

where $\hat{x} = \dfrac{x-m}{M-m} \in [0,1]$ and:

$$\begin{aligned} g_1(y) &= \frac{1}{2}\left[-\frac{y^3}{3} + y^2 - y + \frac{1}{3}\right] \\ g_2(y) &= \frac{y^3}{2} - y^2 + \frac{2}{3} \\ g_3(y) &= \frac{1}{2}\left[-y^3 + y^2 + y + \frac{1}{3}\right] \\ g_4(y) &= \frac{y^3}{6} \end{aligned}$$

Noting that $\boldsymbol{B}(x)$ is in fact only a function of the normalized version $\hat{x}$:

$$\begin{aligned} \widetilde{\boldsymbol{s}}_0 &= (M-m)\int_0^1 \boldsymbol{B}(\hat{x}) \cdot \pi((M-m)\hat{x} + m)\,\mathrm{d}\hat{x} \\ \widetilde{\boldsymbol{s}}_1 &= (M-m) \cdot m \cdot \widetilde{\boldsymbol{s}}_0 + (M-m)^2 \int_0^1 \hat{x} \cdot \boldsymbol{B}(\hat{x}) \cdot \pi((M-m)\hat{x} + m)\,\mathrm{d}\hat{x} \end{aligned}$$

If $X \sim \text{Unif}(m, M)$ so that $\pi(x) = \dfrac{I(m < x < M)}{M - m}$:

$$\begin{aligned}
\widetilde{\boldsymbol{s}}_0 &= \int_0^1 \boldsymbol{B}(\hat{x})\,\mathrm{d}\hat{x} \\
&= \frac{1}{K-3}\left[\frac{1}{24}, \frac{1}{2}, \frac{23}{24}, 1, ..., 1, \frac{23}{24}, \frac{1}{2}, \frac{1}{24}\right]^T, \quad K \geq 7 \\
\widetilde{\boldsymbol{s}}_1 &= m \cdot \widetilde{\boldsymbol{s}}_0 + (M-m)\left[\int_0^1 \hat{x}\boldsymbol{B}(\hat{x})\,\mathrm{d}\hat{x}\right] \\
&= m \cdot \widetilde{\boldsymbol{s}}_0 + (M-m) \cdot \boldsymbol{v}
\end{aligned}$$

where $\boldsymbol{v} = [v_1, ..., v_K]^T$ for $K \geq 7$:

$$\begin{aligned}
v_1 &= \frac{1}{(K-3)^2} \cdot \frac{1}{120} \\
v_2 &= \frac{1}{(K-3)^2} \cdot \frac{7}{30} \\
v_3 &= \frac{1}{(K-3)^2} \cdot \frac{121}{120} \\
v_k &= \frac{k-2}{(K-3)^2} \quad k = 4, ..., K-3 \\
v_{K-2} &= \frac{1}{K-3} \cdot \frac{23}{24} - \frac{1}{(K-3)^2} \cdot \frac{121}{120} \\
v_{K-1} &= \frac{1}{K-3} \cdot \frac{1}{2} - \frac{1}{(K-3)^2} \cdot \frac{7}{30} \\
v_K &= \frac{1}{K-3} \cdot \frac{1}{24} - \frac{1}{(K-3)^2} \cdot \frac{1}{120}
\end{aligned}$$

For $K = 6$:

$$\widetilde{\boldsymbol{s}}_0 = \left[\frac{1}{72}, \frac{1}{6}, \frac{23}{72}, \frac{23}{72}, \frac{1}{6}, \frac{1}{72}\right]^T \quad \boldsymbol{v} = \left[\frac{1}{1080}, \frac{7}{270}, \frac{121}{1080}, \frac{28}{135}, \frac{19}{135}, \frac{7}{540}\right]^T \tag{0.2}$$

For $K = 5$:

$$\widetilde{\boldsymbol{s}}_0 = \left[\frac{1}{48}, \frac{1}{4}, \frac{11}{24}, \frac{1}{4}, \frac{1}{48}\right]^T \quad \boldsymbol{v} = \left[\frac{1}{480}, \frac{7}{120}, \frac{11}{48}, \frac{23}{120}, \frac{3}{160}\right]^T \tag{0.3}$$

For $K = 4$:

$$\widetilde{\boldsymbol{s}}_0 = \left[\frac{1}{24}, \frac{11}{24}, \frac{11}{24}, \frac{1}{24}\right]^T \quad \boldsymbol{v} = \left[\frac{1}{120}, \frac{11}{60}, \frac{11}{40}, \frac{1}{30}\right]^T \tag{0.4}$$

A.10 Equation 3.45 from Example 5

Let $\widetilde{\boldsymbol{R}} = \widetilde{\boldsymbol{G}} - \widetilde{\boldsymbol{W}}$ as in Equation 3.45 and $\widetilde{\boldsymbol{Q}} = (\boldsymbol{\Lambda}\widetilde{\boldsymbol{R}}^*\boldsymbol{\Lambda})^*$. Then $\widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{S}} = \boldsymbol{0}$ if $\widetilde{\boldsymbol{R}}\boldsymbol{\Lambda}\widetilde{\boldsymbol{S}} = \boldsymbol{0}$, i.e. $\widetilde{\boldsymbol{G}}\boldsymbol{\Lambda}\widetilde{\boldsymbol{S}} - \widetilde{\boldsymbol{W}}\boldsymbol{\Lambda}\widetilde{\boldsymbol{S}} = \boldsymbol{0}$ where:

