Received: 28 April 2025 | Revised: 26 September 2025 | Accepted: 24 November 2025

DOI: 10.1111/sjos.70042

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

A standardization procedure to incorporate variance partitioning-based priors in latent Gaussian models

Luisa Ferrari[1] | **Massimo Ventrucci**[2]

[1]Department of Economics, University of Modena & Reggio Emilia, Modena, Italy

[2]Department of Statistical Sciences, University of Bologna, Bologna, Italy

Correspondence
Luisa Ferrari, Department of Economics, University of Modena & Reggio Emilia, Modena, Italy.
Email: luisa.ferrari@unimore.it

Funding information
European Union, Grant/Award Number: 2022PA3BS2 (CUP E53D23007580006)

ABSTRACT

Latent Gaussian models (LGMs) are a subset of Bayesian Hierarchical models where Gaussian priors, conditional on variance parameters, are assigned to all effects in the model. LGMs are employed in many fields for their flexibility and computational efficiency. However, practitioners find prior elicitation on the variance parameters challenging because of a lack of intuitive interpretation for them. Recently, several papers have tackled this issue by representing the model in terms of variance partitioning (VP) and assigning priors to parameters reflecting the relative contribution of each effect to the total variance. So far, the class of priors based on VP has been mainly applied to random effects and fixed effects separately. This work presents a novel standardization procedure that expands the applicability of VP priors to a broader class of LGMs, including both fixed and random effects. The practical advantages of standardization are demonstrated with simulated data and a real dataset on survival analysis.

KEYWORDS

Gaussian Markov Random Fields, hierarchical decomposition priors, P-splines, PC priors, R2D2

Luisa Ferrari and Massimo Ventrucci equally contributing authors.

1 | INTRODUCTION

Latent Gaussian models (LGMs, Rue et al., 2009) are a popular class of Bayesian Hierarchical models used by applied scientists in many fields for their flexibility and ability to include linear and nonlinear effects of continuous covariates, spatial and temporal random effects, and more complex structures such as interactions. A popular subclass of LGMs assumes that the response is linked to the latent parameters with a univariate link function, and that the linear predictor for observations $i = 1,\dots,N$ is structured as:

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

where $\beta_p \sim N(0, \tau_p^2), p = 1,\dots,P$ are linear effects of P covariates $x_1,\dots,x_P$ (usually called fixed effects), and $\boldsymbol{u}_r \sim N_{K_r}(\boldsymbol{0}, \sigma_r^2\boldsymbol{Q}_r^-), r = 1,\dots R$ ($\boldsymbol{Q}_r^-$ denotes here the generalized inverse of $\boldsymbol{Q}_r$) represent other types of effects (usually called random) for additional covariates $z_1,\dots,z_R$. The vector $\boldsymbol{D}_r(\cdot) = [D_{r,1}(\cdot),\dots,D_{r,K_r}(\cdot)]^T$ contains K_r known basis functions. τ_p^2 and σ_r^2 are typically called variance or scale parameters, while $\boldsymbol{Q}_r$ is usually called the structure matrix, and it is either fully known or controlled by additional correlation parameters. The model in Equation (1) can embrace many types of random effects, including group effects or splines. Along with proper Gaussian processes, LGMs can include Intrinsic Gaussian Markov Random Fields (IGMRFs, Rue & Held, 2005; Fahrmeir et al., 2004), which are characterized by a rank-deficient $\boldsymbol{Q}_r$, such as the random walk of order 1 (RW1) and 2 (RW2), and the intrinsic conditional autoregressive process (ICAR, Besag & Kooperberg, 1995), often employed as priors for spatial and temporal random effects.

The traditional approach to prior specification of the variance parameters sets the τ_p^2s to large *fixed* values, to ensure flat priors on the linear coefficients, while the σ_r^2 are usually treated as *random* and assigned independent priors. Prior specification on variance parameters is a long-discussed problem due to their poor interpretability, especially for IGMRFs (Sørbye & Rue, 2014). The conjugate Inverse Gamma priors have been found to lead to overfitting (Gelman, 2006; Lunn et al., 2009; Simpson et al., 2017) and are now being replaced as the default choice by Penalized Complexity (PC) priors (Simpson et al., 2017) in the `R` package `INLA` (Rue et al., 2009), which provides fast approximate inference for LGMs.

1.1 | Variance partitioning-based priors

Domain experts are often more comfortable reasoning about the relative importance of model components (e.g., "the spatial effect should account for about half the variability") than specifying absolute magnitudes of variance parameters. Recently, multiple works have tried to leverage this underlying information by specifying priors on proportions of variance, rather than absolute variance parameters.

Such priors can be specified considering a *variance partitioning* (VP) reparametrization: the original variance parameters are transformed into an overall total variance V, given by the sum of all or some variance parameters, and a set of variance proportions $\boldsymbol{\omega} = [\omega_1, \omega_2,\dots]^T$, given by the ratio between the variance parameters and V. The reparametrization maps the original, often nonintuitive, variance scale to a more interpretable scale of proportions constrained to the (0,1) interval, plus a single total variance. This shift facilitates prior specification as experts usually have

more insight about the relative importance of the different effects, rather than their absolute one. Two main lines of research have so far used the VP reparametrization, namely, the Hierarchical Decomposition (HD) priors (Fuglstad et al., 2020) and the R2D2 priors (Zhang et al., 2022).

The HD approach by Fuglstad et al. (2020) provides a structured and interpretable way to build a joint prior for the random effects' variance parameters, σ_r^2s. It does so by assigning independent priors on the total variance V and the vector of variance proportions $\boldsymbol{\omega}$, where $V = \sum_{r=1}^{R} \sigma_r^2$ includes only the random effects, and the proportion parameters are defined as $\omega_r = \sigma_r^2/V$, for $r = 1,\ldots,R$. One key feature of the HD approach is the possibility of a further reparametrization of $\boldsymbol{\omega}$ into nested sets of proportions through a hierarchical decomposition of the total variance in the form of a tree. This decomposition involves a sequence of splits, structured to reflect the available prior information. The final prior on $\boldsymbol{\omega}$ is then constructed by assigning independent priors (either Dirichlet or PC priors) to the sets of proportions generated at each split. As a prior on V, Fuglstad et al. (2020) suggests a Jeffreys or a PC prior, depending on the response distribution in the observation model. This structured approach improves interpretability and flexibility, allowing users to encode their prior assumptions and knowledge in an intuitive way. Applications of HD priors can be found in Hem et al. (2021) and Franco-Villoria et al. (2022). The R package `makemyprior` (Hem et al., 2024) provides user-friendly tools to design HD priors in a wide range of models. In the same spirit of HD priors, Wakefield (2007) and Riebler et al. (2016) have proposed priors on the proportion of variance explained by spatially structured and unstructured random effects in disease mapping.

The second line of research developed in the context of sparse regression, with the aim of designing a novel shrinkage prior. Following the Dirichlet-Laplace prior of Bhattacharya et al. (2015), the R2D2 prior by Zhang et al. (2022) proposes the use of a VP reparametrization to induce sparsity in multiple linear regression models via prior specification. Considering a model with only linear effects, the total variance is here defined as $V = \sum_{p=1}^{P} \tau_p^2$ and the proportions as $\omega_p = \tau_p^2/V$ for $p = 1,\ldots,P$. The set of proportions is assigned a symmetric Dirichlet distribution with a sparsity-inducing hyperparameter choice. On the other hand, the prior on V is not directly specified, but rather implied by imposing a Beta prior on the measure of goodness-of-fit R^2 (Yanchenko et al., 2024a), defined as a function of V and the additional likelihood parameters. The R2D2 priors have recently been extended to mixed model cases where only a limited class of random effects is allowed, such as group effects (Aguilar & Bürkner, 2023; Yanchenko et al., 2024a, 2024b).

We have seen how both HD and R2D2 approaches make use of a variance partitioning representation of the model. For this reason, we consider them under a common class of priors that, in this paper, we refer to as *VP priors*.

1.2 | The notion of variance contribution in HD and R2D2

The appeal of VP priors lies in the possibility of obtaining parameters that are more easily interpretable than the original ones. Specifically, V can be understood as the *a priori* total variance in the linear predictor due to the summed effects, while the elements in $\boldsymbol{\omega}$ represent the proportional contributions of the individual effects to this total variance. Importantly, this decomposition holds *a priori*: once the model is fitted, different effects will generally be dependent *a posteriori*, and the variance attributable to each term will no longer sum exactly to V. The clear interpretation of V and $\boldsymbol{\omega}$, however, only holds if all effects lie on a comparable scale, so that each variance parameter (either the τ_p^2s or the σ_r^2s) correctly reflects the variance contribution of its corresponding

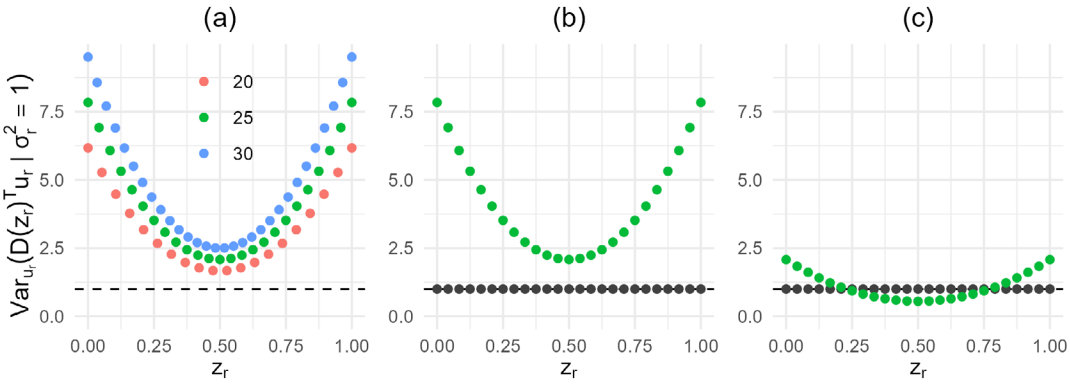

FIGURE 1 Conditional variances over the support of the covariate $z_r \in [0, 1]$ when $\sigma^2_{\mathrm{RW1}} = 1$, $\sigma^2_{\mathrm{IID}} = 1$: (a) RW1 effects observed at 20, 25, or 30 equally spaced points; (b) RW1 effect (green) an IID effect (grey) on 25 equally spaced points; (c) RW1 effect and IID effect after geometric mean scaling. The dashed line indicates variance 1.

effect. We stress that VP priors lose their intuitive advantage if variance parameters are unable to measure variance contributions on a standardized scale.

The literature on VP priors has so far defined the concept of variance contribution of an effect in different ways, either directly or implicitly. In what follows, we compare HD and R2D2 priors regarding the adopted notion of variance contribution and the implied definition of intuitive interpretation of a variance parameter. (Hereafter, $\mathrm{E}_Z[\cdot]$ and $\mathrm{Var}_Z[\cdot]$ will denote, respectively, the expectation and variance taken only with respect to Z, treating all other variables as fixed).

The HD framework explicitly requires that each σ_r^2 must be equal to the *conditional variance* of the effect given the covariate value z_{ir}, i.e., $\mathrm{Var}_{\boldsymbol{u}_r}[\boldsymbol{D}_r^T(z_{ir})\boldsymbol{u}_r|\sigma_r^2, z_{ir}]$. Many popular effects (called heterogeneous by Fuglstad et al., 2020) have a conditional variance that is nonconstant with respect to the covariate support, and potentially not well approximated by the value of σ_r^2. Notable examples include the popular IGMRFs (RW1, R2, ICAR, etc.), which are also affected by the so-called *scaling issue* noted by Sørbye and Rue (2014): the conditional variance of an IGMRF (defined only under appropriate linear constraints) depends on the dimension, i.e., the number of observational points over the domain (see Figure 1, panel a). Sørbye and Rue (2014) proposed a simple procedure to neutralize the effect of the dimension in an IGMRF, which consists of replacing the structure matrix $\boldsymbol{Q}$ with its scaled version $\sigma^2_{\mathrm{ref}}\boldsymbol{Q}$, where σ^2_{ref} is a *reference variance* computed as the geometric mean of the conditional variance, i.e., $\sigma^2_{\mathrm{ref}} = \exp\left\{\frac{1}{n}\sum_{i=1}^n \log \mathrm{Var}_{\boldsymbol{u}_r}\left[\boldsymbol{D}_r^T(z_{ir})\boldsymbol{u}_r|\sigma_r^2 = 1\right]\right\}$. Fuglstad et al. (2020) propose to apply this scaling procedure to each heterogeneous random effect modelled, thus implicitly defining the concept of intuitive interpretation of σ_r^2 as the geometric mean of the conditional variance.

As an example of how scaling issues can limit the use of VP priors, consider a model where the linear predictor is the sum of two random effects, a random walk of order 1 (RW1 effect) and random noise (IID effect); this model will be used in Section 5.1 to demonstrate the benefit of using the standardization procedure proposed in this paper. The user may have prior information on the proportion of variance contributed by the RW1 component, i.e., $\omega = \sigma^2_{\mathrm{RW1}}/V$, where $V = \sigma^2_{\mathrm{RW1}} + \sigma^2_{\mathrm{IID}}$, and hence decide to set priors directly on ω and V instead of on σ^2_{RW1} and σ^2_{IID}. Unfortunately, this model is affected by a scaling issue, shown in Figure 1 panel b: the conditional variances for RW1 and IID for $\sigma_r^2 = 1$ are clearly on different scales (and for the RW1, it

is not constant over the covariate domain $[0, 1]$ and it also depends on the number of observational points). If this mismatch is not fixed, the prior assigned to ω will fail in providing a correct summary of the user's prior information; for instance, a uniform prior on ω will not mean equal importance *a priori* between the RW1 and the IID terms. By applying the geometric mean-based scaling to the RW1, the conditional variance of the RW1 effect gets aligned with the conditional variance of the IID effect (Figure 1 panel c). While the geometric mean-based scaling by Sørbye and Rue (2014) provides an effective solution in this particular example, it exhibits some important limitations in the class of models considered in this paper; notably, it fails for linear effects, which therefore remain outside the scope of application of HD priors.

