Cubic Spline with Spatial Data 2026

Mengyan Jing

January 2026

1 Introduction and Goal

Spatial datasets often exhibit two common features: (i) covariate effects on the response may be nonlinear, and (ii) observations recorded at nearby locations tend to be correlated even after adjusting for covariates. In this report, we build and validate a baseline modeling pipeline that combines additive nonlinear covariate effects with residual spatial dependence for continuous outcomes. Specifically, we represent each nonlinear covariate effect using the Lancaster–Šalkauskas (LS) cubic spline basis and model residual spatial structure using a Matérn Gaussian process. We evaluate the pipeline using a controlled simulation study designed to match the assumed model structure and report accuracy for recovered marginal covariate effects.

2 Model

2.1 Model specification (baby model)

Let $s_r \in \mathbb{R}^2$ denote the spatial location of observation r, for $r = 1, \ldots, n$. At each location s_r, we observe a continuous response $Y(s_r)$ and p covariates $X_1(s_r), \ldots, X_p(s_r)$. Our current model is

$$Y(s_r) = \mu + \sum_{j=1}^{p} f_j\big(X_j(s_r)\big) + b(s_r) + \varepsilon_r, \qquad r = 1, \ldots, n. \tag{1}$$

where

- μ is an intercept
- $f(\cdot)$ are unknown smooth functions representing nonlinear covariate effects,
- $b(\cdot)$ is a mean-zero spatial random effect
- ε_r is independent nugget noise, we assume $\varepsilon_r \overset{iid}{\sim} \mathcal{N}(0, \tau^2)$

2.2 LS-spline representation

For each covariate $j = 1, \ldots, p$, we approximate the unknown smooth function $f_j(\cdot)$ using the Lancaster–Šalkauskas (LS) natural cubic spline basis with $M_j \geq 4$ knots. Let $\tau_{1j} < \cdots < \tau_{M_j j}$ denote the knot locations for covariate j.

For observation r at location s_r, define the LS design row vector

$$\boldsymbol{z}_{rj} \equiv \boldsymbol{z}_j\big(X_j(s_r)\big) \in \mathbb{R}^{M_j},$$

where $\boldsymbol{z}_j(\cdot)$ is the reduced LS basis obtained after eliminating the slope parameters under the natural-spline constraints (Appendix 4). Then the smooth effect can be written as

$$f_j\big(X_j(s_r)\big) = \boldsymbol{z}_{rj}^\top \boldsymbol{\theta}_j, \tag{2}$$

where $\boldsymbol{\theta}_j \in \mathbb{R}^{M_j}$ is the LS coefficient vector which can be interpreted as knot-value coefficients.

Stacking $\boldsymbol{z}_{rj}^\top$ over $r = 1, \ldots, n$ yields the design matrix

$$\mathbf{Z}_j = \begin{pmatrix} \boldsymbol{z}_{1j}^\top \\ \vdots \\ \boldsymbol{z}_{nj}^\top \end{pmatrix} \in \mathbb{R}^{n \times M_j},$$

so that

$$\big(f_j(X_j(s_1)), \ldots, f_j(X_j(s_n))\big)^\top = \mathbf{Z}_j\, \boldsymbol{\theta}_j.$$

The full LS-basis construction (basis functions, knot interpretation, and slope elimination via the $\mathbf{A}_j^{-1}\mathbf{C}_j$ reduction) is provided in Appendix 4.

2.3 Spatial effect and covariance

To account for residual spatial dependence, we include a spatial random effect $b(\cdot)$ and write the observation model as

$$Y(s_r) = \mu + \sum_{j=1}^{p} f_j\big(X_j(s_r)\big) + b(s_r) + \varepsilon_r, \qquad r = 1, \ldots, n.$$

Here $b(\cdot)$ is modeled as a mean-zero Gaussian process on the spatial domain. For any two locations $s, s' \in \mathbb{R}^2$, we assume a Matérn covariance function

$$\mathrm{Cov}\{b(s), b(s')\} = \sigma^2\, K_\nu\!\left(\frac{\|s - s'\|}{\rho}\right), \tag{3}$$

where $\sigma^2 > 0$ is the marginal variance, $\rho > 0$ is a range parameter controlling the decay of correlation with distance, and $\nu > 0$ controls smoothness of the spatial field.

Let

$$\boldsymbol{b} = \big(b(s_1), \ldots, b(s_n)\big)^\top$$

and let $\mathbf{D}$ be the pairwise distance matrix with entries $D_{rs} = \|s_r - s_s\|$.

Define the Matérn correlation matrix $\mathbf{R}$ by

$$R_{rs} = K_\nu\!\left(\frac{D_{rs}}{\rho}\right), \qquad r, s = 1, \ldots, n,$$

with $R_{rr} = 1$. Then the spatial effect vector satisfies

$$\boldsymbol{b} \sim N\big(\mathbf{0},\ \sigma^2 \mathbf{R}\big). \tag{4}$$

The remaining error term ε_r represents nugget noise and is assumed independent across locations:

$$\varepsilon_r \overset{\text{iid}}{\sim} N(0, \tau^2), \qquad r = 1, \ldots, n,$$

independent of $\boldsymbol{b}$. Details of the Matérn correlation function $K_\nu(\cdot)$ and the construction of $\mathbf{R}$ are provided in

Appendix 5.

2.4 Identifiability and reparameterization

Model (1) includes an intercept μ, while each LS spline design matrix $\mathbf{Z}_j$ contains a constant (ones) component in its column space (because cubic spline spaces include constants). Consequently, the overall levels of the spline components are not separately identifiable from μ (and, when $p > 1$, not separately identifiable from each other).

For each covariate $j = 1, \ldots, p$, we impose a sum-to-zero constraint on the LS knot-value coefficients:

$$\mathbf{1}_{M_j}^\top \boldsymbol{\theta}_j = 0, \qquad j = 1, \ldots, p, \tag{5}$$

so that μ captures the overall level of the response and each $f_j(\cdot)$ represents deviations around that level. Constraint (5) implies $\theta_{1j} = -(\theta_{2j} + \cdots + \theta_{M_j j})$.

Define the identified coefficient vector

$$\boldsymbol{\beta}_j = (\theta_{2j}, \ldots, \theta_{M_j j})^\top \in \mathbb{R}^{M_j - 1}.$$

Equivalently, define the contrast matrix $\mathbf{T}_j \in \mathbb{R}^{M_j \times (M_j - 1)}$ by

$$\mathbf{T}_j = \begin{pmatrix} -1 & -1 & \cdots & -1 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

so that $\boldsymbol{\theta}_j = \mathbf{T}_j \boldsymbol{\beta}_j$ enforces $\mathbf{1}_{M_j}^\top \boldsymbol{\theta}_j = 0$ automatically. Define the identified spline design matrix

$$\mathbf{W}_j \equiv \mathbf{Z}_j \mathbf{T}_j \in \mathbb{R}^{n \times (M_j - 1)}.$$

Then

$$\boldsymbol{f}_j = \big(f_j(X_j(s_1)), \ldots, f_j(X_j(s_n))\big)^\top = \mathbf{Z}_j\, \boldsymbol{\theta}_j = \mathbf{W}_j\, \boldsymbol{\beta}_j.$$

For observation r, write the LS row as $\boldsymbol{z}_{rj}^\top = (z_{rj,1}, \ldots, z_{rj,M_j})$. Using $\theta_{1j} = -\sum_{m=2}^{M_j} \theta_{mj}$, we have

$$\boldsymbol{z}_{rj}^\top \boldsymbol{\theta}_j = z_{rj,1}\theta_{1j} + \sum_{m=2}^{M_j} z_{rj,m}\theta_{mj} = \sum_{m=2}^{M_j} \theta_{mj}\,(z_{rj,m} - z_{rj,1}) = \boldsymbol{w}_{rj}^\top \boldsymbol{\beta}_j,$$

where $\boldsymbol{w}_{rj}^\top$ is the rth row of $\mathbf{W}_j = \mathbf{Z}_j \mathbf{T}_j$,

$$\boldsymbol{w}_{rj} = \big(z_{rj,2} - z_{rj,1}, \ldots, z_{rj,M_j} - z_{rj,1}\big)^\top \in \mathbb{R}^{M_j - 1}.$$

Let $\boldsymbol{y} = (Y(s_1), \ldots, Y(s_n))^\top$, $\mathbf{1}_n$ be the $n \times 1$ vector of ones, and $\boldsymbol{b} = (b(s_1), \ldots, b(s_n))^\top$.

Define the block design matrix and parameter vector

$$\mathbf{W} = \big(\mathbf{W}_1, \ldots, \mathbf{W}_p\big), \qquad \boldsymbol{\beta} = \big(\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_p^\top\big)^\top.$$

Then the identified model is

$$\boldsymbol{y} = \mu\, \mathbf{1}_n + \mathbf{W}\, \boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{6}$$

with $\boldsymbol{b} \sim N(\mathbf{0}, \sigma^2\mathbf{R})$ from (4) and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \tau^2\mathbf{I}_n)$ independent of $\boldsymbol{b}$. Equivalently, for each $r = 1, \ldots, n$,

$$Y(s_r) = \mu + \sum_{j=1}^{p} \boldsymbol{w}_{rj}^\top \boldsymbol{\beta}_j + b(s_r) + \varepsilon_r.$$

2.5 Estimation approach

We estimate the mean component parameters $(\mu, \boldsymbol{\beta})$ together with the covariance parameters for the residual spatial effect and nugget variance. Let $\nu > 0$ denote the Matérn smoothness; in our current implementation we fix ν and estimate the remaining covariance parameters (ρ, σ^2, τ^2), where ρ is the Matérn range, σ^2 is the marginal variance of the spatial effect, and τ^2 is the nugget variance.

Let $\mathbf{H} = (\mathbf{1}_n \;\; \mathbf{W})$ denote the fixed-effect design matrix (intercept plus identified LS-basis blocks) and let $\boldsymbol{\eta} = (\mu, \boldsymbol{\beta}^\top)^\top$. The model can be written as

$$\boldsymbol{y} = \mathbf{H}\boldsymbol{\eta} + \boldsymbol{b} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{b} \sim N(\mathbf{0}, \sigma^2\mathbf{R}(\rho, \nu)), \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \tau^2\mathbf{I}_n),$$

with $\boldsymbol{b}$ independent of $\boldsymbol{\varepsilon}$. Equivalently,

$$\boldsymbol{y} \sim N(\mathbf{H}\boldsymbol{\eta},\ \boldsymbol{\Sigma}), \qquad \boldsymbol{\Sigma} = \sigma^2\mathbf{R}(\rho, \nu) + \tau^2\mathbf{I}_n.$$

To estimate covariance parameters, we use a profile restricted maximum likelihood (REML) criterion. Following standard practice, we reparameterize using

$$\lambda = \tau^2/\sigma^2, \qquad \boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Sigma}_0(\rho, \lambda), \qquad \boldsymbol{\Sigma}_0(\rho, \lambda) = \mathbf{R}(\rho, \nu) + \lambda\mathbf{I}_n.$$

We then optimize the profiled REML objective over (ρ, λ), profile σ^2 in closed form, and finally recover $\tau^2 = \lambda\sigma^2$. Given $(\widehat{\rho}, \widehat{\sigma}^2, \widehat{\tau}^2)$, we compute $\widehat{\boldsymbol{\eta}}$ by generalized least squares (GLS).

Special case (no spatial effect). When $b(\cdot) \equiv 0$ (Simulation 1), we set $\sigma^2 = 0$ so that $\boldsymbol{\Sigma} = \tau^2\mathbf{I}_n$, and estimation reduces to ordinary least squares (OLS) for $(\mu, \boldsymbol{\beta})$ under independent errors.

Algorithm 1 Profile-REML for (ρ, σ^2, τ^2) and GLS for $(\mu, \boldsymbol{\beta})$

Require: Response vector $\boldsymbol{y} \in \mathbb{R}^n$; locations $\{s_r\}_{r=1}^n$; identified design $\mathbf{W}$.
Require: Matérn smoothness ν (fixed); parameter search domain for (ρ, λ) where $\lambda = \tau^2/\sigma^2$.
Ensure: Estimates $\widehat{\mu}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2, \widehat{\tau}^2, \widehat{\rho}$.

Set $\mathbf{H} = (\mathbf{1}_n \;\; \mathbf{W})$ and $k = \operatorname{rank}(\mathbf{H})$.
Reparameterize with $\lambda = \tau^2/\sigma^2$ and $\boldsymbol{\Sigma}_0(\rho, \lambda) = \mathbf{R}(\rho, \nu) + \lambda \mathbf{I}_n$.
Optimize over (ρ, λ)**:** for each candidate (ρ, λ):
Compute $\mathbf{R}(\rho, \nu)$ from pairwise distances.
Compute Cholesky $\boldsymbol{\Sigma}_0 = \mathbf{L}\mathbf{L}^\top$.
Whiten: $\tilde{\boldsymbol{y}} = \mathbf{L}^{-1}\boldsymbol{y}$, $\tilde{\mathbf{H}} = \mathbf{L}^{-1}\mathbf{H}$.
Compute $\widehat{\boldsymbol{\eta}}(\rho, \lambda) = (\tilde{\mathbf{H}}^\top \tilde{\mathbf{H}})^{-1} \tilde{\mathbf{H}}^\top \tilde{\boldsymbol{y}}$.
Compute $Q(\rho, \lambda) = \|\tilde{\boldsymbol{y}} - \tilde{\mathbf{H}}\widehat{\boldsymbol{\eta}}(\rho, \lambda)\|^2$.
Profile σ^2: $\widehat{\sigma}^2(\rho, \lambda) = Q(\rho, \lambda)/(n - k)$.
Evaluate profiled REML objective (up to a constant):

$$\ell_{\mathrm{REML}}^{\mathrm{prof}}(\rho, \lambda) = -\frac{1}{2}\left\{\log|\boldsymbol{\Sigma}_0(\rho, \lambda)| + \log|\tilde{\mathbf{H}}^\top \tilde{\mathbf{H}}| + (n - k)\log\big(Q(\rho, \lambda)/(n - k)\big)\right\}.$$

Let $(\widehat{\rho}, \widehat{\lambda}) = \arg\max \ell_{\mathrm{REML}}^{\mathrm{prof}}(\rho, \lambda)$.
Set $\widehat{\sigma}^2 = \widehat{\sigma}^2(\widehat{\rho}, \widehat{\lambda})$ and $\widehat{\tau}^2 = \widehat{\lambda}\,\widehat{\sigma}^2$.
Form $\widehat{\boldsymbol{\Sigma}} = \widehat{\sigma}^2 \mathbf{R}(\widehat{\rho}, \nu) + \widehat{\tau}^2 \mathbf{I}_n$.
Compute final GLS:

$$\widehat{\boldsymbol{\eta}} = (\mathbf{H}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{H})^{-1} \mathbf{H}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{y},$$

and extract $\widehat{\mu}$ and $\widehat{\boldsymbol{\beta}}$.