$$\begin{aligned}\widetilde{\boldsymbol{G}}\boldsymbol{\Lambda}\widetilde{\boldsymbol{S}} - \widetilde{\boldsymbol{W}}\boldsymbol{\Lambda}\widetilde{\boldsymbol{S}} &= \begin{bmatrix} \widetilde{G}_{1,1}\cdot\lambda_1\cdot\widetilde{S}_{1,0} & \widetilde{G}_{1,1}\cdot\lambda_1\cdot\widetilde{S}_{1,1} \\ \widetilde{G}_{2,2}\cdot\lambda_2\cdot\widetilde{S}_{2,0} & \widetilde{G}_{2,2}\cdot\lambda_2\cdot\widetilde{S}_{2,1} \\ \dots & \dots \\ \widetilde{G}_{K,K}\cdot\lambda_K\cdot\widetilde{S}_{K,0} & \widetilde{G}_{K,K}\cdot\lambda_K\cdot\widetilde{S}_{K,1} \end{bmatrix} \\ &- \begin{bmatrix} \sum_{l=1}^K \widetilde{W}_{1,l}\cdot\lambda_l\cdot\widetilde{S}_{l,0} & \sum_{l=1}^K \widetilde{W}_{1,l}\cdot\lambda_l\cdot\widetilde{S}_{l,1} \\ \sum_{l=1}^K \widetilde{W}_{2,l}\cdot\lambda_l\cdot\widetilde{S}_{l,0} & \sum_{l=1}^K \widetilde{W}_{2,l}\cdot\lambda_l\cdot\widetilde{S}_{l,1} \\ \dots & \dots \\ \sum_{l=1}^K \widetilde{W}_{K,l}\cdot\lambda_l\cdot\widetilde{S}_{l,0} & \sum_{l=1}^K \widetilde{W}_{K,l}\cdot\lambda_l\cdot\widetilde{S}_{l,1} \end{bmatrix}\end{aligned}$$

Then $\widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{S}} = \boldsymbol{0}$ if $\forall k = 1, ..., K$:

$$\begin{bmatrix} \widetilde{G}_{k,k}\cdot\lambda_k\cdot\widetilde{S}_{k,0} & \widetilde{G}_{k,k}\cdot\lambda_k\cdot\widetilde{S}_{k,1} \end{bmatrix} = \begin{bmatrix} \sum_{l=1}^K \widetilde{W}_{k,l}\cdot\lambda_l\cdot\widetilde{S}_{l,0} & \sum_{l=1}^K \widetilde{W}_{k,l}\cdot\lambda_l\cdot\widetilde{S}_{l,1} \end{bmatrix}$$

Replacing the entries of $\widetilde{\boldsymbol{G}}$ with their definition from Equation 3.47:

$$\begin{bmatrix} \sum_{l=1}^K \widetilde{W}_{k,l}\lambda_l\widetilde{S}_{l,0} & \frac{\widetilde{S}_{k,1}}{\widetilde{S}_{k,0}}\sum_{l=1}^K \widetilde{W}_{k,l}\lambda_l\widetilde{S}_{l,0} \end{bmatrix} = \begin{bmatrix} \sum_{l=1}^K \widetilde{W}_{k,l}\lambda_l\widetilde{S}_{l,0} & \sum_{l=1}^K \widetilde{W}_{k,l}\lambda_l\widetilde{S}_{l,1} \end{bmatrix}$$

Hence, since the first elements of both vectors are equal, it is only necessary to verify that $\forall k = 1, ..., K$:

$$\begin{gathered}\frac{\widetilde{S}_{k,1}}{\widetilde{S}_{k,0}}\sum_{l=1}^K \widetilde{W}_{k,l}\lambda_l\widetilde{S}_{l,0} - \sum_{l=1}^K \widetilde{W}_{k,l}\lambda_l\widetilde{S}_{l,1} = 0 \\ \Downarrow \\ \sum_{l=1}^K \widetilde{W}_{k,l}\cdot\lambda_l\cdot(\widetilde{S}_{l,0}\cdot\widetilde{S}_{k,1} - \widetilde{S}_{l,1}\cdot\widetilde{S}_{k,0}) = 0\end{gathered}$$

Replacing the entries of $\widetilde{\boldsymbol{W}}$ with their definition from Equation 3.47, these K conditions are transformed into:

$$\lambda_k^{-1}\sum_{l=1}^K (k-l)\cdot W_{k,l} = 0 \quad k = 1, ..., K$$

which is true for all k if $\boldsymbol{W}$ is defined as in Equation 3.44.

A.11 Derivarion of $\widetilde{\mathbf{S}}$ in Example 6

Let $X_1, X_2 \overset{iid}{\sim} \text{Unif}([m_1, M_1] \times [m_2, M_2])$.

$$\begin{aligned}
\widetilde{\boldsymbol{S}} &= \int_{m_1}^{M_1} \int_{m_2}^{M_2} \boldsymbol{B}_{K_1 \times K_2}(x_1, x_2) \pi(x_1, x_2) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \\
&= \frac{1}{(M_1 - m_1)(M_2 - m_2)} \int_{m_1}^{M_1} \int_{m_2}^{M_2} \boldsymbol{B}_{K_1}(x_1) \otimes \boldsymbol{B}_{K_2}(x_2) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \\
&= \frac{1}{(M_1 - m_1)(M_2 - m_2)} \left[\int_{m_1}^{M_1} \boldsymbol{B}_{K_1}(x_1) \, \mathrm{d}x_1 \right] \otimes \left[\int_{m_2}^{M_2} \boldsymbol{B}_{K_2}(x_2) \, \mathrm{d}x_2 \right] \\
&= \left[\frac{1}{M_1 - m_1} \int_{m_1}^{M_1} \boldsymbol{B}_{K_1}(x_1) \, \mathrm{d}x_1 \right] \otimes \left[\frac{1}{M_2 - m_2} \int_{m_2}^{M_2} \boldsymbol{B}_{K_2}(x_2) \, \mathrm{d}x_2 \right]
\end{aligned}$$

