

Estimation and inference in penalized function-on-function regression with dependent residual curves

Fabian Scheipl*

Department of Statistics, LMU Munich

Munich Center for Machine Learning (MCML)

Penalized function-on-function regression, as implemented in the R package *refund*, treats all grid values of all curves as independent. Residual curves are smooth, so values within a curve are dependent. We study what this does to point estimates and pointwise confidence intervals, in a simulation with Gaussian, count and binary responses and in a plasmode study built on nine real datasets. Under within-curve dependence, model-based 95% intervals cover 0.63–0.82. A curve-clustered sandwich covariance with a leverage adjustment for the penalized hat matrix (CL2), computed after the usual REML fit, brings coverage close to nominal in the simulation; in the plasmode study it is a few points short on eight datasets and 0.84–0.91 for the coefficient surface on the ninth. Satterthwaite degrees of freedom recover most of the remaining shortfall except on that dataset, also with only 20 curves, at the cost of wider intervals. REML, however, undersmooths the coefficient surface: its error is two to four times that of a fit whose smoothing parameters are chosen by curve-blocked neighbourhood cross-validation (NCV), and its CL2 intervals for the surface are wide. Intervals centred at the smoother NCV estimate are narrow but undercover, the familiar tension between estimation and pointwise inference in nonparametric regression. We recommend the NCV estimate to describe the shape of the surface and REML + CL2 with Satterthwaite degrees of freedom for inference. For a surface that is zero everywhere, under dependent errors, these intervals exclude zero at about 5% of the grid points, and the NCV estimate is five to seven times closer to zero than REML's. For fitted means, intercepts and scalar effects, REML + CL2 intervals are close to nominal, except for binary intercepts and scalar effects and for misregistered counts. We state which defaults of *refund* the evidence supports.

1 Introduction

Function-on-function regression relates a functional response $Y_i(t)$ to functional and scalar covariates (Ramsey and Silverman 2005; Müller and Yao 2008; J. S. Morris 2015). Its terms are a *functional intercept*, the mean response curve at zero covariates; a *coefficient surface* for each functional covariate, which weights

*All code and text were generated by large language models in collaboration with the author; see the statement on the use of generative AI at the end of the main text.

the covariate's value at position s in its effect on the response at t ; and a *coefficient function* of t for each scalar covariate. In the penalized framework of Ivanescu et al. (2015) and Greven and Scheipl (2017), implemented in `pffr()` of the R package `refund` (Scheipl, Staicu, and Greven 2015; Goldsmith et al. 2024), each term is a penalized spline, and the model is fitted as one generalized additive model (Wood 2017) to the $N = GD$ stacked grid values of G curves with D points each. The fit treats these N values as independent given the covariates.

They are not. Residual curves are smooth deviations from the mean, misaligned features or slow drift, so values at nearby grid points of one curve are correlated. This has two consequences. First, the model-based standard errors count N independent observations where the data carry information closer to G ; the intervals are too narrow, and a finer grid makes them narrower without adding information. Second, smoothing-parameter selection by restricted maximum likelihood (REML) reads the dependent error as signal and undersmooths (Opsomer, Wang, and Yang 2001): the estimates, in particular of the coefficient surface, are noisier than necessary.

An analyst who fits such a model needs an estimate that shows the shape of each effect and intervals whose error rates are close to nominal. We evaluate the available fits and covariances on both counts, for Gaussian, count and binary responses, in a simulation with 152 settings and in a plasmode study that attaches the real residual curves of nine datasets to known truths. The findings, with 100 curves under dependence unless stated:

- Model-based intervals for fitted means and coefficient surfaces cover 0.63–0.82 instead of 0.95; for the functional intercept and the coefficient functions of scalar covariates they fail as well (0.63–0.76).
- A curve-clustered sandwich covariance with a leverage adjustment for the penalized hat matrix (CL2), computed after the usual REML fit, covers 0.94–0.95 in the simulation. On the real residual curves of the plasmode study it covers 0.88–0.95 for the mean and the surface: a few points short on eight datasets and down to 0.84 for the surface on one (weather).
- Satterthwaite degrees of freedom for CL2 recover most of the remaining shortfall, at up to 1.33 times the width. With 20 curves, coverage of the mean, the surface and the scalar effect rises from 0.88–0.93 to 0.94–0.97; on the plasmode data with 40 subjects, that of the surface from 0.914 to 0.940.
- REML's error for the coefficient surface is 2.1–3.7 times that of a fit whose smoothing parameters are chosen by curve-blocked neighbourhood cross-validation (NCV), which leaves out whole curves. The REML + CL2 intervals for the surface are therefore wide.
- Intervals centred at the NCV estimate are narrow but undercover. In the simulation the cause is smoothing bias; on the plasmode data the diagnostics suggest smoothing bias and the variability of the selected smoothing parameters. The corrections we tried did not calibrate them there.
- For a zero surface and smooth errors, REML + CL2 with Satterthwaite degrees of freedom excludes zero at 0.046–0.052 of the grid points, the model-based intervals at 0.23–0.57; the NCV estimate is 5–7 times closer to zero than REML's.

This is the classical tension between estimation and pointwise inference in nonparametric regression. An undersmoothed estimate has small bias relative to its standard error, so its intervals can hold their level; an estimate smoothed for accuracy needs an explicit bias correction, whose own variability then enters the interval (Hall 1992; Calonico, Cattaneo, and Farrell 2018). Here REML undersmooths because of the dependence, not by design, and its intervals cover for that reason. No single fit gave both a good estimate and calibrated intervals for the coefficient surface, so we recommend by goal (Section 6): to learn the shape of the surface, a fit that smooths more than REML, by default curve-blocked NCV; for inference about it, REML + CL2 with Satterthwaite degrees of freedom. The tension is strongest for effects that vary along t and a second argument. REML's error is 2.1–3.7 times NCV's for the coefficient surface and 1.3–1.7 times

for a smooth effect of a scalar covariate that varies along t , against 0.94–1.30 for the functional intercept and the coefficient functions of scalar covariates and 1.11–1.35 for fitted means. For these three REML + CL2 gives both a good estimate and intervals close to nominal, with stated exceptions.

The paper is about practice: what users of `pffr()` should report, and which defaults the package should have. Its contributions are: (i) the CR2 leverage adjustment of Bell and McCaffrey (2002), computed with the non-idempotent hat matrix of penalized fits and combined with the model-based smoothing-bias allowance, with Satterthwaite degrees of freedom, implemented in `refund`; (ii) an evaluation that reports width, detection of non-zero effects, false detections of a zero effect and estimation error next to coverage, in synthetic and real-data-based settings and down to 20 curves; (iii) negative results on intervals centred at the NCV estimate and on other remedies, reported in the supplement.

Related work. Within-curve dependence can be modelled directly: functional mixed models with structured residual covariance (J. S. Morris and Carroll 2006; Meyer et al. 2015), generalized least squares with an estimated residual covariance (Reiss, Huang, and Mennes 2010) and curve-level functional random effects (Scheipl, Gertheiss, and Greven 2016), which underlie two of our comparators. Signal compression (Luo and Qi 2017) is a different estimator of the coefficient surface. Pointwise and simultaneous bands for functional means and for function-on-scalar and scalar-on-function models are well developed (Degras 2011; Goldsmith, Greven, and Crainiceanu 2013; Choi and Reimherr 2018; Liebl and Reimherr 2023). Working-independence estimators for clustered data are consistent but can be inefficient (Lin and Carroll 2000; Welsh, Lin, and Carroll 2002). Cluster-robust covariances need only independence across clusters (Liang and Zeger 1986; Cameron and Miller 2015; Zeileis, Köll, and Graham 2020; Hansen and Lee 2019). With few clusters the leverage adjustment of Bell and McCaffrey (2002) and its extensions (McCaffrey and Bell 2006; Pustejovsky and Tipton 2018) reduce the downward bias of the sandwich (Kauermann and Carroll 2001). Here the number of coefficients (about 100) is comparable to the number of curves, so per-curve leverage is not negligible (MacKinnon, Nielsen, and Webb 2023; Carter, Schnepel, and Steigerwald 2017). Korolev (2026) adjusts the cluster-robust covariance for the penalty in semiparametric panel models fitted with `mgcv`. Fast univariate inference (Cui et al. 2022) and penalized functional generalized estimating equations (Loewinger et al. 2025) address longitudinal functional data with many subjects. Correlated errors distort data-driven smoothing in scalar smoothing (Opsomer, Wang, and Yang 2001; Caron Parte, Dedeker, and Michel 2026); likelihood-based selection is less sensitive than generalized cross-validation (Krivobokova and Kauermann 2007). Leaving out whole curves avoids modelling the dependence (Rice and Silverman 1991); neighbourhood cross-validation (Wood 2024) makes this cheap.

2 Model, fits and covariances

2.1 Model and penalized fit

For curves $i = 1, \dots, G$ observed at $t_1, \dots, t_D$, the model is

$$g\{E[Y_i(t) \mid X_i, z_i]\} = \eta_i(t) = \alpha(t) + \int_{\mathcal{S}} X_i(s) \beta(s, t) ds + z_i \gamma(t), \quad (1)$$

with link g , functional intercept $\alpha(t)$, coefficient surface $\beta(s, t)$ of the functional covariate $X_i(s)$ and coefficient function $\gamma(t)$ of the scalar covariate z_i . A smooth effect $f(x_i, t)$ of a scalar covariate can replace the functional term. The responses follow an exponential family, or an extended family such as the scaled t , with mean $g^{-1}\{\eta_i(t)\}$ and variance $\phi V(\mu)$.