3 Simulation Design and Results

3.1 Setup

Locations. For each run we generate spatial coordinates $s_r = (x_r, y_r) \in [0, 1]^2$, $r = 1, \ldots, n$.

Mean function. All simulations share the same additive nonlinear mean:

$$\eta_r = \mu + 2\sin(\pi X_{r1}) + 1.5\exp(X_{r2} - 0.5) + 0.7X_{r3}^2 + 0.5\sin(2\pi X_{r4}), \qquad \mu = 0. \tag{7}$$

Fitted model. We fit

$$y = \eta + b + \varepsilon, \qquad \eta = \beta_0 \mathbf{1} + W\beta, \qquad \varepsilon \sim N(\mathbf{0}, \tau^2 I),$$

where W is the identified LS-basis block design matrix (LS natural cubic splines with $M = 6$ knots per covariate, and each spline block is centered via a sum-to-zero constraint to avoid confounding with the intercept). The spatial effect is modeled as a mean-zero Matérn GP:

$$b \sim N\big(\mathbf{0}, \sigma^2 R(\rho, \nu)\big),$$

with $\nu = 1.5$ fixed and (ρ, σ^2, τ^2) estimated by profile-REML using the parameterization $\mathrm{Var}(y \mid \beta) = \sigma^2(R + \lambda I)$, $\lambda = \tau^2/\sigma^2$.

Metrics. We report mean recovery via $\mathrm{RMSE}(\hat{\eta}, \eta)$ and $\mathrm{cor}(\hat{\eta}, \eta)$; spatial recovery via $\mathrm{cor}(\hat{b}, b)$ (when b exists); and an overlap/leakage diagnostic $\mathrm{cor}(\hat{\eta}, b)$.

Marginal diagnostics. For covariate X_j $(j = 1, \ldots, p)$, let $w_j(x)^\top$ denote the identified LS-basis row vector for the jth spline block evaluated at $x \in [0,1]$ (consistent with the block columns in $\mathbf{W}_j$). The fitted marginal component curve is

$$\widehat{f}_j(x) = w_j(x)^\top \widehat{\beta}_j, \qquad x \in [0,1].$$

At observed covariate values, define component contributions

$$\widehat{f}_{rj} = \widehat{f}_j(X_{rj}), \qquad r = 1, \ldots, n.$$

When the true component function $f_j(x)$ is available, we apply the same centering convention used in fitting and compare to the truth-centered target

$$f_j^c(x) = f_j(x) - \frac{1}{n}\sum_{r=1}^{n} f_j(X_{rj}).$$

Let $x_1, \ldots, x_G$ be an evaluation grid in $[0,1]$ (here $G = 101$ equally spaced points), and write $\widehat{f}_{jg} = \widehat{f}_j(x_g)$ and $f_{jg}^c = f_j^c(x_g)$. We report grid-based curve metrics:

$$\mathrm{RMSE}_{\mathrm{curve},j} = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\big(\widehat{f}_{jg} - f_{jg}^c\big)^2}, \qquad \mathrm{cor}_{\mathrm{curve},j} = \mathrm{cor}\Big(\{\widehat{f}_{jg}\}_{g=1}^{G},\ \{f_{jg}^c\}_{g=1}^{G}\Big).$$

We also summarize fitted component magnitudes at observed points:

$$\mathrm{sd}_j = \mathrm{sd}\big(\{\widehat{f}_{rj}\}_{r=1}^{n}\big), \qquad \mathrm{meanabs}_j = \frac{1}{n}\sum_{r=1}^{n}\big|\widehat{f}_{rj}\big|, \qquad \mathrm{range}_j = \max_{1\le r\le n}\widehat{f}_{rj} - \min_{1\le r\le n}\widehat{f}_{rj}.$$

3.2 Simulation 1: no spatial effect in truth

Goal. Validate the LS-basis mean component when $b(s) \equiv 0$ (while still fitting the spatial model for consistent code).

Data generating model.

$$Y_r = \eta_r + \varepsilon_r, \qquad \varepsilon_r \overset{\text{iid}}{\sim} N(0, \tau^2),$$

with η_r given by (7) and $X_{rj} \overset{\text{iid}}{\sim} \mathrm{Unif}(0,1)$.

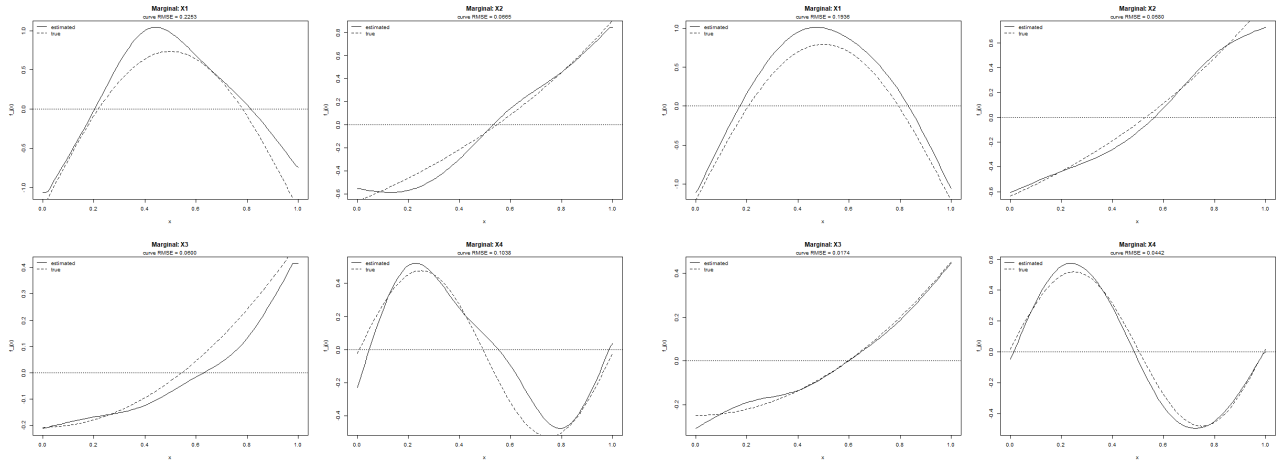
Table 1: Sim1 results. Since a short-range GP can behave like i.i.d. noise, we also report $\hat{\sigma}^2 + \hat{\tau}^2$ as the stable overall noise-scale summary.

n	RMSE_η	cor_η	$\hat{\rho}$	$\hat{\sigma}^2$	$\hat{\tau}^2$	$\hat{\sigma}^2 + \hat{\tau}^2$
100	0.1773	0.9815	1.2924	0.07544	0.10216	0.17760
400	0.08449	0.9959	0.001603	0.14135	0.0000004	0.14135
1000	0.06824	0.9968	0.4482	0.003986	0.14424	0.14822
10000	0.02220	0.9997	0.01727	0.0003036	0.15213	0.15243

Interpretation. First, the LS-basis mean component is recovered increasingly well as sample size grows: RMSE_η decreases from 0.177 at $n = 100$ to 0.022 at $n = 10{,}000$, while cor_η increases from 0.982 to 0.9997. Second, because the truth has no spatial effect, the REML decomposition between $\hat{\sigma}^2$ and $\hat{\tau}^2$ is not reliably identifiable in a single run. In particular, the fit may allocate most variability to $\hat{\sigma}^2$ with a very short estimated range (e.g., $\hat{\rho} = 0.0016$ and $\hat{\tau}^2 \approx 0$ at $n = 400$), or allocate most variability to $\hat{\tau}^2$ with $\hat{\sigma}^2 \approx 0$ (e.g., $n = 10{,}000$).

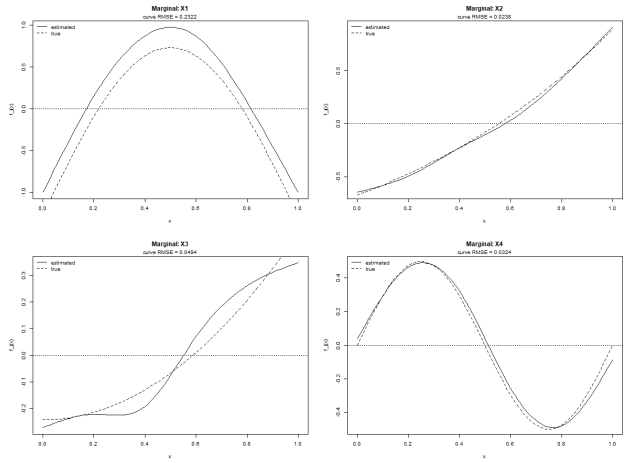
In this setting, the more stable diagnostic is the overall noise scale $\hat{\sigma}^2 + \hat{\tau}^2$, which is close to the injected nugget level for large n (e.g., 0.148 at $n = 1000$ and 0.152 at $n = 10{,}000$, compared to $\tau^2 = 0.15$).

Sim1: marginal curves across sample sizes

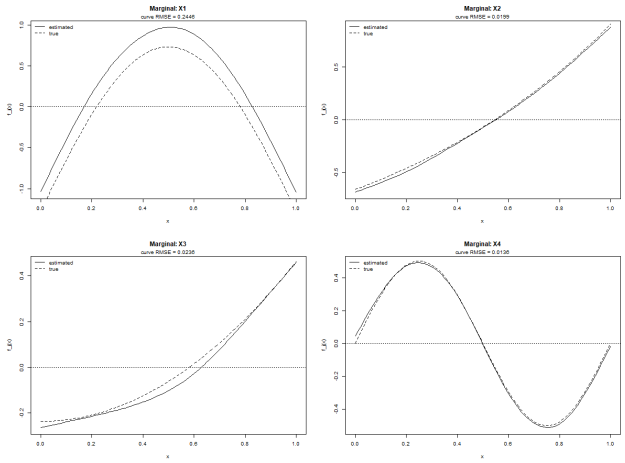


(a) $n = 100$

(b) $n = 400$



(c) $n = 1000$



(d) $n = 10{,}000$

Figure 1: Sim1 marginal component curves for $X_1, \ldots, X_4$ across sample sizes. Each panel overlays $\widehat{f}_j(x)$ with the truth-centered $f_j^c(x)$ and reports grid RMSE.

Table 2: (Sim1) Grid-based marginal curve accuracy across sample sizes. $\text{RMSE}_{\text{curve}}$ and $\text{cor}_{\text{curve}}$ compare $\widehat{f}_j(x)$ to the truth-centered target $f_j^c(x)$ on a grid in $[0, 1]$.

	$n = 100$		$n = 400$		$n = 1000$		$n = 10{,}000$	
Var	RMSE	cor	RMSE	cor	RMSE	cor	RMSE	cor
X_1	0.2253	0.9762	0.1936	0.9989	0.2322	0.9991	0.2446	0.9999
X_2	0.0665	0.9917	0.0580	0.9943	0.0238	0.9991	0.0199	0.9999
X_3	0.0600	0.9915	0.0174	0.9968	0.0494	0.9757	0.0236	0.9980
X_4	0.1038	0.9660	0.0442	0.9947	0.0324	0.9959	0.0136	0.9994

Table 3: (Sim1) Magnitude of fitted component contributions at observed covariate values across sample sizes. For each j, we summarize $\widehat{f}_j(X_{rj})$ over $r = 1, \ldots, n$ by sd, mean absolute value, and range.

	$n = 100$			$n = 400$			$n = 1000$			$n = 10{,}000$		
Var	sd	mean$\lvert\cdot\rvert$	range	sd	mean$\lvert\cdot\rvert$	range	sd	mean$\lvert\cdot\rvert$	range	sd	mean$\lvert\cdot\rvert$	range
X_1	0.6674	0.6097	2.1060	0.6857	0.6363	2.1186	0.6117	0.5765	1.9673	0.6218	0.5942	2.0131
X_2	0.4861	0.4325	1.4277	0.4194	0.3739	1.3313	0.4580	0.3948	1.5515	0.4566	0.3943	1.5581
X_3	0.1735	0.1536	0.6280	0.2153	0.1873	0.7532	0.2207	0.2038	0.6175	0.2148	0.1896	0.7211
X_4	0.3366	0.2918	0.9958	0.3823	0.3411	1.0705	0.3509	0.3202	0.9822	0.3564	0.3237	1.0010

Table 2 shows that the fitted marginal curves recover the *shape* of the true components well: $\mathrm{cor}_{\mathrm{curve},j}$ is consistently high (especially for $n \geq 400$), indicating close agreement between $\widehat{f}_j(x)$ and the truth-centered target $f_j^c(x)$ over the evaluation grid. In contrast, $\mathrm{RMSE}_{\mathrm{curve},j}$ is not strictly monotone in n for all components (notably for X_1 and X_3), which is consistent with a combination of (i) fixed-basis approximation error under $M = 6$ and (ii) single-realization sampling variability in the covariate draws and noise.

Table 3 summarizes the fitted component magnitudes at observed points via $\{\widehat{f}_j(X_{rj})\}_{r=1}^n$. These magnitude summaries are stable across sample sizes and align with the expected effect scales from the data-generating mean: the fitted range for X_1 is about 2 and for X_4 about 1, while X_2 shows an intermediate range and X_3 is the smallest. Overall, these results support that the marginal extraction based on the identified LS-basis blocks and the centering convention is implemented consistently in Sim1.

3.3 Simulation 2: spatial residual with i.i.d. covariates

Goal. Validate the full pipeline when $b(s)$ is present but confounding is minimized (covariates are independent of space).

Data generating model.

$$Y_r = \eta_r + b(s_r) + \varepsilon_r, \qquad \varepsilon_r \overset{\text{iid}}{\sim} N(0, \tau^2),$$

with η_r as in (7), $X_{rj} \overset{\text{iid}}{\sim} \mathrm{Unif}(0,1)$, and $b \sim N(\mathbf{0}, \sigma^2 R(\rho, \nu))$.

Table 4: Sim2 results (truth: $\rho = 0.2$, $\sigma^2 = 0.8$, $\tau^2 = 0.15$; $\nu = 1.5$ fixed).