Using the results from the previous section:

$$\begin{aligned}
\widetilde{\boldsymbol{S}} \;&=\; \frac{1}{K_1 - 3} \left[\frac{1}{24}, \frac{1}{2}, \frac{23}{24}, 1, ..., 1, \frac{23}{24}, \frac{1}{2}, \frac{1}{24} \right]^T \\
&\otimes\; \frac{1}{K_2 - 3} \left[\frac{1}{24}, \frac{1}{2}, \frac{23}{24}, 1, ..., 1, \frac{23}{24}, \frac{1}{2}, \frac{1}{24} \right]^T \qquad K_1 \geq 7, K_2 \geq 7
\end{aligned}$$

In the case of $3 < K_1 < 7$ or $3 < K_2 < 7$, the correct entries are found considering the Kronecker product for $\widetilde{\boldsymbol{s}}_0$ for K_1 and $\widetilde{\boldsymbol{s}}_0$ for K_2 as derived in the Section A.9 in Equations 0.2 and following.

A.12 Equation 3.54 from Example 6

Let $\widetilde{\boldsymbol{R}} = \widetilde{\boldsymbol{G}} - \widetilde{\boldsymbol{W}}$ as in Equation 3.54 and $\widetilde{\boldsymbol{Q}} = (\boldsymbol{\Lambda} \widetilde{\boldsymbol{R}}^* \boldsymbol{\Lambda})^*$. Then $\widetilde{\boldsymbol{Q}} \widetilde{\boldsymbol{S}} = \mathbf{0}$ if $\widetilde{\boldsymbol{R}} \boldsymbol{\Lambda} \widetilde{\boldsymbol{S}} = \mathbf{0}$, i.e. $\widetilde{\boldsymbol{G}} \boldsymbol{\Lambda} \widetilde{\boldsymbol{S}} - \widetilde{\boldsymbol{W}} \boldsymbol{\Lambda} \widetilde{\boldsymbol{S}} = \mathbf{0}$ where:

$$\widetilde{\boldsymbol{G}} \boldsymbol{\Lambda} \widetilde{\boldsymbol{S}} - \widetilde{\boldsymbol{W}} \boldsymbol{\Lambda} \widetilde{\boldsymbol{S}} = \begin{bmatrix} \widetilde{G}_{1,1} \cdot \lambda_1 \cdot \widetilde{S}_1 \\ ... \\ \widetilde{G}_{K_1 \times K_2, K_1 \times K_2} \cdot \lambda_{K_1 \times K_2} \cdot \widetilde{S}_{K_1 \times K_2} \end{bmatrix} - \begin{bmatrix} \sum_{l=1}^{K_1 \times K_2} \widetilde{W}_{1,l} \cdot \lambda_l \cdot \widetilde{S}_l \\ ... \\ \sum_{l=1}^{K_1 \times K_2} \widetilde{W}_{K_1 \times K_2, l} \cdot \lambda_l \cdot \widetilde{S}_l \end{bmatrix}$$

Replacing $\widetilde{\boldsymbol{G}}$ and $\widetilde{\boldsymbol{W}}$ with their definitions from Equations 3.55-3.56 and remem-

bering Equation 3.53, it is found that:

$$\widetilde{\boldsymbol{G}}\boldsymbol{\Lambda}\widetilde{\boldsymbol{S}} - \widetilde{\boldsymbol{W}}\boldsymbol{\Lambda}\widetilde{\boldsymbol{S}} = \begin{bmatrix} \frac{G_{1,1}}{\lambda_1 \cdot \widetilde{S}_1} \\ ... \\ \frac{G_{K_1\times K_2, K_1\times K_2}}{\lambda_{K_1\times K_2} \cdot \widetilde{S}_{K_1\times K_2}} \end{bmatrix} - \begin{bmatrix} \frac{\sum_{l=1}^{K_1\times K_2} W_{1,l}}{\lambda_1 \cdot \widetilde{S}_1} \\ ... \\ \frac{\sum_{l=1}^{K_1\times K_2} W_{K_1\times K_2,l}}{\lambda_{K_1\times K_2} \cdot \widetilde{S}_{K_1\times K_2}} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{G_{1,1}}{\lambda_1 \cdot \widetilde{S}_1} \\ ... \\ \frac{G_{K_1\times K_2, K_1\times K_2}}{\lambda_{K_1\times K_2} \cdot \widetilde{S}_{K_1\times K_2}} \end{bmatrix} - \begin{bmatrix} \frac{G_{1,1}}{\lambda_1 \cdot \widetilde{S}_1} \\ ... \\ \frac{G_{K_1\times K_2, K_1\times K_2}}{\lambda_{K_1\times K_2} \cdot \widetilde{S}_{K_1\times K_2}} \end{bmatrix} = \mathbf{0}$$

A.13 Implied prior on $E[s^2]$ from Section 3.5.1

Recalling that $E[s^2] = \sigma^2[1 - \text{tr}(\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Q}^*)]$ from Section 3.3.5, we find that:

$$\pi(E[s^2]) \;=\; \pi_{\sigma^2}\left(\frac{E[s^2]}{1 - \text{tr}[\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Q}^*]}\right) \cdot \frac{1}{1 - \text{tr}[\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Q}^*]}$$