Each term is expanded in a penalized spline basis, with tensor products for the bivariate terms. The stacked predictor is $\eta = \mathbf{X}\theta$ with penalty $\sum_j \lambda_j \theta^\top \mathbf{S}_j \theta$. Penalized iteratively reweighted least squares gives $\hat{\theta}$ for given smoothing parameters λ . Let $\mathbf{W}$ be the diagonal matrix of working weights $(d\mu/d\eta)^2 / \{\phi V(\mu)\}$ at convergence and $\mathbf{S} = \sum_j (\lambda_j / \phi) \mathbf{S}_j$. The Bayesian and the frequentist covariance under the working model are

$$\mathbf{V}_p = (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \mathbf{S})^{-1}, \quad \mathbf{V}_e = \mathbf{V}_p \mathbf{X}^\top \mathbf{W} \mathbf{X} \mathbf{V}_p. \quad (2)$$

Their difference $\mathbf{V}_p - \mathbf{V}_e = \mathbf{V}_p \mathbf{S} \mathbf{V}_p$ is a prior-averaged allowance for smoothing bias (Wahba 1983; Nychka 1988; Marra and Wood 2012). For REML fits mgcv adds a term for the uncertainty of $\hat{\lambda}$ and returns $\mathbf{V}_c$ (Wood, Pya, and Säfken 2016). The *model-based* intervals below use $\mathbf{V}_c$ for REML fits and $\mathbf{V}_p$ for NCV fits, because the $\mathbf{V}_c$ that mgcv returns for NCV fits in the version used here is not a valid covariance (Supplement S6). The two model-based intervals therefore contain different uncertainty terms. All of these assume that the N rows are independent.

2.2 Smoothing-parameter selection

REML maximizes the Laplace-approximate restricted likelihood of the working model (Wood 2011). Under within-curve dependence a smooth residual curve looks like many independent observations of a smooth signal, and REML selects too little smoothing.

Curve-blocked NCV. Neighbourhood cross-validation (Wood 2024) minimizes

$$\mathcal{E}(\lambda) = \sum_{k=1}^m \sum_{i \in \mathcal{N}_k} \text{dev}(y_i, \hat{\theta}_\lambda^{-\mathcal{N}_k}), \quad (3)$$

the deviance of predictions for the points in $\mathcal{N}_k$ from the fit without them, with $\hat{\theta}^{-\mathcal{N}_k}$ approximated by a Newton step from the full fit. We take one neighbourhood per curve, $\mathcal{N}_k = \{\text{all grid points of curve } k\}$, so $m = G$ and the criterion measures prediction of a new curve (Rice and Silverman 1991). Leaving out single grid points instead predicts each point from its correlated neighbours; it did not solve the problem in our settings (Supplement S1.5). In Gaussian fits with $G = 100$ and $D = 61$, a curve-blocked NCV fit took a median 8.6 s against 2.0 s for REML.

2.3 Curve-clustered covariances

The curve-clustered sandwich keeps the fit and replaces the variance part of $\mathbf{V}_p$ by one that does not model the dependence within curves. It needs independence across curves and a correct mean model. Write $\tilde{\mathbf{X}}_g = \mathbf{W}_g^{1/2} \mathbf{X}_g$ for the weighted design rows of curve g , $\mathbf{z}_g$ for its Pearson residuals, and $\tilde{\mathbf{H}} = \tilde{\mathbf{X}} \mathbf{V}_p \tilde{\mathbf{X}}^\top$ for the penalized hat matrix. The CL2 covariance is

$$\hat{\mathbf{V}}_{\text{CL2}} = \frac{G}{G-1} \mathbf{V}_p \left(\sum_{g=1}^G \mathbf{U}_g \mathbf{U}_g^\top \right) \mathbf{V}_p + (\mathbf{V}_p - \mathbf{V}_e), \quad \mathbf{U}_g = \tilde{\mathbf{X}}_g^\top \mathbf{R}_g^{-1/2} \mathbf{z}_g, \quad (4)$$

with $\mathbf{R}_g = [(\mathbf{I} - \tilde{\mathbf{H}})(\mathbf{I} - \tilde{\mathbf{H}})^\top]_{gg}$, the block of curve g in the covariance of the residuals under the working model. CR1, the usual cluster-robust covariance (Liang and Zeger 1986), sets $\mathbf{R}_g = \mathbf{I}$. The adjustment $\mathbf{R}_g^{-1/2}$ is the bias-reduced linearization of Bell and McCaffrey (2002) (CR2), with the penalized hat in place of a projection. Because $\tilde{\mathbf{H}}$ is not idempotent, $\mathbf{R}_g \neq \mathbf{I} - \tilde{\mathbf{H}}_{gg}$; the full block is computed. We keep the factor

$G/(G - 1)$ and floor the eigenvalues of $\mathbf{R}_g$ at 10^{-6} , and write CL2 to mark these departures from CR2. The second term of Equation 4 keeps the model-based allowance for smoothing bias, so CL2 replaces only the variance part of $\mathbf{V}_p$. The variance part estimates the sampling variance of $\hat{\theta}$ around its penalized target; the added $\mathbf{V}_p - \mathbf{V}_e$ is a prior-averaged allowance for the distance between that target and the truth, not a correction that guarantees coverage of every fixed truth (Appendix A.3).

The sandwich must use the penalized bread $\mathbf{V}_p$ and the penalized hat. General-purpose cluster-robust software uses the unpenalized hat or bread and gives invalid covariances for penalized fits (Zeileis et al. 2024; Pustejovsky 2024). An asymptotic result for fixed smoothing parameters exists (Appendix A.5), but it is not a guarantee in the regime studied here, with about as many coefficients as curves, a non-negligible penalty and estimated smoothing parameters. CL2 is a finite-sample procedure whose behaviour is established empirically below.

Satterthwaite degrees of freedom. For a contrast $\mathbf{a}^\top \theta$ the variance part of Equation 4 is a sum of squares $\sum_g (\mathbf{q}_g^\top \mathbf{z}_g)^2$ with $\mathbf{q}_g = \mathbf{R}_g^{-1/2} \tilde{\mathbf{X}}_g \mathbf{V}_p \mathbf{a}$, up to the factor $G/(G - 1)$. Under the working model with fixed λ and no smoothing bias, the residuals are $(\mathbf{I} - \tilde{\mathbf{H}})\varepsilon$ with independent standard normal ε , and matching the first two moments of the quadratic form to a scaled χ^2_ν gives $\nu = (\text{tr } \Gamma)^2 / \text{tr}(\Gamma^2)$ with $\Gamma_{gh} = \mathbf{q}_g^\top [(\mathbf{I} - \tilde{\mathbf{H}})(\mathbf{I} - \tilde{\mathbf{H}})^\top]_{gh} \mathbf{q}_h$ (see Appendix A.7 and Bell and McCaffrey 2002; Pustejovsky and Tipton 2018). The interval uses the t_ν quantile at each grid point. The degrees of freedom ignore the bias allowance, the dependence and the selection of λ ; they are a reference distribution, not an exact pivot.

The same constructions apply to NCV fits: $\text{NCV} + \text{CL2}$ is Equation 4 at the NCV estimate.

2.4 Estimands and performance measures

Four estimands are evaluated on fixed grids: the conditional mean $E(Y \mid X)$ for a fixed set of 50 subjects' covariates, on the response scale; $\beta(s, t)$ (or $f(x, t)$); $\alpha(t)$; and $\gamma(t)$, on the link scale. Pointwise intervals are estimate $\pm$ critical value $\times$ standard error, transformed by g^{-1} for the mean. Every method is judged on accuracy of the estimate, calibration and informativeness (T. P. Morris, White, and Crowther 2019):

- *Coverage*: the share of replicates whose interval contains the truth at a grid point, averaged over the grid. Its Monte Carlo standard error is the standard deviation over replicates of the grid-averaged coverage divided by the square root of the number of replicates. The *lower tail* is the 5% quantile of pointwise coverage over the grid. It is a noisy summary: a single grid point's coverage near 0.95 has a Monte Carlo standard error of about 0.015 with 200 replicates. It does not show where misses occur; the supplement's coverage maps do.
- *Relative error*: $\{\sum_j (\hat{\vartheta}_j - \vartheta_j)^2 / \sum_j \tilde{\vartheta}_j^2\}^{1/2}$ over grid points j , for estimate $\hat{\vartheta}$ and truth ϑ , with $\tilde{\vartheta} = \vartheta$ for β and γ and $\tilde{\vartheta}$ the truth centred at its grid mean for the mean and $\alpha(t)$, whose level is not the effect. A value of 1 means the error is as large as the effect. We report its median over replicates.
- *Relative half-width*: the root-mean-square half-width of the intervals over the grid, divided by $(\sum_j \hat{\vartheta}_j^2 / D)^{1/2}$, the root-mean-square size of the (centred) truth.
- *Detection rate*: among grid points where the absolute true value exceeds 25% of its maximum, $|\vartheta_j| > 0.25 \max_k |\vartheta_k|$, the share whose interval excludes zero on the side of the truth.
- *Interval score* (Gneiting and Raftery 2007): for an interval $[l, u]$ and true value ϑ , $(u - l) + (2/0.05)\{(l - \vartheta)_+ + (\vartheta - u)_+\}$, averaged over the grid and the replicates; lower is better. It is a proper scoring rule for predictive quantile intervals; against fixed truths over repeated data sets it rewards an accurate

Table 1: Principal interval constructions. Covariance, bias-allowance and critical-value variants are specified in Appendix A.6; comparator and NCV-centred constructions in Supplement S3 and S4. All CL2 variants cluster by curve, use the penalized bread $\mathbf{V}_p$ and include the allowance $\mathbf{V}_p - \mathbf{V}_e$ unless called frequentist. Full block: $\mathbf{R}_g$ of Equation 4; shortcut: $\mathbf{R}_g \approx \mathbf{I} - \tilde{\mathbf{H}}_{gg}$.

interval	fit	covariance	critical value	evaluated in
model-based	REML	mgcv's $\mathbf{V}_e$	z	all settings, plasmode
model-based	NCV	$\mathbf{V}_p$	z	all settings (supplement)
REML + CL2	REML	CL2, full block, + $(\mathbf{V}_p - \mathbf{V}_e)$	z	all settings, plasmode
REML + CL2, Satterthwaite	REML	as REML + CL2	t_v (Satterthwaite)	main factorial (default signal), extensions, plasmode
NCV + CL2	NCV	CL2, full block, + $(\mathbf{V}_p - \mathbf{V}_e)$	z	all settings, plasmode
CR1, CL2 shortcut, frequentist CL2, t_{G-1}	REML	variants of CL2	z or t_{G-1}	main factorial, default signal

centre and a narrow width together and is not a calibration criterion. We report it relative to REML + CL2 and read it with the other measures.

The supplement adds wrong-sign rates and studentised errors. Comparisons of estimators use the relative error; where ratios of mean squared errors are given, they are labelled as such. Table 1 lists the principal interval constructions and where each was evaluated.