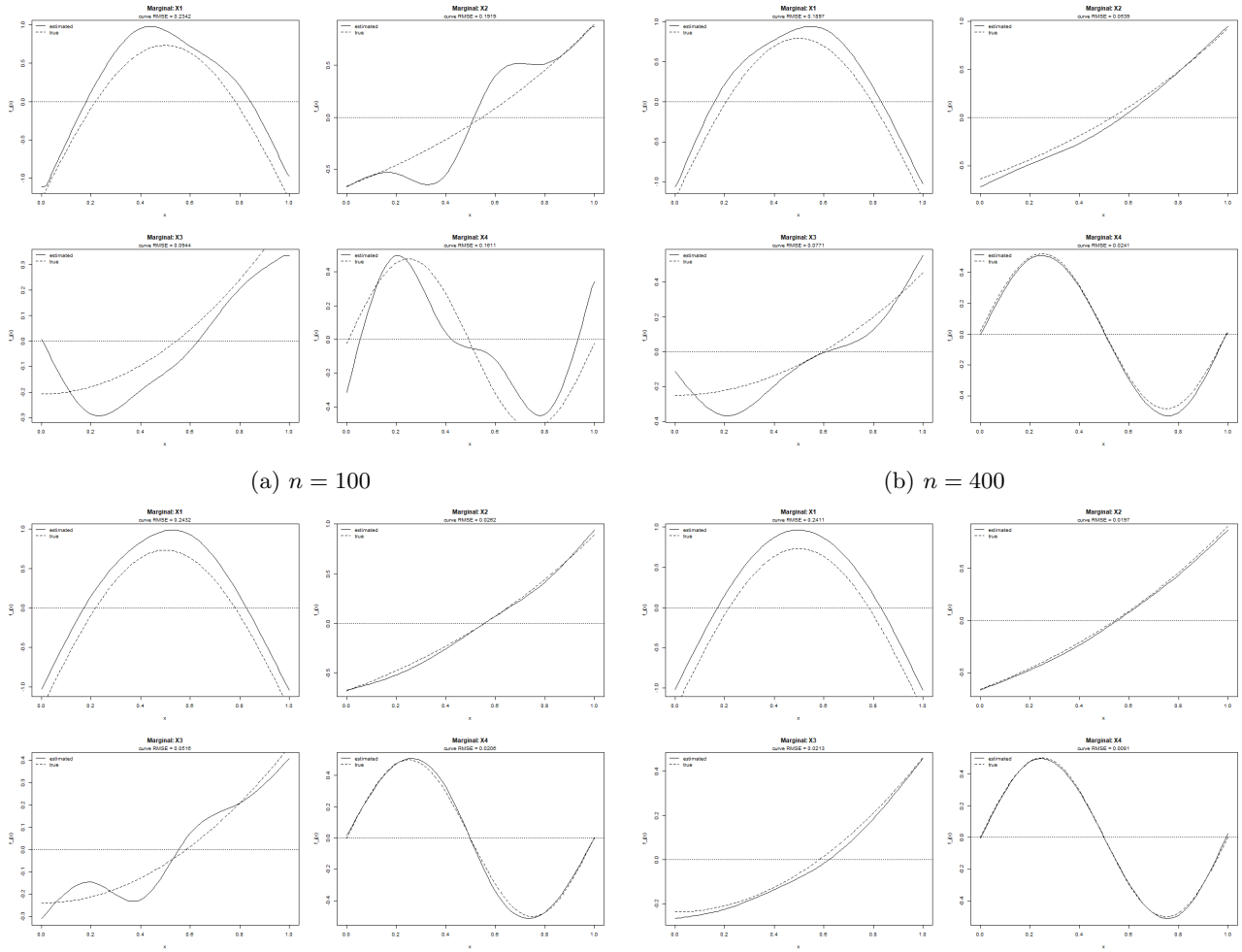
n	RMSE_η	cor_η	cor_b	$\mathrm{cor}(\hat{\eta}, b)$	$\hat{\rho}$ (0.2)	$\hat{\sigma}^2$ (0.8)	$\hat{\tau}^2$ (0.15)
100	0.2715	0.9672	0.8949	-0.0850	0.1819	0.8350	0.1363
400	0.4818	0.9954	0.8851	0.0529	0.2670	0.9909	0.1651
1000	0.4464	0.9973	0.8889	0.0105	0.2537	1.2744	0.1579
10000	0.3444	0.9999	0.7785	0.0114	0.1753	0.7200	0.1460

Interpretation. In Sim2, covariates are generated independently of space, so the mean component $\eta(X)$ and spatial residual $b(s)$ should be well separated. The results largely match this expectation. Across all n, the leakage diagnostic $\mathrm{cor}(\hat{\eta}, b)$ remains close to zero (ranging from -0.085 to 0.011), indicating minimal confounding. The Matérn range is recovered reasonably well, with $\hat{\rho}$ between 0.175 and 0.267 (true $\rho = 0.2$), and the nugget is also close to truth, with $\hat{\tau}^2$ between 0.136 and 0.165 (true $\tau^2 = 0.15$). Spatial recovery is strong for moderate sizes, with $\mathrm{cor}_b \approx 0.89$ for $n \leq 1000$. Meanwhile the mean shape improves with n (cor_η increases from 0.967 at $n = 100$ to 0.9999 at $n = 10{,}000$), supporting the correctness of the end-to-end fitting pipeline when confounding is weak.

Comment on RMSE_η. RMSE_η evaluates *mean recovery only*, i.e., how well the spline mean $\hat{\eta} = X_{\mathrm{fix}}\hat{\beta}$ matches the true η. In a joint spatial fit, some signal can be traded off between the mean component $\hat{\eta}$ and

the spatial BLUP $\hat{b}$ even when confounding is weak, which can inflate RMSE_η while cor_η remains near one. To complement mean recovery, we will also report an *overall signal* metric such as $\mathrm{RMSE}(\hat{\eta} + \hat{b},\ \eta + b)$ (or $\mathrm{RMSE}(\hat{y},\ y)$).

Sim2: marginal curves across sample sizes



(a) $n = 100$ (b) $n = 400$

(c) $n = 1000$ (d) $n = 10{,}000$

Figure 2: Sim2 marginal component curves for $X_1, \ldots, X_4$ across sample sizes. Each panel overlays $\widehat{f_j}(x)$ with the truth-centered $f_j^c(x)$ and reports grid RMSE.

Table 5: (Sim2) Grid-based marginal curve accuracy across sample sizes. $\mathrm{RMSE}_{\mathrm{curve}}$ and $\mathrm{cor}_{\mathrm{curve}}$ compare $\widehat{f_j}(x)$ to the truth-centered target $f_j^c(x)$ on a grid in $[0, 1]$.

	$n = 100$		$n = 400$		$n = 1000$		$n = 10{,}000$	
Var	RMSE	cor	RMSE	cor	RMSE	cor	RMSE	cor
X_1	0.2342	0.9902	0.1897	0.9969	0.2432	0.9991	0.2411	1.0000
X_2	0.1919	0.9461	0.0539	0.9994	0.0262	0.9994	0.0197	0.9999
X_3	0.0944	0.9395	0.0771	0.9678	0.0516	0.9700	0.0213	0.9992
X_4	0.1611	0.8956	0.0241	0.9997	0.0206	0.9988	0.0081	0.9998

Table 6: (Sim2) Magnitude of fitted component contributions at observed covariate values across sample sizes. For each j, we summarize $\widehat{f}_j(X_{rj})$ over $r = 1, \ldots, n$ by sd, mean absolute value, and range.

	$n = 100$			$n = 400$			$n = 1000$			$n = 10{,}000$		
Var	sd	mean$\lvert\cdot\rvert$	range	sd	mean$\lvert\cdot\rvert$	range	sd	mean$\lvert\cdot\rvert$	range	sd	mean$\lvert\cdot\rvert$	range
X_1	0.6622	0.6133	2.0912	0.6381	0.5958	2.0017	0.6209	0.5903	2.0263	0.6155	0.5876	1.9920
X_2	0.5630	0.5327	1.5415	0.4599	0.3997	1.6616	0.4705	0.4089	1.6021	0.4476	0.3890	1.5288
X_3	0.1983	0.1853	0.6270	0.2518	0.2128	0.9163	0.2074	0.1899	0.7158	0.2112	0.1839	0.7242
X_4	0.3090	0.2650	0.9467	0.3688	0.3311	1.0342	0.3617	0.3258	1.0209	0.3514	0.3146	1.0056

Table 5 indicates that the fitted marginal curves improve substantially with sample size when a spatial residual is present but covariates are i.i.d. and independent of space. In particular, for X_2 and X_4 the grid RMSE drops sharply from $n = 100$ to $n \geq 400$ and the corresponding $\text{cor}_{\text{curve},j}$ increases to essentially one (e.g., X_4 reaches cor ≈ 0.9997 at $n = 400$). For X_3, the curve RMSE decreases more gradually and the correlation is somewhat lower at moderate n (cor ≈ 0.97 for $n = 400$–1000), but by $n = 10{,}000$ it also attains excellent agreement (RMSE $= 0.0213$, cor $= 0.9992$). For X_1, $\text{cor}_{\text{curve},1}$ is already high across all n, while $\text{RMSE}_{\text{curve},1}$ remains relatively stable around 0.19–0.24, consistent with a combination of fixed-basis approximation error under $M = 6$ and single-realization variability rather than purely sampling noise.

Table 6 summarizes the distribution of fitted component contributions $\{\widehat{f}_j(X_{rj})\}_{r=1}^n$. These magnitude summaries are stable across sample sizes and preserve the expected ordering of effect scales from the data-generating mean: X_1 shows the largest variability (sd ≈ 0.62–0.66, range ≈ 2), X_2 is intermediate (range ≈ 1.5–1.7), and X_3 remains the smallest on average. Overall, the marginal diagnostics support that, when covariates are not spatially structured (Sim2), the identified LS-spline mean components can be recovered accurately even in the presence of a Matérn residual term.

3.4 Simulation 3: spatial covariates

Goal. Assess separation between $\eta(X(s))$ and $b(s)$ when covariates themselves are spatially structured.

Data generating model. For $r = 1, \ldots, n$ at locations $s_r \in [0, 1]^2$, we generate

$$Y_r = \eta_r + b(s_i) + \varepsilon_r, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \tau^2),$$

with η_r as in (7) and a Matérn residual

$$b \sim N(\mathbf{0}, \sigma^2 R(\rho, \nu)).$$

Unlike Simulation 2, the covariates are spatially structured. For each $j = 1, \ldots, 4$, we draw an independent latent Matérn GP field $z_j(s)$ on the same locations $\{s_i\}$ with parameters (ρ_X, ν_X) (unit-scale correlation matrix $R(\rho_X, \nu_X)$). We add a tiny diagonal nugget $10^{-8}I$ to the covariance matrix for numerical stability in Cholesky sampling. We then apply a rank transform to obtain approximately Unif$(0, 1)$ marginals while preserving spatial dependence:

$$X_{ij} = \frac{\text{rank}\{z_j(s_r)\} - 0.5}{n}, \qquad i = 1, \ldots, n.$$

(Analogously, we add the same $10^{-8}I$ nugget when sampling the residual spatial effect b for numerical stability.)

Interpretation. Sim3 is intentionally challenging because the covariates are spatially structured, so the mean surface $\eta(X(s))$ can inherit spatial smoothness and overlap with the residual spatial effect $b(s)$. This overlap is visible even under the truth: cor(η, b) is not near zero (e.g., 0.133 at $n = 400$ and -0.389 at $n = 1000$ in this sweep). As a consequence, the estimated mean can partially absorb spatial residual structure (or vice versa),

Table 7: Sim3 results (truth: covariates $\rho_X = 0.10$, $\nu_X = 1.0$; residual $\rho = 0.2$, $\sigma^2 = 0.8$, $\tau^2 = 0.15$; $\nu = 1.5$ fixed). We report $\mathrm{cor}(\eta, b)$ under the truth to quantify inherent overlap.

n	$\mathrm{cor}(\eta, b)$	RMSE_η	cor_η	cor_b	$\mathrm{cor}(\hat{\eta}, b)$	$\hat{\rho}$ (0.2)	$\hat{\sigma}^2$ (0.8)	$\hat{\tau}^2$ (0.15)
100	0.2409	0.2516	0.9607	0.7935	0.3138	0.2338	1.1097	0.1098
400	0.1335	0.2323	0.9898	0.7839	0.1250	0.1254	0.2310	0.1476
1000	-0.3889	0.09527	0.9979	0.8227	-0.3824	0.1945	0.8757	0.1445
10000	0.1039	0.03335	0.9995	0.7441	0.1105	0.1614	0.3930	0.1487

which is reflected by non-negligible $\mathrm{cor}(\hat{\eta}, b)$ (e.g., 0.125 at $n = 400$ and -0.382 at $n = 1000$). Despite this confounding, mean recovery still improves markedly with sample size: RMSE_η decreases from 0.252 ($n = 100$) to 0.033 ($n = 10{,}000$), and cor_η increases from 0.961 to 0.9995. The nugget estimate $\hat{\tau}^2$ remains close to the truth (near 0.15), while the process variance estimate $\hat{\sigma}^2$ can be unstable in a single realization (e.g., $\hat{\sigma}^2 = 0.231$ at $n = 400$ versus true $\sigma^2 = 0.8$), consistent with the identifiability challenges induced by overlap.

Comment (inherent overlap and variance-component instability). Sim3 induces spatially structured covariates, so the true mean surface $\eta(X(s))$ can itself be spatially smooth and overlap with the residual spatial effect $b(s)$. We therefore report $\mathrm{cor}(\eta, b)$ to quantify this inherent overlap in the data-generating mechanism. A natural leakage diagnostic is $\mathrm{cor}(\hat{\eta}, b)$: values away from zero do not necessarily indicate model failure here, but rather reflect the confounding built into the truth. In this regime, REML variance decomposition becomes less identifiable, and $\hat{\sigma}^2$ (and sometimes $\hat{\rho}$) can be unstable in a single realization because spatial-looking variation can be attributed to either $\hat{\eta}$ or $\hat{b}$.
We will map confounding severity by varying ρ_X (and/or ν_X) and tracking $\mathrm{cor}(\eta, b)$, $\mathrm{cor}(\hat{\eta}, b)$, and the stability of $(\hat{\rho}, \hat{\sigma}^2)$ over repeated seeds.

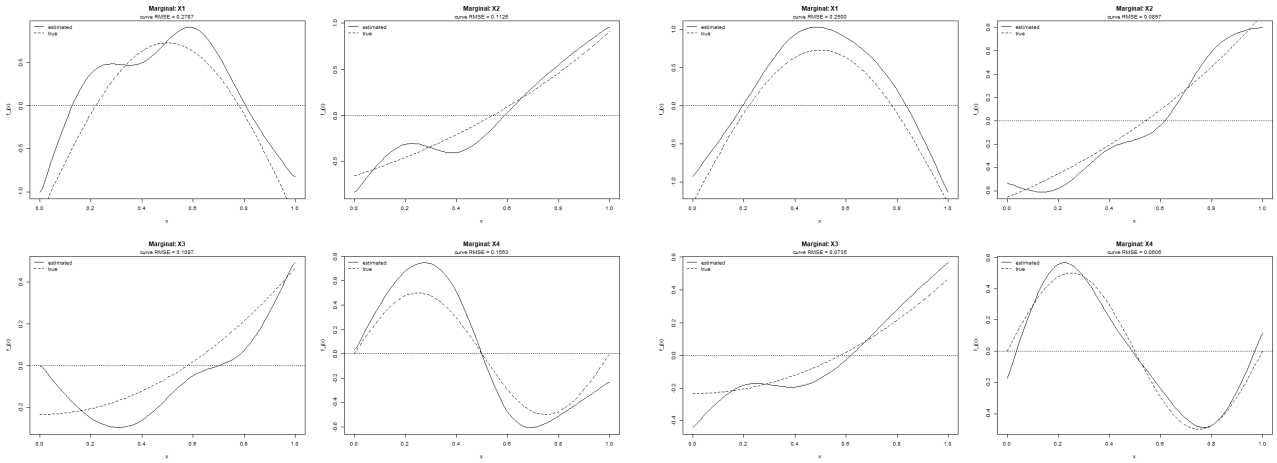
Table 8: (Sim3) Grid-based marginal curve accuracy across sample sizes. $\mathrm{RMSE}_{\mathrm{curve}}$ and $\mathrm{cor}_{\mathrm{curve}}$ compare $\widehat{f}_j(x)$ to the truth-centered target $f_j^c(x)$ on a grid in $[0, 1]$.