Alternatively, the R2D2 framework mostly focuses on linear effects, which are also heterogeneous. The original work of Zhang et al. (2022) considers the covariates as random variables and simply requires that they are standardized (i.e., $\mathrm{E}_{X_p}[X_p] = 0, \mathrm{Var}_{X_p}[X_p] = 1$ for $p = 1,\dots,P$), so that $\tau_p^2 = \mathrm{E}_{X_p}\{\mathrm{Var}_{\beta_p}[X_p\beta_p|X_p, \tau_p^2]\}$ for $p = 1,\dots,P$. This expression is proven in Zhang et al. (2022) to correspond to the *marginal variance*, i.e., $\mathrm{Var}_{X_p,\beta_p}[X_p\beta_p|\sigma_p^2]$: hence, we consider this to be the implicit definition of the intuitive interpretation of τ_p^2 from the R2D2 framework. While standardization is sufficient for linear effects, it is not straightforward to generalize it to random effects with general structures, such as those from Equation (1).

In conclusion, the notion of variance contribution has been defined in different ways in the VP literature. In models with both fixed and random effects, a unique definition of what we mean by variance contribution is currently missing. Scaling methods, to align variance parameters to their intuitive interpretation, have been proposed in specific models, but a general procedure valid for a large class of LGMs is currently unavailable. We believe that filling these gaps may greatly increase the use of HD priors and R2D2 priors in LGM contexts.

1.3 | Contribution & outline

This paper proposes a standardization procedure that enables the consistent and interpretable application of existing VP priors (such as HD and R2D2) within a broad class of LGMs, defined to include both fixed and random effects, including IGMRFs. We start by defining the concept of *expected variance of interest* based on the inferential goal of the user. We argue that the expected variance of interest formalizes a notion of variance contribution that reflects the intuitive interpretation of the user about the variance parameters for both fixed and random effects. We then derive the conditions under which the variance parameters match the expected variance of interest, which leads to a general standardization procedure that guarantees an exact match between variance parameters and variance contributions. We support our methodology with a theoretical result, proving that the proportions of variance parameters are correctly interpreted after standardization (Remark 1), and simulations. On the practical side, our primary interest is to enable the use of VP priors in realistic settings. For this reason, we illustrate the proposed standardization through various examples frequently encountered in LGMs, such as linear and nonlinear effects of continuous covariates, group effects (i.e., effects of categorical covariates), and spatial and temporal random effects. We put emphasis on the challenges behind standardization of IGMRFs, as these are frequently used not only to model spatial and temporal dependence but also to construct nonlinear effects via P-splines (Eilers & Marx, 1996). We argue that a preliminary adjustment is necessary for P-splines before standardization. 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.

The remainder of the paper is organized as follows. Section 2 introduces the class of LGMs under study, our definitions of fixed/random effects, and variance contributions. Section 3 explains in great detail the standardization procedure and its applications to linear effects and group effects. Section 4 discusses standardization for IGMRFs, focusing on the P-spline case. Section 5 reports simulation studies and a real-data application, while Section 6 concludes.

2 | DEFINING VARIANCE CONTRIBUTION FOR FIXED AND RANDOM EFFECTS

In this section, we first define the class of LGMs considered in this paper, we then formally adopt a modern perspective on the fixed/random effects classification, and finally, we discuss how the intuitive interpretation of variance contribution varies between the two categories. In contrast to the common formulation of LGMs, our definition treats the covariates as random variables, as in the R2D2 proposal of Zhang et al. (2022). This choice is not intended to model the true generative distribution of the covariates, but rather to allow the user to formalize how variability in each covariate must be weighted when computing the variance of its associated effect.

Definition 1 (Latent Gaussian Model). Let $\boldsymbol{X} = [X_1, \ldots, X_J]^T$ be a random vector of dimension J, with marginal distributions $X_j \sim \pi_j(x_j)$ on support $\mathcal{X}_j \subseteq \mathbb{R}$ for $j = 1, \ldots, J$. Let a response $Y \sim \text{Dist}(\eta, \boldsymbol{\psi})$ have an exponential family distribution, where η corresponds to a transformation of the expected value of Y given $\boldsymbol{\psi}$ and:

$$\eta = \mu + \sum_{j=1}^{J} f_j(X_j).$$

Each $f_j(X_j)$ for $j = 1, \ldots, J$ is defined as: $f_j(X_j) = \boldsymbol{D}_j^T(X_j)\boldsymbol{u}_j$ where $\boldsymbol{D}_j(X_j)$ is a column vector containing K_j known functions evaluated at X_j, and $\boldsymbol{u}_j = [u_{j,1}, \ldots, u_{j,K_j}]^T$ is a random vector of dimension K_j distributed as: $\boldsymbol{u}_j | \sigma_j^2 \sim N_{K_j}(\boldsymbol{0}, \sigma_j^2 \boldsymbol{Q}_j^-)$ where $\boldsymbol{Q}_j$ is a known matrix of dimension $K_j \times K_j$, and σ_j^2 is a scalar parameter. We shall call $\boldsymbol{D}_j(X_j)$ the basis, $\boldsymbol{u}_j$ the coefficients, σ_j^2 the variance parameter, $\boldsymbol{Q}_j$ the *structure matrix* for *effect* $f_j(X_j)$.

Implementing VP priors on models from Definition 1 consists in specifying a joint prior on $\boldsymbol{\sigma}^2 = [\sigma_1^2, \ldots, \sigma_J^2]^T$. Specification is then completed with priors on μ and $\boldsymbol{\psi}$. The class of models from Definition 1 is very flexible as it can accommodate a wide variety of effects. For instance, linear effects can be specified by setting $K_j = 1$ and the basis to a single identity function, i.e., $D_j(X_j) = X_j$, hence $u_j \sim N_1(0, \sigma_j^2)$ is the slope coefficient. Various IGMRFs can be accommodated by making precise assumptions on the basis and precision matrix $\boldsymbol{D}_j(\cdot)$ and $\boldsymbol{Q}_j$. Effects whose $\boldsymbol{Q}_j$ depends on additional correlation parameters (e.g., Matern) can still be included after conditioning on reasonable values, as suggested in Fuglstad et al. (2020).

2.1 | Redefinition of fixed and random effects

In a Bayesian LGM, common practice is to label the effects as fixed or random according to whether σ_j^2 is treated as fixed (typically, a large value) or random (hence, assigned a prior). Following this perspective, all the effects in our LGM would be classified as random. In this paper, we focus on a different categorization of the effects, which recognizes the different inferential

goals of the user or, in other words, the different ways in which the model estimates can be used by practitioners. This classification is not novel, and it is inspired by the discussion in Gelman et al. (2013) and Hodges (2013).

Gelman et al. (2013) in Section 15.6 states 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". Similarly, Hodges (2013) proposes to categorize the random effects (i.e., effects where σ_j^2 is treated as random) between old-style random effect, i.e., when the levels "[…] are draws from a population, and the draws are not of interest in themselves […]", and new-style random effects, i.e., when the "[…] levels themselves are of interest". Following these ideas, we find it convenient to define a simple rule that classifies each effect as fixed or random on the basis of the concept of inferential interest.

> **Definition 2** (Fixed and random effects). Consider each effect $f_j(X_j)$ from the model in Definition 1. $f_j(X_j)$ is called "fixed" if the inferential interest is on $\boldsymbol{u}_j$ (or on some transformation of them). $f_j(X_j)$ is called "random" if the inferential interest is on σ_j^2 (or some transformation of it).

Hereafter, we use the terms 'fixed' and 'random' according to Definition 2. As an example of the application of the above definition, linear effects will typically be considered fixed because the inferential focus is always on the slope coefficients u (i.e., the effect estimate). Other effects, such as nonlinear, spatial, temporal, interactions, or group effects, may be classified as fixed or random according to the researcher's specific inferential objectives. Consider an interaction between a continuous (X_1) and a categorical (X_2) covariate, where the linear effect of X_1 is let vary across the levels of X_2, i.e., $f(X_1, X_2) = u(X_2)X_1$. If the interest is not on the slope coefficients themselves but on their variation across the different levels, then the parameter of interest is σ_j^2 and not $u(\cdot)$, hence the effect will be classified as random. On the contrary, if the interest is in the slope coefficient estimates, the interaction effect will be classified as fixed. As an example of a group effect, consider a medical study in which we want to account for the impact of the hospital where patients receive treatment. If the goal is to characterize the variability across hospitals, then the hospital effect must be considered random. In contrast, if interest lies in comparing specific hospitals, for example, to evaluate performance, the hospital effect should be considered fixed. The group effect will be examined further in Example 2.

In summary, we find it reasonable to let the user specify each effect as fixed or random according to their specific inferential goal. For convenience, we assume that the effects from Definition 1 are ordered such that the first L effects are fixed, and the last $J - L$ are random. Hence, we can denote with:

$$\boldsymbol{\theta} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L, \sigma_{L+1}^2, \ldots, \sigma_J^2]^T \tag{2}$$

the vector containing all the parameters of interest for the user.

2.2 | Intuitive interpretation of σ^2

We now formally define the concept of intuitive interpretation of σ^2, based on the consideration that the user has a different intuition of variance contribution for fixed and random effects. For example, the variance contribution of a fixed linear effect is typically measured conditioning

on the slope coefficient to obtain the so-called *explained variance* $\mathrm{Var}_X[Xu]$, which quantifies the variability in the response due to the variation in the covariate, while the coefficient is fixed. Instead, the amount of variability imputed to random group effects is usually estimated directly through the variance parameter σ^2, which captures the heterogeneity across the groups or levels themselves, considered as draws from a larger population distribution. More generally, we adopt the terminology of Gelman et al. (2013) to argue that the user intends variance contribution as the:

- *finite-population variance* for fixed effects, i.e., $\mathrm{Var}_{X_j}[f_j(X_j)|\boldsymbol{u}_j]$;
- *super-population variance* for random effects, i.e., $\mathrm{Var}_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2]$.

This distinction can be neatly summarized using the concept of *variance of interest*, defined as the variance of the effect conditional on the parameters of interest, i.e., $\mathrm{Var}_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\boldsymbol{\theta}]$ with $\boldsymbol{\theta}$ as in Equation (2). Note, the variance of interest of a random effect (i.e., $\theta_j = \sigma_j^2$) corresponds to the super-population variance, and the variance of interest of a fixed effect (i.e., $\theta_j = u_j$) equals the finite-population variance. However, the variance of interest cannot be directly used for the definition of the intuitive interpretation of σ_j^2, as it is not a function of σ_j^2 for fixed effects. Hence, we propose to use instead the *expected variance of interest*, defined as: $\mathrm{E}_{\boldsymbol{\theta}}\{\mathrm{Var}_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\boldsymbol{\theta}]|\boldsymbol{\sigma}^2\}$.

Definition 3 (Intuitive interpretation of $\boldsymbol{\sigma}^2$ parameters). Consider the model from Definition 1 with parameters of interest $\boldsymbol{\theta} = [\boldsymbol{u}_1,\ldots,\boldsymbol{u}_L,\sigma_{L+1}^2,\ldots,\sigma_J^2]^T$. We say that the variance parameter $\sigma_j^2, j = 1,\ldots,J$, has an intuitive interpretation if it is equal to the expected variance of interest of the corresponding effect, defined as $\sigma_j^2 = \mathrm{E}_{\theta_j}\{\mathrm{Var}_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\theta_j]|\boldsymbol{\sigma}^2\}$.

Deriving the expected variance of interest separately for fixed and random effects, we find out the conditions under which we obtain an intuitive interpretation of the σ_j^2 parameters for all effects in our model. For fixed effects ($j = 1,\ldots,L$), the condition for an intuitive interpretation of σ_j^2 is

$$\sigma_j^2 = \mathrm{E}_{\boldsymbol{u}_j}\{\mathrm{Var}_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\}, \tag{3}$$

which is the expected finite-population variance marginalizing with respect to $\boldsymbol{u_j}$. For random effects ($j = L+1,\ldots,J$), the condition for an intuitive interpretation of σ_j^2 is

$$\sigma_j^2 = \mathrm{Var}_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2], \tag{4}$$

which is the super-population variance.

Definition 3 is consistent with the suggestion in Fuglstad et al. (2020) of considering the explained variance for the introduction of fixed effects in the HD framework, since Equation (3) corresponds to the *expected explained variance*. The definition is also coherent with the R2D2 approach since Equation (4) is equivalent to the (implicit) definition of variance contribution in the R2D2 literature (see Section 1).

Note that Definition 3 relies on the probability assumption made in Definition 1 about the covariates' marginal distributions $\pi_j(x_j)$. By explicitly specifying the distribution of the covariates, the user can express their view about what constitutes the contribution of each effect. For multiple reasons, a convenient default option consists of assuming marginal uniform distributions

(either discrete or continuous) over sensible supports. First, a Uniform distribution treats all values within the support as equally important in the variance computation: this property ensures that the variance contributions are not biased toward specific regions of the support space. Secondly, it leads to analytically tractable expressions for the variance contributions of popular effects (see Examples 2 to 4).

3 | STANDARDIZATION PROCEDURE

Based on the considerations from the previous section and inspired by Sørbye and Rue (2014), we propose a standardization procedure that ensures that the condition of Definition 3 is satisfied for all effects that can be specified in the model class from Definition 1.

Proposition 1 (Standardization procedure). *Consider the class of models from Definition 1 with parameters of interest* $\boldsymbol{\theta} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L, \sigma^2_{L+1}, \ldots, \sigma^2_J]^T$. *The condition in Definition 3 is satisfied after the implementation of the following two steps.*

1. *0-mean constraint on fixed effects. Each fixed effect* $j = 1, \ldots, L$ *is redesigned, either by appropriately modifying the basis or constraining the coefficients, so that*:

$$\mathrm{E}_{X_j}[f_j(X_j)|\boldsymbol{u}_j] = 0.$$

2. *Scaling. Each effect* $j = 1, \ldots, J$ *is replaced by* $\widetilde{f}_j(X_j) = \dfrac{f_j(X_j)}{\sqrt{C_j}}$, *where* C_j *is a constant that depends on the specific choice of* $\pi_j(x_j)$ *and is given by*:

$$\begin{aligned} C_j &= \mathrm{Var}_{X_j, \boldsymbol{u}_j}[f_j(X_j)|\sigma^2_j = 1] \\ &= \mathrm{E}_{X_j}\left[\boldsymbol{D}_j^T(X_j)\boldsymbol{Q}_j^- \boldsymbol{D}_j(X_j)\right]. \end{aligned} \tag{5}$$

See the proof in Section A1 of the Appendix.