3 Simulation design

The simulation follows model Equation 1 with a functional covariate whose spectrum matches the corpus callosum profiles of the DTI data in refund (152 settings, 200 replicates each, 30,400 fits per selector; Appendix B.1 lists all blocks of settings). Every fit ran to completion; mgcv flagged 78% of the NCV fits of scaled- t responses as not converged (a flat optimum in the five fits checked; Appendix B.1). The truths are fixed smooth functions, projected onto the fitted bases so that they are representable; a rough variant adds higher-frequency components. Responses are Gaussian (signal-to-noise $R^2 = 0.2, 0.5, 0.8$), Poisson and Bernoulli at each grid point (two signal levels on the link scale), with scaled t , beta and negative-binomial responses as extensions. Within-curve errors are a zero-mean process with correlation kernel R , used directly for Gaussian responses and as the latent layer of a Gaussian copula otherwise, so that every marginal is exactly the fitted family. For count and binary data, therefore, the dependence is a copula construction, and the support for CL2 with these families is empirical within that construction. Three error processes form the main factorial:

- *independent*: $R(s, t) = \mathbf{1}\{s = t\}$;
- *rough stationary*: $R(s, t) = 0.8 \exp(-|s - t|/0.089) + 0.2 \mathbf{1}\{s = t\}$;
- *smooth*: the residual covariance of the DTI regression in refund, as a correlation kernel.

The main factorial crosses these processes with the family, $G \in \{40, 100\}$ and the signal level at $D = 61$ grid points. Extensions vary one factor at a time: the main factorial with $G = 20$, a four times denser grid, a rough truth, a coefficient surface that is zero everywhere ($\beta \equiv 0$; Gaussian and binary responses, independent and smooth errors), a smooth effect $f(x, t)$ in place of the functional term, sign-changing and heteroskedastic errors, misregistration (each curve observed on its own clock), a low-rank functional covariate and larger bases. The fitted bases are cubic B-splines with first-order difference penalties (mgcv's ps basis with $m = c(2, 1)$), so the penalty null space contains the constants along each direction: $k = 12$ along t for α and γ and an 8×10 tensor product for β , about 100 coefficients in total; the larger bases halve and quarter the knot spacing. Results with second-order penalties were not computed. Each replicate

is fitted by REML and by curve-blocked NCV. Random draws depend only on the replicate, so contrasts between methods and settings are paired. The Monte Carlo standard error of a grid-averaged coverage is 0.21–0.33 pp for REML + CL2 in the dependent main-factorial settings with 100 curves. Fewer than 20 curves, 20 curves outside the main factorial, and error kernels from other sources than the DTI data were not part of the design.

4 Simulation results

4.1 Intervals for the REML fit

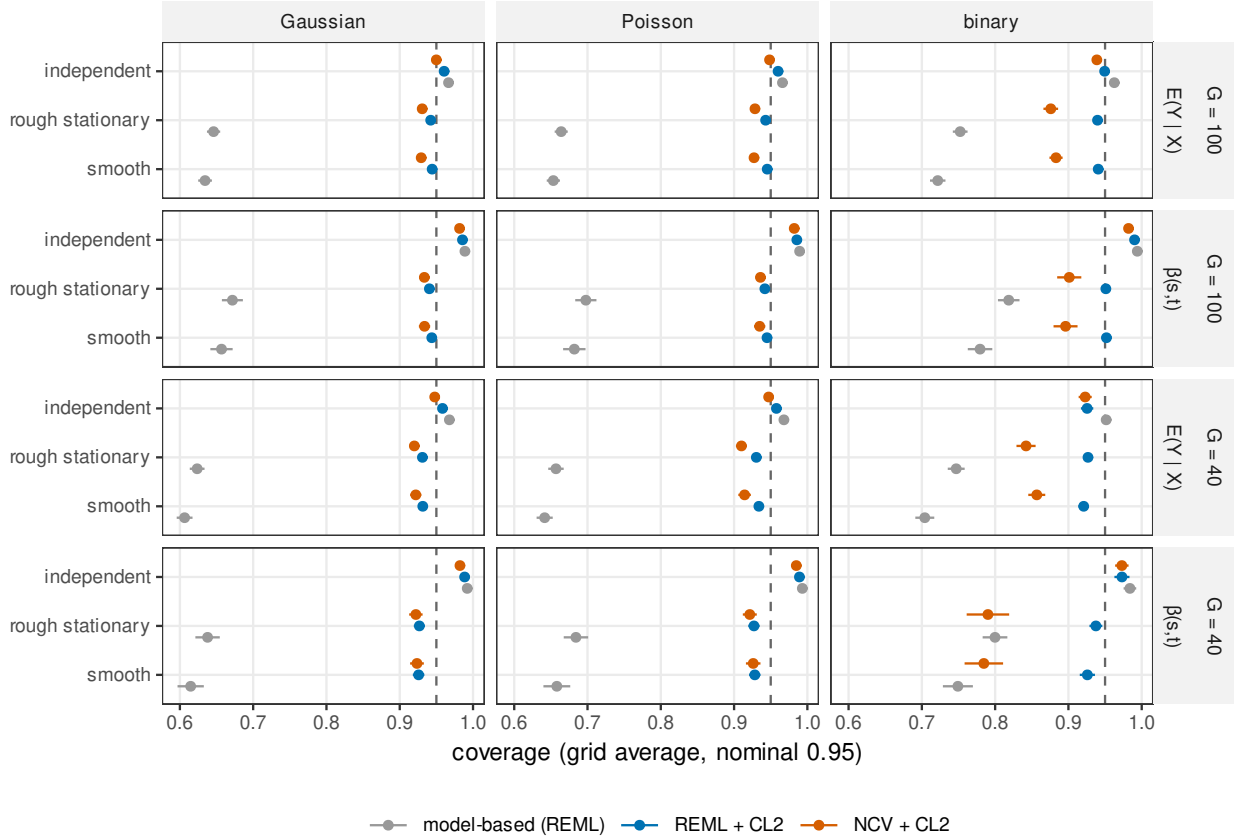


Figure 1: Coverage of nominal 95% pointwise intervals for the mean and for $\beta(s, t)$, main factorial at the default signal level, by family (columns), error process, estimand and number of curves G (rows). Points: grid-averaged coverage; bars: ± 2 Monte Carlo SE; dashed: 0.95.

Model-based intervals undercover in every dependent setting and every family (Figure 1): 0.63–0.82 for the mean and $\beta(s, t)$ with 100 curves. On a four times denser grid ($D = 241$) their coverage of β drops from 0.66–0.82 to 0.35–0.45. REML + CL2 covers the mean and β at 0.94–0.95 with 100 curves and does not change with the grid (0.94 for β at $D = 241$). The same holds for $\alpha(t)$ and $\gamma(t)$ of Gaussian and count responses, for scaled t , beta and negative-binomial responses, for the smooth-effect model, for sign-changing and heteroskedastic errors, and for misregistered Gaussian curves (mean and β 0.93–0.95; Supplement S1.1–S1.3). Under independent errors with 100 curves CL2 gives up 0.3–1.1 pp of coverage for the mean and β

Table 2: REML + CL2 with the normal (z) and the Satterthwaite (Satt.) critical value, main factorial at the default signal, dependent errors (mean over the two processes): grid-averaged coverage and width of the Satterthwaite interval relative to z .

family	estimand	G = 20			G = 40			G = 100		
		coverage			coverage			coverage		
		z	Satt.	width	z	Satt.	width	z	Satt.	width
Gaussian	E(Y X)	0.909	0.958	1.25	0.931	0.953	1.10	0.943	0.951	1.04
	$\beta(s,t)$	0.913	0.961	1.26	0.926	0.950	1.10	0.942	0.951	1.04
	$\alpha(t)$	0.905	0.945	1.16	0.938	0.952	1.05	0.951	0.953	1.02
	$\gamma(t)$	0.918	0.964	1.25	0.942	0.961	1.11	0.943	0.953	1.04
Poisson	E(Y X)	0.908	0.956	1.33	0.932	0.954	1.12	0.944	0.953	1.05
	$\beta(s,t)$	0.912	0.960	1.25	0.928	0.953	1.11	0.943	0.953	1.04
	$\alpha(t)$	0.904	0.941	1.15	0.934	0.945	1.05	0.949	0.952	1.01
	$\gamma(t)$	0.915	0.962	1.26	0.940	0.963	1.12	0.941	0.951	1.05
binary	E(Y X)	0.895	0.945	1.19	0.924	0.946	1.10	0.940	0.949	1.04
	$\beta(s,t)$	0.910	0.958	1.25	0.932	0.954	1.11	0.951	0.960	1.04
	$\alpha(t)$	0.856	0.897	1.14	0.894	0.909	1.05	0.923	0.928	1.01
	$\gamma(t)$	0.895	0.950	1.26	0.921	0.945	1.11	0.931	0.941	1.04

against the model-based intervals (averaged over signal levels within each family); for binary responses at the default signal it gives up 1.3 pp (mean), 1.9 pp ($\alpha(t)$) and 2.2 pp ($\gamma(t)$). With 40 curves it gives up 2.6 pp for the binary mean (0.925 against 0.952, where the model-based intervals do not overcover) and 4.4 pp and 4.1 pp for its $\alpha(t)$ and $\gamma(t)$ at the default signal (Appendix A.6).

Lower tails. The 5% quantile of pointwise coverage for the mean and β is 0.91–0.93 for REML + CL2 with 100 curves and 0.89–0.91 with 40; for the model-based intervals it is 0.54–0.76.

Fewer curves. With 40 curves REML + CL2 covers the mean and β at 0.92–0.94, 1.1–1.8 pp less than with 100 curves; with 20 curves it covers the mean, β and γ at 0.88–0.93 and $\alpha(t)$ at 0.83–0.91 (per setting, all signal levels). Satterthwaite degrees of freedom recover most of this shortfall (Table 2; default signal): with 20 curves, coverage of the mean, β and γ moves from 0.89–0.92 to 0.94–0.96, at 1.19–1.33 times the width; with 100 curves the change is small. The binary intercept reaches only 0.897 with 20 and 0.909 with 40 curves. The per-point degrees of freedom are about a third of $G - 1$ for the mean, β and γ (medians 0.22–0.47 times $G - 1$ in every synthetic setting), larger for $\alpha(t)$. In the extensions they add 2.2 pp of coverage with 40 curves and 0.7 pp with 100 on average and never lower it; coverage stays below 0.93 in 24 of 315 combinations of setting and estimand, all under misregistration or for binary intercepts and scalar effects (Supplement S1.7).