This result translates to the following implied priors on $E[s_1^2]$ for specific choices of $\pi(\sigma_1^2)$:

- **IG priors**: $\sigma_1^2 \sim \text{IG}(\alpha, \beta) \rightarrow E[s_1^2] \sim \text{IG}(\alpha, \beta[1 - \text{tr}(\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Q}^*)])$
- **PC priors**: $\sigma_1^2 \sim \text{PC}_0(U, \alpha) \rightarrow E[s_1^2] \sim \text{PC}\left(U \cdot \sqrt{1 - \text{tr}(\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Q}^*)}, \alpha\right)$

For the **VP prior case**, we first need to compute the prior implied on σ_1^2:

$$\begin{aligned} \pi(\sigma_1^2) &\propto \int_0^\infty \pi_V(\sigma_1^2, \sigma_\epsilon^2)\pi_\omega(\sigma_1^2, \sigma_\epsilon^2)\frac{1}{\sigma_1^2 + \sigma_\epsilon^2}\,\mathrm{d}\sigma_\epsilon^2 \\ &\propto \int_0^\infty \frac{1}{(\sigma_1^2 + \sigma_\epsilon^2)^2}\,\mathrm{d}\sigma_\epsilon^2 \\ &\propto \frac{1}{\sigma_1^2} \end{aligned}$$

Hence, we find that:

$$\begin{aligned} \pi(E[s_1^2]) &\propto \frac{1}{E[s_1^2]} \cdot \frac{1 - \text{tr}[\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Q}^*]}{1 - \text{tr}[\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Q}^*]} \\ &\propto \frac{1}{E[s_1^2]} \end{aligned}$$

A.14 Implied prior on φ from Section 3.5.2

Recall the definitions of V, ω, T, φ from Section 3.5.2. For a given prior specification on $\sigma_1^2, \sigma_\epsilon^2$, the implied prior on V and ω is:

$$\begin{aligned}\pi(V,\omega) &= \pi_{\sigma_1^2}(V\cdot\omega)\cdot\pi_{\sigma_\epsilon^2}(V-V\cdot\omega)\cdot\left|\det\begin{bmatrix}\frac{\mathrm{d}V\omega}{\mathrm{d}\omega} & \frac{\mathrm{d}V\omega}{\mathrm{d}V}\\ \frac{\mathrm{d}V(1-\omega)}{\mathrm{d}\omega} & \frac{\mathrm{d}V(1-\omega)}{\mathrm{d}V}\end{bmatrix}\right| \\ &= \pi_{\sigma_1^2}(V\cdot\omega)\cdot\pi_{\sigma_\epsilon^2}(V-V\cdot\omega)\cdot V\end{aligned}$$

The marginal of φ implied by a prior specification on V, ω can be found in two steps. First, the marginal of ω is found marginalizing out V:

$$\pi(\omega) = \int_0^\infty \pi(V,\omega)\,\mathrm{d}V$$

Secondly, the marginal of φ is found through a change of variable formula using the transformation $\omega = \dfrac{\varphi}{\varphi + C - \varphi C}$:

$$\pi(\varphi) = \pi_\omega\left(\frac{\varphi}{\varphi + C - \varphi C}\right)\cdot\frac{C}{[\varphi + C - \varphi C]^2}$$

For the 3 prior specifications from Section 3.5.1, we can derive the implied prior on φ.

- **IG priors**: $\sigma_1^2, \sigma_\epsilon^2 \overset{iid}{\sim} \mathrm{IG}(1,\beta)$ First, we derive the implied prior on V, ω.

$$\begin{aligned}\pi(V,\omega) &= \beta^2 V^{-4}\omega^{-2}(1-\omega)^{-2}\exp\left[-\frac{\beta}{V}\cdot\left(\frac{1}{\omega}+\frac{1}{1-\omega}\right)\right]\cdot V \\ &= \beta^2 V^{-3}\omega^{-2}(1-\omega)^{-2}\exp\left[-\frac{\beta}{V}\cdot\left(\frac{1}{\omega}+\frac{1}{1-\omega}\right)\right]\end{aligned}$$

  Secondly, we marginalize out V.

$$\begin{aligned}\pi(\omega) &= \beta^2\omega^{-2}(1-\omega)^{-2}\int_0^\infty \exp\left[-\frac{\beta}{V}\cdot\left(\frac{1}{\omega}+\frac{1}{1-\omega}\right)\right]V^{-3}\,\mathrm{d}V \\ &= \left[\frac{\beta}{\omega(1-\omega)}\right]^2\left[\frac{\beta}{\omega(1-\omega)}\right]^{-2} \\ &= 1\end{aligned}$$

  Finally, we find the implied prior on φ.

$$\pi(\varphi) = \frac{C}{[\varphi + C - \varphi C]^2}$$

- **PC priors**: $\sigma_1^2, \sigma_\epsilon^2 \overset{iid}{\sim} \text{PC}_0(\delta)$

$$\begin{aligned}\pi(V,\omega) &= \delta^2 \exp(-\delta\sqrt{V\omega}) \exp(-\delta\sqrt{V-V\omega}) \frac{1}{4V\sqrt{\omega(1-\omega)}} \cdot V \\ &= \frac{\delta^2}{4\sqrt{\omega(1-\omega)}} \exp[-\delta\sqrt{V}(\sqrt{\omega}+\sqrt{1-\omega})]\end{aligned}$$