In other words, Proposition 1 states that scaling by the appropriate constant C_j is sufficient for random effects, as this guarantees the equality between σ^2_j and the super-population variance (Equation 4). On the other hand, both steps of Proposition 1 are necessary for fixed effects to satisfy Equation (3). The procedure is called *standardization* as it can be proven that classical standardization is a special case that arises for linear effects.

Example 1 (Linear effects). Consider a linear effect $f(X) = Xu$ where $u|\sigma^2 \sim N(0, \sigma^2)$. Assuming the effect is fixed, Proposition 1 requires the application of both steps of the standardization procedure. The 0-mean constraint can be simply imposed by redefining the single basis function as $X - \mathrm{E}_X[X]$, while the scaling constant is found to be $C = \mathrm{Var}_X[X]$. As a consequence, a linear effect after the application of Proposition 1 becomes:

$$\widetilde{f}(X) = \frac{X - \mathrm{E}_X[X]}{\sqrt{\mathrm{Var}_X[X]}} u$$

which corresponds to a linear effect on the standardized version of covariate X. This result shows the coherence of Proposition 1 with the approach taken in the R2D2 literature.

Linear effects represent a special case in which the standardization is straightforward. The correct application of more complex effects requires special care in considering some technical aspects affecting both steps of the procedure.

3.1 | 0-mean constraint

The 0-mean constraint can be imposed in two ways. One method consists in redefining the basis as $\boldsymbol{D}_j(X_j) - \mathrm{E}_{X_j}[\boldsymbol{D}_j(X_j)]$: this is the only option when the process has a single basis function $D_j(X_j)$, as in Example 1. For a general process with more than one basis function, the constraint can also be imposed by replacing the prior on $\boldsymbol{u}_j$ with its distribution conditional on the linear constraint $\boldsymbol{a}_j^T \boldsymbol{u}_j = 0$, where $\boldsymbol{a}_j = \mathrm{E}_{X_j}[\boldsymbol{D}_j(X_j)]$. This latter solution is preferable since it avoids modifying the basis and therefore preserves the original interpretation of the coefficients. Additionally, popular IGMRFs inherently require the imposition of linear constraints to ensure propriety (see Section 4): it is therefore convenient to directly exploit this natural property to impose the 0-mean constraint whenever possible, rather than introducing unnecessary transformations of the basis.

Example 2 (Group effect, i.e., effect of a categorical variable). Consider a categorical variable X with support $\{1,\dots,K\}$ and $\pi(X = k) = p_k$ for $k = 1,\dots,K$. An effect for X can be expressed using a vector of K coefficients $\boldsymbol{u} = [u_1,\dots,u_K]^T$, each of them being linked to a value of the support through basis functions $D_k(X) = \mathbb{I}[X = k]$:

$$f(X) = \sum_{k=1}^{K} \mathbb{I}(X = k) u_k; \qquad \boldsymbol{u}|\sigma^2 \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

If the effect is considered random, the standardization procedure has no impact as $C = 1$. However, there might be cases where $u_1,\dots,u_K$ are directly of interest, i.e., the effect should be treated as fixed. In this second scenario, the 0-mean constraint is such that $\boldsymbol{a} = [p_1,\dots,p_K]^T$, and the scaling constant, defined as in Equation (5), is found to be: $C = 1 - \frac{\sum_{k=1}^{K} p_k^3}{\sum_{k=1}^{K} p_k^2}$. If X is uniformly distributed with $p_k = 1/K$, $\forall k$, then $C = (K-1)K^{-1}$. As the number of levels K increases, C converges to 1: this happens because the variance of the mean decreases and therefore the 0-mean constraint becomes increasingly less relevant. See the proof in the Supporting Information.

3.2 | Scaling constant

The scaling step of Proposition 1 can be implemented in practice either by dividing the basis matrix by $\sqrt{C_j}$ or by multiplying the structure matrix by C_j. The computation of the scaling constants C_j requires taking into account different aspects. First, all linear constraints imposed on the effect (including the ones from the 0-mean constraint step) must be considered before the application of the scaling procedure (see Section 4). Second, there is no guarantee that C_j will be non-null and finite for all potential models, as it is a function of $\pi_j(x_j), \boldsymbol{D}_j(X_j), \boldsymbol{Q}_j$: hence, the couples $f_j(X_j)$ and $\pi_j(x_j)$ must be designed to ensure that $0 < C_j < \infty$, $\forall j$. Thirdly, it might

be easier to approximate the values of C_j using a Monte Carlo simulation, sampling N values $x_{j,1},\dots,x_{j,N}$ from $\pi_j(x_j)$ to compute an estimate $\widehat{C}_j = N^{-1}\sum_{i=1}^{N} \boldsymbol{D}_j^T(x_i)\boldsymbol{Q}_j^- \boldsymbol{D}_j(x_i)$. The scaling method from Proposition 1 can be viewed as an extension of Sørbye and Rue (2014) to the case where the covariates are treated as random variables. In this generalized setting, the *reference variance* can be redefined as $\sigma^2_{\text{ref}} = \text{GM}_{X_j}\left\{\text{Var}_{\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2 = 1, X_j]\right\}$, where $\text{GM}_X[\cdot] = \exp[\text{E}_X(\log\cdot)]$. Noting that Equation (5) can be expressed as $C_j = \text{E}_{X_j}\{\text{Var}_{\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2 = 1, X_j]\}$, it is easy to see that the new scaling constant simply replaces the geometric mean (GM) with the expected value (E). Importantly, expectation-based scaling has advantages over the geometric mean approach. For example, geometric mean scaling would return a null σ^2_{ref} constant for linear effects and, more generally, for all effects for which $\exists x \in \mathcal{X}_j$ such that $\pi_j(x_j) > 0$ and $\text{Var}_{\boldsymbol{u}_j}[f_j(x)|\sigma_j^2, X_j = x] = 0$. Additionally, the useful result in Remark 1 from the following section is only achievable thanks to the linearity property of expectation, and therefore does not hold under geometric mean scaling.

3.3 | Interpretation of VP parameters

We formally define VP priors for the class of models in Definition 1, and show how the corresponding reparametrization offers nicely interpretable parameters after the standardization.

Definition 4 (VP priors). Consider the model from Definition 1. We define the VP parameters as:

$$\begin{aligned} V &= \sum_{j=1}^{J} \sigma_j^2, \\ \boldsymbol{\omega} &= \left[\omega_1 = \frac{\sigma_1^2}{V}, \dots, \omega_{J-1} = \frac{\sigma_{J-1}^2}{V}, \omega_J = 1 - \sum_{j=1}^{J-1}\omega_j\right]^T. \end{aligned} \tag{6}$$

Let $\boldsymbol{J}$ represent the Jacobian associated with the transformation $\boldsymbol{\sigma}^2 \to [V, \omega_1, \dots, \omega_{J-1}]^T$. We call VP those priors implied on the original variance parameters $\boldsymbol{\sigma}^2$ by the specification of independent priors on the VP parameters, i.e.,

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

As defined, the class of VP priors is extremely broad, and can cover most of the works cited in Section 1, in particular HD and R2D2 priors. In the case of a Gaussian likelihood, Definition 4 technically excludes the residual observation variance from the partitioning. It is possible to include the residual variance simply by specifying the model so that $Y \sim N(\eta, 0)$ and add an IID effect in the linear predictor (see Section 5.1): we recommend this option only when it is meaningful to include prior beliefs about the signal–to-noise ratio.

Proposition 1 enables us to derive interpretable expressions for the VP parameters, as detailed in the following remark.

Remark 1 (Interpretation of the VP parameters). Consider the class of models from Definition 1 with parameters of interest $\boldsymbol{\theta} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_L, \sigma^2_{L+1}, \dots, \sigma^2_J]^T$. After the application of standardization (Proposition 1), the V parameter from Equation (6) can be rewritten as the sum of the expected variances of interest:

$$V = \sum_{j=1}^{L} \mathrm{E}_{\boldsymbol{u}_j}\{\mathrm{Var}_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\} + \sum_{k=L+1}^{J} \mathrm{Var}_{X_k,\boldsymbol{u}_k}[f_k(X_k)|\sigma_k^2]$$
$$= \sum_{j=1}^{J} \mathrm{E}_{\boldsymbol{\theta}}\{\mathrm{Var}_{\boldsymbol{X},\boldsymbol{u}_1,\ldots,\boldsymbol{u}_J}[f_j(X_j)|\boldsymbol{\theta}]|\boldsymbol{\sigma}^2\}.$$

More importantly, the VP parameters from Equation (6) become equal to:

$$V = \mathrm{E}_{\boldsymbol{\theta}}\{\mathrm{Var}_{\boldsymbol{X},\boldsymbol{u}_1,\ldots,\boldsymbol{u}_J}[\eta|\mu,\boldsymbol{\theta}]|\boldsymbol{\sigma}^2\},$$
$$\omega_j = \frac{\mathrm{E}_{\boldsymbol{\theta}}\{\mathrm{Var}_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\boldsymbol{\theta}]|\boldsymbol{\sigma}^2\}}{V} \quad j = 1,\ldots,J.$$

See the proof in Section A2 of the Appendix.

Remark 1 shows that, after standardization, the parameter V is not only equal to the sum of the expected variances of interest of the individual effects, but also exactly corresponds to the expected variance of interest of the entire linear predictor η. It is important to note that this interpretation of the VP parameters always holds a priori but not necessarily a posteriori, as Remark 1 relies on the prior assumption of conditional independence among the model coefficients given the variance parameters (not on the independence between the covariates which is not required), an assumption violated a posteriori.

4 | IGMRFS

In this section, we discuss the steps required for a sensible application of the standardization procedure to effects with IGMRF priors, which are often present in LGMs.

IGMRFs are improper multivariate Gaussian distributions, i.e., their structure matrix has a rank-deficiency $d > 0$ and therefore a non-empty null space, denoted by matrix $\boldsymbol{S}$ with d columns, such that $\boldsymbol{QS} = \boldsymbol{0}$. Rue and Held (2005) defined various classes of IGMRFs: in this paper, we focus on a subclass of Definition 3.5 of Rue and Held (2005).

Definition 5 (IGMRFs on regular locations on the line). Consider a vector of regular locations denoted by $\boldsymbol{x} = [1,\ldots,K]^T$. A random vector $\boldsymbol{u} \in \mathbb{R}^K$ is an IGMRF of order d with parameters $\boldsymbol{\mu}$ and $\boldsymbol{Q}$ on regular locations on the line if it has (improper) density:

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

and $\boldsymbol{QS}_{(d-1)} = \boldsymbol{0}$ where $\boldsymbol{S}_{(d-1)} = [\boldsymbol{x}^0, \boldsymbol{x}^1,\ldots,\boldsymbol{x}^{d-1}]$ is called a polynomial matrix (note that the exponentiation is to be interpreted element-wise).

An effect $f(X)$ from Definition 1 with an IGMRF prior of order d from Definition 5 can be represented as:

$$f(X) = \boldsymbol{D}^T(X)\boldsymbol{u}; \quad \boldsymbol{u}|\sigma^2 \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{Q}^-) \text{ where } \boldsymbol{QS}_{(d-1)} = \boldsymbol{0}. \tag{7}$$

To understand the role of σ^2 in such effects, we make use of the decomposition from Rue and Held (2005, Section 3.4.1):

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

where $\boldsymbol{u}$ is decomposed in a polynomial trend of degree $d-1$ and a residual part through the *hat matrix* $\boldsymbol{H}_{(d-1)} = \boldsymbol{S}_{(d-1)}[\boldsymbol{S}^T_{(d-1)}\boldsymbol{S}_{(d-1)}]^{-}\boldsymbol{S}^T_{(d-1)}$. Since $\boldsymbol{Q}\boldsymbol{H}_{(d-1)} = \boldsymbol{0}$, the density of $\boldsymbol{u}$ (Equation 5) when $\boldsymbol{\mu} = \boldsymbol{0}$) can be rewritten as:

$$\pi(\boldsymbol{u}) \propto \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{u} - \boldsymbol{H}_{(d-1)}\boldsymbol{u})^T\boldsymbol{Q}(\boldsymbol{u} - \boldsymbol{H}_{(d-1)}\boldsymbol{u})\right]. \tag{8}$$

Equation (8) clearly shows how σ^2 does not control the dispersion of the IGMRF process around its mean but rather only the deviation from its polynomial trend of degree $d-1$. Hence, the σ^2 parameter can only be correctly interpreted as the variance of the process after the removal of the polynomial trend, which can be achieved by imposing the *null space constraints* $\boldsymbol{S}^T_{(d-1)}\boldsymbol{u} = \boldsymbol{0}$. The new process $\boldsymbol{u}|\boldsymbol{S}^T_{(d-1)}\boldsymbol{u} = \boldsymbol{0}$ corresponds to a proper GMRF under constraints, whose variance parameter can be correctly interpreted as a measure of dispersion of the process from its mean, i.e., $\boldsymbol{0}$. However, the constrained process $\boldsymbol{u}|\boldsymbol{S}^T_{(d-1)}\boldsymbol{u} = \boldsymbol{0}$ is not equivalent to the original (unconstrained) process $\boldsymbol{u}$, since it "has lost" its polynomial trend of degree $d-1$. As a consequence, $f(X)$ from Equation (7) will also be different from $f(X)|\boldsymbol{S}^T_{(d-1)}\boldsymbol{u} = \boldsymbol{0}$. Hence, σ^2 alone is unable to capture the variance contribution of the original effect, only that one of the constrained version.

In order to specify processes from Equation (7) so that variance parameters can measure their variance contribution, a single effect is therefore insufficient, and it becomes necessary to use a representation of the process through multiple separate effects, each with its own associated variance parameter. In what follows, we describe how such an alternative representation, denoted by $f^{\text{new}}(X)$, can be conveniently built for two classes of effects frequently employed in LGMs: Section 4.1 considers IGMRF effects for discrete supports (e.g., areal/grid data, time series observed at regular time points); Section 4.2 focuses on P-splines effects, which are used to model nonlinear effects of continuous covariates.