Where CL2 falls short. For binary responses the functional intercept and the scalar effect are covered at 0.88–0.94 under dependence and 0.86–0.95 under independent errors. Under independent errors with 40 curves the studentized error of the intercept is centred (mean 0.04) but has standard deviation 1.28 (model-based 1.15; averages over the two signal levels); under dependence it is also shifted (mean -0.45 to -0.39 with 40 curves and -0.24 to -0.23 with 100, per dependent error process and averaged over the signal levels). This is compatible with a CL2 standard error that is too small for this term; pooled studentized errors cannot exclude local biases of opposite sign. Fixing the smoothing parameters at their typical value raises coverage of the binary $\gamma(t)$ under independent errors from 0.890 to 0.929 and of $\alpha(t)$ under smooth errors from 0.885 to

0.925 (40 curves), so part of the short standard error looks like smoothing-parameter variability; $\alpha(t)$ under independent errors stays at 0.867 (Appendix A.6, Supplement S1.4). With 20 curves and independent errors every interval undercovers binary $\alpha(t)$ and $\gamma(t)$, the model-based ones included (0.80–0.93 per setting), so this shortfall is not specific to the sandwich. For misregistered counts the working model for the mean is wrong: REML + CL2 covers the mean of misregistered Poisson responses at 0.90–0.91 (model-based 0.56–0.63).

4.2 Estimation: REML versus NCV

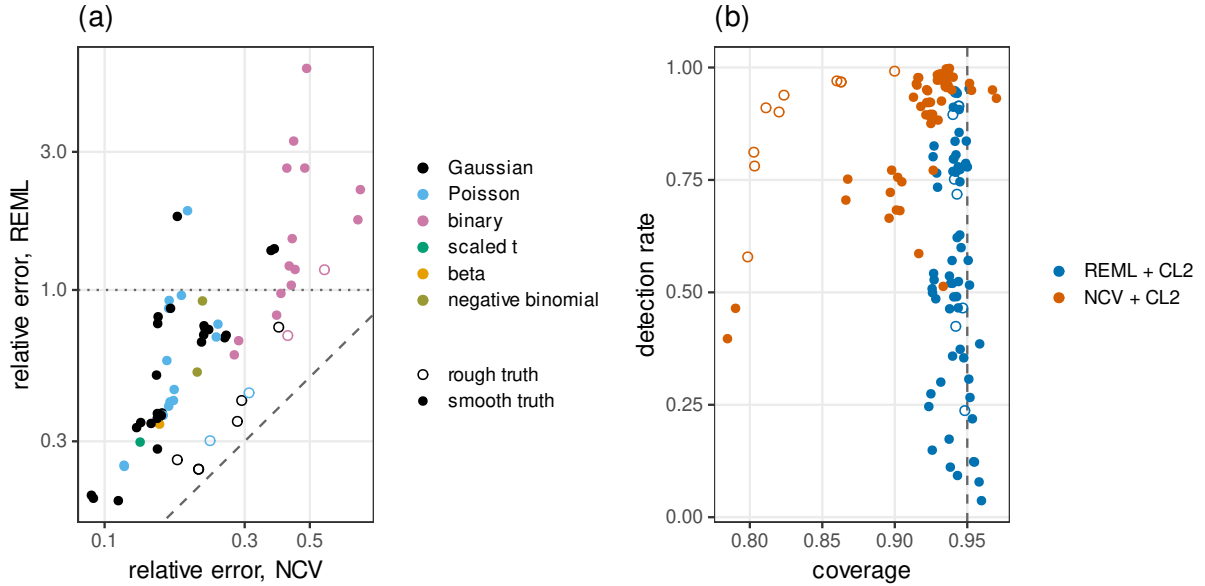


Figure 2: Estimation and intervals for $\beta(s, t)$ in all simulation settings with dependent errors, 40 or 100 curves and a non-zero surface. (a) Relative error of the REML and the NCV estimate (median over replicates, log scales; dashed: equal error, dotted: error as large as the effect). (b) Coverage and detection rate of REML + CL2 and NCV + CL2 (dashed: 0.95); open symbols: rough truth.

Under independent errors both selectors estimate β well: median relative error 0.16 (REML) and 0.12 (NCV) over the settings. Under dependence REML's error rises to 0.67 (maximum 5.8), NCV's to 0.21 (Figure 2 a, Table 3). In the main factorial the ratio of the two is 2.1–3.7 under dependence and 1.1–1.4 under independence, at every signal level; with 20 curves it is 2.5–4.7 and 1.0–1.3. NCV estimates β better also under independent Gaussian errors (MSE ratio 0.55–0.94), so part of the gap reflects the smooth truths rather than the dependence. REML's surface uses 1.9–2.3 times NCV's effective degrees of freedom (EDF) under dependence (1.1–1.3 under independence). Its excess error is variance, not bias: squared bias is 0.01 of REML's mean squared error for β and 0.12–0.20 of NCV's (100 curves, smooth truth). Larger bases make it worse: with the largest basis, REML's surface EDF under smooth errors grows from 27–42 to 113–176 across families, NCV's from 12–21 to 18–41 (Supplement S5).

The intercept and the scalar effect are estimated about equally well by both selectors under dependence, the mean modestly better by NCV (ratio of mean squared errors NCV/REML 0.59–0.80 in the dependent main-factorial settings with 100 curves; relative errors in Table 3). For the mean REML has a heavier tail: 4.3% of replicates have a relative error above 1, against 0.7% for NCV.

Table 3: All simulation settings with dependent errors, 40 or 100 curves and a non-zero surface: median over settings of the relative error of the estimate (rel. error), grid-averaged coverage, 5% quantile of pointwise coverage (q05), relative half-width, detection rate, and interval score relative to REML + CL2 (IS ratio; geometric mean over settings), per estimand and interval. The number of settings differs by estimand: $\beta(s, t)$ and $\alpha(t)$ are not scored in the smooth-effect model, and only the mean is scored for misregistered count and binary responses. Detection is defined for $\beta(s, t)$ and $\gamma(t)$ only.

estimand	interval	settings	rel. error	coverage	q05	half-width	detection	IS ratio
$\beta(s, t)$	model-based (REML)	69	0.67	0.671	0.575	0.57	0.83	1.96
	REML + CL2	69	0.67	0.943	0.915	1.35	0.54	1.00
	NCV + CL2	69	0.21	0.924	0.840	0.37	0.95	0.41
$E(Y X)$	model-based (REML)	83	0.28	0.652	0.560	0.27		1.95
	REML + CL2	83	0.28	0.941	0.910	0.58		1.00
	NCV + CL2	83	0.23	0.918	0.845	0.42		0.85
$\alpha(t)$	model-based (REML)	69	0.31	0.655	0.583	0.31		1.97
	REML + CL2	69	0.31	0.947	0.920	0.65		1.00
	NCV + CL2	69	0.28	0.924	0.857	0.57		1.00
$\gamma(t)$	model-based (REML)	75	0.22	0.663	0.587	0.23	0.99	1.96
	REML + CL2	75	0.22	0.940	0.915	0.46	0.94	1.00
	NCV + CL2	75	0.22	0.920	0.860	0.42	0.96	1.03

Other fits that smooth more. Modelling the residual dependence also reduces REML’s surface error. Two Gaussian methods were compared with NCV on the same data, in 4 Gaussian main-factorial settings with dependent errors and 12 plasmode combinations (Supplement S3.2): a functional random effect for each curve on the eigenfunctions of the REML residual curves (`pcr` in `refund`), and generalized least squares (GLS) with a residual covariance estimated from the REML residual curves by functional principal components (FPCA) plus white noise. Their mean squared error for β , relative to NCV’s (geometric mean over settings), is 1.75 (`pcr`) and 1.79 (GLS) in the simulation, where REML’s is 7.4, and 0.72 and 0.68 on the plasmode data (midpoint truth, defined in Section 5, with REML residuals; the GLS covariance is estimated from the same REML residual curves that generate the plasmode errors): much better than REML, worse than NCV in the simulation and comparable to NCV in that plasmode comparison. A `pcr` fit took a median 18 s and up to 1342 s. The GLS algorithm is given in Supplement S3.2.

4.3 The trade-off for the coefficient surface

REML + CL2 covers β but says little about it. Its median half-width under dependence is 1.35 times the size of the effect, and it detects 0.54 of the grid points where the true value exceeds 25% of its maximum (Table 3, Figure 2 b). NCV + CL2 is 3.7 times narrower and detects 0.95, but covers less (median 0.924, lower tail 0.84). Its interval score is 0.41 times REML + CL2’s: the score rewards its accurate, narrow intervals more than it charges the extra misses, while the model-based REML intervals score 1.96 times. We nevertheless do not recommend NCV + CL2 for inference: its coverage is short on the real residual curves (Section 5). In the simulation the shortfall of NCV + CL2 is smoothing bias: subtracting the bias of the NCV estimate, estimated from the replicates, raises its coverage from 0.90–0.94 to 0.93–0.96 in the main factorial with 100 curves. With 20 curves and dependent errors NCV + CL2 covers the four estimands at 0.71–0.92, with lower tails down to 0.50. With a rough truth, NCV + CL2 covers 0.80–0.90 with lower tails 0.42–0.78 (REML + CL2: 0.94–0.95), and for binary responses with 40 curves 0.78–0.90. The trade-off grows with the basis:

with the largest basis REML + CL2's half-width for β reaches 3.78–12.44 times the effect and its detection rate 0.04–0.12, while its coverage stays at 0.95–0.96 (smooth errors). Wrong-sign exclusions are rare for REML + CL2 (at most 2.3% of the grid in any dependent setting); false detections of a zero surface are in Section 4.4.

For the mean, $\alpha(t)$ and $\gamma(t)$ the trade-off is mild. REML + CL2's intervals are wider than NCV + CL2's (mean: 1.37 times), but they cover, the estimates of $\alpha(t)$ and $\gamma(t)$ are about as accurate, and $\gamma(t)$ is detected at 0.94.