	$n = 100$		$n = 400$		$n = 1000$		$n = 10{,}000$	
Var	RMSE	cor	RMSE	cor	RMSE	cor	RMSE	cor
X_1	0.2787	0.9649	0.2500	0.9960	0.2659	0.9984	0.2464	0.9999
X_2	0.1126	0.9765	0.0897	0.9872	0.0527	0.9995	0.0251	0.9991
X_3	0.1097	0.8995	0.0735	0.9844	0.0385	0.9941	0.0260	0.9970
X_4	0.1563	0.9881	0.0606	0.9854	0.0225	0.9985	0.0178	0.9991

Table 9: (Sim3) Magnitude of fitted component contributions at observed covariate values across sample sizes. For each j, we summarize $\widehat{f}_j(X_{rj})$ over $r = 1, \ldots, n$ by sd, mean absolute value, and range.

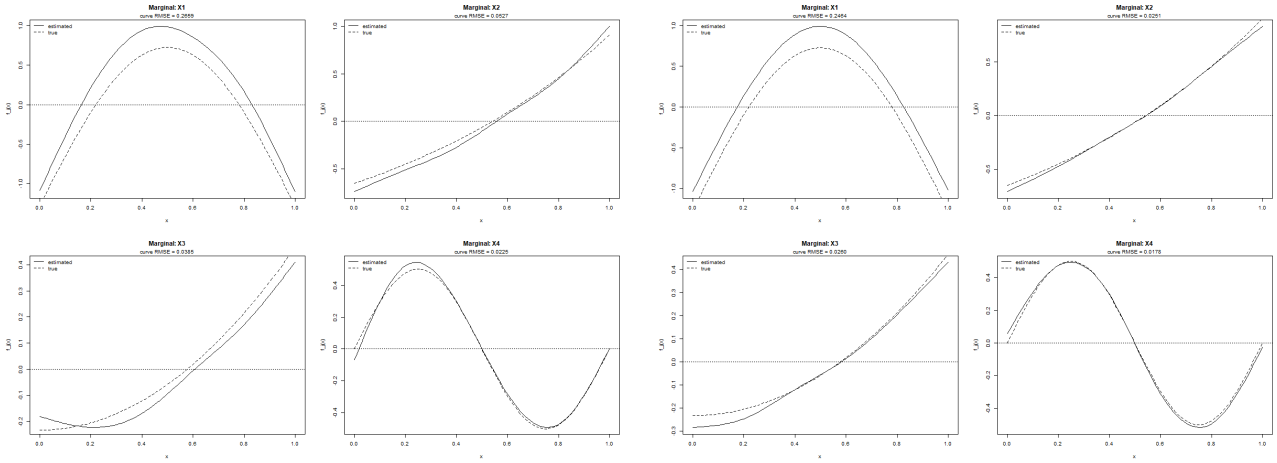
	$n = 100$			$n = 400$			$n = 1000$			$n = 10{,}000$		
Var	sd	mean$\lvert\cdot\rvert$	range	sd	mean$\lvert\cdot\rvert$	range	sd	mean$\lvert\cdot\rvert$	range	sd	mean$\lvert\cdot\rvert$	range
X_1	0.5313	0.5128	1.9066	0.6423	0.6087	2.1573	0.6360	0.6148	2.0894	0.6228	0.5951	2.0263
X_2	0.4945	0.4346	1.7836	0.4889	0.4310	1.4081	0.4863	0.4216	1.7372	0.4483	0.3875	1.5424
X_3	0.2021	0.1724	0.7887	0.2668	0.2314	1.0064	0.1940	0.1741	0.6321	0.2158	0.1877	0.7173
X_4	0.4965	0.4551	1.3540	0.3468	0.3005	1.0541	0.3604	0.3210	1.0346	0.3622	0.3301	1.0093

Sim3: marginal curves across sample sizes



(a) $n = 100$

(b) $n = 400$



(c) $n = 1000$

(d) $n = 10{,}000$

Figure 3: Sim3 marginal component curves for $X_1, \ldots, X_4$ across sample sizes. Each panel overlays $\widehat{f}_j(x)$ with the truth-centered $f_j^c(x)$ and reports grid RMSE.

Table 8 shows that marginal curve recovery improves with n despite the intentionally challenging setting with spatially structured covariates. For X_2, X_3, and X_4, $\mathrm{RMSE}_{\mathrm{curve}}$ decreases substantially as n grows and $\mathrm{cor}_{\mathrm{curve}}$ approaches one by $n \geq 1000$. The main non-monotone pattern is for X_1, where $\mathrm{RMSE}_{\mathrm{curve}}$ remains around 0.25–0.28 even as $\mathrm{cor}_{\mathrm{curve}}$ is near one for $n \geq 400$, consistent with a stable shape match but a persistent approximation/realization-level discrepancy under fixed $M = 6$.

Table 9 indicates that fitted component magnitudes are broadly stable across sample sizes and remain on the expected scale for the true signals. Some variability across n is visible (e.g., larger range for X_3 at $n = 400$ and a downward shift in X_4 magnitude from $n = 100$ to $n \geq 400$), which is plausible in a single realization when spatial structure in $X(s)$ can increase overlap between the mean surface and the residual spatial component.

3.5 Simulation 4: adding irrelevant covariates (negative control)

Goal. Assess robustness of the additive LS-spline mean estimator to irrelevant predictors by augmenting the covariate set with two "garbage" variables. We expect the marginal components for X_5 and X_6 to be approximately flat (near zero), while the recovered components for $X_1, \ldots, X_4$ remain comparable to Simulation 3.

Data generating model. The data generating mechanism matches Simulation 3 for $(X_1, \ldots, X_4)$ and the residual spatial effect $b(s)$: for $r = 1, \ldots, n$ at locations $s_r \in [0,1]^2$,

$$Y_r = \eta_r + b(s_r) + \varepsilon_r, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \tau^2),$$

with η_r as in (7), $(X_1, \ldots, X_4)$ generated via spatially structured latent Matérn GP fields with (ρ_X, ν_X) followed by the rank transform in Sim3, and $b \sim N(\mathbf{0}, \sigma^2 R(\rho, \nu))$.

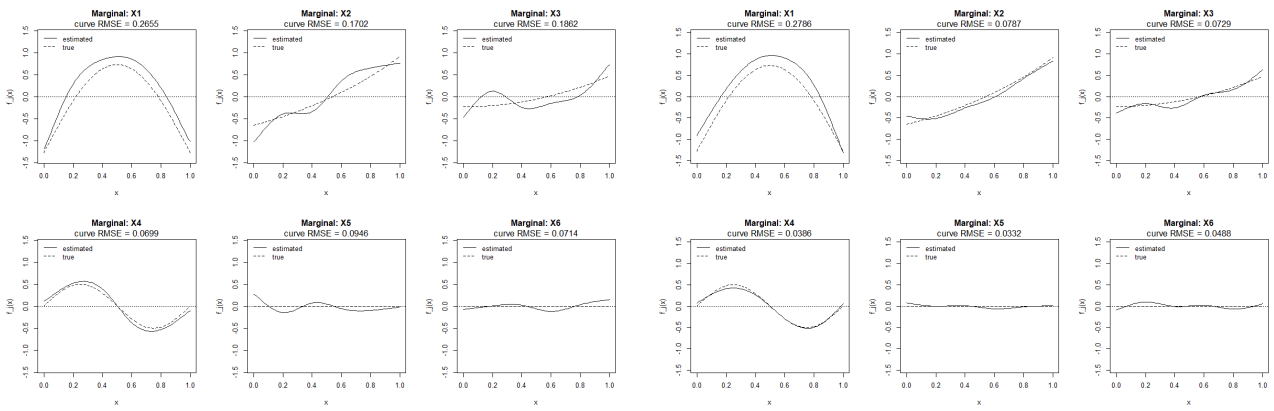
We then append two irrelevant covariates

$$X_{r5} \overset{\text{iid}}{\sim} \text{Unif}(0,1), \qquad X_{r6} \overset{\text{iid}}{\sim} \text{Unif}(0,1),$$

independent of $(X_1, \ldots, X_4)$, $b(\cdot)$, and ε_i. Importantly, the true mean depends only on $X_1, \ldots, X_4$, so $f_5 \equiv 0$ and $f_6 \equiv 0$ under the truth.

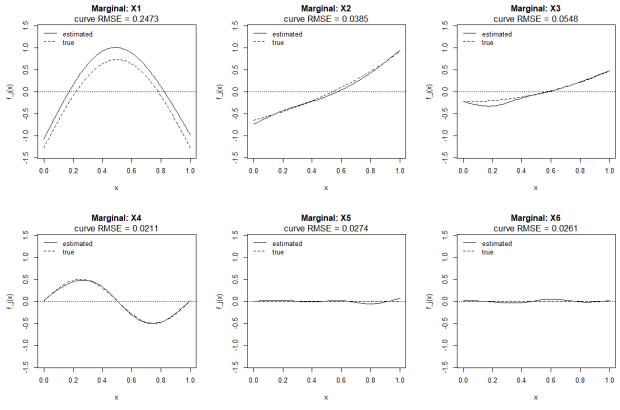
Fitted model. We fit the same additive LS-spline + Matérn residual model as in Sim3, but include all six covariates in the mean: $\eta(s) = \mu + \sum_{j=1}^{6} f_j(X_j(s))$, with the same identifiability constraint and REML profiling (with ν fixed).

Sim4: marginal curves across sample sizes

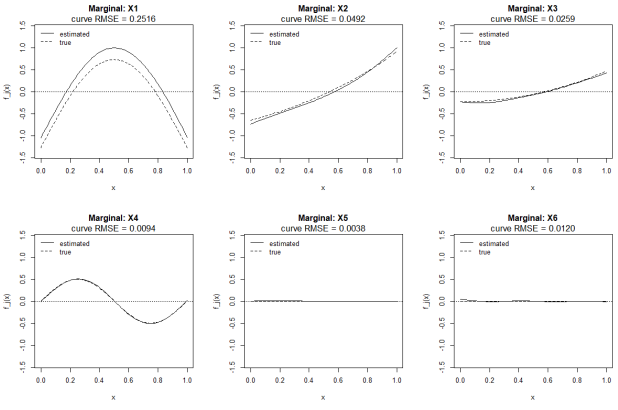


(a) $n = 100$

(b) $n = 400$



(c) $n = 1000$



(d) $n = 10{,}000$

Figure 4: Sim4 marginal component curves for $X_1, \ldots, X_6$ across sample sizes. Each panel overlays $\widehat{f}_j(x)$ with the truth-centered target $f_j^c(x)$ and reports grid RMSE. For the irrelevant covariates (X_5, X_6), the truth-centered target is identically zero.

Table 10: (Sim4) Grid-based marginal curve accuracy across sample sizes. $\mathrm{RMSE}_{\mathrm{curve}}$ and $\mathrm{cor}_{\mathrm{curve}}$ compare $\widehat{f}_j(x)$ to the truth-centered target $f_j^c(x)$ on a grid in $[0,1]$. For X_5 and X_6, $\mathrm{cor}_{\mathrm{curve}}$ is undefined because $f_j^c(x) \equiv 0$ has zero variance; we report "–".

	$n = 100$		$n = 400$		$n = 1000$		$n = 10{,}000$	
Var	RMSE	cor	RMSE	cor	RMSE	cor	RMSE	cor
X_1	0.2655	0.9943	0.2786	0.9922	0.2473	0.9991	0.2516	0.9998
X_2	0.1702	0.9624	0.0787	0.9904	0.0385	0.9977	0.0492	0.9988
X_3	0.1862	0.6885	0.0729	0.9716	0.0548	0.9920	0.0259	0.9985
X_4	0.0699	0.9963	0.0386	0.9962	0.0211	0.9983	0.0094	0.9997
X_5	0.0946	–	0.0332	–	0.0274	–	0.0038	–
X_6	0.0714	–	0.0488	–	0.0261	–	0.0120	–

Interpretation. Table 10 shows the expected negative-control behavior after adding two irrelevant covariates. The fitted marginal curves for the true signals (X_1–X_4) remain accurate and broadly comparable to Sim3, with $\mathrm{cor}_{\mathrm{curve}}$ close to one for moderate and large n. Meanwhile, for the irrelevant covariates X_5 and X_6 (true $f_5 \equiv 0$, $f_6 \equiv 0$), the grid RMSE values decrease with n (e.g., from 0.0946 and 0.0714 at $n = 100$ to 0.0038 and 0.0120 at $n = 10{,}000$), indicating that the fitted components shrink toward a flat zero function as sample size grows.

For X_5 and X_6, the truth-centered target is identically zero on the grid: $f_j^c(x) \equiv 0$. Therefore the vector $\{f_{jg}^c\}_{g=1}^G$ has zero variance, so the correlation $\mathrm{cor}(\{\widehat{f}_{jg}\}, \{f_{jg}^c\})$ is mathematically undefined (division by zero in the correlation formula). Hence we report "–" (NA) for $\mathrm{cor}_{\mathrm{curve}}$ while still reporting $\mathrm{RMSE}_{\mathrm{curve}}$, which remains well-defined and is the appropriate accuracy measure for a zero target.

3.6 Simulation 5: scaling to many covariates with irrelevant variables

Goal. Extend the negative-control design of Simulation 4 to a higher-dimensional covariate space: ten covariates with genuine nonlinear effects and ten irrelevant "garbage" covariates ($p = 20$ total). This tests whether the additive LS-spline mean estimator can (i) recover a richer set of nonlinear components, and (ii) correctly suppress the irrelevant ones, as the number of spline blocks in the design matrix grows.

Data generating model. The spatial structure matches Simulations 3–4. For $r = 1, \ldots, n$ at locations $s_r \in [0,1]^2$,

$$Y_r = \eta_r + b(s_r) + \varepsilon_r, \qquad \varepsilon_r \overset{\text{iid}}{\sim} N(0, \tau^2), \qquad b \sim N\big(\mathbf{0},\, \sigma^2 R(\rho, \nu)\big),$$

with the same spatial parameters as before ($\sigma^2 = 0.8,\ \rho = 0.2,\ \nu = 1.5,\ \tau^2 = 0.15$).