4.1 | IGMRFs for discrete effects

Consider an IGMRF effect for a discrete covariate X with support $\{1,\dots,K\}$ with finite K defined as:

$$f(X) = \sum_{k=1}^{K}\mathbb{I}(X=k)u_k; \quad \boldsymbol{u}|\sigma^2 \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{Q}^{-}) \text{ where } \boldsymbol{Q}\boldsymbol{S}_{(d-1)} = \boldsymbol{0}. \tag{9}$$

Equation (9) covers various discrete space and time effects, as the ICAR for areal data (Besag & Kooperberg, 1995) and the RW1/RW2 processes for temporal data. Since X usually represents space (e.g., grid cell/county) or time (e.g., day/year) in most applications of Equation (9), we define $X \sim \text{Unif}\{1,\dots,K\}$ to reflect that each location or time point is equally represented in the population-level model: this assumption simplifies the derivation of theoretical results and avoids giving preferences to some values of the support over others.

From the previous discussion, it is clear that the specification of the effect from Equation (9) is inadequate whenever the interpretation of σ^2 is desirable for prior specification, e.g., for VP priors. In order to build an appropriate alternative representation, we start by considering the effect under the null space constraints, which we formally denote by $f^{(r)}(X)$ where r stands for *residual*:

$$f^{(r)}(X) = \sum_{k=1}^{K}\mathbb{I}(X=k)u_k; \quad \boldsymbol{u}|\boldsymbol{S}^T_{(d-1)}\boldsymbol{u} = \boldsymbol{0}, \sigma^2_r \sim N(\boldsymbol{0}, \sigma^{2(r)}\boldsymbol{Q}^{-}).$$

We know that the null space constraints $\boldsymbol{S}_{(d-1)}^T\boldsymbol{u} = \boldsymbol{0}$ guarantee that the coefficients $\boldsymbol{u}$ have a null polynomial trend of degree $d-1$. In this particular case, this is also true for $f^{(r)}(X)$ due to the special choice of basis, i.e.,

$$\mathrm{E}_X[X^m f^{(r)}(X)] = 0 \quad m = 0,\dots,d-1. \tag{10}$$

See the proof in the Supporting Information. The result in Equation (10) proves that imposing the null space constraints on the original process removes the polynomial trend. Therefore, $f(X)$ can be nicely reconstructed defining $f^{\text{new}}(X)$ as:

$$f^{\text{new}}(X) = f^{(t)}(X) + f^{(r)}(X) \tag{11}$$

where $f^{(t)}(X)$ (t stands for *trend*) is designed to model the polynomial effect of degree $d-1$ for covariate X, and $\sigma^{2(t)}$ is the associated variance parameter.

In the case of $d = 1$, $f^{(t)}(X)$ is redundant as the polynomial trend would simply be a constant effect with respect to X, whose inclusion would clash with the intercept parameter μ in the linear predictor and cause an identifiability issue. Hence, $f^{\text{new}}(X) = f^{(r)}(X)$, so that the single $\sigma^{2(r)}$ is sufficient to measure the variance contribution of the effect. For $d = 2$ instead, the polynomial term must be a linear effect $f^{(t)}(X) = X\beta^{(t)}$ where $\beta|\sigma^{2(t)} \sim N(0, \sigma^{2(t)})$, so that $f^{\text{new}}(X) = f^{(t)}(X) + f^{(r)}(X)$. In general, $f^{(t)}(X)$ will be designed using a vector of coefficients $\boldsymbol{\beta}$ of dimension $d-1$ (since IGMRFs of order $d > 2$ are rarely used in practice, the design of $f^{(t)}(X)$ for $d > 2$ is described in the Supporting Information). We denote the coefficients of the trend effect $f^{(t)}(X)$ as $\boldsymbol{\beta}$, instead of $\boldsymbol{u}$, simply to emphasize the linear/polynomial nature of these effects.

In order to obtain the interpretability of $\sigma^{2(t)}$ and $\sigma^{2(r)}$, the two effects must be separately standardized according to Proposition 1. Note how the null space constraints already imply $\mathrm{E}_X[f^{(r)}(X)] = 0$ (i.e., a 0-mean constraint on $f^{(r)}(X)$) for every $d > 0$: hence, the standardization of $f^{(r)}(X)$ is invariant to the fixed/random effect classification.

Remark 2. After standardizing both $f^{(t)}(X)$ and $f^{(r)}(X)$ terms from Equation (11), it can be proved that:

$$\mathrm{Var}_{X,\boldsymbol{\beta},\boldsymbol{u}}[f^{\text{new}}(X)|\sigma^{2(t)}, \sigma^{2(r)}] = \sigma^{2(t)} + \sigma^{2(r)}, \tag{12}$$

where $\sigma^{2(r)}$ controls the polynomial contribution, i.e., the deviation of the polynomial trend from the null mean of the process, while $\sigma^{2(r)}$ controls the residual contribution, i.e., the deviation of the process from its polynomial trend. The process $f^{\text{new}}(X)$ converges to the original $f(X)$ when $\sigma^{2(t)} \to \infty$ (i.e., any possible realization of the polynomial trend is equally likely). See the proof in the Supporting Information.

Example 3 (Standardization of RW1/RW2/ICAR). The effect of a time covariate with regularly spaced observations is often modelled in LGMs assuming a RW1 or RW2 on the basis coefficients $\boldsymbol{u}$ in Equation (9). An RW1 is an IGMRF of order $d = 1$ with structure matrix:

$$\boldsymbol{Q}_{\text{RW1}} = \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \cdots & \cdots & \cdots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}. \tag{13}$$

Since $d = 1$, it is sufficient to impose the null space constraint $\mathbf{1}^T\boldsymbol{u} = 0$ to obtain a measure of the variance contribution of the effect. After standardization, an RW1 effect can be written as:

$$f^{\text{new}}(X) = \sum_{k=1}^{K} \mathbb{I}(X = k)u_k; \qquad \boldsymbol{u}|\mathbf{1}^T\boldsymbol{u} = 0, \sigma^2 \sim N\left(\mathbf{0}, \frac{\sigma^{2(r)}}{C}\boldsymbol{Q}_{\text{RW1}}^{-}\right) \tag{14}$$

where C depends on the choice of the distribution on X according to Equation (5) and is such that $\text{Var}_{X,\boldsymbol{u}}[f^{\text{new}}(X)|\sigma^{2(r)}] = \sigma^{2(r)}$. Being an IGMRF of order 1, an ICAR effect (Besag & Kooperberg, 1995) can also be introduced using Equation (14), replacing the structure matrix with $\text{diag}(\boldsymbol{W1}) - \boldsymbol{W}$ where $\boldsymbol{W}$ must be the adjacency matrix for the given lattice. An RW2 is instead an IGMRF of order $d = 2$ with structure matrix:

$$\boldsymbol{Q}_{\text{RW2}} = \begin{bmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \dots & \dots & \dots & \dots & \dots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{bmatrix}. \tag{15}$$

In this case, it is necessary to decompose the effect into a linear term and a residual one subject to the null space constraints. After standardization of both terms, the effect can be written as:

$$\begin{aligned} f^{\text{new}}(X) &= \frac{X - \text{E}_X[X]}{\sqrt{\text{Var}_X[X]}}\beta + \sum_{k=1}^{K} \mathbb{I}(X = k)u_k, \\ \beta|\sigma^{2(t)} &\sim N(0, \sigma^{2(t)}), \\ \boldsymbol{u}|\boldsymbol{S}_{(1)}^T\boldsymbol{u} = \mathbf{0}, \sigma^{2(r)} &\sim N\left(\mathbf{0}, \frac{\sigma^{2(r)}}{C}\boldsymbol{Q}_{\text{RW2}}^{-}\right), \end{aligned}$$

so that $\text{Var}_{X,\beta,\boldsymbol{u}}[f^{\text{new}}(X)|\sigma^{2(t)}, \sigma^{2(r)}] = \sigma^{2(t)} + \sigma^{2(r)}$. The null space is $\boldsymbol{S}_{(1)} = [\mathbf{1}_K, (1,\dots,K)^T]$. Under the assumption of $X \sim \text{Unif}\{1,\dots,K\}$, the scaling constants C for different values of K for both RW1 and RW2 are reported in the Supporting Information.

4.2 | IGMRFs for P-splines

We now discuss standardization of P-splines (Eilers & Marx, 1996; Lang & Brezger, 2004), a popular approach to model the nonlinear effect of a continuous covariate X. Formally, a P-spline is a special case of the IGMRF effect in Equation (7),

$$f(X) = \boldsymbol{B}(X)^T\boldsymbol{u}; \quad \boldsymbol{u}|\sigma^2 \sim N(\mathbf{0}, \sigma^2\boldsymbol{Q}^{-}) \text{ where } \boldsymbol{Q}\boldsymbol{S}_{(d-1)} = \mathbf{0}, \tag{16}$$

where the basis $\boldsymbol{D}(X)$ is replaced by $\boldsymbol{B}(X) = [B_1(X),\ldots, B_K(X)]^T$, with $B_k(X), k = 1,\ldots, K$, a B-spline of any degree evaluated at X, and $\boldsymbol{u}$ being the vector of spline coefficients with an IGMRF prior on it, often a RW2. Note, the notation $\boldsymbol{D}(X)$ is replaced by $\boldsymbol{B}(X)$ to explicitly highlight that we now focus on the special case of B-spline bases. Calculating B-splines requires the choice of a finite interval $[m, M]$ over the support of X. As a distribution for X, we find convenient to assume equal weight over $[m, M]$, thus we take $X \sim \text{Unif}(m, M)$.

Defining the concept of variance contribution for a P-spline is challenging as the variance parameter σ^2 of the IGMRF, see Equation (7), does not measure variance contribution directly. A measure of variance contribution is given by the conditional variance, i.e., the variance of $f(X)$ computed under null space constraints at each value of X (see Figure 3). The challenge here is that the latter not only depends on σ^2 but also on the user-defined number of basis functions K (Ventrucci & Rue, 2016).

The alternative representation required for the interpretability of the variance parameter could be achieved using Equation (11) again, with $f^{(t)}(X)$ being a polynomial trend effect (see Section 4.1), and $f^{(r)}(X)$ being equal to Equation (16) subject to $\boldsymbol{S}_{(d-1)}^T \boldsymbol{u} = \boldsymbol{0}$. However, this solution is not viable as the use of a more complex basis is responsible for an important difference with respect to Section 4.1: the null space constraints do not guarantee a null polynomial trend on $f_r(X)$. As a consequence, the null space constraints are also no longer sufficient to guarantee the 0-mean constraint from Proposition 1 on $f_r(X)$: if $f_r(X)$ is to be treated as fixed, it would be necessary to further constrain the process during standardization to guarantee the 0-mean requirement (step 1 of Proposition 1). Additionally, an identifiability issue arises between $f_t(X)$ and $f_r(X)$ for $d \geq 2$, since both terms can potentially capture the polynomial trend in the effect of X, as $f_r(X)$ is not appropriately constrained. Consequently, the variance parameters σ_t^2 and σ_r^2 can no longer be neatly interpreted as the polynomial and residual contributions to the variance.

In order to solve the issue raised by the more complex basis choice, we propose a procedure whose implementation can guarantee identifiability between the trend and polynomial terms, $f^{(t)}(X)$ and $f^{(r)}(X)$, and the consequent interpretability of their associated variance parameters.

The procedure presented below, named *Q modification*, solves the problem by replacing the structure matrix with a new $\widetilde{\boldsymbol{Q}}$ whose null space $\widetilde{\boldsymbol{S}}$ is such that the null space constraints $\widetilde{\boldsymbol{S}}^T \boldsymbol{u} = \boldsymbol{0}$ guarantee that $f^{(r)}(X)$ has a null polynomial trend of degree $d-1$, i.e., Equation (10). The adequate null space $\widetilde{\boldsymbol{S}}$ is:

$$\widetilde{\boldsymbol{S}} = \left[\text{E}_X[X^0 \boldsymbol{B}(X)],\ldots, \text{E}_X[X^{(d-1)} \boldsymbol{B}(X)]\right]. \tag{17}$$

$\widetilde{\boldsymbol{Q}}$ must then be found as the solution of $\widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{S}} = \boldsymbol{0}$. We propose a solution that preserves sparsity through the following decomposition:

$$\widetilde{\boldsymbol{Q}} = (\boldsymbol{\Lambda}\widetilde{\boldsymbol{R}}^{-}\boldsymbol{\Lambda})^{-}. \tag{18}$$

$\boldsymbol{\Lambda}$ must be a positive diagonal matrix of entries $\lambda = [\lambda_1,\ldots, \lambda_K]$ and $\widetilde{\boldsymbol{R}}$ a square matrix with the same sparsity and sign structure of $\boldsymbol{Q}$. Finding a solution that guarantees $\widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{S}} = \boldsymbol{0}$ is now equivalent to finding $\widetilde{\boldsymbol{R}}$ such that $\widetilde{\boldsymbol{R}}\boldsymbol{\Lambda}\widetilde{\boldsymbol{S}} = \boldsymbol{0}$. The correct entries of $\widetilde{\boldsymbol{R}}$ are functions of the known elements of $\boldsymbol{Q}, \widetilde{\boldsymbol{S}}$, and the unknown λ. Hence, the new structure matrix $\widetilde{\boldsymbol{Q}}$ is known up to λ. In order to obtain the closest possible solution to the original $\boldsymbol{Q}$, we propose to find λ that minimize the Kullback-Leibler divergence (Rue & Tjelmeland, 2002):

$$\hat{\lambda} = \underset{\lambda > \boldsymbol{0}}{\arg\min}\, D_{\text{KL}}\left(\mathcal{N}_{\widetilde{\boldsymbol{Q}}(\lambda)} \,||\, \mathcal{N}_{\boldsymbol{Q}}\right). \tag{19}$$

After the Q modification, the modified P-spline effect can be defined as $f^{\text{new}} = f^{(t)}(X) + f^{(r)}(X)$, where $f^{(t)}(X)$ is defined as detailed in Section 4.1 and:

$$f^{(r)}(X) = \boldsymbol{B}^T(X)\boldsymbol{u}; \quad \boldsymbol{u}|\widetilde{\boldsymbol{S}}^T\boldsymbol{u} = \boldsymbol{0}, \sigma^{2(r)} \sim N(\boldsymbol{0}, \sigma^{2(r)}\widetilde{\boldsymbol{Q}}^-)$$

As in Section 4.1, $\sigma^{2(t)}$ can be interpreted, after standardization, as the polynomial contribution to the variance, while $\sigma^{2(r)}$ measures the residual contribution. The decomposition in Equation (18) preserves sparsity since $f^{(r)}(X)$ can be equivalently defined as $f^{(r)}(X) = \boldsymbol{B}^T(X)\boldsymbol{\Lambda}\boldsymbol{u}$ and $\boldsymbol{u}|\widetilde{\boldsymbol{S}}^T\boldsymbol{\Lambda}\boldsymbol{u} = \boldsymbol{0}, \sigma^{2(r)} \sim N(\boldsymbol{0}, \sigma^{2(r)}\widetilde{\boldsymbol{R}}^-)$, where $\widetilde{\boldsymbol{R}}$ is, by design, as sparse as the original structure matrix $\boldsymbol{Q}$.