Secondly, we marginalize out V.

$$\begin{aligned}\pi(\omega) &= \frac{\delta^2}{4\sqrt{\omega(1-\omega)}} \int_0^\infty \exp[-\delta\sqrt{V}(\sqrt{\omega}+\sqrt{1-\omega})]\, \mathrm{d}V \\ &= \frac{\delta^2}{4\sqrt{\omega(1-\omega)}} \frac{2}{\delta^2(\sqrt{\omega}+\sqrt{1-\omega})^2} \\ &= \left[2\sqrt{\omega(1-\omega)}(\sqrt{\omega}+\sqrt{1-\omega})^2\right]^{-1}\end{aligned}$$

Finally, we find the implied prior on φ.

$$\pi(\varphi) = \left[2 \cdot \sqrt{C\varphi(1-\varphi)} \cdot \left(\sqrt{\frac{\varphi}{C}} + \sqrt{1-\varphi}\right)^2\right]^{-1}$$

- **VP priors**: $\omega \sim$Beta(α, β) (generalization of the Uniform case)

In this case, the prior is simply found through the change of variable formula.

$$\pi(\varphi) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{C \cdot (\varphi)^{\alpha-1} \cdot (C-\varphi C)^{\beta-1}}{[\varphi + C - \varphi C]^{\alpha+\beta}}$$

B Code for Chapter 3

Here, the code to obtain the precision matrices $\widetilde{\boldsymbol{Q}}$ for Example 5 and 6 is written using the R language. Each of the 3 main functions takes as argument the number of basis functions, either as K, or as $K1$ and $K2$ for the 2D case. For convenience, the optimization of the $\boldsymbol{\lambda}$ vector is constrained such that $\boldsymbol{\lambda}$ has symmetric entries for the univariates cases, and it is equal to the Kronecker product of vectors $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, both having symmetric entries, for the 2D case.