The true mean function is an additive combination of ten nonlinear components,

$$\eta_r = \mu + \sum_{j=1}^{10} f_j(X_{rj}), \qquad \mu = 0, \tag{8}$$

with the first four components identical to the earlier simulations and six new terms:

$$\begin{aligned}
f_1(x) &= 2\sin(\pi x), & f_2(x) &= 1.5\exp(x - 0.5),\\
f_3(x) &= 0.7\,x^2, & f_4(x) &= 0.5\sin(2\pi x),\\
f_5(x) &= (x - 0.5)^3, & f_6(x) &= 0.8\cos(\pi x),\\
f_7(x) &= 1.2\log(x + 0.1), & f_8(x) &= 0.6\,|x - 0.5|,\\
f_9(x) &= 0.9\sin(3\pi x), & f_{10}(x) &= 0.4\,(x^2 - x).
\end{aligned}$$

This set includes a variety of shapes: smooth periodic (f_1, f_4, f_6, f_9), monotone (f_2, f_7), polynomial (f_3, f_5, f_{10}), and non-smooth (f_8, a V-shape). The components span a wide range of signal amplitudes: f_1 has the largest range (≈ 4.0 on $[0,1]$) while f_{10} has the smallest (≈ 0.1).

Covariates $X_1, \ldots, X_{10}$ are generated via independent latent Matérn GP fields followed by the rank transform (as in Sim 3–4), with $(\rho_X, \nu_X) = (0.10, 1.0)$. Ten irrelevant covariates are appended:

$$X_{r,j} \overset{\text{iid}}{\sim} \text{Unif}(0,1), \qquad j = 11, \ldots, 20,$$

independent of everything else. The true mean depends only on $X_1, \ldots, X_{10}$, so $f_j \equiv 0$ for $j = 11, \ldots, 20$.

Fitted model. We fit the additive LS-spline + Matérn residual model with all 20 covariates: $\eta(s) = \mu + \sum_{j=1}^{20} f_j(X_j(s))$, with $M = 6$ knots per covariate (giving $20 \times 5 = 100$ identified spline coefficients plus the intercept). REML profiling and BLUP extraction follow the same procedure as before.

Results. Table 11 reports global recovery metrics for $n = 400$ and $n = 1000$. The spatial parameters are well estimated: $\hat{\rho}$ is close to the true value 0.2, and $\hat{\sigma}^2$ and $\hat{\tau}^2$ are reasonable, though with some attenuation in $\hat{\sigma}^2$ at the smaller sample size. The mean function correlation cor_η exceeds 0.98 at both sample sizes, indicating that the additive mean is well recovered despite having 20 covariates in the model.

Table 11: Sim5 global recovery metrics. True parameters: $\rho = 0.2$, $\sigma^2 = 0.8$, $\tau^2 = 0.15$.

n	$\hat{\rho}$	$\hat{\sigma}^2$	$\hat{\tau}^2$	RMSE_η	cor_η	cor_b	$\text{cor}(\hat{\eta}, b)$
400	0.187	0.636	0.140	0.261	0.983	0.677	0.038
1000	0.179	0.721	0.153	0.476	0.991	0.725	0.055

Table 12 reports the per-covariate marginal curve accuracy. Several patterns emerge.

First, the original four components (X_1–X_4) are recovered with accuracy comparable to Simulations 3 and 4, confirming that the additional covariates do not degrade estimation of the strong signals.

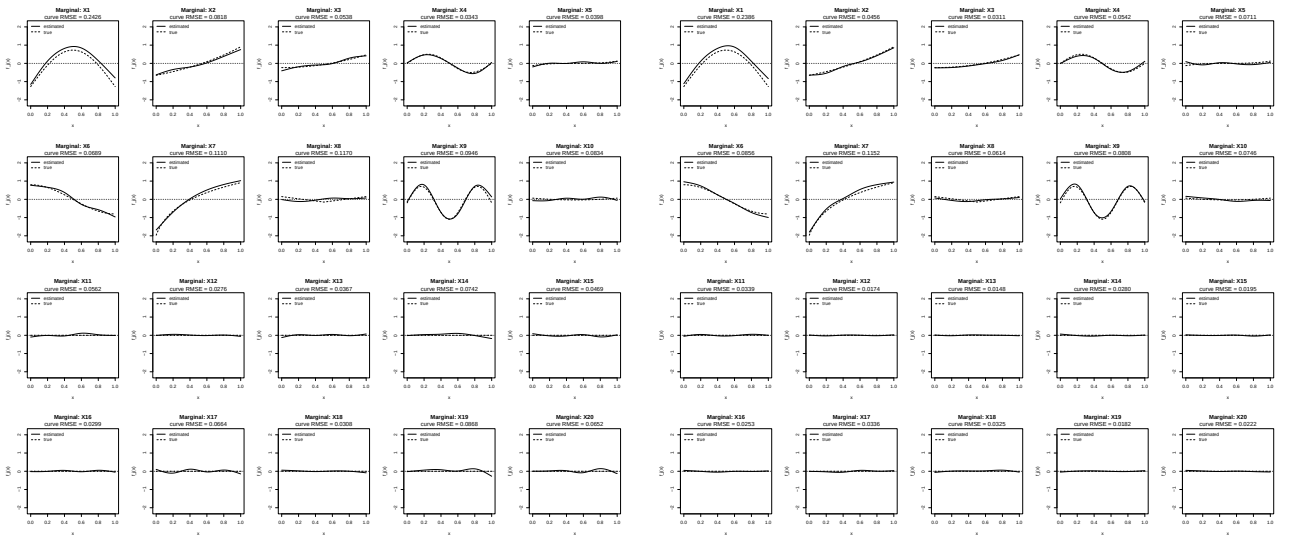
Second, among the six new real components, recovery quality varies with signal amplitude and shape complexity. Components with large amplitude and smooth shape—f_6 (cor = 0.999 at $n = 1000$), f_7 (cor = 0.999), f_9 (cor = 0.996)—are recovered well at both sample sizes. In contrast, $f_5 = (x-0.5)^3$ has a very small effective range (≈ 0.125 on $[0,1]$) and is poorly recovered (cor = 0.775 at n=400, dropping to -0.076 at n=1000). Similarly, $f_8 = 0.6\,|x-0.5|$ (a non-smooth V-shape with range ≈ 0.3) and $f_{10} = 0.4\,(x^2-x)$ (range ≈ 0.1) show weak or unstable correlations. These components have signal magnitudes comparable to or smaller than the residual noise, making them inherently difficult to separate from the spatial effect and the garbage covariates.

Third, the garbage covariates (X_{11}–X_{20}) show the expected negative-control behavior: their RMSE values are small and decrease with n (e.g., from a range of 0.028–0.087 at n=400 to 0.015–0.034 at n=1000), and correlation is undefined (reported as "–") because the true target is identically zero.

Sim5: marginal curves across sample sizes

Table 12: (Sim5) Grid-based marginal curve accuracy across sample sizes. $\mathrm{RMSE}_{\mathrm{curve}}$ and $\mathrm{cor}_{\mathrm{curve}}$ compare $\widehat{f}_j(x)$ to the truth-centered target $f_j^c(x)$ on a 101-point grid in $[0,1]$. For garbage covariates (X_{11}–X_{20}), $f_j^c(x) \equiv 0$ so $\mathrm{cor}_{\mathrm{curve}}$ is undefined; we report "–".

	$n = 400$		$n = 1000$	
Var	RMSE	cor	RMSE	cor
Real covariates (original four)				
X_1	0.2426	0.9944	0.2386	0.9965
X_2	0.0818	0.9966	0.0456	0.9968
X_3	0.0538	0.9767	0.0311	0.9980
X_4	0.0343	0.9960	0.0542	0.9919
Real covariates (new six)				
X_5	0.0398	0.775	0.0711	−0.076
X_6	0.0689	0.9938	0.0856	0.9986
X_7	0.1110	0.9965	0.1152	0.9986
X_8	0.1170	−0.108	0.0614	0.756
X_9	0.0946	0.9939	0.0808	0.9965
X_{10}	0.0834	−0.438	0.0746	0.478
Garbage covariates				
X_{11}	0.0562	–	0.0339	–
X_{12}	0.0276	–	0.0174	–
X_{13}	0.0367	–	0.0148	–
X_{14}	0.0742	–	0.0280	–
X_{15}	0.0469	–	0.0195	–
X_{16}	0.0299	–	0.0253	–
X_{17}	0.0664	–	0.0336	–
X_{18}	0.0308	–	0.0325	–
X_{19}	0.0869	–	0.0182	–
X_{20}	0.0652	–	0.0222	–



(a) $n = 400$ (b) $n = 1000$

Figure 5: Sim5 marginal component curves for $X_1, \ldots, X_{20}$ at two sample sizes. Each panel overlays $\widehat{f}_j(x)$ (solid) with the truth-centered target $f_j^c(x)$ (dashed) and reports grid RMSE. For the real covariates (X_1–X_{10}), the dashed curves show the true nonlinear component; for the garbage covariates (X_{11}–X_{20}), the target is identically zero. Components with small effective signal amplitude (X_5, X_8, X_{10}) show visibly noisier fits, consistent with the low correlations in Table 12.

Interpretation. Simulation 5 demonstrates that the LS-spline additive model scales gracefully from $p = 6$ (Sim4) to $p = 20$ covariates. The key findings are:

1. **Strong signals are robust to model expansion.** The original four components (X_1–X_4) and the new components with substantial amplitude (X_6, X_7, X_9) are recovered with $\text{cor}_{\text{curve}} > 0.99$ at $n = 1000$, comparable to Sim3–4. Including additional covariates (real or garbage) does not materially degrade estimation of strong nonlinear effects.

2. **Weak signals are inherently hard to detect.** Components X_5, X_8, and X_{10} have effective signal amplitudes (ranges of approximately 0.125, 0.3, and 0.1 respectively) that are small relative to the residual noise scale ($\tau^2 = 0.15$, $\sigma^2 = 0.8$). Their marginal curves are poorly recovered even at $n = 1000$, with $\text{cor}_{\text{curve}}$ values that are low or negative. This is not a failure of the method per se, but reflects a fundamental signal-to-noise limitation: the fitted curves for these components are comparable in magnitude to the garbage covariates' spurious fits, making them statistically indistinguishable from noise.

3. **Garbage covariates are correctly suppressed.** The fitted marginals for X_{11}–X_{20} converge toward zero with increasing n, mirroring the Sim4 behavior. The RMSE values at $n = 1000$ (range 0.015–0.034) are comparable to or smaller than the weakest real signals (X_5, X_{10}), which foreshadows the variable selection challenge addressed in Section 3.8.

3.7 Variable selection via block Wald test

The results in Simulations 4–5 show that garbage covariates' fitted marginals shrink toward zero with increasing n, while weak real signals may remain comparable in magnitude to noise. This raises a natural question: given an automated test, how many irrelevant covariates can the model tolerate before real signals become undetectable?

We address this through the block Wald F-test, a standard tool from GLS inference applied to the identified spline coefficient blocks from our fitted model.

3.7.1 Method

Recall that the GLS estimator from Algorithm 1 yields $\hat{\beta} = (H^\top \hat{\Sigma}_0^{-1} H)^{-1} H^\top \hat{\Sigma}_0^{-1} y$, where $H = (1_n \;\; W)$ is the full design matrix and $\hat{\Sigma}_0 = R(\hat{\rho}, \nu) + \hat{\lambda}\, I_n$. The covariance of $\hat{\beta}$ is

$$\text{Cov}(\hat{\beta}) = \hat{\sigma}^2 \big(H^\top \hat{\Sigma}_0^{-1} H\big)^{-1}.$$

For covariate j, the identified LS-spline block has $q_j = M_j - 1$ coefficients $\hat{\beta}_j \in \mathbb{R}^{q_j}$. To test $H_0 \colon \beta_j = \mathbf{0}$ (i.e., $f_j \equiv 0$), we form the block Wald statistic

$$W_j = \hat{\beta}_j^\top \big[\text{Cov}(\hat{\beta}_j)\big]^{-1} \hat{\beta}_j,$$

where $\text{Cov}(\hat{\beta}_j)$ is the $q_j \times q_j$ diagonal sub-block of $\text{Cov}(\hat{\beta})$ corresponding to covariate j. Under H_0 with Gaussian errors,

$$F_j = \frac{W_j}{q_j} \;\sim\; F(q_j,\, n - k), \tag{9}$$

where $k = 1 + \sum_{j=1}^{p} q_j$ is the total number of fixed-effect parameters (intercept plus all spline blocks). With $M = 6$ knots, $q_j = 5$ for all j and $k = 1 + 5p$.

For simultaneous testing across p covariates, we apply Benjamini–Hochberg (BH) correction [1] at level $\alpha = 0.05$ to control the false discovery rate. A covariate is declared "selected" (non-null) if its BH-adjusted p-value falls below α.

All quantities needed for this test—$\hat{\Sigma}_0$, H, $\hat{\sigma}^2$, and the coefficient vector $\hat{\beta}$—are already available from the fitted model, so the test requires only one additional matrix inversion beyond the existing GLS fit.

3.7.2 Three boundary experiments

We investigate three interacting factors that limit the ability to distinguish real from garbage covariates: algebraic feasibility of the GLS system, statistical detection power as the number of garbage covariates grows, and the role of signal amplitude in determining the detection boundary.

Experiment 1: matrix boundary. The GLS estimator requires $k < n$; with $M = 6$ knots this gives $p < (n-1)/5$. However, practical reliability degrades well before this algebraic limit because $\hat{\sigma}^2 = \text{RSS}/(n-k)$ becomes unstable when $\text{df}_{\text{resid}} = n - k$ is small.

We fix $p_{\text{real}} = 5$ strong covariates (f_1–f_4 as in the earlier simulations plus $f_5(x) = 0.8\cos(\pi x)$) and increase p_{garbage} toward the algebraic boundary for $n \in \{100, 200, 400\}$. Table 16 reports the results.

Table 13: Experiment 1: matrix boundary. True parameters: $\sigma^2 = 0.8$, $\tau^2 = 0.15$. "Power" is the fraction of the 5 real covariates selected (BH-adjusted $\alpha = 0.05$); "FPR" is the fraction of garbage covariates falsely selected.

n	p_{garb}	p	k	df	Status	Power	FPR	$\hat{\sigma}^2$	$\hat{\tau}^2$
100	1	6	31	69	OK	0.80	0.000	0.76	0.172
100	5	10	51	49	OK	0.60	0.000	0.84	0.169
100	8	13	66	34	OK	0.60	0.000	0.21	0.201
100	11	16	81	19	Low df	0.40	0.000	2.96	0.099
100	13	18	91	9	Low df	0.40	0.000	73.9	0.219
100	14	19	96	4	Low df	0.00	0.000	1.25	≈ 0
200	7	12	61	139	OK	1.00	0.000	0.67	0.199
200	15	20	101	99	OK	1.00	0.133	1.31	0.127
200	22	27	136	64	OK	0.80	0.182	0.77	0.077
200	28	33	166	34	OK	0.60	0.000	0.42	0.127
200	32	37	186	14	Low df	0.80	0.000	0.12	0.100
200	34	39	196	4	Low df	0.00	0.000	20.8	≈ 0
400	19	24	121	279	OK	1.00	0.000	2.01	0.163
400	35	40	201	199	OK	1.00	0.114	0.25	0.111
400	50	55	276	124	OK	0.80	0.000	0.78	0.206
400	62	67	336	64	OK	0.80	0.000	0.96	0.150
400	70	75	376	24	OK	0.00	0.000	2.00	0.233
400	74	79	396	4	Low df	0.00	0.000	0.20	≈ 0

Two clear patterns emerge. First, at $\text{df}_{\text{resid}} = 4$ across all three sample sizes, $\hat{\sigma}^2$ either explodes (73.9, 20.8) or $\hat{\tau}^2$ collapses to zero, and power drops to 0. The variance estimator $\hat{\sigma}^2 = \text{RSS}/(n-k)$ is dividing by a near-zero denominator, making all downstream inference unreliable. Second, the collapse occurs well before df = 0: at $n = 400$ with df = 24, power already drops to 0 despite the status being nominally OK. An intermediate regime is also visible at $n = 200$: when df is in the range 64–99, FPR rises to 0.13–0.18, indicating that garbage covariates are falsely selected when variance estimates are mildly destabilized.