The model-based intervals of pcre and GLS with an FPCA residual covariance show the trade-off from the other side. In the 4 Gaussian main-factorial settings with dependent errors they cover β at 0.93–0.97, at 0.51–0.66 times REML + CL2's width and 0.52–0.57 times its interval score, but pcre's intervals cover the mean at only 0.88–0.91 and $\alpha(t)$ at 0.51–0.59; on real residual curves both cover β at only 0.59–0.86 (Section 5). GLS whitens the data within curves, so curves stay independent clusters and CL2 applies to the whitened fit; pcre + CL2 is CL2 on the pcre fit (Supplement S3.2; replicates 1–100, where all constructions exist). Under dependent synthetic errors GLS-FPCA + CL2 covers β at 0.959 at 0.61 times REML + CL2's width (interval score 0.55 times), and pcre + CL2 repairs pcre's intervals for $\alpha(t)$ (0.56 to 0.95). On the real residual curves GLS-FPCA + CL2 covers β at 0.823 (0.65–0.90 by dataset; REML + CL2 0.931), pcre + CL2 at 0.831. There its standard error is close to the spread of the estimate on average (ratio 0.98), consistent with residual smoothing bias, but an average ratio does not exclude local or tail errors in the standard error. GLS with the raw residual covariance fails with or without CL2 (0.65 under dependent synthetic errors). Fits that smooth more thus estimate better, and the sandwich repairs much, but not all, of their undercoverage.

4.4 A zero coefficient surface

With $\beta \equiv 0$ the share of grid points at which an interval excludes zero is the pointwise false-positive rate, nominally 0.05 (Table 4). Under smooth errors it is 0.054–0.076 for REML + CL2 with the normal critical value and 0.046–0.052 with Satterthwaite degrees of freedom, against 0.11–0.15 for NCV + CL2 and 0.23–0.57 for the model-based intervals of either fit. Under independent errors REML + CL2 has 0.042–0.056, except 0.095 for Gaussian responses with 40 curves, where REML shrinks the surface to near the null space of its penalty in most replicates and in some of these the interval excludes zero over much of the grid (0.18 of replicates have more than 5% of the grid excluding zero), a known weakness of intervals for terms shrunk to their null space (Marra and Wood 2012).

Under smooth errors REML's estimate of the zero surface has 0.39–2.25 times the root-mean-square size of the non-zero surface in the corresponding setting, a structure that is not there; NCV's estimate is 4.8–6.8 times smaller. Under independent errors both are close to zero. Pointwise intervals do not answer whether β is zero everywhere: under smooth errors 0.72–0.83 of the replicates have at least one grid point where the Satterthwaite interval excludes zero. That question needs simultaneous bands or global tests, which this study does not evaluate.

4.5 Other remedies

- *Intervals centred at the NCV estimate.* Widening by the NCV–REML difference, the NCV estimate with REML + CL2's half-width, a one-step bias correction of the NCV estimate, and a cluster-robust term for the variability of the selected smoothing parameters, alone and combined. The combination is a frozen-map approximation: it propagates the smoothing-parameter term through the bias-correction map but ignores how that map changes with the smoothing parameters. It covers β at 0.971 (smooth

Table 4: Zero surface ($\beta \equiv 0$), default signal: pointwise false-positive rate (nominal 0.05) of the model-based REML interval, REML + CL2 with the normal (z) and the Satterthwaite (Satt.) critical value and NCV + CL2; root-mean-square size of the REML and NCV estimates relative to the non-zero surface of the corresponding setting.

family	errors	G	model	false-positive rate			size	
				REML + CL2		NCV + CL2	REML	NCV
				z	Satt.			
Gaussian	smooth	40	0.399	0.076	0.052	0.130	0.77	0.11
		100	0.361	0.058	0.049	0.151	0.39	0.06
	independent	40	0.024	0.095	0.075	0.096	0.02	0.03
		100	0.009	0.050	0.046	0.047	0.01	0.02
binary	smooth	40	0.250	0.071	0.048	0.111	2.25	0.45
		100	0.228	0.054	0.046	0.118	1.11	0.23
	independent	40	0.009	0.056	0.029	0.060	0.09	0.16
		100	0.015	0.042	0.031	0.048	0.06	0.09

truths) and 0.941 (rough truths) in Gaussian settings, but undercovers on the plasmode data (Section 5.2; Supplement S4.1).

- *mgcv's neighbourhood covariance for NCV fits.* For curve blocks the mgcv version used here returns the Bayesian V_p or, with `jackknife = TRUE`, essentially $V_p - V_e$; it computes this covariance correctly only for per-point neighbourhoods. Computed correctly from the block fit, it is conservative in the simulation for β (0.96–0.98 with smooth errors and 100 curves, at 0.46–0.51 times REML + CL2's width) and for Gaussian and Poisson responses, but covers binary $\alpha(t)$ and $\gamma(t)$ at only 0.89. On the plasmode data it undercovers β : 0.87 on average over the 42 combinations of the six datasets with independent source curves (0.86 over datasets; NCV + CL2 0.82, REML + CL2 0.93), and 0.82, 0.875 and 0.91 under the REML-fitted, midpoint and NCV-fitted truths. Residual smoothing bias of the NCV estimate is a plausible contributor; variance and tail errors are not excluded (SE/SD 0.85 on ECG 8-lead, 0.93 on running; Supplement S4.2).
- *A smoothing-parameter uncertainty term for REML + CL2* raises coverage of β by 1.2–1.6 pp without improving the interval score (Supplement S4.3).
- *An AR(1) working model* (`bam()` with AR(1) errors restarting at each curve, correlation profiled by the restricted-likelihood criterion; Appendix B.1) covers 0.44–0.93 on the nine plasmode datasets (Supplement S3.1).
- *The curve bootstrap* (resampling curves and refitting) covers 0.93–0.97 on real residual curves, with intervals 1.42 times as wide as REML + CL2's and a median of 507 s per analysed data set (Supplement S3.2).
- *Penalized functional GEE* [Loewinger et al. (2025); R package fastFGEE] improves the surface over REML by less than NCV (mean squared error 1.32–3.70 times NCV's); Wald intervals from its sandwich covariance cover β at 0.67–0.89. Its calibrated bootstrap is implemented for intercepts and scalar effects only, so this is not a test of that procedure in its own setting (Supplement S3.3).

5 Plasmode study

5.1 Design

A plasmode study simulates from real data with a known truth (Franklin et al. 2014). Each of nine datasets is fitted with model Equation 1 by REML and by curve-blocked NCV. A fit’s fitted means serve as the truth: the *REML-fitted truth*, the *NCV-fitted truth*, and a *midpoint truth* refitted with every smoothing parameter at the geometric mean of the two fits’ values. A replicate adds each subject’s real residual curve, from the REML or the NCV fit, with a random sign to the truth at its own covariates. This keeps the covariates, the residual shapes and their within-curve dependence, and makes the truth known. Because the signs are independent, the simulated errors are independent across curves, and each subject’s error takes only two values, plus or minus its residual curve: the construction is a Rademacher wild cluster bootstrap around the fitted residuals, not a model of new real curves. Its truths are exactly representable by the fitted model, so it cannot evaluate misspecification of the mean. In six datasets the source curves are independent: ECG-derived myocardial strain, atrial strain from a randomised trial, knee moments during running and walking, the DTI tract profiles and an eight-lead ECG dataset. In three the source curves are serially or spatially dependent (monthly ocean profiles, weather stations, weekly electricity demand); their simulated errors are still independent across curves, except in block-flip runs reported in the supplement. The midpoint fit lies between the REML and NCV fits for every term on only 3 of the 9 datasets, and its surface is rougher than both on running, ocean, weather; it is a third smoothing level, not a compromise. The datasets have 73–138 curves on 30–101 grid points, design effects of the residual curves of 6–40 and functional covariates of rank 4–24 (Appendix B.2). Each combination of dataset, truth and residual source is run with all subjects and with 40, with 200 replicates; the mean is scored for a fixed set of 50 subjects.

5.2 Results

Intervals. With curve flips the simulated errors are independent across curves on all nine datasets, so CL2’s assumption holds on all of them. Model-based intervals cover 0.34–0.75 for the mean and β . REML + CL2 covers 0.88–0.94 for β and 0.92–0.95 for the mean (Figure 3): 0.92–0.94 for β on eight datasets, and 0.88 (0.84–0.91 over truth and residual sources) on the weather data, which combine few curves (73), a functional covariate of rank 4 and rough residuals (lag-one correlation 0.50, 19 principal components for 90% of the residual variance). REML + CL2’s coverage of β is below 0.95 in all 84 combinations of truth, residual source and number of subjects on the six datasets with independent source curves, and in 130 of 130 over all nine datasets and construction variants; the combinations share data and are not independent confirmations. Its standard error is somewhat smaller than the replicate spread of the estimate (median ratio 0.89–1.01 per dataset over all nine datasets). A ratio of 0.90 alone would give a centred normal estimator coverage of about 0.92, so these ratios are consistent with a standard error a little short, plus further error on some datasets. The lower tail for β is lower under the REML-fitted truth (0.85–0.90) than under the NCV-fitted truth (0.88–0.90). With 40 subjects REML + CL2 covers β 1.4–4.8 pp below nominal and the mean 0.6–3.2 pp below on the six datasets with independent source curves, and β at 0.77–0.87 on the weather data.