Example 4 (Standardization of a cubic P-spline effect). The most popular choice of P-spline effects consists in the use of a cubic B-spline basis and an RW2 process on the coefficients, so that the structure matrix is $\boldsymbol{Q}_{\text{RW2}}$ from Equation (15) and $\boldsymbol{Q}_{\text{RW2}}\boldsymbol{S}_{(1)} = \boldsymbol{0}$. Since it cannot be proven that the null space constraints imply a null polynomial trend on the constrained process $f^{(r)}(X)$, the Q modification procedure is implemented to achieve interpretability of the variance parameters. First, the new null space $\widetilde{\boldsymbol{S}}$ must be found applying Equation (17):

$$\underset{K\times 2}{\widetilde{\boldsymbol{S}}} = \begin{bmatrix} \widetilde{S}_{1,0} & \widetilde{S}_{1,1} \\ \widetilde{S}_{2,0} & \widetilde{S}_{2,1} \\ \dots & \dots \\ \widetilde{S}_{K,0} & \widetilde{S}_{K,1} \end{bmatrix} = \Big[\mathrm{E}_X[\boldsymbol{B}(X)], \mathrm{E}_X[X\boldsymbol{B}(X)]\Big].$$

Then, a valid solution for $\widetilde{\boldsymbol{Q}}$ from Equation (18) is found if $\widetilde{\boldsymbol{R}} = \widetilde{\boldsymbol{G}} - \widetilde{\boldsymbol{W}}$ where:

$$\widetilde{W}_{k,l} = \begin{cases} \dfrac{(l-k)W_{k,l}}{\lambda_k\lambda_l(\widetilde{S}_{k,0}\widetilde{S}_{l,1} - \widetilde{S}_{k,1}\widetilde{S}_{l,0})} & k \neq l \\ 0 & k = l \end{cases}, \tag{20}$$

$$\widetilde{G}_{k,l} = \mathbb{I}[k = l]\frac{1}{\lambda_k\widetilde{S}_{k,0}}\left[\sum_{j=1}^{K}\widetilde{W}_{k,j}\lambda_j\widetilde{S}_{j,0}\right] \tag{21}$$

and $\boldsymbol{W} = \text{diag}(\text{diag}(\boldsymbol{Q}_{\text{RW2}})) - \boldsymbol{Q}_{\text{RW2}}$. See the Supporting Information, which also contains the solution for an RW1 on the coefficients. Finally, the matrix $\widetilde{\boldsymbol{Q}}$ that best approximates the original $\boldsymbol{Q}$ is found optimizing λ according to Equation (19). After standardization, the P-spline effect can be written as:

$$f^{\text{new}}(X) = f^{(t)}(X) + f^{(r)}(X); \tag{22}$$

$$f^{(t)}(X) = \frac{X - \mathrm{E}_X[X]}{\sqrt{\text{Var}_X[X]}}\beta, \quad \beta|\sigma^{2(t)} \sim N(0, \sigma^{2(t)}); \tag{23}$$

$$f^{(r)}(X) = \boldsymbol{B}^T(X)\boldsymbol{u}, \quad \boldsymbol{u}|\widetilde{\boldsymbol{S}}^T\boldsymbol{u} = \boldsymbol{0}, \sigma^{2(r)} \sim N\left(\boldsymbol{0}, \frac{\sigma^{2(r)}}{C}\widetilde{\boldsymbol{Q}}^-\right). \tag{24}$$

Under the assumption of $X \sim \text{Unif}\{1,\dots,K\}$, the scaling constants C for different values of K are reported in the Supporting Information.

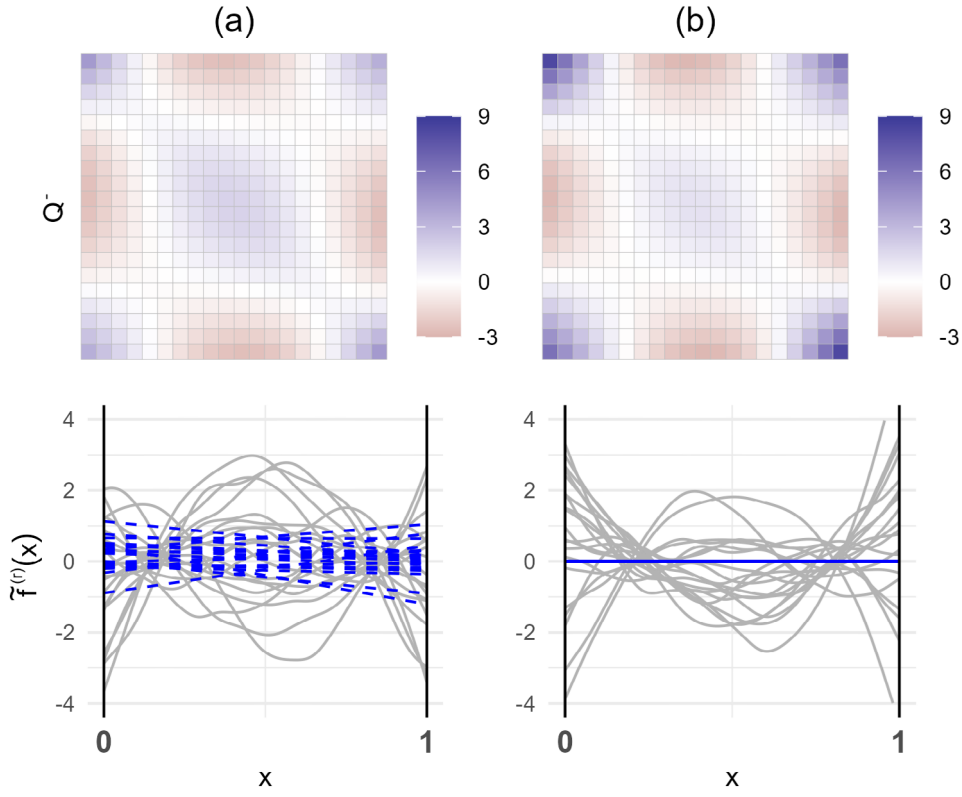

FIGURE 2 Properties of $f^{(r)}(X)$ using (a) $\boldsymbol{Q}_{\mathrm{RW2}}$ and (b) $\widetilde{\boldsymbol{Q}}$ for $K = 20$. Top panel: generalized inverse of the structure matrix on the coefficients (each cell represents an element of the matrix). Bottom panel: realizations of $f^{(r)}(x)$ (grey) with corresponding linear trends (blue) when $\sigma^{2(r)} = 1$ after scaling.

In order to appreciate the impact of the Q modification, the behaviour of $f^{(r)}(X)$ with $K = 20$ from Example 4 is compared under the use of $\boldsymbol{Q}_{\mathrm{RW2}}$ and $\widetilde{\boldsymbol{Q}}$ in Figure 2. First, the top panel shows how the generalized inverse of $\widetilde{\boldsymbol{Q}}$ displays a similar pattern to the one of the original $\boldsymbol{Q}$. The bottom panel of Figure 2 shows realizations of $f^{(r)}(x)$ and their linear trends when $\sigma^{2(r)} = 1$: the original model has non-null trends despite the null space constraints, while the new model by design removes the linear trends.

Figure 3 shows the impact of the Q modification on the shape of the conditional variance of $f^{(r)}(X)$, i.e., $\mathrm{Var}_{\boldsymbol{u}}[f^{(r)}(X)|\sigma^2 = 1, X = x]$. It is important to note that the shape of the conditional variance of the process after the Q modification tends to approximate that one of the original model when $K \to \infty$. Because of this, the conditional variances after the modification (and the standardization) for different values of K are extremely similar (panel b), while this is not the case for the original model (panel a), where we can see that K impacts the conditional variance. Therefore, the Q modification has the important advantage of neutralizing the impact of the choice of K on the conditional variance function: in the proposed modified P-spline model, K still controls the local flexibility of the process (i.e., its wiggliness) but not the global shape of the smooth function, which is controlled by the shape of the conditional variance. Finally, note that the impact of the Q modification is less relevant as K grows, since the difference between $\boldsymbol{S}_{(1)}$ and $\widetilde{\boldsymbol{S}}$ decreases.

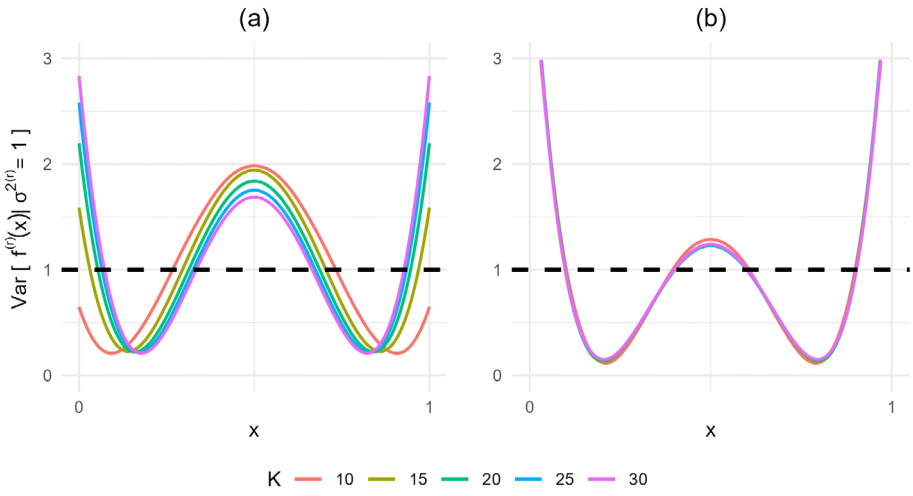

FIGURE 3 Conditional variance of $f^{(r)}(X)$ given $\sigma^{2(r)} = 1$ (after scaling) for different values of K: (a) before the Q modification, (b) after the Q modification.

5 | EMPIRICAL RESULTS

In this section, we investigate the practical benefits of the standardization procedure and the Q modification through two simulation studies and a survival analysis application on real data. Along with traditional independent priors, we consider VP priors, designed according to the HD and/or R2D2 frameworks, depending on what the context allows. The models are fitted in R using INLA (Rue et al., 2009).

5.1 | Simulation study: Impact of standardization

The aim of the first simulation study is to assess the practical impact of the application of the standardization procedure from Proposition 1.

Consider the following simple model:

$$Y \sim N(\mu + f(X), \sigma^2_{\text{IID}}) \tag{25}$$

where $X \sim \text{Unif}[1, 25]$ and $f(X) = \sum_{k=1}^{25} \mathbb{I}(X = k)u_k$ is to be considered random (i.e., standardization simplifies to scaling). The coefficients are specified as a first-order random walk (Example 3), under the appropriate constraint and three different scaling strategies:

- Expectation scaling (Proposition 1): $\boldsymbol{u}|\mathbf{1}^T\boldsymbol{u} = 0, \sigma^2_{\text{RW1}} \sim N\left(\mathbf{0}, \frac{\sigma^2_{\text{RW1}}}{C}\boldsymbol{Q}^-_{\text{RW1}}\right)$
- No scaling: $\boldsymbol{u}|\mathbf{1}^T\boldsymbol{u} = 0, \sigma^2_{\text{RW1}} \sim N\left(\mathbf{0}, \sigma^2_{\text{RW1}}\boldsymbol{Q}^-_{\text{RW1}}\right)$
- Geometric mean scaling (Sørbye & Rue, 2014): $\boldsymbol{u}|\mathbf{1}^T\boldsymbol{u} = 0, \sigma^2_{\text{RW1}} \sim N\left(\mathbf{0}, \frac{\sigma^2_{\text{RW1}}}{\sigma^2_{\text{ref}}}\boldsymbol{Q}^-_{\text{RW1}}\right)$

where the scaling constants are equal to $C = 4.16$ and $\sigma^2_{\text{ref}} = 3.77$. The parameters of the model in Equation (25) are therefore $\mu, \boldsymbol{\sigma}^2 = [\sigma^2_{\text{RW1}}, \sigma^2_{\text{IID}}]$. The only way to apply a VP reparameterization in this case is to include both $\sigma^2_{\text{RW1}}, \sigma^2_{\text{IID}}$: hence, in order to comply with Definition 4, the model must technically be redefined as $Y \sim N(\mu + f(X) + f_{\text{IID}}(X), 0)$ where $f_{\text{IID}}(X) = \sum_{x=1}^{25} \mathbb{I}(X = k) u_{\text{IID},k}$ and $\boldsymbol{u}_{\text{IID}} \sim N(0, \sigma^2_{\text{IID}} \boldsymbol{I})$. Therefore:

$$V = \sigma^2_{\text{RW1}} + \sigma^2_{\text{IID}}; \qquad \omega = \frac{\sigma^2}{\sigma^2_{\text{RW1}} + \sigma^2_{\text{IID}}}.$$

From the model in Equation (25), we also denote by T and φ, respectively, the total variance in Y given the model parameters and the proportion of variance due to $f(X)$:

$$T = \text{Var}_{X,\boldsymbol{u}}[Y|\mu, \boldsymbol{\sigma}^2]; \qquad \varphi = \frac{\text{Var}_{X,\boldsymbol{u}}[f(X)|\boldsymbol{\sigma}^2]}{\text{Var}_{X,\boldsymbol{u}}[Y|\mu, \boldsymbol{\sigma}^2]}.$$

Using VP priors, assumptions about T and φ can be introduced through priors on V and ω. This is correct under expectation scaling since $V = T$ and $\omega = \varphi$. Instead, without scaling we have $T = \sigma^2_{\text{RW1}} C + \sigma^2_{\text{IID}}$ and $\varphi = \frac{\sigma^2_{\text{RW1}} C}{\sigma^2_{\text{RW1}} C + \sigma^2_{\text{IID}}}$, and the VP parameters V and ω are no longer equal to their intuitive interpretations, i.e., T and φ. The implied priors on T and φ would then be distorted in the sense that they would not correctly reflect the user's assumptions about these parameters. For example:

$$\pi(\varphi) = \pi_\omega\left(\frac{\varphi}{\varphi + C - \varphi C}\right)\frac{C}{[\varphi + C - \varphi C]^2} \tag{26}$$

since $\omega = \frac{\varphi T}{\varphi T + (1-\varphi) TC}$ (see the Supporting Information). Under geometric mean scaling, the implied priors would also be distorted, since $T = \sigma^2_{\text{RW1}} \frac{C}{\sigma^2_{\text{ref}}} + \sigma^2_{\text{IID}}$ and $\varphi = \frac{\sigma^2_{\text{RW1}}}{\sigma^2_{\text{RW1}} + \sigma^2_{\text{IID}} \sigma^2_{\text{ref}}}$.