A practical guideline is $\text{df}_{\text{resid}} \gtrsim 50$, i.e., $p < (n - 50)/(M-1)$, for the block Wald test to produce reliable results.

Experiment 2: statistical boundary (garbage sweep). We fix $p_{\text{real}} = 10$ with the Sim5 truth function (strong signals) and increase p_{garbage} from 10 to 160, well within the matrix-feasible region, for $n \in \{1000, 2000\}$. Table 17 reports the results.

The most striking feature is that X_5 and X_{10} are missed at virtually every setting, including $p_{\text{garbage}} = 10$. Recall from Table 12 that these components have effective signal ranges of approximately 0.125 and 0.1 on $[0, 1]$, respectively—comparable to or smaller than the noise floor. Their absence is therefore a signal-to-noise limitation, not an artifact of the garbage covariates.

Table 14: Experiment 2: statistical boundary. Ten real covariates (Sim5 strong signals), $n \in \{1000, 2000\}$, $\alpha = 0.05$ with BH correction. "Missed" lists real covariates not selected.

n	p_{garb}	p	k	df	Power	FPR	Missed
1000	10	20	101	899	0.70	0.000	X_5, X_8, X_{10}
1000	20	30	151	849	0.80	0.000	X_5, X_{10}
1000	40	50	251	749	0.80	0.000	X_5, X_{10}
1000	60	70	351	649	0.80	0.000	X_8, X_{10}
1000	80	90	451	549	0.80	0.000	X_5, X_{10}
1000	100	110	551	449	0.70	0.000	X_5, X_8, X_{10}
1000	130	140	701	299	0.80	0.000	X_5, X_{10}
1000	160	170	851	149	0.60	0.006	X_3, X_5, X_8, X_{10}
2000	10	20	101	1899	0.80	0.000	X_5, X_{10}
2000	20	30	151	1849	0.80	0.000	X_5, X_{10}
2000	40	50	251	1749	0.90	0.000	X_5
2000	60	70	351	1649	0.80	0.000	X_5, X_{10}
2000	80	90	451	1549	0.80	0.013	X_5, X_{10}
2000	100	110	551	1449	0.80	0.010	X_5, X_{10}
2000	130	140	701	1299	0.80	0.008	X_5, X_{10}
2000	160	170	851	1149	0.80	0.000	X_5, X_{10}

For the remaining strong signals, power is remarkably stable: it stays at 0.70–0.80 from $p_{\text{garbage}} = 10$ all the way to 130 at $n = 1000$. The statistical boundary, defined as the point where previously detectable signals begin to be missed, appears at $p_{\text{garbage}} = 160$ (df = 149), where X_3 ($f_3(x) = 0.7\,x^2$, a moderate signal) is lost and the first false positive appears.

At $n = 2000$, power is even more stable: it holds at 0.80 across the entire garbage sweep, and X_{10} is occasionally recovered (power reaches 0.90 at $p_{\text{garbage}} = 40$). More data extends the statistical boundary, consistent with the increased residual degrees of freedom.

Experiment 3: signal strength boundary. To isolate the effect of signal amplitude, we fix $n = 1000$ and $p_{\text{real}} = 10$, and compare two covariate sets with identical functional shapes but different amplitudes:

- **Strong signals**: the Sim5 truth function, with amplitudes ranging from 0.4 to 2.0 (Table 12).
- **Weak signals**: the same ten shapes with amplitudes scaled down by approximately 3×, ranging from 0.12 to 0.6.

We sweep p_{garbage} from 10 to 130 for each set. Table 18 reports the comparison.

Table 15: Experiment 3: signal strength boundary at $n = 1000$. "Strong" and "Weak" refer to the two signal amplitude sets described in the text.

	Strong		Weak		
p_{garb}	Power	FPR	Power	FPR	Missed (weak)
10	0.70	0.000	0.60	0.000	X_3, X_5, X_8, X_{10}
20	0.80	0.000	0.60	0.000	X_3, X_5, X_8, X_{10}
40	0.80	0.000	0.50	0.000	$X_3, X_4, X_5, X_8, X_{10}$
60	0.80	0.000	0.60	0.000	X_3, X_6, X_8, X_{10}
80	0.80	0.000	0.50	0.000	$X_3, X_5, X_6, X_8, X_{10}$
100	0.70	0.000	0.50	0.000	$X_3, X_4, X_5, X_8, X_{10}$
130	0.80	0.000	0.20	0.000	$X_1, X_2, X_3, X_4, X_5, X_6, X_8, X_{10}$

The results confirm that the garbage boundary depends strongly on signal amplitude. Strong signals maintain power at 0.70–0.80 across all garbage levels tested, with only the inherently weak components (X_5, X_8, X_{10}) being missed. In contrast, weak signals begin with power 0.60 at $p_{\text{garbage}} = 10$ and deteriorate steadily to power 0.20

at $p_{\text{garbage}} = 130$, where 8 of 10 components are missed—only X_7 ($0.4\log(x{+}0.1)$) and X_9 ($0.3\sin(3\pi x)$) survive, having the most detectable shapes among the weak set (a monotone function and a high-frequency oscillation, respectively).

Notably, FPR remains exactly zero throughout both signal sets. The BH-corrected Wald test is conservative: it never falsely selects garbage covariates, even when it fails to detect weak real signals. This means the test errs on the side of omission rather than false inclusion.

3.7.3 Summary of boundary analysis

The three experiments reveal three interacting constraints on variable selection in the additive LS-spline spatial model:

1. **Matrix boundary.** The GLS system requires $k = 1 + p(M{-}1) < n$. In practice, reliable variance estimation requires $n - k \gtrsim 50$; below this threshold, $\hat{\sigma}^2$ becomes unstable and the Wald test collapses. This gives a practical ceiling of $p < (n - 50)/(M{-}1)$.
2. **Statistical boundary.** Within the matrix-feasible region, strong signals (range $\gtrsim 0.5$ on $[0, 1]$) are robustly detected even when the number of garbage covariates greatly exceeds the number of real ones (e.g., 130 garbage vs. 10 real at $n = 1000$). The detection bottleneck is signal amplitude, not garbage count: weak components with effective ranges below ≈ 0.15 are missed regardless of how few garbage covariates are present.
3. **Signal strength boundary.** The garbage tolerance of the test depends on the weakest signal the analyst wishes to detect. For the strong signal set, power remains above 0.70 at $p_{\text{garbage}} = 130$; for the weak signal set (amplitudes reduced $3\times$), power drops to 0.20 at the same garbage level. The "boundary" is therefore not a single number but a function of the minimum detectable signal amplitude relative to the noise scale.

These findings are consistent with classical power analysis for the F-test: the noncentrality parameter scales with $\|\beta_j\|^2/\mathrm{Var}(\hat{\beta}_j)$, so weaker signals require either larger n or fewer nuisance parameters (smaller k) to maintain the same detection probability.

3.8 Variable selection via block Wald test

The results in Simulations 4–5 show that garbage covariates' fitted marginals shrink toward zero with increasing n, while weak real signals may remain comparable in magnitude to noise. This raises a natural question: given an automated test, how many irrelevant covariates can the model tolerate before real signals become undetectable?

We address this through the block Wald F-test, a standard tool from GLS inference applied to the identified spline coefficient blocks from our fitted model.

3.8.1 Method

Recall that the GLS estimator from Algorithm 1 yields $\hat{\beta} = (H^\top \hat{\Sigma}_0^{-1} H)^{-1} H^\top \hat{\Sigma}_0^{-1} y$, where $H = (1_n \;\; W)$ is the full design matrix and $\hat{\Sigma}_0 = R(\hat{\rho}, \nu) + \hat{\lambda}\, I_n$. The covariance of $\hat{\beta}$ is

$$\mathrm{Cov}(\hat{\beta}) = \hat{\sigma}^2 \big(H^\top \hat{\Sigma}_0^{-1} H\big)^{-1}.$$

For covariate j, the identified LS-spline block has $q_j = M_j - 1$ coefficients $\hat{\beta}_j \in \mathbb{R}^{q_j}$. To test $H_0 \colon \beta_j = \mathbf{0}$ (i.e., $f_j \equiv 0$), we form the block Wald statistic

$$W_j = \hat{\beta}_j^\top \big[\mathrm{Cov}(\hat{\beta}_j)\big]^{-1} \hat{\beta}_j,$$

where $\mathrm{Cov}(\hat{\beta}_j)$ is the $q_j \times q_j$ diagonal sub-block of $\mathrm{Cov}(\hat{\beta})$ corresponding to covariate j. Under H_0 with Gaussian errors,

$$F_j = \frac{W_j}{q_j} \;\sim\; F(q_j,\, n-k), \tag{10}$$

where $k = 1 + \sum_{j=1}^{p} q_j$ is the total number of fixed-effect parameters (intercept plus all spline blocks). With $M = 6$ knots, $q_j = 5$ for all j and $k = 1 + 5p$.

For simultaneous testing across p covariates, we apply Benjamini–Hochberg (BH) correction [1] at level $\alpha = 0.05$ to control the false discovery rate. A covariate is declared "selected" (non-null) if its BH-adjusted p-value falls below α.

All quantities needed for this test—$\hat{\Sigma}_0$, H, $\hat{\sigma}^2$, and the coefficient vector $\hat{\beta}$—are already available from the fitted model, so the test requires only one additional matrix inversion beyond the existing GLS fit.

3.8.2 Three boundary experiments

We investigate three interacting factors that limit the ability to distinguish real from garbage covariates: algebraic feasibility of the GLS system, statistical detection power as the number of garbage covariates grows, and the role of signal amplitude in determining the detection boundary.

Experiment 1: matrix boundary. The GLS estimator requires $k < n$; with $M = 6$ knots this gives $p < (n-1)/5$. However, practical reliability degrades well before this algebraic limit because $\hat{\sigma}^2 = \mathrm{RSS}/(n-k)$ becomes unstable when $\mathrm{df}_{\mathrm{resid}} = n - k$ is small.

We fix $p_{\mathrm{real}} = 5$ strong covariates (f_1–f_4 as in the earlier simulations plus $f_5(x) = 0.8\cos(\pi x)$) and increase p_{garbage} toward the algebraic boundary for $n \in \{100, 200, 400\}$. Table 16 reports the results.

Table 16: Experiment 1: matrix boundary. True parameters: $\sigma^2 = 0.8$, $\tau^2 = 0.15$. "Power" is the fraction of the 5 real covariates selected (BH-adjusted $\alpha = 0.05$); "FPR" is the fraction of garbage covariates falsely selected.

n	p_{garb}	p	k	df	Status	Power	FPR	$\hat{\sigma}^2$	$\hat{\tau}^2$
100	1	6	31	69	OK	0.80	0.000	0.76	0.172
100	5	10	51	49	OK	0.60	0.000	0.84	0.169
100	8	13	66	34	OK	0.60	0.000	0.21	0.201
100	11	16	81	19	Low df	0.40	0.000	2.96	0.099
100	13	18	91	9	Low df	0.40	0.000	73.9	0.219
100	14	19	96	4	Low df	0.00	0.000	1.25	≈ 0
200	7	12	61	139	OK	1.00	0.000	0.67	0.199
200	15	20	101	99	OK	1.00	0.133	1.31	0.127
200	22	27	136	64	OK	0.80	0.182	0.77	0.077
200	28	33	166	34	OK	0.60	0.000	0.42	0.127
200	32	37	186	14	Low df	0.80	0.000	0.12	0.100
200	34	39	196	4	Low df	0.00	0.000	20.8	≈ 0
400	19	24	121	279	OK	1.00	0.000	2.01	0.163
400	35	40	201	199	OK	1.00	0.114	0.25	0.111
400	50	55	276	124	OK	0.80	0.000	0.78	0.206
400	62	67	336	64	OK	0.80	0.000	0.96	0.150
400	70	75	376	24	OK	0.00	0.000	2.00	0.233
400	74	79	396	4	Low df	0.00	0.000	0.20	≈ 0

Two clear patterns emerge. First, at $\mathrm{df}_{\mathrm{resid}} = 4$ across all three sample sizes, $\hat{\sigma}^2$ either explodes (73.9, 20.8) or $\hat{\tau}^2$ collapses to zero, and power drops to 0. The variance estimator $\hat{\sigma}^2 = \mathrm{RSS}/(n-k)$ is dividing by a near-zero denominator, making all downstream inference unreliable. Second, the collapse occurs well before $\mathrm{df} = 0$: at $n = 400$ with $\mathrm{df} = 24$, power already drops to 0 despite the status being nominally OK. An intermediate regime

is also visible at $n = 200$: when df is in the range 64–99, FPR rises to 0.13–0.18, indicating that garbage covariates are falsely selected when variance estimates are mildly destabilized.

A practical guideline is $\mathrm{df}_{\mathrm{resid}} \gtrsim 50$, i.e., $p < (n-50)/(M-1)$, for the block Wald test to produce reliable results.

Experiment 2: statistical boundary (garbage sweep). We fix $p_{\mathrm{real}} = 10$ with the Sim5 truth function (strong signals) and increase p_{garbage} from 10 to 160, well within the matrix-feasible region, for $n \in \{1000, 2000\}$. Table 17 reports the results.

Table 17: Experiment 2: statistical boundary. Ten real covariates (Sim5 strong signals), $n \in \{1000, 2000\}$, $\alpha = 0.05$ with BH correction. "Missed" lists real covariates not selected.

n	p_{garb}	p	k	df	Power	FPR	Missed
1000	10	20	101	899	0.70	0.000	X_5, X_8, X_{10}
1000	20	30	151	849	0.80	0.000	X_5, X_{10}
1000	40	50	251	749	0.80	0.000	X_5, X_{10}
1000	60	70	351	649	0.80	0.000	X_8, X_{10}
1000	80	90	451	549	0.80	0.000	X_5, X_{10}
1000	100	110	551	449	0.70	0.000	X_5, X_8, X_{10}
1000	130	140	701	299	0.80	0.000	X_5, X_{10}
1000	160	170	851	149	0.60	0.006	X_3, X_5, X_8, X_{10}
2000	10	20	101	1899	0.80	0.000	X_5, X_{10}
2000	20	30	151	1849	0.80	0.000	X_5, X_{10}
2000	40	50	251	1749	0.90	0.000	X_5
2000	60	70	351	1649	0.80	0.000	X_5, X_{10}
2000	80	90	451	1549	0.80	0.013	X_5, X_{10}
2000	100	110	551	1449	0.80	0.010	X_5, X_{10}
2000	130	140	701	1299	0.80	0.008	X_5, X_{10}
2000	160	170	851	1149	0.80	0.000	X_5, X_{10}

The most striking feature is that X_5 and X_{10} are missed at virtually every setting, including $p_{\mathrm{garbage}} = 10$. Recall from Table 12 that these components have effective signal ranges of approximately 0.125 and 0.1 on $[0, 1]$, respectively—comparable to or smaller than the noise floor. Their absence is therefore a signal-to-noise limitation, not an artifact of the garbage covariates.