Two corrections that helped in the simulation help only partly here. With Satterthwaite degrees of freedom (Supplement S2.1), REML + CL2 covers β at 0.939 on average over the nine datasets with all subjects and at 0.940 with 40 (normal critical value: 0.927 and 0.914), at 1.05 and 1.13 times the width, and the mean, $\alpha(t)$ and $\gamma(t)$ at 0.944–0.955; the weather surface stays short (0.899 with all subjects, 0.863 with 40). Adding the smoothing-parameter uncertainty term and using the t_{G-1} critical value helps about as much (Supplement

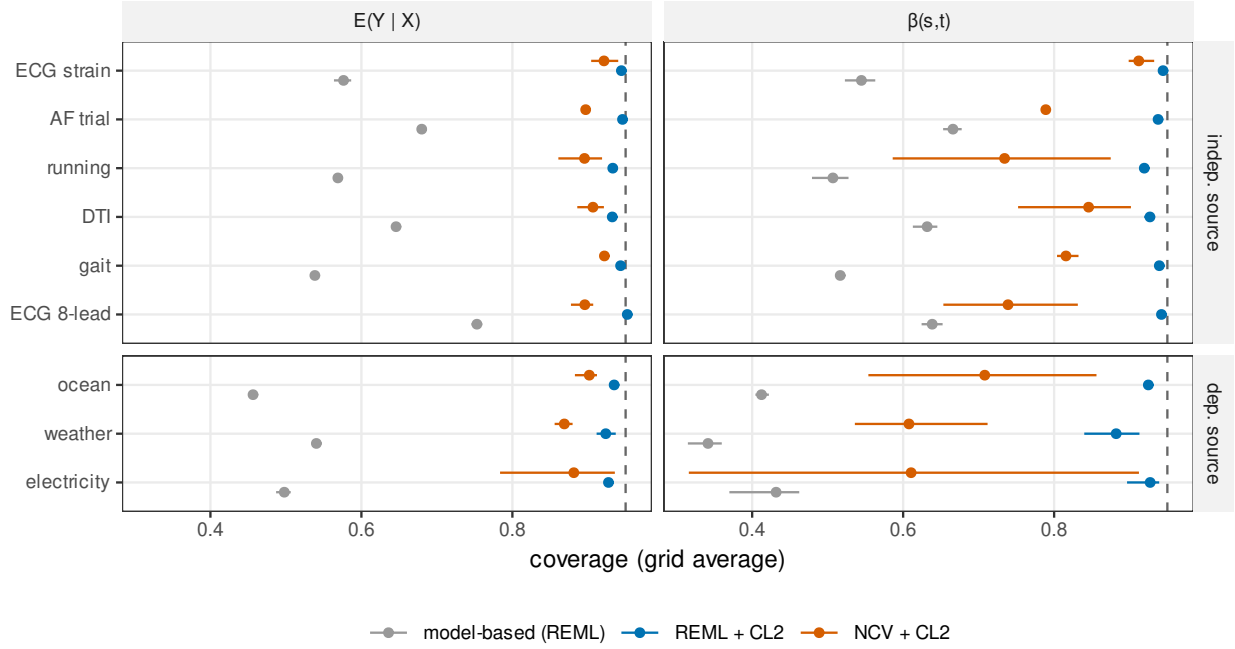


Figure 3: Plasmode coverage of nominal 95% pointwise intervals for the mean and $\beta(s, t)$, all subjects: mean over truth and residual sources (points) and their range (bars). Dashed: 0.95.

S4.3): β at 0.93–0.96 on eight datasets with all subjects (0.92–0.96 with 40), but only 0.90 (0.85) on the weather data.

On the six datasets with independent source curves NCV + CL2 covers β at 0.73–0.91 with lower tails 0.44–0.81 and loses up to 18.6 pp with 40 subjects. The diagnostics suggest two contributions. Its standard error is short (median ratio to the replicate standard deviation 0.68–0.98 across datasets), which a refit with the smoothing parameters fixed at their median restores (1.03–1.15), but its coverage then still stays at 0.64–0.92. Fixing the smoothing parameters changes both the amount and the variability of smoothing, so this is a diagnostic, not an exact decomposition; it suggests smoothing bias as the larger part (Supplement S4.1).

Estimation. On real covariates β is weakly determined by either fit: median relative errors of 0.63–3.42 for REML and 0.47–0.81 for NCV across datasets (Figure 4, Table 5). Which fit is better depends on the truth. The ratio of NCV’s to REML’s relative error is 0.07–0.87 under the NCV-fitted truth, 0.24–0.97 under the midpoint truth and 0.55–1.69 under the REML-fitted truth, where NCV is worse on 4 of the 9 datasets. A fit used as truth is shaped like the selector that produced it, so the truth source favours that selector; the synthetic study, whose truths come from neither fit, is the cleaner comparison. For the mean the ratio of NCV’s to REML’s relative error is 0.71–1.00 per dataset.

Informativeness. As in the simulation, REML + CL2 is close to nominal but wide for β : its half-width is 1.2–5.7 times the effect and it detects 0.08–0.64 of the grid points where the true value exceeds 25% of its maximum. NCV + CL2 detects more (0.46–0.77) but does not hold its level, and here its interval score is 0.75–1.84 times REML + CL2’s: higher on 7 of the 9 datasets. Intervals centred at the NCV estimate with a bias correction and a smoothing-parameter term (Section 4.5) cover β at 0.83–0.94 on the six datasets with independent source curves, against 0.91–0.94 for REML + CL2, at 0.74–1.10 times its width and with lower tails as low as 0.38 in single combinations, and at 0.66–0.84 on the three with dependent source curves (Supplement S4.1). These are descriptive comparisons on different replicate windows (replicates 1–50 for

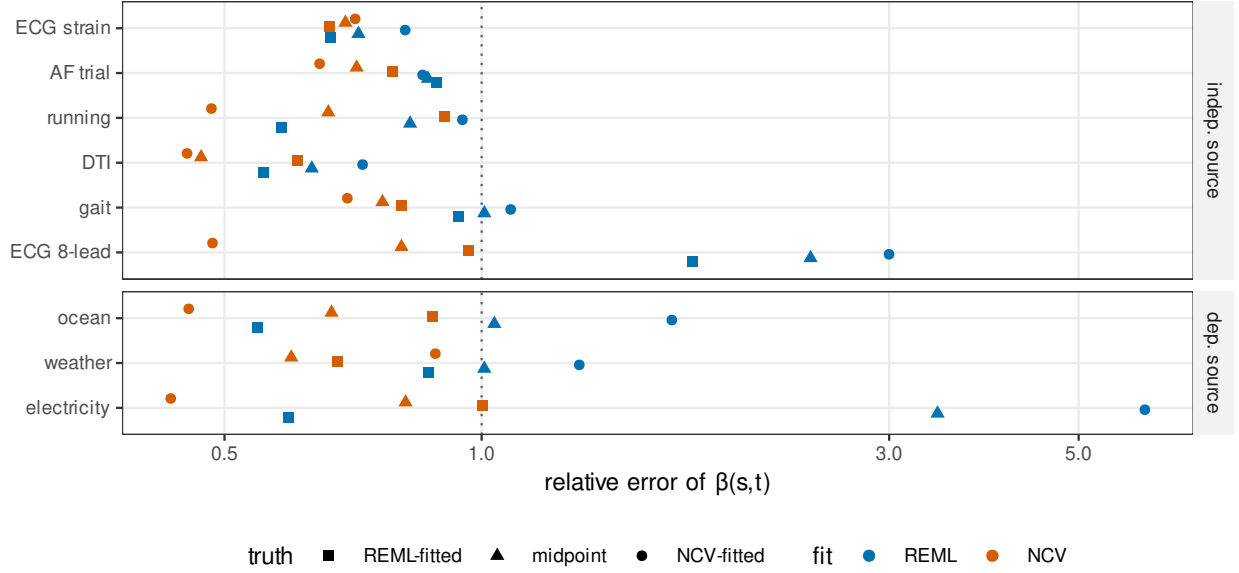


Figure 4: Plasmode relative error of the REML and NCV estimates of $\beta(s, t)$, all subjects, by dataset and truth source (median over replicates and the two residual sources; log scale; dotted: error as large as the effect).

Table 5: Plasmode study, $\beta(s, t)$, all subjects: median relative error of the REML and NCV estimates; REML + CL2 coverage, 5% quantile of pointwise coverage (q05), relative half-width and detection rate; NCV + CL2 coverage, detection rate and interval score relative to REML + CL2 (IS ratio) (means, medians or geometric means over truth and residual sources). The last three datasets have dependent source curves.

dataset	rel. error		REML + CL2				NCV + CL2		
	REML	NCV	coverage	q05	half-width	detection	coverage	detection	IS ratio
ECG strain	0.72	0.69	0.944	0.90	1.5	0.42	0.904	0.46	1.04
AF trial	0.87	0.71	0.938	0.88	1.8	0.37	0.789	0.53	1.19
running	0.82	0.66	0.919	0.87	1.9	0.64	0.734	0.73	1.50
DTI	0.63	0.47	0.927	0.88	1.2	0.64	0.846	0.77	1.03
gait	1.01	0.77	0.939	0.90	2.4	0.55	0.816	0.65	1.28
ECG 8-lead	2.43	0.81	0.942	0.87	4.9	0.22	0.739	0.57	0.75
ocean	1.03	0.67	0.925	0.89	2.2	0.45	0.708	0.53	1.43
weather	1.01	0.68	0.882	0.83	2.0	0.32	0.608	0.54	1.84
electricity	3.42	0.81	0.927	0.88	5.7	0.08	0.611	0.51	0.86

the combination, 1–100 for REML + CL2), without a paired measure of precision. The NCV estimate with REML + CL2's half-width covers 0.906 under the REML-fitted truth, with a lower tail of 0.67 against 0.87 for REML + CL2.

6 Recommendations and software

i Recommendations for applied work with `pffr()`

1. **Assume within-curve dependence.** Do not use model-based intervals of the working-independence fit; they undercover, more so on dense grids.
2. **Inference: REML + CL2 with Satterthwaite degrees of freedom.** Fit by REML and use the curve-clustered CL2 covariance with the full block and the Satterthwaite critical value. This was evaluated for all four estimands in the main factorial at the default signal with 20, 40 and 100 curves, in all extensions and in the plasmode study. Under dependence it covered the mean, β and γ at 0.94–0.96 in the main factorial, and β at 0.939 on average in the plasmode study, but at 0.90 on one dataset with few curves and a low-rank covariate. The price is width: up to 1.33 times the width with the normal critical value with 20 curves, about 1.04 times with 100. The intervals for the surface are wide; report them as they are. Only pointwise intervals were evaluated: a pointwise band answers “is there an effect, and where” poorly (detection rates of 0.08–0.64 in the plasmode study) and “is β zero everywhere” not at all, which needs simultaneous bands or global tests.
3. **Shape of the surface: a fit that smooths more than REML.** By default, fit the same model with `method = "NCV"` and show its surface estimate as the description of the effect. For Gaussian responses, a functional random effect per curve (`pcr`) or GLS with an FPCA residual covariance estimated the surface much better than REML; they were worse than NCV in the four synthetic settings compared and comparable to it in a plasmode comparison with six datasets. When the surface was zero, the NCV estimate stayed close to zero where REML's showed structure. Do not attach the REML + CL2 band to these estimates, and do not use NCV + CL2 intervals for inference. The same holds, more weakly, for smooth effects of scalar covariates that vary along t (REML's error 1.3–1.7 times NCV's); for the intercept, scalar effects and fitted means, REML + CL2 gives both the estimate and the intervals.
4. **Exceptions.** Treat intervals for the intercept and scalar effects of binary responses and for mis-registered counts as approximate; they undercover, also with Satterthwaite degrees of freedom.
5. **Few curves.** Evaluated down to 20 curves in the main factorial. With 20 curves and dependent errors REML + CL2 with Satterthwaite degrees of freedom covered the mean, β and γ at 0.94–0.97 and binary intercepts at 0.89–0.91; under independent errors every interval undercovered binary intercepts and scalar effects. Nothing below 20 curves, or with 20 curves outside the main factorial, was evaluated; treat intervals there as approximate. The weather data with block flips have 16 independent blocks; block-clustered CL2 covered β at 0.89 there (Supplement S2.3). With five curves per subject in a function-on-scalar design, subject-clustered CL2 covered 0.92–0.93 with 25 subjects (Supplement S3.3).
6. **Several curves per subject.** Cluster by subject (`cluster =`); in the function-on-scalar design just mentioned this covered 0.94–0.95 with 100 or 500 subjects. It leaves fewer clusters and does not address dependence between subjects.
7. **Basis and identification.** Refit with a richer basis and compare. A richer basis protects against approximation bias but widens REML + CL2's intervals. It does not help where the functional