The unconstrained optimization returns the same results but it is slower, especially for the 2D case.

```
######################  MODIFIED IGMRFs for P-SPLINE USE  ######################
rm(list=ls())

# LIBRARIES ------------------------------------------------------------------
library(spam)
```

```
# Function to get the generalized inverse of matrix M with given rank-deficiency
gen_inverse_func <- function(M,rank_def=1) {
  # Eigendecomposition in U eigenvectors and V eigenvalues
  M_eigen <- eigen(M,symmetric = T)
  U <- M_eigen$vectors
  V <- M_eigen$values
  # Generalized inverse
  gen_inverse <- U%*%diag(c(1/(V[1:(nrow(M)-rank_def)]),rep(0,rank_def)))%*%t(U)
  return(gen_inverse)
}

# Modified version of IGMRF of order 1 ----------------------------------------
mod_IGMRF_1_prec_mat <- function(K) {
  # Original Q, G, W matrices and generalized inverse of Q
  Q <- as.matrix(precmat.RW1(n = K))
  G <- diag(diag(Q))
  W <- G-Q
  Sigma <- gen_inverse_func(Q,rank_def = 1)
  # Computation of the S tilde matrix for different values of K
  if (K==4) {
    S_tilde <- c(1/24,11/24,11/24,1/24)/(K-3)
  } else if (K==5) {
    S_tilde <- c(1/24,1/2,22/24,1/2,1/24)/(K-3)
  } else if (K>5) {
    S_tilde <- c(1/24,1/2,23/24,rep(1,K-6),23/24,1/2,1/24)/(K-3)
  }
  # Function computing the KLD for a given choice of lambdas
  optim_function <- function(lambdas) {
      lambdas <- abs(lambdas)
    # Creation of the diagonal matrix Lambda
    if (K%%2==0) {
      Lambda <- diag(c(lambdas,rev(lambdas)))
    } else {
      Lambda <- diag(c(lambdas,rev(lambdas)[-1]))
    }
    # Null space for the R matrix
     new_S <- Lambda %*% S_tilde
    # New W, G, R matrix and generalized inverse of new Q
    W_tilde <- W/(new_S%*%t(new_S))
    G_tilde <- diag(as.vector(W_tilde%*%new_S/new_S))
    R_tilde <- G_tilde-W_tilde
    Sigma_tilde <-  Lambda %*% gen_inverse_func(R_tilde,rank_def = 1) %*% Lambda
    # Computation of the KLD (only the non-constant part wrt to lambda)
    kld <- sum(colSums(Q*Sigma_tilde))-
      sum(log(eigen(Sigma_tilde)$values[1:(K-1)]))
    return(kld)
  }
  # Optimization of the KLD function with symmetric entries for lambda
  results <- nlm(optim_function,rep(1,ceiling(K/2)),print.level = 2)
  # Save the lambda values that minimize the KLD
  lambdas <- abs(results$estimate)
  # Lambda matrix
  if (K%%2==0) {
    Lambda <- diag(c(lambdas,rev(lambdas)))
  } else {
    Lambda <- diag(c(lambdas,rev(lambdas)[-1]))
```

```
  }
  # Null space of R matrix
  new_S <- Lambda %*% S_tilde
  # New W,G,R,Q matrices
  W_tilde <- W/(new_S%*%t(new_S))
  G_tilde <- diag(as.vector(W_tilde%*%new_S/new_S))
  R_tilde <- G_tilde-W_tilde
  Q_tilde <- gen_inverse_func(
    Lambda %*% gen_inverse_func(R_tilde,rank_def=1) %*% Lambda,rank_def = 1)
  return(list("Q_tilde"=Q_tilde,"R_tilde"=R_tilde,"Lambda"=Lambda))
}

# Modified version of IGMRF of order 2 ----------------------------------------
mod_IGMRF_2_prec_mat <- function(K) {
  # Original Q, G, W matrices and generalized inverse of Q
  Q <- as.matrix(precmat.RW2(n = K))
  G <- diag(diag(Q))
  W <- G-Q
  Sigma <- gen_inverse_func(Q,rank_def = 2)
  # Computation of the S tilde matrix for different values of K
  Delta <- 1/(K-3)
  if (K==4) {
    S_tilde <- cbind(
      c(1/24,11/24,11/24,1/24)*Delta,
      c(1/120,11/60,11/40,1/30))
  } else if (K==5) {
    S_tilde <- cbind(
      c(1/24,1/2,22/24,1/2,1/24)*Delta,
      c(1/480,7/120,11/48,23/120,3/160))
  } else if (K==6) {
    S_tilde <- cbind(
      c(1/24,1/2,23/24,rep(1,K-6),23/24,1/2,1/24)*Delta,
      c(Delta/120,
        14*Delta/60,
        121*Delta/120,
        23/24-121*Delta/120,
        1/2-14*Delta/60,
        1/24-Delta/120)*Delta)
  } else if (K>=7) {
    S_tilde <- cbind(
      c(1/24,1/2,23/24,rep(1,K-6),23/24,1/2,1/24)*Delta,
      c(Delta/120,
        14*Delta/60,
        121*Delta/120,
        c(2:(K-5))*Delta,
        23/24-121*Delta/120,
        1/2-14*Delta/60,
        1/24-Delta/120)*Delta)
  }
  # Optimization of the KLD function with symmetric entries for lambdas
  optim_function <- function(lambdas) {
      lambdas <- abs(lambdas)
    # Creation of the diagonal matrix Lambda
    if (K%%2==0) {
      Lambda <- diag(c(lambdas,rev(lambdas)))
    } else {
      Lambda <- diag(c(lambdas,rev(lambdas)[-1]))
```

```
    }
    # Null space for the R matrix
    new_S <- Lambda %*% S_tilde
    # New W, G, R matrix and generalized inverse of new Q
    W_tilde <- matrix(0,nrow = K,ncol = K)
    G_tilde <- matrix(0,nrow = K,ncol = K)
    for (k in 1:K) {
      for (l in 1:K) {
        W_tilde[k,l]<-(l-k)*W[k,l]/(new_S[k,1]*new_S[l,2]-new_S[k,2]*new_S[l,1])
      }
      W_tilde[k,k] <- 0
      G_tilde[k,k] <- W_tilde[k,]%*%new_S[,1]/new_S[k,1]
    }
    R_tilde <- G_tilde-W_tilde
    Sigma_tilde <- Lambda %*% gen_inverse_func(R_tilde,rank_def = 2) %*% Lambda
    # Computation of the KLD (only the non-constant part wrt to lambda)
    kld <- sum(colSums(Q*Sigma_tilde))-
      sum(log(eigen(Sigma_tilde)$values[1:(K-2)]))
    return(kld)
  }
  # Optimization of the KLD function with symmetric entries for lambda
  results <- nlm(optim_function,rep(1,ceiling(K/2)),print.level = 2)
  # Save the lambda values that minimize the KLD
  lambdas <- abs(results$estimate)
  # Lambda matrix
  if (K%%2==0) {
    Lambda <- diag(c(lambdas,rev(lambdas)))
  } else {
    Lambda <- diag(c(lambdas,rev(lambdas)[-1]))
  }
  # Null space for R
  new_S <- Lambda %*% S_tilde
  # New W,G,R,Q matrices
  W_tilde <- matrix(0,nrow = K,ncol = K)
  G_tilde <- matrix(0,nrow = K,ncol = K)
  for (k in 1:K) {
    for (l in 1:K) {
      W_tilde[k,l] <- (l-k)*W[k,l]/(new_S[k,1]*new_S[l,2]-new_S[k,2]*new_S[l,1])
    }
    W_tilde[k,k] <- 0
    G_tilde[k,k] <- W_tilde[k,]%*%new_S[,1]/new_S[k,1]
  }
  R_tilde <- G_tilde-W_tilde
  Q_tilde <- gen_inverse_func(
    Lambda %*% gen_inverse_func(R_tilde,rank_def=2) %*% Lambda,rank_def = 2)
  return(list("Q_tilde"=Q_tilde,"R_tilde"=R_tilde,"Lambda"=Lambda))
}

# Modified version of 2D IGMRF of order 1 ------------------------------------
mod_IGMRF_2D_prec_mat <- function(K1,K2) {
  # Original Q, G, W matrices and generalized inverse of Q
  Q <- as.matrix(precmat.IGMRFreglat(K1,K2,order = 1))
  G <- diag(diag(Q))
  W <- G-Q
  Sigma <- gen_inverse_func(Q,rank_def = 1)
  # Computation of the S tilde matrix
  Delta1 <- 1/(K1-3)
```

```
  if (K1==4) {
    s1 <- c(1/24,11/24,11/24,1/24)*Delta1
  } else if (K1==5) {
    s1 <- c(1/24,1/2,22/24,1/2,1/24)*Delta1
  } else if (K1>5) {
    s1 <- c(1/24,1/2,23/24,rep(1,K1-6),23/24,1/2,1/24)*Delta1
  }
  Delta2 <- 1/(K2-3)
  if (K2==4) {
    s2 <- c(1/24,11/24,11/24,1/24)*Delta2
  } else if (K2==5) {
    s2 <- c(1/24,1/2,22/24,1/2,1/24)*Delta2
  } else if (K2>5) {
    s2 <- c(1/24,1/2,23/24,rep(1,K2-6),23/24,1/2,1/24)*Delta2
  }
  S_tilde <- kronecker(s2,s1)
  # Optimization of the KLD function with symmetric entries for lambda
  optim_function <- function(lambdas) {
      lambdas <- abs(lambdas)
    # Creation of the diagonal matrix Lambda
    lambdas1 <- lambdas[1:ceiling(K2/2)]
    lambdas2 <- lambdas[(ceiling(K2/2)+1):(ceiling(K2/2)+ceiling(K1/2))]
    Lambda <- diag(kronecker(c(lambdas2,rev(lambdas2)),
                        c(lambdas1,rev(lambdas1))))
    # Null space for the R matrix
    new_S <- Lambda %*% S_tilde
    # New W,G,R,Q matrices
    W_tilde <- W/(new_S%*%t(new_S))
    G_tilde <- diag(diag(G)/c(new_S^2))
    R_tilde <- G_tilde-W_tilde
    Sigma_tilde <-  Lambda %*% gen_inverse_func(R_tilde,rank_def = 1) %*% Lambda
    # Computation of the KLD (only non-constant part)
    kld <- sum(colSums(Q*Sigma_tilde))-
      sum(log(eigen(Sigma_tilde,only.values = T)$values[1:((K1*K2)-1)]))
    return(kld)
  }
  # Optimization of the KLD function with symmetric entries for lambda
  results <- nlm(optim_function,
                 c(rep(1,ceiling(K2/2)),rep(1,ceiling(K1/2))),print.level = 2)
  # Save the lambda values that minimize the KLD
  lambdas <- abs(results$estimate)
  # Lambda matrix
  lambdas1 <- lambdas[1:ceiling(K2/2)]
  lambdas2 <- lambdas[(ceiling(K2/2)+1):(ceiling(K2/2)+ceiling(K1/2))]
  Lambda <- diag(kronecker(c(lambdas2,rev(lambdas2)),
                      c(lambdas1,rev(lambdas1))))
  # Null space for the R matrix
  new_S <- Lambda %*% S_tilde
  # New W,G,R,Q matrices
  W_tilde <- W/(new_S%*%t(new_S))
  G_tilde <- diag(diag(G)/c(new_S^2))
  R_tilde <- D_tilde-W_tilde
  Sigma_tilde <- Lambda %*% gen_inverse_func(R_tilde,rank_def=1) %*% Lambda
  Q <- gen_inverse_func(Sigma_tilde,rank_def = 1)
  return(list("Q_tilde"=Q_tilde,"R_tilde"=R_tilde,"Lambda"=Lambda))
}
```