To assess the impact of the lack of appropriate standardization, we first compare the distortion of the implied prior on T and φ for three different prior choices, listed below.

(a) *IG priors*: $\sigma^2_{\text{RW1}}, \sigma^2_{\text{IID}} \overset{iid}{\sim} \text{IG}(1, 5e-5)$;
(b) *PC priors*: $\sigma^2_{\text{RW1}}, \sigma^2_{\text{IID}} \overset{iid}{\sim} \text{PC}_0(3, 0.05)$ where PC_0 is a PC prior (Simpson et al., 2017) with base model $\sigma^2 = 0$;
(c) *HD prior*: $V \sim$ Jeffreys, $\omega \sim \text{Unif}(0, 1)$.

Note that all chosen priors imply a symmetric density on ω. In particular, prior (c) is an HD prior that assumes a Uniform density for ω, which means that all values are equally likely a priori. In this context, it is not meaningful to use an R2D2 prior since there is only one effect in the linear predictor.

Figure 4 reports the density of the implied priors on φ for the different scaling strategies. The distortion caused by inappropriate scaling is assessed by comparing the black lines, i.e., the desired priors on φ, to the other coloured lines, which report the actual implied priors on φ. First, it can be noted that the IG and HD prior choices (which return identical results for φ) seem more sensitive to appropriate scaling than PC priors, since the distortion appears much larger. Secondly, we note that large values of φ are favoured under no scaling more than in the desired

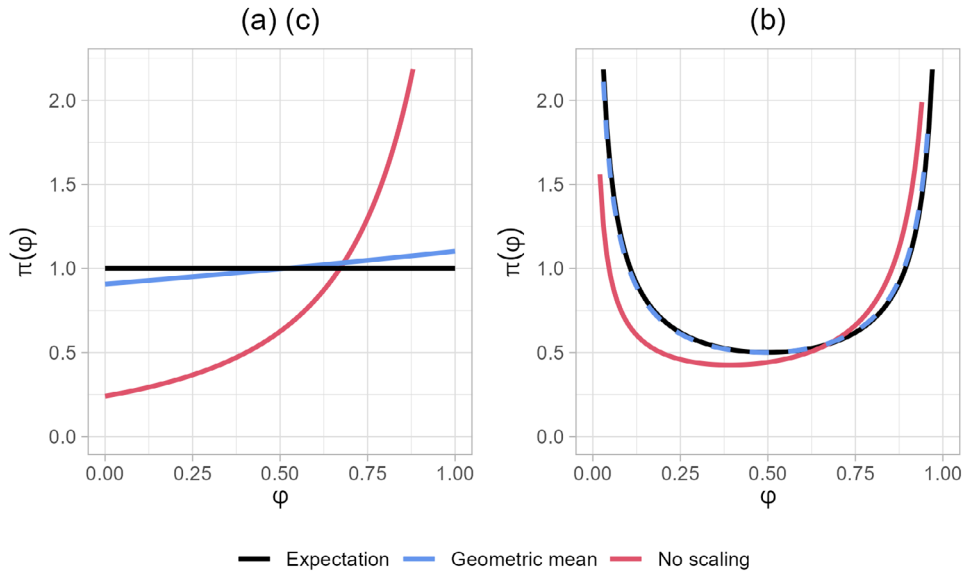

FIGURE 4 Implied prior on φ: (a) IG priors; (b) PC priors; (c) HD prior. The results for (a) and (c) are reported in the same panel as they are identical.

prior, due to the fact that $C > 1$ (the contrary would be true for $C < 1$). Thirdly, geometric mean scaling always favours larger values more than the desired prior since by construction $C > \sigma^2_{\text{ref}}$. Nevertheless, the difference between expectation and geometric mean scaling methods appears small, especially for PC priors: this result suggests that geometric mean scaling already removes most of the distortion.

Since we have found non-negligible distortion in the prior, it is necessary to assess whether this distortion affects posterior inference. To do so, a single observation of Y is simulated for each of the $K = 25$ locations on X, i.e., $x_i = i$ for $i = 1,\ldots, 25$. 200 datasets with $N = 25$ are generated in this manner using $\mu = 0$, $T = 1$ and three different values for $\varphi = 0.2, 0.5, 0.8$. The datasets are then fitted under the three prior choices and the three scaling strategies.

Figure 5 reports the bias of the posterior mean of φ: there is evidence for a non-negligible difference between the results of a scaled model (either geometric mean or expectation scaling) and an unscaled one. As expected, this difference is larger in the case of the HD prior (c), while the PC priors' choice (b) appears more robust, and even more so the IG prior. In most scenarios, there is no relevant difference between the geometric mean and the expectation scaling strategy. Finally, note that the HD prior performs at least as well as the PC prior specification, as it exhibits the smallest variation across simulations and fairly limited estimation bias for all values of φ. Similar conclusions can be drawn looking at the estimates for T, reported in the Supporting Information, along with coverage rates.

In summary, the simulation further confirmed that VP priors can perform at least as well as the popular PC priors. However, the former seems to suffer more severely from the distortion due to inappropriate standardization than the latter. In conclusion, the study just presented highlighted the importance of the scaling step of the standardization procedure; the impact of the 0-mean constraint step has been tested in contexts of Example 2 and found to be practically negligible.

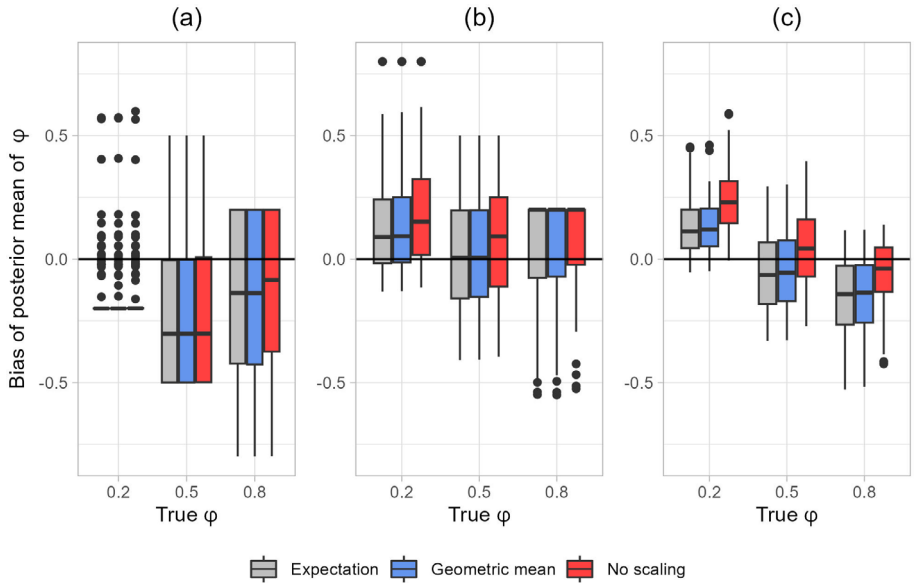

FIGURE 5 Bias (estimate minus true value) of the posterior mean of φ under the following prior choices: (a) IG priors; (b) PC priors; (c) HD prior.

5.2 | Simulation study: Impact of Q modification

The goal of this second simulation study is to investigate the identifiability issue found in the context of P-splines and assess the practical relevance of applying the proposed Q modification (Section 4.2).

Consider a Gaussian likelihood model with a nonlinear effect for continuous covariate $X \sim \text{Unif}(0,1)$ where $Y \sim N(f^{(t)}(X) + f^{(r)}(X), \sigma_\epsilon^2)$, $f^{(t)}(X) = (X - 0.5)\sqrt{12}\beta$ is the linear part and $f^{(r)}(X) = \cos(2\pi X)$ is the nonlinear one. Let φ denote the proportional contribution of the nonlinear effect to the variance in the linear predictor, $\varphi = \frac{\text{Var}_X[f^{(r)}(X)]}{\text{Var}_X[f^{(t)}(X)+f^{(r)}(X)]}$. 200 datasets are generated with $N = 300$ observations and $\sigma_\epsilon^2 = 1, \beta = \sqrt{0.5}$ so that the variance contribution of $f^{(t)}(X)$ and $f^{(r)}(X)$ is equal to 0.5, and so is φ.

The response is then fitted using a Gaussian likelihood model with the modified version of the P-spline effect from Equations (23) and (24) with $K = 10$ (hence, $C = 0.835$). The practical impact of the Q modification is evaluated by comparing the results to a model using the original prior on the coefficients, that is, $\boldsymbol{u}|\boldsymbol{S}_{(1)}^T\boldsymbol{u} = \boldsymbol{0}, \sigma^{2(r)} \sim N\left(\boldsymbol{0}, \frac{\sigma^{2(r)}}{C}\boldsymbol{Q}_{\text{RW2}}^-\right)$ for $C = 1.432$.

Both models are fitted with four different priors, including two VP priors, one based on the approach of the HD framework and the other on the R2D2 literature. Contrary to the previous example, the VP reparametrization is only applied to the linear predictor, as our focus is on the partition between the linear and nonlinear contributions.

(a) *IG priors*: $\sigma^{2(t)}, \sigma^{2(r)}, \sigma_\epsilon^2 \overset{iid}{\sim} \text{IG}(1, 5e-5)$;
(b) *PC priors*: $\sigma^{2(t)}, \sigma^{2(r)}, \sigma_\epsilon^2 \overset{iid}{\sim} \text{PC}_0(3, 0.05)$;
(c) *HD prior*: $V = \sigma^{2(t)} + \sigma^{2(r)} \sim \text{Jeffreys}, \omega = \sigma^{2(t)}/V \sim \text{Unif}(0,1), \sigma_\epsilon^2 \sim \text{IG}(1, 5e-5)$;
(d) *R2D2 prior*: $R^2 = V/(V + \sigma_\epsilon^2) \sim \text{Beta}(1,1), \omega = \sigma^{2(t)}/V \sim \text{Unif}(0,1), \sigma_\epsilon^2 \sim \text{IG}(1, 5e-5)$.

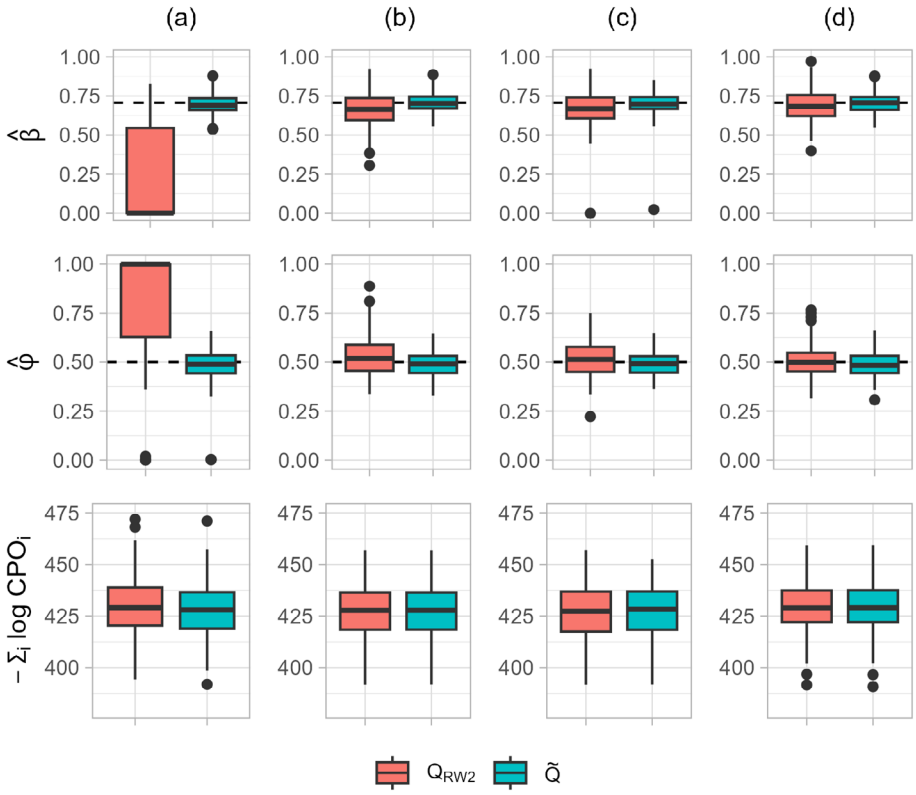

FIGURE 6 Posterior means of β, φ (along with their true values in dashed lines), and summary of the CPO values for the original $\boldsymbol{Q}_{\text{RW2}}$ and its modified version $\widetilde{\boldsymbol{Q}}$: (a) IG priors; (b) PC priors; (c) HD prior; (d) R2D2 prior.

To evaluate the impact on the estimation of the linear/nonlinear contributions, Figure 6 reports the posterior means of β and φ, along with their true values reported via dashed black lines. The results for prior choice (a) highlight the identifiability issue in the original P-spline effect: without the Q modification, the estimates can be highly biased since the linear contribution can be partially or totally absorbed by the nonlinear term. For all prior choices, the Q modification improves the estimates for β and φ, both in terms of bias and variance: the gap between the two models is particularly relevant under prior (a), but the improvement is still evident for priors (b), (c), and (d). Additionally, Figure 6 also reports a measure of predictive performance in terms of the negative sum of the log conditional predictive ordinates (CPO) of all observations, defined as $\text{CPO}_i = \pi(y_i|\boldsymbol{y}_{-i})$. The model after the Q modification performs as well as the original version in terms of predictive ability.

The bias in the estimates of β and φ is caused by the misattribution of the signal among the effects in the linear predictor: specifically, the lack of orthogonality between the two effects before the Q modification causes the P-spline nonlinear term to absorb part of the linear trend. The Q modification removes this identifiability issue by enforcing orthogonality between the linear and nonlinear effects, thus ensuring more accurate estimation of the model parameters and a correct interpretation of the respective contributions of the effects.