For the remaining strong signals, power is remarkably stable: it stays at 0.70–0.80 from $p_{\mathrm{garbage}} = 10$ all the way to 130 at $n = 1000$. The statistical boundary, defined as the point where previously detectable signals begin to be missed, appears at $p_{\mathrm{garbage}} = 160$ (df = 149), where X_3 ($f_3(x) = 0.7\,x^2$, a moderate signal) is lost and the first false positive appears.

At $n = 2000$, power is even more stable: it holds at 0.80 across the entire garbage sweep, and X_{10} is occasionally recovered (power reaches 0.90 at $p_{\mathrm{garbage}} = 40$). More data extends the statistical boundary, consistent with the increased residual degrees of freedom.

Experiment 3: signal strength boundary. To isolate the effect of signal amplitude, we fix $n = 1000$ and $p_{\mathrm{real}} = 10$, and compare two covariate sets with identical functional shapes but different amplitudes:

- **Strong signals**: the Sim5 truth function, with amplitudes ranging from 0.4 to 2.0 (Table 12).
- **Weak signals**: the same ten shapes with amplitudes scaled down by approximately 3×, ranging from 0.12 to 0.6.

We sweep p_{garbage} from 10 to 130 for each set. Table 18 reports the comparison.

The results confirm that the garbage boundary depends strongly on signal amplitude. Strong signals maintain power at 0.70–0.80 across all garbage levels tested, with only the inherently weak components (X_5, X_8, X_{10}) being

Table 18: Experiment 3: signal strength boundary at $n = 1000$. "Strong" and "Weak" refer to the two signal amplitude sets described in the text.

	Strong		Weak		
p_{garb}	Power	FPR	Power	FPR	Missed (weak)
10	0.70	0.000	0.60	0.000	X_3, X_5, X_8, X_{10}
20	0.80	0.000	0.60	0.000	X_3, X_5, X_8, X_{10}
40	0.80	0.000	0.50	0.000	$X_3, X_4, X_5, X_8, X_{10}$
60	0.80	0.000	0.60	0.000	X_3, X_6, X_8, X_{10}
80	0.80	0.000	0.50	0.000	$X_3, X_5, X_6, X_8, X_{10}$
100	0.70	0.000	0.50	0.000	$X_3, X_4, X_5, X_8, X_{10}$
130	0.80	0.000	0.20	0.000	$X_1, X_2, X_3, X_4, X_5, X_6, X_8, X_{10}$

missed. In contrast, weak signals begin with power 0.60 at $p_{\text{garbage}} = 10$ and deteriorate steadily to power 0.20 at $p_{\text{garbage}} = 130$, where 8 of 10 components are missed—only X_7 ($0.4\log(x{+}0.1)$) and X_9 ($0.3\sin(3\pi x)$) survive, having the most detectable shapes among the weak set (a monotone function and a high-frequency oscillation, respectively).

Notably, FPR remains exactly zero throughout both signal sets. The BH-corrected Wald test is conservative: it never falsely selects garbage covariates, even when it fails to detect weak real signals. This means the test errs on the side of omission rather than false inclusion.

3.8.3 Summary of boundary analysis

The three experiments reveal three interacting constraints on variable selection in the additive LS-spline spatial model:

1. **Matrix boundary.** The GLS system requires $k = 1 + p(M{-}1) < n$. In practice, reliable variance estimation requires $n - k \gtrsim 50$; below this threshold, $\hat{\sigma}^2$ becomes unstable and the Wald test collapses. This gives a practical ceiling of $p < (n - 50)/(M{-}1)$.

2. **Statistical boundary.** Within the matrix-feasible region, strong signals (range $\gtrsim 0.5$ on $[0, 1]$) are robustly detected even when the number of garbage covariates greatly exceeds the number of real ones (e.g., 130 garbage vs. 10 real at $n = 1000$). The detection bottleneck is signal amplitude, not garbage count: weak components with effective ranges below ≈ 0.15 are missed regardless of how few garbage covariates are present.

3. **Signal strength boundary.** The garbage tolerance of the test depends on the weakest signal the analyst wishes to detect. For the strong signal set, power remains above 0.70 at $p_{\text{garbage}} = 130$; for the weak signal set (amplitudes reduced $3\times$), power drops to 0.20 at the same garbage level. The "boundary" is therefore not a single number but a function of the minimum detectable signal amplitude relative to the noise scale.

These findings are consistent with classical power analysis for the F-test: the noncentrality parameter scales with $\|\beta_j\|^2/\text{Var}(\hat{\beta}_j)$, so weaker signals require either larger n or fewer nuisance parameters (smaller k) to maintain the same detection probability.

4 LS-basis details

This appendix presents the Lancaster–Šalkauskas (LS) natural cubic spline construction used for each unknown function $f_j(\cdot)$ in the additive model.

4.1 Knots and indexing

Fix a covariate index $j \in \{1, \dots, p\}$.

Let

$$\tau_{1j} < \tau_{2j} < \cdots < \tau_{M_j j}$$

denote the ordered knot locations for covariate j, where $M_j \geq 4$.

Define knot spacings

$$h_{mj} = \tau_{mj} - \tau_{m-1,j}, \qquad m = 2, \dots, M_j.$$

For any scalar input $x \in \mathbb{R}$, the LS basis consists of two collections of compactly supported cubic spline basis functions,

$$\{\Phi_{mj}(x)\}_{m=1}^{M_j} \quad \text{and} \quad \{\Psi_{mj}(x)\}_{m=1}^{M_j}.$$

4.2 LS basis functions Φ_{mj} and Ψ_{mj}

For interior indices $m = 2, \dots, M_j - 1$, define

$$\Phi_{mj}(x) = \begin{cases} 0, & x < \tau_{m-1,j}, \\ -\dfrac{2}{h_{mj}^3}\,(x - \tau_{m-1,j})^2\,(x - \tau_{mj} - 0.5h_{mj}), & \tau_{m-1,j} \leq x < \tau_{mj}, \\ \dfrac{2}{h_{m+1,j}^3}\,(x - \tau_{m+1,j})^2\,(x - \tau_{mj} + 0.5h_{m+1,j}), & \tau_{mj} \leq x < \tau_{m+1,j}, \\ 0, & x \geq \tau_{m+1,j}, \end{cases} \tag{11}$$

$$\Psi_{mj}(x) = \begin{cases} 0, & x < \tau_{m-1,j}, \\ \dfrac{1}{h_{mj}^2}\,(x - \tau_{m-1,j})^2\,(x - \tau_{mj}), & \tau_{m-1,j} \leq x < \tau_{mj}, \\ \dfrac{1}{h_{m+1,j}^2}\,(x - \tau_{m+1,j})^2\,(x - \tau_{mj}), & \tau_{mj} \leq x < \tau_{m+1,j}, \\ 0, & x \geq \tau_{m+1,j}. \end{cases} \tag{12}$$

Boundary basis functions use only one adjacent interval:

$$\Phi_{1j}(x) = \begin{cases} \dfrac{2}{h_{2j}^3}\,(x - \tau_{2j})^2\,(x - \tau_{1j} + 0.5h_{2j}), & \tau_{1j} \leq x < \tau_{2j}, \\ 0, & \text{otherwise}, \end{cases}$$

$$\Phi_{M_j j}(x) = \begin{cases} -\dfrac{2}{h_{M_j j}^3}\,(x - \tau_{M_j-1,j})^2\,(x - \tau_{M_j j} - 0.5h_{M_j j}), & \tau_{M_j-1,j} \leq x < \tau_{M_j j}, \\ 0, & \text{otherwise}, \end{cases}$$

$$\Psi_{1j}(x) = \begin{cases} \dfrac{1}{h_{2j}^2}\,(x - \tau_{2j})^2\,(x - \tau_{1j}), & \tau_{1j} \leq x < \tau_{2j}, \\ 0, & \text{otherwise}, \end{cases}$$

$$\Psi_{M_j j}(x) = \begin{cases} \dfrac{1}{h_{M_j j}^2}\,(x - \tau_{M_j-1,j})^2\,(x - \tau_{M_j j}), & \tau_{M_j-1,j} \leq x < \tau_{M_j j}, \\ 0, & \text{otherwise}. \end{cases}$$

Endpoint convention. Following the note in Lancaster–Šalkauskas that the strict inequality at the upper limit should be replaced by a weak inequality, we include the right endpoint at the final knot. In implementation, we treat the last interval as $[\tau_{M_j-1,j},\, \tau_{M_jj}]$ (i.e., $x = \tau_{M_jj}$ is included), while using left-closed/right-open intervals elsewhere to avoid ambiguity.

Let

$$\boldsymbol{\theta}_j = (\theta_{1j}, \ldots, \theta_{M_jj})^\top, \qquad \boldsymbol{\gamma}_j = (\gamma_{1j}, \ldots, \gamma_{M_jj})^\top.$$

The LS spline representation of $f_j(\cdot)$ is

$$f_j(x) = \sum_{m=1}^{M_j} \Big[\Phi_{mj}(x)\, \theta_{mj} + \Psi_{mj}(x)\, \gamma_{mj} \Big]. \tag{13}$$

This construction yields the knot interpretation

$$f_j(\tau_{mj}) = \theta_{mj}, \qquad f_j'(\tau_{mj}) = \gamma_{mj}, \qquad m = 1, \ldots, M_j.$$

For interior indices $m = 2, \ldots, M_j - 1$, the derivatives are quadratic splines given by

$$\Phi'_{mj}(x) = \begin{cases} 0, & x < \tau_{m-1,j}, \\ -\dfrac{2}{h_{mj}^3}\Big[(x-\tau_{m-1,j})^2 + 2(x-\tau_{mj}-0.5h_{mj})(x-\tau_{m-1,j})\Big], & \tau_{m-1,j} \le x < \tau_{mj}, \\ \dfrac{2}{h_{m+1,j}^3}\Big[(x-\tau_{m+1,j})^2 + 2(x-\tau_{mj}+0.5h_{m+1,j})(x-\tau_{m+1,j})\Big], & \tau_{mj} \le x < \tau_{m+1,j}, \\ 0, & x \ge \tau_{m+1,j}, \end{cases} \tag{14}$$

$$\Psi'_{mj}(x) = \begin{cases} 0, & x < \tau_{m-1,j}, \\ \dfrac{1}{h_{mj}^2}\Big[(x-\tau_{m-1,j})^2 + 2(x-\tau_{m-1,j})(x-\tau_{mj})\Big], & \tau_{m-1,j} \le x < \tau_{mj}, \\ \dfrac{1}{h_{m+1,j}^2}\Big[(x-\tau_{m+1,j})^2 + 2(x-\tau_{mj})(x-\tau_{m+1,j})\Big], & \tau_{mj} \le x < \tau_{m+1,j}, \\ 0, & x \ge \tau_{m+1,j}. \end{cases} \tag{15}$$

For the boundary indices $m = 1$ and $m = M_j$, the derivatives are obtained by restricting (14)–(15) to the support of the boundary basis functions. We also record explicit formulas for the boundary derivative basis functions.

Let $h_{2j} = \tau_{2j} - \tau_{1j}$ and $h_{M_j j} = \tau_{M_j j} - \tau_{M_j-1,j}$. we have,

$$\Phi'_{1j}(x) = \begin{cases} \dfrac{2}{h_{2j}^3}\Big[(x-\tau_{2j})^2 + 2(x-\tau_{1j}+0.5h_{2j})(x-\tau_{2j})\Big], & \tau_{1j} \le x < \tau_{2j}, \\ 0, & \text{otherwise}, \end{cases}$$

$$\Psi'_{1j}(x) = \begin{cases} \dfrac{1}{h_{2j}^2}\Big[(x-\tau_{2j})^2 + 2(x-\tau_{1j})(x-\tau_{2j})\Big], & \tau_{1j} \le x < \tau_{2j}, \\ 0, & \text{otherwise}. \end{cases}$$

$$\Phi'_{M_j j}(x) = \begin{cases} 0, & x < \tau_{M_j-1,j}, \\ -\dfrac{2}{h_{M_j j}^3}\Big[(x-\tau_{M_j-1,j})^2 + 2(x-\tau_{M_j j}-0.5h_{M_j j})(x-\tau_{M_j-1,j})\Big], & \tau_{M_j-1,j} \le x < \tau_{M_j j}, \\ 0, & x \ge \tau_{M_j j}, \end{cases}$$

$$\Psi'_{M_j j}(x) = \begin{cases} 0, & x < \tau_{M_j-1,j}, \\ \dfrac{1}{h_{M_j j}^2}\Big[(x-\tau_{M_j-1,j})^2 + 2(x-\tau_{M_j-1,j})(x-\tau_{M_j j})\Big], & \tau_{M_j-1,j} \le x < \tau_{M_j j}, \\ 0, & x \ge \tau_{M_j j}. \end{cases}$$

Evaluating the representation (13) at $x = \tau_{mj}$ and using the cardinal properties of the LS basis yields

$$\Phi_{mj}(\tau_{mj}) = 1, \qquad \Phi_{\ell j}(\tau_{mj}) = 0 \ (\ell \neq m), \qquad \Psi_{\ell j}(\tau_{mj}) = 0 \ (\forall \ell),$$

which implies $f_j(\tau_{mj}) = \theta_{mj}$.

Similarly, evaluating derivatives at $x = \tau_{mj}$ gives

$$\Psi'_{mj}(\tau_{mj}) = 1, \qquad \Psi'_{\ell j}(\tau_{mj}) = 0 \ (\ell \neq m), \qquad \Phi'_{\ell j}(\tau_{mj}) = 0 \ (\forall \ell),$$

which confirms $f'_j(\tau_{mj}) = \gamma_{mj}$.