C Proofs of Chapter 4

C.1 Equation 4.14

The KLD-based distance between two multivariate Gaussian distributions simplifies to Equation 4.14 through the following steps:

$$
\begin{aligned}
d(\omega;\omega_0) &= \sqrt{2 \cdot KLD[\pi(\boldsymbol{y};\omega)||\pi(\boldsymbol{y};\omega_0)]} \\
&= \sqrt{2 \int \log \left[\frac{\sqrt{(2\pi)^{R(\omega_0)} \cdot |\boldsymbol{\Sigma}(\omega_0)|} \exp\left(-\frac{1}{2}\boldsymbol{y}^T \boldsymbol{\Sigma}^*(\omega)\boldsymbol{y}\right)}{\sqrt{(2\pi)^{R(\omega)} \cdot |\boldsymbol{\Sigma}(\omega)|} \exp\left(-\frac{1}{2}\boldsymbol{y}^T \boldsymbol{\Sigma}^*(\omega_0)\boldsymbol{y}\right)} \right] \pi(\boldsymbol{y};\omega)\, \mathrm{d}\boldsymbol{y}} \\
&= \sqrt{[R(\omega_0) - R(\omega)] \log(2\pi) + \log \frac{|\boldsymbol{\Sigma}(\omega_0)|}{|\boldsymbol{\Sigma}(\omega)|} + \int \boldsymbol{y}^T [\boldsymbol{\Sigma}^*(\omega_0) - \boldsymbol{\Sigma}^*(\omega)]\boldsymbol{y} \cdot \pi(\boldsymbol{y};\omega)\, \mathrm{d}\boldsymbol{y}} \\
&= \sqrt{[R(\omega_0) - R(\omega)] \log(2\pi) + \log \frac{|\boldsymbol{\Sigma}(\omega_0)|}{|\boldsymbol{\Sigma}(\omega)|} + E_{\pi(\boldsymbol{y};\omega)} \left[\boldsymbol{y}^T [\boldsymbol{\Sigma}^*(\omega_0) - \boldsymbol{\Sigma}^*(\omega)]\boldsymbol{y} \cdot \pi(\boldsymbol{y};\omega)\right]}
\end{aligned}
$$

Recalling the formula for the expectation of a quadratic form from the proof in Kendrick 1981, it is found that:

$$
\begin{aligned}
E_{\pi(\boldsymbol{y};\omega)} \left[\boldsymbol{y}^T [\boldsymbol{\Sigma}^*(\omega_0) - \boldsymbol{\Sigma}^*(\omega)]\boldsymbol{y} \cdot \pi(\boldsymbol{y};\omega)\right] &= \mathrm{tr}\{[\boldsymbol{\Sigma}^*(\omega_0) - \boldsymbol{\Sigma}^*(\omega)]\boldsymbol{\Sigma}(\omega)\} \\
&= \mathrm{tr}[\boldsymbol{\Sigma}^*(\omega_0)\boldsymbol{\Sigma}(\omega)] - R(\omega)
\end{aligned}
$$