Further simulations with different linear/nonlinear contribution ratios show similar results. As K grows, the Q modification becomes less important and the difference in estimation performance becomes negligible for $K > 25$.

5.3 | Case study: Leukaemia in North West England

We consider the dataset analysed by Henderson et al. (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 between 1982 and 1998 in North West England (UK). The following covariates are reported for each patient: *Age*, *Wbc* (white blood cell count at diagnosis), *Tdi* (Townsend social deprivation index), *Sex*, *District* (district of residence). Martino et al. (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:

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

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

The effects $f_1(Age)$, $f_2(Wbc)$, $f_3(Tdi)$ are modelled as P-spline effects with $K = 50$ basis functions (Equations 23 and 24). We denote by $\sigma_1^{2(t)}, \sigma_2^{2(t)}, \sigma_3^{2(t)}$ the variance parameters of the trend terms, and by $\sigma_1^{2(r)}, \sigma_2^{2(r)}, \sigma_3^{2(r)}$ the parameters of the residual terms. $f_4(Sex)$ is set to a group effect (Example 2), while an ICAR model (Besag & 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 $f_T(T)$ (Equation 14): these effects are, respectively, associated with variance parameters $\sigma_4^2, \sigma_S^2, \sigma_T^2$. All effects are treated as fixed; for convenience, a discrete Uniform distribution is assumed for *Sex*, *Time*, *District*, and a continuous one for *Age*, *Wbc*, *Tdi*, on their respective empirical ranges (we checked that slight modification to the ranges does not affect the results). On the basis of this distributional choice, the necessary 0-mean and null space constraints are imposed on the effects.

In terms of VP priors, we implement here the HD framework by applying the VP reparameterization from Definition 4 to all the 9 variance parameters with priors $V \sim$ 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 $\sigma_p^{2(r)}(\sigma_p^{2(t)} + \sigma_p^{2(r)})^{-1}$ for $p = 1, 2, 3$ to penalize nonlinearity. An application of R2D2 in this example would also be possible by following Yanchenko et al. (2024a).

In order to assess the impact of our proposal in this application, we compare the model results before (*No scaling*) and after (*Scaling*) the application of scaling to all effects, since it is the most relevant step of the standardization procedure. Note that, for the linear effects, the basis function is always centered as in Equation (23), but not scaled by the standard deviation of the covariate in the *No scaling* case. Figure 7 reports the posterior mean of the effects for *Wbc*, *Tdi* in both scenarios. 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 & Fahrmeir, 2007). To assess the hyperparameter sensitivity of the results, we examined the impact of scaling under several common choices of the symmetric Dirichlet hyperparameter, namely, 1, 1/2, and 1/9 (i.e., the reciprocal of the number of effects). Our results show that, in the absence of scaling, smaller hyperparameter values improve model performance. In contrast, the results remain robust across all hyperparameter choices when scaling is applied, further highlighting the benefits of scaling (see Supporting Information). Finally, Table 1 compares the goodness-of-fit and the predictive ability of the model, before and after scaling, under different prior choices: the model performance always improves after scaling.

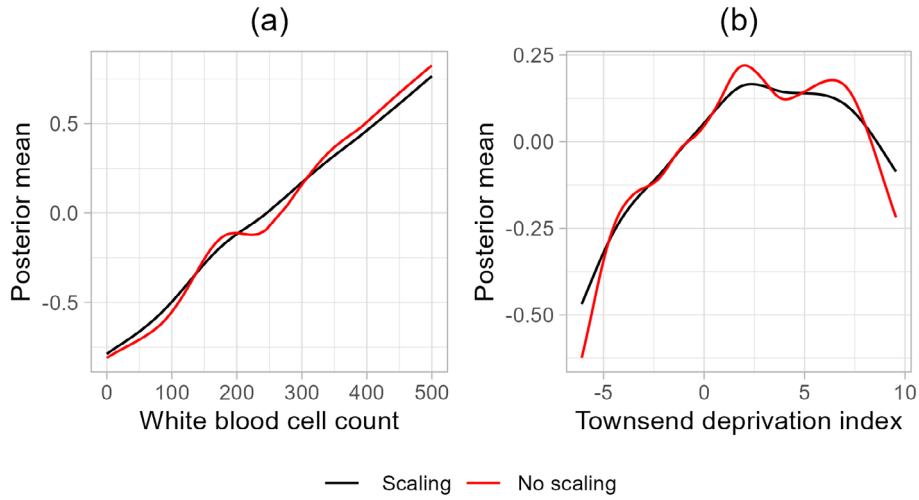

FIGURE 7 Posterior means of (a) $f_2(Wbc)$ and (b) $f_3(Tdi)$ before (red) and after expectation scaling (black).

TABLE 1 Comparison of model performance using the Deviance Information Criterion (DIC) and a summary of the Conditional Predictive Ordinates (CPO_i). The PC and IG prior choices are defined as i.i.d. priors on the original variance parameters, with the same hyperparameter values of Section 5.2.

	DIC			$-\sum_{i=1}^{N} \log CPO_i$		
Prior	**HD**	**PC**	**IG**	**HD**	**PC**	**IG**
No scaling	5655.67	5654.43	5656.25	2841.73	2858.31	2847.96
Scaling	5654.35	5654.38	5654.38	2838.77	2838.91	2838.91

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

6 | DISCUSSION

This paper introduced a standardization method that allows VP priors to be applied for general LGMs, incorporating both fixed and random effects, including IGMRFs. Via a combination of centering and scaling, the standardization procedure aligns the variance parameters with their intuitive interpretation as the variance contribution of the corresponding effects. By bringing together fixed and random effects under a common framework, this work removes a key obstacle to broader adoption of VP priors in applied Bayesian modelling. We proved the interpretation of the VP parameters analytically after standardization. Simulation studies further confirmed the importance of this standardization, especially of its scaling step, for the correct and interpretable use of VP priors.

The proposed expectation-based scaling method showed negligible practical differences from geometric mean-based scaling in our simulation settings. We expect the two methods to yield similar results in most applied contexts. Nonetheless, we found that the scaling constants derived under the two approaches can diverge for certain models, such as RW1 and RW2 processes, as the model dimension increases (i.e., as the process is more densely observed). In these cases,

however, the effect on inference may be limited, since the prior influence on the corresponding variance parameters decreases with increasing data. Despite small practical differences, we argue that our expectation-based approach is preferable due to theoretical properties, specifically, the interpretability result in Remark 1.

We also explored how the concept of variance contribution can be extended to IGMRFs. For these models, variance alignment is achieved by separating the polynomial and residual components into distinct effects with their own variance parameters. While this decomposition is straightforward in simpler cases, P-splines require an additional modification of the structure matrix, which we called the *Q modification*, to resolve identifiability issues. This modification can also be generalized to other structured IGMRFs, such as those formed via Kronecker products.

We believe that providing a general standardization procedure will stimulate further development within the VP framework. While the primary focus of this paper is not on prior specification itself, our goal is to facilitate the incorporation of prior beliefs about relative variance contributions in a principled and interpretable way. After Fuglstad et al. (2020) and Yanchenko et al. (2024a), our simulations further support the claim that VP priors offer a robust and intuitive alternative to conventional priors, since they perform competitively both in terms of parameter estimation and predictive accuracy. Considerable work remains to fully explore the potential of VP priors in this broader context, for example, to investigate the impact of the choice of the probability distributions on the covariates and to assess the sensitivity to hyperparameter choices.

In the future, it would be meaningful to investigate how the VP framework can be effectively integrated into a broader prior specification strategy that accounts for potential correlation parameters in the structure matrices of the effects. Developing a framework for specifying an overall joint prior on variance and correlation parameters would acknowledge that prior beliefs about variance contributions may depend on the magnitude of correlation (see Fuglstad et al., 2019 for an example of a joint prior on the marginal variance and range parameters of the Matern covariance). Another promising direction is opened up by the possibility to apply VP priors to models including smooth nonlinear effects on covariates via P-splines, which may lead to the extension of the use of the R2D2 approach for sparse regression in the context of generalized additive models (Wei et al., 2020).

Finally, we highlight that the proposed standardization procedure is a valuable tool *regardless of the chosen prior*, by ensuring appropriate scaling and maintaining the conceptual distinction between fixed and random effects. In this sense, our contribution extends beyond VP priors and offers a broadly applicable, principled approach to variance parameter scaling in LGMs.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the valuable support and feedback provided by Professor Garritt Page, which greatly contributed to the development of this article.

FUNDING INFORMATION

The authors were supported by the European Union under the NextGeneration EU Programme within the Plan "PNRR—Missione 4" "Istruzione e Ricerca"—Componente C2 Investimento 1.1 "Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)" by the Italian Ministry of University and Research (MUR), Project title: "METAbarcoding for METAcommunities: towards a genetic approach to community ecology (META2)", Project code: 2022PA3BS2 (CUP E53D23007580006), MUR D.D. financing decree n. 1015 of 07/07/2023.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the R package spBayesSurv at https://CRAN.R-project.org/package=spBayesSurv. These data were derived from the following resources available in the public domain:—LeukSurv, https://search.r-project.org/CRAN/refmans/spBayesSurv/html/LeukSurv.html.

ORCID

Luisa Ferrari https://orcid.org/0000-0002-7922-6795
Massimo Ventrucci https://orcid.org/0000-0001-6903-2076

REFERENCES

Aguilar, J. E., & Bürkner, P.-C. (2023). Intuitive joint priors for Bayesian linear multilevel models: The r2d2m2 prior. *Electronic Journal of Statistics*, *17*, 1711–1767.

Besag, J., & Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, *82*, 733–746.

Bhattacharya, A., Pati, D., Pillai, N. S., & Dunson, D. B. (2015). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, *110*, 1479–1490.

Breslow, N. (1972). Disussion of regression models and life-tables by cox, dr. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *34*, 216–217.

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

Fahrmeir, L., Kneib, T., & Lang, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica*, *14*, 731–761.

Franco-Villoria, M., Ventrucci, M., & Rue, H. (2022). Variance partitioning in spatio-temporal disease mapping models. *Statistical Methods in Medical Research*, *31*, 1566–1578.

Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H., & Riebler, A. (2020). Intuitive joint priors for variance parameters. *Bayesian Analysis*, *15*, 1109–1137.

Fuglstad, G.-A., Simpson, D., Lindgren, F., & Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, *114*, 445–452. https://doi.org/10.1080/01621459.2017.1415907

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*, 515–534.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Hem, I., Fuglstad, G.-A., & Riebler, A. (2024). Makemyprior: Intuitive construction of joint priors for variance parameters in R. *Journal of Statistical Software*, *110*, 1–39.

Hem, I. G., Selle, M. L., Gorjanc, G., Fuglstad, G.-A., & Riebler, A. (2021). Robust modeling of additive and nonadditive variation with intuitive inclusion of expert knowledge. *Genetics*, *217*(3), 1–14.

Henderson, R., Shimakura, S., & Gorst, D. (2002). Modeling spatial variation in leukemia survival data. *Journal of the American Statistical Association*, *97*, 965–972.

Hodges, J. S. (2013). *Richly parameterized linear models: Additive, time series, and spatial models using random effects*. CRC Press.

Kneib, T., & Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, *34*, 207–228.

Lang, S., & Brezger, A. (2004). Bayesian p-splines. *Journal of computational and graphical statistics*, *13*, 183–212.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). Rejoinder to commentaries on: The bugs project: Evolution, critique and future directions. *Statistics in Medicine*, *28*, 3081–3082.

Martino, S., Akerkar, R., & Rue, H. (2011). Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics*, *38*, 514–528.

Riebler, A., Sørbye, S. H., Simpson, D., & Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, *25*, 1145–1165.

Rue, H., & Held, L. (2005). *Gaussian Markov random fields: Theory and applications*. CRC press.

Rue, H., Martino, S., & Chopin, N. (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*, 319–392.

Rue, H., & Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian journal of Statistics*, *29*, 31–49.
Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, *32*, 1–28.
Sørbye, S. H., & Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, *8*, 39–51.
Ventrucci, M., & Rue, H. (2016). Penalized complexity priors for degrees of freedom in Bayesian p-splines. *Statistical Modelling*, *16*, 429–453.
Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, *8*, 158–183.
Wei, R., Reich, B. J., Hoppin, J. A., & Ghosal, S. (2020). Sparse Bayesian additive nonparametric regression with application to health effects of pesticides mixtures. *Statistica Sinica*, *30*, 55–79.
Yanchenko, E., Bondell, H. D., & Reich, B. J. (2024a). The r2d2 prior for generalized linear mixed models. *The American Statistician*, *79*(1), 10–40.
Yanchenko, E., Bondell, H. D., & Reich, B. J. (2024b). Spatial regression modeling via the r2d2 framework. *Environmetrics*, *35*, 1–13.
Zhang, Y. D., Naughton, B. P., Bondell, H. D., & Reich, B. J. (2022). Bayesian regression using a prior on the model fit: The r2-d2 shrinkage prior. *Journal of the American Statistical Association*, *117*, 862–874.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ferrari, L., & Ventrucci, M. (2026). A standardization procedure to incorporate variance partitioning-based priors in latent Gaussian models. *Scandinavian Journal of Statistics*, *53*(1), 364–394. https://doi.org/10.1111/sjos.70042

APPENDIX A

A1 Proof of Proposition 1

Proposition 1 can be proved in multiple steps. Recall that the j^{th} effect from Definition 1 is defined as $f_j(X_j) = \mathbf{D}_j(X_j)^T \mathbf{u}_j$. First, it can be noticed that the definition of intuitive interpretation requirement for the fixed effects (Equation 3) can be rewritten as a difference between two terms if the variance is written in terms of difference of expectations:

$$\begin{aligned}\mathrm{E}_{\boldsymbol{u}_j}\{\mathrm{Var}_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\} &= \mathrm{E}_{\boldsymbol{u}_j}\{\mathrm{E}_{X_j}[f_j^2(X_j)|\boldsymbol{u}_j] - \mathrm{E}^2_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\} \\ &= \mathrm{E}_{\boldsymbol{u}_j}\{\mathrm{E}_{X_j}[f_j^2(X_j)|\boldsymbol{u}_j]|\sigma_j^2\} - \mathrm{E}_{\boldsymbol{u}_j}\{\mathrm{E}^2_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\}\end{aligned}$$