4.3 Eliminating slope parameters via natural-spline constraints

For $m = 2, \ldots, M_j - 1$, define

$$\omega_{mj} = \frac{h_{mj}}{h_{mj} + h_{m+1,j}}, \qquad \bar{\omega}_{mj} = 1 - \omega_{mj}.$$

Then $\mathbf{A}_j \in \mathbb{R}^{M_j \times M_j}$ is tridiagonal with

$$(\mathbf{A}_j)_{11} = 2, \ (\mathbf{A}_j)_{12} = 1; \qquad (\mathbf{A}_j)_{M_j, M_j-1} = 1, \ (\mathbf{A}_j)_{M_j, M_j} = 2;$$

and for each interior row $m = 2, \ldots, M_j - 1$,

$$(\mathbf{A}_j)_{m,m-1} = \omega_{mj}, \quad (\mathbf{A}_j)_{m,m} = 2, \quad (\mathbf{A}_j)_{m,m+1} = \bar{\omega}_{mj},$$

with all other entries zero. Equivalently,

$$\mathbf{A}_j = \begin{pmatrix} 2 & 1 & 0 & 0 & \cdots & 0 \\ \omega_{2j} & 2 & \bar{\omega}_{2j} & 0 & \cdots & 0 \\ 0 & \omega_{3j} & 2 & \bar{\omega}_{3j} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \omega_{M_j-1,j} & 2 & \bar{\omega}_{M_j-1,j} \\ 0 & \cdots & 0 & 0 & 1 & 2 \end{pmatrix}.$$

The matrix $\mathbf{C}_j \in \mathbb{R}^{M_j \times M_j}$ is sparse and depends on $\{h_{mj}\}$ and $\{\omega_{mj}, \bar{\omega}_{mj}\}$. In the LS construction (Lancaster–Šalkauskas, Sec. 4.2),

$$\mathbf{C}_j = 3 \begin{pmatrix} -\dfrac{1}{h_{2j}} & \dfrac{1}{h_{2j}} & 0 & 0 & \cdots & 0 & 0 \\ -\dfrac{\omega_{2j}}{h_{2j}} & \dfrac{\omega_{2j}}{h_{2j}} - \dfrac{\bar{\omega}_{2j}}{h_{3j}} & \dfrac{\bar{\omega}_{2j}}{h_{3j}} & 0 & \cdots & 0 & 0 \\ 0 & -\dfrac{\omega_{3j}}{h_{3j}} & \dfrac{\omega_{3j}}{h_{3j}} - \dfrac{\bar{\omega}_{3j}}{h_{4j}} & \dfrac{\bar{\omega}_{3j}}{h_{4j}} & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 & 0 \\ 0 & \cdots & 0 & -\dfrac{\omega_{M_j-1,j}}{h_{M_j-1,j}} & \dfrac{\omega_{M_j-1,j}}{h_{M_j-1,j}} - \dfrac{\bar{\omega}_{M_j-1,j}}{h_{M_jj}} & \dfrac{\bar{\omega}_{M_j-1,j}}{h_{M_jj}} & 0 \\ 0 & \cdots & 0 & 0 & 0 & -\dfrac{1}{h_{M_jj}} & \dfrac{1}{h_{M_jj}} \end{pmatrix}.$$

Because f_j is a natural cubic spline, it satisfies

$$f_j''(\tau_{1j}) = 0, \qquad f_j''(\tau_{M_jj}) = 0,$$

and $f_j''(\cdot)$ is continuous at interior knots. These constraints imply a linear relation

$$\mathbf{C}_j\,\boldsymbol{\theta}_j = \mathbf{A}_j\,\boldsymbol{\gamma}_j, \qquad \text{so that} \qquad \boldsymbol{\gamma}_j = \mathbf{A}_j^{-1}\mathbf{C}_j\,\boldsymbol{\theta}_j. \tag{16}$$

4.4 Reduced form and LS design vectors

Define

$$\boldsymbol{\Phi}_j(x) = \big(\Phi_{1j}(x), \ldots, \Phi_{M_jj}(x)\big)^\top, \qquad \boldsymbol{\Psi}_j(x) = \big(\Psi_{1j}(x), \ldots, \Psi_{M_jj}(x)\big)^\top.$$

Substituting $\boldsymbol{\gamma}_j = \mathbf{A}_j^{-1}\mathbf{C}_j\boldsymbol{\theta}_j$ into (13) gives

$$\begin{aligned} f_j(x) &= \boldsymbol{\Phi}_j(x)^\top\boldsymbol{\theta}_j + \boldsymbol{\Psi}_j(x)^\top\boldsymbol{\gamma}_j \\ &= \Big(\boldsymbol{\Phi}_j(x)^\top + \boldsymbol{\Psi}_j(x)^\top\mathbf{A}_j^{-1}\mathbf{C}_j\Big)\boldsymbol{\theta}_j \;=\; \boldsymbol{z}_j(x)^\top\boldsymbol{\theta}_j, \end{aligned} \tag{17}$$

where

$$\boldsymbol{z}_j(x)^\top = \boldsymbol{\Phi}_j(x)^\top + \boldsymbol{\Psi}_j(x)^\top\mathbf{A}_j^{-1}\mathbf{C}_j, \qquad \boldsymbol{z}_j(x) \in \mathbb{R}^{M_j}.$$

4.5 Design matrix for observed inputs

For covariate j, suppose we observe inputs $x_{1j}, \ldots, x_{nj} \in \mathbb{R}$. Define

$$\boldsymbol{z}_{rj} \equiv \boldsymbol{z}_j(x_{rj}) \in \mathbb{R}^{M_j}, \qquad r = 1, \ldots, n.$$

Stacking $\boldsymbol{z}_{rj}^\top$ yields the LS design matrix

$$\mathbf{Z}_j = \begin{pmatrix} \boldsymbol{z}_{1j}^\top \\ \vdots \\ \boldsymbol{z}_{nj}^\top \end{pmatrix} \in \mathbb{R}^{n \times M_j}.$$

Let $\boldsymbol{f}_j = (f_j(x_{1j}), \ldots, f_j(x_{nj}))^\top$. Then

$$\boldsymbol{f}_j = \mathbf{Z}_j \, \boldsymbol{\theta}_j.$$

5 Spatial covariance details

This appendix records the Matérn correlation function used for the spatial effect $b(\cdot)$ and the construction of the correlation matrix $\mathbf{R}$ for the observed locations $\{s_r\}_{r=1}^n$.

5.1 Matérn correlation function

Let $d = \|s - s'\| \geq 0$ denote Euclidean distance between locations $s, s' \in \mathbb{R}^2$. We use the Matérn correlation function parameterized by range $\rho > 0$ and smoothness $\nu > 0$:

$$K_\nu\left(\frac{d}{\rho}\right) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{d}{\rho}\right)^\nu \mathcal{K}_\nu\left(\frac{d}{\rho}\right), \tag{18}$$

where $\Gamma(\cdot)$ is the Gamma function and $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of the second kind. The Matérn covariance function is then

$$\mathrm{Cov}\{b(s), b(s')\} = \sigma^2 \, K_\nu\left(\frac{\|s - s'\|}{\rho}\right).$$

At $d = 0$, the Matérn correlation satisfies $K_\nu(0) = 1$ by continuity, so the diagonal entries of the correlation matrix are set to one.

5.2 Distance matrix and correlation matrix construction

Given observed locations $s_r = (s_{r1}, s_{r2})^\top \in \mathbb{R}^2$, $r = 1, \ldots, n$, define the pairwise distance matrix $\mathbf{D} \in \mathbb{R}^{n \times n}$ by

$$D_{rs} = \|s_r - s_s\| = \sqrt{(s_{r1} - s_{s1})^2 + (s_{r2} - s_{s2})^2}.$$

The Matérn correlation matrix $\mathbf{R} \in \mathbb{R}^{n \times n}$ is computed entrywise as

$$R_{rs} = K_\nu\left(\frac{D_{rs}}{\rho}\right), \qquad r, s = 1, \ldots, n,$$

and we set $R_{rr} = 1$. In computation, it is common to add a small jitter term for numerical stability when forming covariance matrices. Accordingly, the covariance matrix of $\boldsymbol{b}$ may be implemented as

$$\sigma^2 \mathbf{R} + \delta \mathbf{I}_n,$$

with a very small $\delta > 0$ (for example, $\delta = 10^{-8}$).

5.3 Marginal covariance of the observation vector

Stacking observations yields

$$\mathbf{y} = \mu \mathbf{1} + \mathbf{f} + \mathbf{b} + \boldsymbol{\varepsilon},$$

where $\mathbf{f} = (f(x_1), \ldots, f(x_n))^\top$ denotes the mean component, $\mathbf{b} \sim N(\mathbf{0}, \sigma^2\mathbf{R})$, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \tau^2\mathbf{I}_n)$ is nugget noise, independent of $\mathbf{b}$. Conditioning on $(\mu, \mathbf{f})$ treats $\mu\mathbf{1} + \mathbf{f}$ as fixed, so

$$\mathrm{Var}(\mathbf{y} \mid \mu, \mathbf{f}) = \mathrm{Var}(\mathbf{b} + \boldsymbol{\varepsilon}) = \mathrm{Var}(\mathbf{b}) + \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{R} + \tau^2\mathbf{I}_n.$$

Equivalently, for $r \neq s$, $\mathrm{Cov}(y_r, y_s) = \sigma^2 R_{rs}$, while $\mathrm{Var}(y_r) = \sigma^2 + \tau^2$ since $R_{rr} = 1$.

Full R Code (plug-and-play)

```r
# ============================
# Spatial nonlinear simulator
# ============================

# -- dependencies
# install.packages(c("mvtnorm","ggplot2"))  # if needed
library(mvtnorm)
library(ggplot2)

# Matern correlation (returns correlation matrix given distance matrix)
matern_cor <- function(D, rho = 0.25, nu = 1.5) {
  # D: distance matrix
  D <- as.matrix(D)
  # Avoid division by zero at 0 distance
  eps <- .Machine$double.eps
  Z <- pmax(D / rho, eps)
  Knu <- besselK(Z, nu)
  cor <- (2^(1 - nu) / gamma(nu)) * (Z^nu) * Knu
  diag(cor) <- 1
  cor
}

# Pairwise Euclidean distance
pairdist <- function(coords) {
  n <- nrow(coords)
  out <- matrix(0, n, n)
  for (i in 1:n) {
    for (j in i:n) {
      d <- sqrt(sum((coords[i,] - coords[j,])^2))
      out[i,j] <- out[j,i] <- d
    }
  }
  out
}

# Nonlinear signal f (modifiable!)
f_nonlinear <- function(X) {
  # X: n x p matrix in [0,1]
  x1 <- X[,1]; x2 <- X[,2]; x3 <- X[,3]; x4 <- X[,4]
  2*sin(pi * x1) + 1.5*exp(x2 - 0.5) + 0.7*(x3^2)*x4 + 0.5*sin(2*pi*x1*x2)
}
```

```r

# Main simulator
simulate_spatial <- function(n = 500,
                             domain = c(0,1,0,1),     # [xmin, xmax, ymin, ymax]
                             design = c("grid","random")["random"],
                             p = 4,                   # number of covariates
                             covariate_structure = c("indep","spatial")["indep"],
                             # GP parameters
                             sigma2 = 1.0, rho = 0.2, nu = 1.5,
                             tau2 = 0.2,
                             seed = 123) {
  set.seed(seed)

  # 1) Sample coordinates
  if (design == "grid") {
    m <- ceiling(sqrt(n)); n <- m*m
    gx <- seq(domain[1], domain[2], length.out = m)
    gy <- seq(domain[3], domain[4], length.out = m)
    coords <- as.matrix(expand.grid(gx, gy))
    colnames(coords) <- c("x","y")
  } else {
    x <- runif(n, min = domain[1], max = domain[2])
    y <- runif(n, min = domain[3], max = domain[4])
    coords <- cbind(x = x, y = y)
  }

  # 2) Generate covariates X(s)
  if (covariate_structure == "indep") {
    X <- matrix(runif(n*p, 0, 1), n, p)
  } else {
    # spatially correlated covariates via GP transforms
    D <- pairdist(coords)
    R <- matern_cor(D, rho = rho*1.2, nu = 1.0)
    # draw p GP fields and squash to [0,1] via rank
    X <- sapply(1:p, function(j) {
      z <- as.numeric(rmvnorm(1, sigma = R))
      (rank(z, ties.method = "average") - 0.5) / n
    })
  }
  colnames(X) <- paste0("X", 1:p)

  # 3) Nonlinear signal f(X)
  eta <- f_nonlinear(X)

  # 4) Spatial random effect b(s)
  D <- pairdist(coords)
  R <- matern_cor(D, rho = rho, nu = nu)
  # Covariance of b: sigma2 * R (with a tiny jitter for stability)
  b <- as.numeric(rmvnorm(1, sigma = sigma2 * R + diag(1e-8, nrow(R))))

  # 5) Nugget
```

```r
  eps <- rnorm(n, mean = 0, sd = sqrt(tau2))

  # 6) Response
  y <- eta + b + eps

  # Return
  data.frame(
    x = coords[,1],
    y_coord = coords[,2],
    Y = y,
    f = eta,
    b = b,
    eps = eps,
    X
  )
}

# ============================
# Example usage
# ============================

dat <- simulate_spatial(
  n = 400,
  design = "random",
  p = 4,
  covariate_structure = "spatial",  # try "indep" or "spatial"
  sigma2 = 0.8, rho = 0.2, nu = 1.5,
  tau2 = 0.15, seed = 42
)

# Quick visuals
ggplot(dat, aes(x, y_coord, color = Y)) +
  geom_point(size = 1.8) +
  coord_fixed() +
  scale_color_viridis_c() +
  labs(title = "Simulated spatial response Y", x = "x", y = "y") +
  theme_minimal()

ggplot(dat, aes(X1, f)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
  theme_minimal() +
  labs(title = "Nonlinear signal vs X1 (truth)", x = "X1", y = "f(X)")
```

Tailoring the Simulation (optional snippets)

Custom nonlinearity. Replace `f_nonlinear()`:

```r
f_nonlinear <- function(X){
  x1 <- X[,1]; x2 <- X[,2]; x3 <- X[,3]
  1.2*(x1 > 0.6) + 2*sin(2*pi*x2) + 0.8*(x3 - 0.5)^2
```

```
}
```

Anisotropy. Transform coordinates before distances:

```
A <- matrix(c(1, 0.4,
              0, 1.5), 2, 2)  # shear & stretch
coords <- as.matrix(dat[,c("x","y_coord")]) %*% A
D_aniso <- pairdist(coords)
R_aniso <- matern_cor(D_aniso, rho=0.2, nu=1.5)
```

Nonstationarity. Vary variance by location:

```
w <- 0.5 + (dat$x > 0.5)            # higher variance where x>0.5
Sigma_b_ns <- diag(w) %*% (0.8 * matern_cor(
                    pairdist(dat[,c("x","y_coord")]), rho=0.2, nu=1.5)) %*% diag(w)
b_ns <- as.numeric(rmvnorm(1, sigma = Sigma_b_ns + diag(1e-8, nrow(dat))))
```

What This Provides

- A continuous response Y composed of a *nonlinear covariate signal* $f(X)$, a *spatial GP effect* $b(s)$, and *nugget noise*.
- Covariates that can be independent or spatially structured.
- Easy hooks to add anisotropy, nonstationarity, and alternate nonlinearities to stress-test spatial regression, GAMs, GP regression, spatial GLMMs, random forests, boosting, etc.

References

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.