C.2 Equation 4.16

Under the assumption of a common eigenbasis $\boldsymbol{V}$ between the covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$, it is possible to simplify the distance following the proof provided in Franco-Villoria, Ventrucci, and Rue 2022. First, note that $\boldsymbol{\Sigma}(\omega)$ and its generalized inverse can be rewritten as:

$$
\begin{aligned}
\boldsymbol{\Sigma}(\omega) &= \boldsymbol{V} \left[(1-\omega)\boldsymbol{\Lambda}_0 + \omega\boldsymbol{\Lambda}_1\right] \boldsymbol{V}^T \\
\boldsymbol{\Sigma}^*(\omega) &= \boldsymbol{V} \left[(1-\omega)\boldsymbol{\Lambda}_0 + \omega\boldsymbol{\Lambda}_1\right]^* \boldsymbol{V}^T
\end{aligned}
$$

Then, it can be noticed that the first term of Equation 4.14 simplifies and be-

comes only a function of the eigenvalues in $\mathbf{\Lambda}_0$ and $\mathbf{\Lambda}_1$:

$$\begin{aligned}\operatorname{tr}\left[\mathbf{\Sigma}^*(\omega_0)\mathbf{\Sigma}(\omega)\right] &= \operatorname{tr}[\boldsymbol{V}\left[(1-\omega_0)\mathbf{\Lambda}_0+\omega_0\mathbf{\Lambda}_1\right]^*\boldsymbol{V}^T\boldsymbol{V}\left[(1-\omega)\mathbf{\Lambda}_0+\omega\mathbf{\Lambda}_1\right]\boldsymbol{V}^T] \\ &= \operatorname{tr}[\left[(1-\omega_0)\mathbf{\Lambda}_0+\omega_0\mathbf{\Lambda}_1\right]^*\left[(1-\omega)\mathbf{\Lambda}_0+\omega\mathbf{\Lambda}_1\right]] \\ &= \sum_{n:\lambda_{0,n}+\lambda_{1,n}>0}\frac{(1-\omega)\lambda_{0,n}+\omega\lambda_{1,n}}{(1-\omega_0)\lambda_{0,n}+\omega_0\lambda_{1,n}}\end{aligned}$$

Finally, also the second term of Equation 4.14 can be rewritten simly in terms of eigenvalues under the common eigenbasis assumption:

$$\begin{aligned}\log\frac{|\mathbf{\Sigma}(\omega)|}{|\mathbf{\Sigma}(\omega_0)|} &= \log\frac{|(1-\omega)\mathbf{\Lambda}_0+\omega\mathbf{\Lambda}_1|}{|(1-\omega_0)\mathbf{\Lambda}_0+\omega_0\mathbf{\Lambda}_1|} \\ &= \sum_{n:\lambda_{0,n}+\lambda_{1,n}>0}\log\frac{(1-\omega)\lambda_{0,n}+\omega\lambda_{1,n}}{(1-\omega_0)\lambda_{0,n}+\omega_0\lambda_{1,n}}\end{aligned}$$

At this point, the distance function can be further simplified because under the sum-of-ranks condition, for each $n \in [1, N]$, at least one eigenvalue between $\lambda_{0,n}$ and $\lambda_{1,n}$ will be 0. This leads to the following simplification:

$$\frac{(1-\omega)\lambda_{0,n}+\omega\lambda_{1,n}}{(1-\omega_0)\lambda_{0,n}+\omega_0\lambda_{1,n}} = \begin{cases}\frac{1-\omega}{1-\omega_0}, & \lambda_{0,n}>0, \lambda_{1,n}=0 \\ \frac{\omega}{\omega_0}, & \lambda_{0,n}=0, \lambda_{1,n}>0\end{cases}$$

The final form of the distance function is the following:

$$d(\omega;\omega_0) = \sqrt{\begin{aligned}&R(0)\frac{1-\omega}{1-\omega_0}+R(1)\frac{\omega}{\omega_0}-R(0)\log\frac{1-\omega}{1-\omega_0}-R(1)\log\frac{\omega}{\omega_0} \\ &-R(\omega)+[R(\omega_0)-R(\omega)]\log(2\pi)\end{aligned}}$$

The function does not include the eigenvalues of $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ but it still depends on $R(\omega)$, which is computationally inefficient. However, noting that the distance is not finite for $\omega_0 = 0$, it is necessary to compute it as a limit. Consider first $d(\omega;\omega_0)^2\omega_0$:

$$\begin{aligned}d(\omega;\omega_0)^2\omega_0 &= \omega_0 R(0)\frac{1-\omega}{1-\omega_0}-\omega_0 R(0)\log\frac{1-\omega}{1-\omega_0} \\ &\quad +R(1)\omega-\omega_0 R(1)\log\frac{\omega}{\omega_0} \\ &\quad -\omega_0 R(\omega)+\omega_0[R(\omega_0)-R(\omega)]\log(2\pi)\end{aligned}$$

Computing the limit for $\omega_0 \to 0$, it is found that:

$$\lim_{\omega_0 \to 0} d(\omega; \omega_0)^2 \omega_0 = R(1)\omega$$