At this stage, the order of integration in the first term on the right-hand side can be changed as long as $\mathrm{E}_{\boldsymbol{u}_j}\{\mathrm{E}_{X_j}[f_j^2(X_j)|\boldsymbol{u}_j]|\sigma_j^2\}$ is finite (Fubini-Tonelli theorem). Inverting the expectations, the first term becomes equal to the marginal variance given σ_j^2. The second term on the right-hand side can also be rewritten as a variance, noting that $\mathrm{E}_{\boldsymbol{u}_j}\{\mathrm{E}_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\} = 0$:

$$\begin{aligned}\mathrm{E}_{\boldsymbol{u}_j}[\mathrm{Var}_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2] &= \mathrm{E}_{X_j}\{\mathrm{E}_{\boldsymbol{u}_j}[f_j^2(X_j)|X_j]|\sigma_j^2\} - \mathrm{E}_{\boldsymbol{u}_j}\{\mathrm{E}^2_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\} \\ &= \mathrm{Var}_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2] - \mathrm{Var}_{\boldsymbol{u}_j}\{\mathrm{E}_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\}\end{aligned}$$

Hence, a 0-mean constraint on $j = 1,\ldots,L$ (step 1 of Proposition 1) guarantees that $\mathrm{Var}_{\boldsymbol{u}_j}\{\mathrm{E}_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\} = 0,\ j = 1,\ldots,L$ so that:

$$\mathrm{E}_{\boldsymbol{u}_j}[\mathrm{Var}_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2] = \mathrm{Var}_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2] \quad j = 1,\ldots,L$$

Finally, if the original effect is replaced with $\widetilde{f}_j(X_j) = f_j(X_j)/\sqrt{C_j}$ (step 2 of Proposition 1), then we obtain that:

$$\begin{aligned}
\mathrm{Var}_{X_j,\boldsymbol{u}_j}\left[\widetilde{f}_j(X_j)|\sigma_j^2\right] &= \frac{1}{C_j}\mathrm{Var}_{X_j,\boldsymbol{u}_j}\left[f_j(X_j)|\sigma_j^2\right] \\
&= \frac{1}{C_j}\mathrm{E}_{X_j}\left\{\mathrm{Var}_{\boldsymbol{u}_j}\left[f_j(X_j)|\sigma_j^2\right]\right\} \\
&= \frac{\sigma_j^2}{C_j}\mathrm{E}_{X_j}\left[\boldsymbol{D}_j^T(X_j)\boldsymbol{Q}_j^-\boldsymbol{D}_j(X_j)\right] \\
&= \frac{\sigma_j^2 C_j}{C_j} = \sigma_j^2.
\end{aligned}$$

Hence, we have found that Proposition 1 ensures that:

$$\sigma_j^2 = \mathrm{Var}_{X_j,\boldsymbol{u}_j}[\widetilde{f}_j(X_j)|\sigma_j^2] \quad j = 1,\ldots,J$$

but also that:

$$\sigma_j^2 = \mathrm{E}_{\boldsymbol{u}_j}\{\mathrm{Var}_{X_j}[\widetilde{f}_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\} \quad j = 1,\ldots,L$$

which proves that the two steps of Proposition 1 are sufficient to guarantee Definition 3.

A2 Proof of Remark 1

Proposition 1 guarantees that Equations (3–4) are respected. The first part of Remark 1 can be proven by showing that both expressions can be rewritten as the expected variance of interest, i.e., $\sigma_j^2 = \mathrm{E}_{\boldsymbol{\theta}}\{\mathrm{Var}_{\boldsymbol{X},\boldsymbol{u}_1,\ldots,\boldsymbol{u}_J}[f_j(X_j)|\boldsymbol{\theta}]|\boldsymbol{\sigma}^2\}$. For fixed effects $j = 1,\ldots,L$, the expression becomes:

$$\mathrm{E}_{\boldsymbol{\theta}}\{\mathrm{Var}_{\boldsymbol{X},\boldsymbol{u}_1,\ldots,\boldsymbol{u}_J}[f_j(X_j)|\boldsymbol{\theta}]|\boldsymbol{\sigma}^2\} = \mathrm{E}_{\boldsymbol{u}_j}\{\mathrm{Var}_{X_j}[f_j(X_j)|\boldsymbol{u}_j]|\sigma_j^2\}$$

which corresponds to Equation (3). For random effects $j = L+1,\ldots,J$, the expression becomes:

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\theta}}\{\mathrm{Var}_{\boldsymbol{X},\boldsymbol{u}_1,\ldots,\boldsymbol{u}_J}[f_j(X_j)|\boldsymbol{\theta}]|\boldsymbol{\sigma}^2\} &= \mathrm{E}_{\sigma_j^2}\{\mathrm{Var}_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2]|\sigma_j^2\} \\
&= \mathrm{Var}_{X_j,\boldsymbol{u}_j}[f_j(X_j)|\sigma_j^2]
\end{aligned}$$

which corresponds to Equation (4). Hence, if $V = \sum_{j=1}^J \sigma_j^2$ as in Equation (6) and the conditions of Definition 3 hold, then V becomes the sum of the expected variances of interest of all the J effects:

$$V = \sum_{j=1}^{J} \mathrm{E}_{\boldsymbol{\theta}}\{\mathrm{Var}_{\boldsymbol{X},\boldsymbol{u}_1,\ldots,\boldsymbol{u}_J}[f_j(X_j)|\boldsymbol{\theta}]|\boldsymbol{\sigma}^2\}$$

Secondly, it is necessary to prove that the expected variance of interest of the linear predictor, i.e., $\mathrm{E}_\theta\{\mathrm{Var}_{\boldsymbol{X},\boldsymbol{u}_1,\ldots,\boldsymbol{u}_J}[\eta|\mu,\boldsymbol{\theta}]|\boldsymbol{\sigma}^2\}$, is equal to the sum of the expected variances of interest of all the J effects, or after standardization, the sum of σ_j^2. First, we can note that:

$$\mathrm{E}_\theta\{\mathrm{Var}_{\boldsymbol{X},\boldsymbol{u}_1,\ldots,\boldsymbol{u}_J}[\eta|\mu,\boldsymbol{\theta}]|\boldsymbol{\sigma}^2\} = \mathrm{E}_\theta\left\{\mathrm{Var}_{\boldsymbol{X},\boldsymbol{u}_1,\ldots,\boldsymbol{u}_J}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{\theta}\right]|\boldsymbol{\sigma}^2\right\}$$

Secondly, we can rewrite this expression as:

$$\mathrm{E}_{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_L}\left\{\mathrm{Var}_{\boldsymbol{X},\boldsymbol{u}_{L+1},\ldots,\boldsymbol{u}_J}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{u}_1,\ldots,\boldsymbol{u}_L,\sigma_{L+1}^2,\ldots,\sigma_J^2\right]|\sigma_1^2,\ldots,\sigma_L^2\right\}$$

This expression can be rewritten more concisely using the notation $\boldsymbol{U}_F = [\boldsymbol{u}_1,\ldots,\boldsymbol{u}_L]$, $\boldsymbol{U}_R = [\boldsymbol{u}_{L+1},\ldots,\boldsymbol{u}_J]$, $\boldsymbol{\sigma}_F^2 = [\sigma_1^2,\ldots,\sigma_L^2]$, $\boldsymbol{\sigma}_R^2 = [\sigma_{L+1}^2,\ldots,\sigma_J^2]$ as:

$$\mathrm{E}_{\boldsymbol{U}_F}\left\{\mathrm{Var}_{\boldsymbol{X},\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{U}_F,\boldsymbol{\sigma}_R^2\right]|\boldsymbol{\sigma}_F^2\right\}$$

Then, we can write the argument of the expectation using the law of total variance as:

$$\begin{aligned}
\mathrm{Var}_{\boldsymbol{X},\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{U}_F,\boldsymbol{\sigma}_R^2\right] &= \mathrm{E}_{\boldsymbol{X}}\left\{\mathrm{Var}_{\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{X},\boldsymbol{U}_F,\boldsymbol{\sigma}_R^2\right]|\boldsymbol{U}_F,\boldsymbol{\sigma}_R^2\right\} \\
&\quad + \mathrm{Var}_{\boldsymbol{X}}\left\{\mathrm{E}_{\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{X},\boldsymbol{U}_F,\boldsymbol{\sigma}_R^2\right]|\boldsymbol{U}_F,\boldsymbol{\sigma}_R^2\right\} \\
&= \mathrm{E}_{\boldsymbol{X}}\left\{\mathrm{Var}_{\boldsymbol{U}_R}\left[\sum_{j=L+1}^{J} f_j(X_j)|\boldsymbol{X},\boldsymbol{\sigma}_R^2\right]|\boldsymbol{\sigma}_R^2\right\} \\
&\quad + \mathrm{Var}_{\boldsymbol{X}}\left[\sum_{j=1}^{L} f_j(X_j)|\boldsymbol{U}_F\right] \\
&= \sum_{j=L+1}^{J} \mathrm{E}_{\boldsymbol{X}}\left\{\mathrm{Var}_{\boldsymbol{U}_R}\left[f_j(X_j)|\boldsymbol{X},\boldsymbol{\sigma}_R^2\right]|\boldsymbol{\sigma}_R^2\right\} \\
&\quad + \mathrm{Var}_{\boldsymbol{X}}\left[\sum_{j=1}^{L} f_j(X_j)|\boldsymbol{U}_F\right] \\
&= \sum_{j=L+1}^{J} \mathrm{E}_{\boldsymbol{X}}\left\{\mathrm{E}_{\boldsymbol{U}_R}\left[f_j^2(X_j)|\boldsymbol{X},\boldsymbol{\sigma}_R^2\right]\right\} \\
&\quad + \mathrm{Var}_{\boldsymbol{X}}\left[\sum_{j=1}^{L} f_j(X_j)|\boldsymbol{U}_F\right]
\end{aligned}$$

If a 0-mean constraint is imposed on the $j = 1,\ldots,L$ effects, then:

$$\mathrm{Var}_{\boldsymbol{X}}\left[\sum_{j=1}^{L} f_j(X_j)|\boldsymbol{U}_F\right] = \mathrm{E}_{\boldsymbol{X}}\left\{\left[\sum_{j=1}^{L} f_j(X_j)\right]^2|\boldsymbol{U}_F\right\}$$

so that:

$$\mathrm{Var}_{\boldsymbol{X},\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{U}_F,\sigma_R^2\right] = \sum_{j=L+1}^{J}\mathrm{E}_{\boldsymbol{X}}\left\{\mathrm{E}_{\boldsymbol{U}_R}\left[f_j^2(X_j)|\boldsymbol{X},\sigma_R^2\right]\right\} + \mathrm{E}_{\boldsymbol{X}}\left\{\left[\sum_{j=1}^{L} f_j(X_j)\right]^2|\boldsymbol{U}_F\right\}$$

If we consider again $\mathrm{E}_{\boldsymbol{U}_F}\left\{\mathrm{Var}_{\boldsymbol{X},\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{U}_F,\sigma_R^2\right]|\sigma_F^2\right\}$, it can be written as:

$$\mathrm{E}_{\boldsymbol{U}_F}\left\{\mathrm{Var}_{\boldsymbol{X},\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{U}_F,\sigma_R^2\right]|\sigma_F^2\right\} = \sum_{j=L+1}^{J}\mathrm{E}_{\boldsymbol{X}}\left\{\mathrm{E}_{\boldsymbol{U}_R}\left[f_j^2(X_j)|\boldsymbol{X},\sigma_R^2\right]\right\} + \mathrm{E}_{\boldsymbol{U}_F}\left\{\mathrm{E}_{\boldsymbol{X}}\left\{\left[\sum_{j=1}^{L} f_j(X_j)\right]^2|\boldsymbol{U}_F\right\}|\sigma_F^2\right\}.$$

Inverting the order of expectation, we get:

$$\begin{aligned}\mathrm{E}_{\boldsymbol{U}_F}\left\{\mathrm{Var}_{\boldsymbol{X},\boldsymbol{U}_R}\left[\sum_{j=1}^{J} f_j(X_j)|\boldsymbol{U}_F,\sigma_R^2\right]|\sigma_F^2\right\} &= \sum_{j=L+1}^{J}\mathrm{E}_{\boldsymbol{X}}\left\{\mathrm{E}_{\boldsymbol{U}_R}\left[f_j^2(X_j)|\boldsymbol{X},\sigma_R^2\right]\right\} + \mathrm{E}_{\boldsymbol{X}}\left\{\mathrm{E}_{\boldsymbol{U}_F}\left\{\left[\sum_{j=1}^{L} f_j(X_j)\right]^2|\sigma_F^2\right\}\right\}\\ &= \sum_{j=L+1}^{J}\mathrm{E}_{\boldsymbol{X}}\left\{\mathrm{E}_{\boldsymbol{U}_R}\left[f_j^2(X_j)|\boldsymbol{X},\sigma_R^2\right]\right\} + \sum_{j=1}^{L}\mathrm{E}_{\boldsymbol{X}}\left\{\mathrm{E}_{\boldsymbol{U}_F}\left[f_j^2(X_j)\right]|\sigma_F^2\right\}\\ &= \sum_{j=1}^{J}\mathrm{E}_{\boldsymbol{X}}\left\{\mathrm{E}_{\boldsymbol{U}_F,\boldsymbol{U}_R}\left[f_j^2(X_j)|\sigma_F^2,\sigma_R^2\right]\right\}\\ &= \sum_{j=1}^{J}\mathrm{E}_{\boldsymbol{X}}\left\{\mathrm{Var}_{\boldsymbol{u}_1,..,\boldsymbol{u}_J}\left[f_j(X_j)|\boldsymbol{\sigma}\right]\right\}\\ &= \sum_{j=1}^{J}\sigma_j^2\mathrm{E}_{X_j}\left[\boldsymbol{D}_j^T(X_j)\boldsymbol{Q}_j^{-}\boldsymbol{D}_j(X_j)\right]\end{aligned}$$

If scaling has been applied as in Proposition 1, then we know that:

$$\sum_{j=1}^{J}\sigma_j^2\mathrm{E}_{X_j}\left[\boldsymbol{D}_j^T(X_j)\boldsymbol{Q}_j^{-}\boldsymbol{D}_j(X_j)\right] = \sum_{j=1}^{J}\sigma_j^2$$

which completes the proof.