covariate carries little information about a direction of the surface: the surface is then weakly identified (Scheipl and Greven 2016). Refrain from interpreting the surface’s shape when the estimate or its interval changes substantially with the basis, or when the covariate’s principal components leave the relevant directions poorly covered given the penalty’s null space; agreement between two penalized fits does not certify identification. Fitted means or well-supported contrasts are then safer targets, though not protected against misspecification.

8. **Covariates with few principal components.** If the functional covariate’s variation is concentrated in a few principal components, an `ffpc()` term with a small number of components can replace `ff()`. With the low-rank covariate of one synthetic setting, the 3 components that explain 95% of its variance estimated β with a relative root-mean-square error of 0.20, against 0.22 for NCV and 0.76 for REML with `ff()`, and their CL2 intervals with the normal critical value covered 0.915 (0.948 with five components; Supplement S1.9). For covariates with a richer spectrum keep `ff()`: few components truncate the surface and the intervals undercover, and a high variance-explained threshold, such as refund’s default 0.99, gives unstable, very noisy estimates.

Forthcoming releases of refund (version X.Y or later) and mgcv (version X.Y or later) will implement these recommendations, with REML + CL2 and Satterthwaite degrees of freedom as the default intervals of `pffr()` fits, and contain the corrections found in this study, among them those of mgcv’s covariances for NCV fits (Supplement S4.2). Supplement S6 gives example code, the software used in the study and a reproduction route.

7 Discussion

Under within-curve dependence the working-independence REML fit of a penalized function-on-function regression gives model-based intervals that are far too narrow and an undersmoothed, noisy estimate of the coefficient surface. A curve-clustered sandwich with a leverage adjustment for the penalized hat and Satterthwaite degrees of freedom brings the intervals close to nominal for Gaussian, count and binary responses in the simulation, down to 20 curves, and for the surface on eight of the nine plasmode datasets, but it cannot improve the estimate it is centred on. A fit that smooths more improves the estimate but not the intervals: the undersmoothing-versus-bias-correction tension of nonparametric inference (Hall 1992; Calonico, Cattaneo, and Farrell 2018), here produced by the dependence, hence the recommendation by goal. Intermediate amounts of smoothing between REML and NCV, global tests of $\beta \equiv 0$ and simultaneous bands are the natural next steps; they were not studied here. The evidence has the following limits.

- *Binary responses.* REML + CL2 undercovers intercepts and scalar effects of binary responses, also under independent errors (as low as 0.859) and with Satterthwaite degrees of freedom; with 20 curves and independent errors the model-based intervals undercover them too. The diagnostics are compatible with a standard error that is too small for these terms, partly through smoothing-parameter variability; local bias is not excluded. Every interval centred at the NCV estimate undercovers for binary responses (Supplement S4.1).
- *Misregistration.* When curves are observed on their own clocks, the observed-clock mean of a count response is not a generalized linear predictor, and no method covers it; for Gaussian responses REML + CL2 is unaffected.
- *Design.* The synthetic design has 20 curves only in the main factorial, a zero surface only for Gaussian and binary responses with independent and smooth errors, one empirically derived smooth error

kernel (from the DTI data) beside parametric kernels, and first-order difference penalties throughout. In the plasmode study with 40 subjects REML + CL2 with Satterthwaite degrees of freedom covers β at 0.931–0.963 on the six datasets with independent source curves.

- *Truths outside the spline space.* All synthetic truths in the main study are representable. For a truth with components the covariates barely excite and the default basis cannot represent (Supplement S5), REML + CL2 covers β at only 0.44–0.56 under smooth errors and 0.30 under independent errors. A basis with four times finer knot spacing restores 0.93 under smooth errors, at the price of wider intervals, but only 0.86 under independent errors; NCV + CL2 stays short even then. These coverages refer to the exact truth, which mixes approximation bias with weak identification of directions the covariates barely excite.
- *Plasmode study.* Fits used as truths favour their own selector, the errors are sign-flipped fitted residuals with independent signs across curves, the truths are representable by the model, and all nine datasets are Gaussian and were chosen, not sampled.
- *Intervals around the NCV estimate.* Their shortfall is smoothing bias in the simulation; on the plasmode data the diagnostics suggest smoothing bias and the variability of the selected smoothing parameters, but fixing the smoothing parameters changes both. Neither a one-step bias correction nor a variance term for the smoothing parameters, nor their combination as a frozen-map approximation, gave a narrower calibrated interval in the plasmode study. Calibrated intervals after data-driven heavier smoothing remain open.

Use of generative AI

All code and all text of this paper and its supplement were generated by large language models: Claude Opus 5.5 and Claude Fable 5.1 (Anthropic) and ChatGPT models Sol and Astra, versions 5.6 and 6 (OpenAI). The study design, structural decisions and research directions were developed collaboratively between the author and these models. The author takes full responsibility for the content.

References

- Bell, Robert M., and Daniel F. McCaffrey. 2002. “Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples.” *Survey Methodology* 28 (2): 169–81.
- Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. 2018. “On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference.” *Journal of the American Statistical Association* 113 (522): 767–79. <https://doi.org/10.1080/01621459.2017.1285776>.
- Cameron, A. Colin, and Douglas L. Miller. 2015. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources* 50 (2): 317–72. <https://doi.org/10.3368/jhr.50.2.317>.
- Caron Parte, Emmanuel, Jérôme Dedecker, and Bertrand Michel. 2026. “Choosing the Penalty in Nonparametric Regression: Short and Long-Range Dependence.” <https://arxiv.org/abs/2609.29231>.
- Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald. 2017. “Asymptotic Behavior of a t -Test Robust to Cluster Heterogeneity.” *The Review of Economics and Statistics* 99 (4): 698–709. https://doi.org/10.1162/REST_a_00639.
- Choi, Hyunphil, and Matthew Reimherr. 2018. “A Geometric Approach to Confidence Regions and Bands for Functional Parameters.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80 (1): 239–60. <https://doi.org/10.1111/rssb.12239>.

- Cui, Erjia, Andrew Leroux, Ekaterina Smirnova, and Ciprian M. Crainiceanu. 2022. “Fast Univariate Inference for Longitudinal Functional Models.” *Journal of Computational and Graphical Statistics* 31 (1): 219–30. <https://doi.org/10.1080/10618600.2021.1950006>.
- Degras, David A. 2011. “Simultaneous Confidence Bands for Nonparametric Regression with Functional Data.” *Statistica Sinica* 21 (4): 1735–65. <https://doi.org/10.5705/ss.2009.207>.
- Eilers, Paul H. C., and Brian D. Marx. 1996. “Flexible Smoothing with B-Splines and Penalties.” *Statistical Science* 11 (2): 89–121. <https://doi.org/10.1214/ss/1038425655>.
- Franklin, Jessica M., Sebastian Schneeweiss, Jennifer M. Polinski, and Jeremy A. Rassen. 2014. “Plasmode Simulation for the Evaluation of Pharmacoepidemiologic Methods in Complex Healthcare Databases.” *Computational Statistics & Data Analysis* 72: 219–26. <https://doi.org/10.1016/j.csda.2013.10.018>.
- Gneiting, Tilmann, and Adrian E. Raftery. 2007. “Strictly Proper Scoring Rules, Prediction, and Estimation.” *Journal of the American Statistical Association* 102 (477): 359–78. <https://doi.org/10.1198/016214506000001437>.
- Goldsmith, Jeff, Sonja Greven, and Ciprian Crainiceanu. 2013. “Corrected Confidence Bands for Functional Data Using Principal Components.” *Biometrics* 69 (1): 41–51. <https://doi.org/10.1111/j.1541-0420.2012.01808.x>.
- Goldsmith, Jeff, Fabian Scheipl, Lei Huang, Julia Wrobel, Chongzhi Di, Jonathan Gellar, Jaroslaw Harezlak, et al. 2024. *Refund: Regression with Functional Data*. <https://CRAN.R-project.org/package=refund>.
- Greven, Sonja, and Fabian Scheipl. 2017. “A General Framework for Functional Regression Modelling.” *Statistical Modelling* 17 (1–2): 1–35. <https://doi.org/10.1177/1471082X16681317>.
- Hall, Peter. 1992. “Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density.” *The Annals of Statistics* 20 (2): 675–94. <https://doi.org/10.1214/aos/1176348651>.
- Hansen, Bruce E., and Seojeong Lee. 2019. “Asymptotic Theory for Clustered Samples.” *Journal of Econometrics* 210 (2): 268–90. <https://doi.org/10.1016/j.jeconom.2019.02.001>.
- Ivanescu, Andra E., Ana-Maria Staicu, Fabian Scheipl, and Sonja Greven. 2015. “Penalized Function-on-Function Regression.” *Computational Statistics* 30 (2): 539–68. <https://doi.org/10.1007/s00180-014-0548-4>.
- Kauermann, Göran, and Raymond J. Carroll. 2001. “A Note on the Efficiency of Sandwich Covariance Matrix Estimation.” *Journal of the American Statistical Association* 96 (456): 1387–96. <https://doi.org/10.1198/016214501753382309>.
- Korolev, Ivan. 2026. “Estimating Semiparametric and Nonparametric Fixed Effects Panel Data Models with MgcV.” <https://arxiv.org/abs/2606.12739>.
- Krivobokova, Tatyana, and Göran Kauermann. 2007. “A Note on Penalized Spline Smoothing with Correlated Errors.” *Journal of the American Statistical Association* 102 (480): 1328–37. <https://doi.org/10.1198/016214507000000978>.
- Liang, Kung-Yee, and Scott L. Zeger. 1986. “Longitudinal Data Analysis Using Generalized Linear Models.” *Biometrika* 73 (1): 13–22. <https://doi.org/10.1093/biomet/73.1.13>.
- Liebl, Dominik, and Matthew Reimherr. 2023. “Fast and Fair Simultaneous Confidence Bands for Functional Parameters.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 85 (3): 842–68. <https://doi.org/10.1093/jrsssb/qqad026>.
- Lin, Xihong, and Raymond J. Carroll. 2000. “Nonparametric Function Estimation for Clustered Data When the Predictor Is Measured Without/with Error.” *Journal of the American Statistical Association* 95 (450): 520–34. <https://doi.org/10.1080/01621459.2000.10474229>.
- Loewinger, Gabriel, Alex W. Levis, Erjia Cui, and Francisco Pereira. 2025. “Fast Penalized Generalized Estimating Equations for Large Longitudinal Functional Datasets.” <https://arxiv.org/abs/2506.20437v3>.
- Luo, Ruiyan, and Xin Qi. 2017. “Function-on-Function Linear Regression by Signal Compression.” *Journal of the American Statistical Association* 112 (518): 690–705. <https://doi.org/10.1080/01621459.2016.1164053>.

- MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb. 2023. "Cluster-Robust Inference: A Guide to Empirical Practice." *Journal of Econometrics* 232 (2): 272–99. <https://doi.org/10.1016/j.jeconom.2022.04.001>.
- Marra, Giampiero, and Simon N. Wood. 2012. "Coverage Properties of Confidence Intervals for Generalized Additive Model Components." *Scandinavian Journal of Statistics* 39 (1): 53–74. <https://doi.org/10.1111/j.1467-9469.2011.00760.x>.
- McCaffrey, Daniel F., and Robert M. Bell. 2006. "Improved Hypothesis Testing for Coefficients in Generalized Estimating Equations with Small Samples of Clusters." *Statistics in Medicine* 25 (23): 4081–98. <https://doi.org/10.1002/sim.2502>.
- Meyer, Mark J., Brent A. Coull, Francesco Versace, Paul Cinciripini, and Jeffrey S. Morris. 2015. "Bayesian Function-on-Function Regression for Multilevel Functional Data." *Biometrics* 71 (3): 563–74. <https://doi.org/10.1111/biom.12299>.
- Morris, Jeffrey S. 2015. "Functional Regression." *Annual Review of Statistics and Its Application* 2: 321–59. <https://doi.org/10.1146/annurev-statistics-010814-020413>.
- Morris, Jeffrey S., and Raymond J. Carroll. 2006. "Wavelet-Based Functional Mixed Models." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68 (2): 179–99. <https://doi.org/10.1111/j.1467-9868.2006.00539.x>.
- Morris, Tim P., Ian R. White, and Michael J. Crowther. 2019. "Using Simulation Studies to Evaluate Statistical Methods." *Statistics in Medicine* 38 (11): 2074–2102. <https://doi.org/10.1002/sim.8086>.
- Müller, Hans-Georg, and Fang Yao. 2008. "Functional Additive Models." *Journal of the American Statistical Association* 103 (484): 1534–44. <https://doi.org/10.1198/016214508000000751>.
- Nychka, Douglas. 1988. "Bayesian Confidence Intervals for Smoothing Splines." *Journal of the American Statistical Association* 83 (404): 1134–43. <https://doi.org/10.1080/01621459.1988.10478711>.
- Opsomer, Jean D., Yuedong Wang, and Yuhong Yang. 2001. "Nonparametric Regression with Correlated Errors." *Statistical Science* 16 (2): 134–53. <https://doi.org/10.1214/ss/1009213287>.
- Pustejovsky, James E. 2024. *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections*. <https://CRAN.R-project.org/package=clubSandwich>.
- Pustejovsky, James E., and Elizabeth Tipton. 2018. "Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models." *Journal of Business & Economic Statistics* 36 (4): 672–83. <https://doi.org/10.1080/07350015.2016.1247004>.
- Ramsay, James O., and Bernard W. Silverman. 2005. *Functional Data Analysis*. 2nd ed. Springer Series in Statistics. New York: Springer. <https://doi.org/10.1007/b98888>.
- Reiss, Philip T., Lei Huang, and Maarten Mennes. 2010. "Fast Function-on-Scalar Regression with Penalized Basis Expansions." *The International Journal of Biostatistics* 6 (1). <https://doi.org/10.2202/1557-4679.1246>.
- Rice, John A., and B. W. Silverman. 1991. "Estimating the Mean and Covariance Structure Nonparametrically When the Data Are Curves." *Journal of the Royal Statistical Society: Series B (Methodological)* 53 (1): 233–43. <https://doi.org/10.1111/j.2517-6161.1991.tb01821.x>.
- Scheipl, Fabian, Jan Gertheiss, and Sonja Greven. 2016. "Generalized Functional Additive Mixed Models." *Electronic Journal of Statistics* 10 (1): 1455–92. <https://doi.org/10.1214/16-EJS1145>.
- Scheipl, Fabian, and Sonja Greven. 2016. "Identifiability in Penalized Function-on-Function Regression Models." *Electronic Journal of Statistics* 10 (1): 495–526. <https://doi.org/10.1214/16-EJS1123>.
- Scheipl, Fabian, Ana-Maria Staicu, and Sonja Greven. 2015. "Functional Additive Mixed Models." *Journal of Computational and Graphical Statistics* 24 (2): 477–501. <https://doi.org/10.1080/10618600.2014.901914>.
- Wahba, Grace. 1983. "Bayesian 'Confidence Intervals' for the Cross-Validated Smoothing Spline." *Journal of the Royal Statistical Society: Series B (Methodological)* 45 (1): 133–50.
- Welsh, Alan H., Xihong Lin, and Raymond J. Carroll. 2002. "Marginal Longitudinal Nonparametric Regression: Locality and Efficiency of Spline and Kernel Methods." *Journal of the American Statistical*

- Association 97 (458): 482–93. <https://doi.org/10.1198/016214502760047131>.
- Wood, Simon N. 2011. “Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (1): 3–36. <https://doi.org/10.1111/j.1467-9868.2010.00749.x>.
- . 2017. *Generalized Additive Models: An Introduction with R*. 2nd ed. Chapman; Hall/CRC. <https://doi.org/10.1201/9781315370279>.
- . 2024. “On Neighbourhood Cross Validation.” <https://arxiv.org/abs/2404.16490>.
- Wood, Simon N., Natalya Pya, and Benjamin Säfken. 2016. “Smoothing Parameter and Model Selection for General Smooth Models.” *Journal of the American Statistical Association* 111 (516): 1548–63. <https://doi.org/10.1080/01621459.2016.1180986>.
- Zeileis, Achim, Susanne Köll, and Nathaniel Graham. 2020. “Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R.” *Journal of Statistical Software* 95 (1): 1–36. <https://doi.org/10.18637/jss.v095.i01>.
- Zeileis, Achim, Thomas Lumley, Nathaniel Graham, and Susanne Köll. 2024. *Sandwich: Robust Covariance Matrix Estimators*. <https://CRAN.R-project.org/package=sandwich>.

A The CL2 covariance

Notation follows the main text: $\mathbf{W}$ holds the working weights $w_{gd} = (d\mu/d\eta)_{gd}^2 / \{\phi V(\mu_{gd})\}$, $\mathbf{S} = \sum_j (\lambda_j / \phi) \mathbf{S}_j$, $\mathbf{V}_p = (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \mathbf{S})^{-1}$, $\mathbf{V}_e = \mathbf{V}_p \mathbf{X}^\top \mathbf{W} \mathbf{X} \mathbf{V}_p$, and for curve g $\tilde{\mathbf{X}}_g = \mathbf{W}_g^{1/2} \mathbf{X}_g$ and $\mathbf{z}_{gd} = (y_{gd} - \hat{\mu}_{gd}) / \{\phi V(\hat{\mu}_{gd})\}^{1/2}$. The stacked whitened design is $\tilde{\mathbf{X}}$, and $\tilde{\mathbf{H}} = \tilde{\mathbf{X}} \mathbf{V}_p \tilde{\mathbf{X}}^\top$ has blocks $\tilde{\mathbf{H}}_{gh}$.

A.1 Scores and the penalized bread

For an exponential-family response the score of the log-likelihood contribution of y_{gd} is

$$\mathbf{s}_{gd}(\theta) = \frac{(y_{gd} - \mu_{gd})(d\mu/d\eta)_{gd}}{\phi V(\mu_{gd})} \mathbf{x}_{gd} = \mathbf{z}_{gd} \tilde{\mathbf{x}}_{gd}, \quad \tilde{\mathbf{x}}_{gd} = \mathbf{w}_{gd}^{1/2} \mathbf{x}_{gd}, \quad (5)$$

so the score of curve g is $\mathbf{u}_g = \sum_d \mathbf{s}_{gd} = \tilde{\mathbf{X}}_g^\top \mathbf{z}_g$. The fit maximizes $\sum_{g,d} \ell_{gd}(\theta) - \frac{1}{2} \theta^\top \mathbf{S} \theta$, whose estimating equation is

$$\sum_{g=1}^G \mathbf{u}_g(\hat{\theta}) = \hat{\mathbf{S}} \hat{\theta}. \quad (6)$$

The negative Hessian of the penalized log-likelihood is $\mathbf{X}^\top \mathbf{W} \mathbf{X} + \mathbf{S} = \mathbf{V}_p^{-1}$ (exactly for canonical links, in expectation otherwise). Linearizing Equation 6 around the target θ^* at which its expectation vanishes gives

$$\hat{\theta} - \theta^* \approx \mathbf{V}_p \sum_{g=1}^G \{\mathbf{u}_g(\theta^*) - \mathbf{E} \mathbf{u}_g(\theta^*)\}, \quad \text{Var}(\hat{\theta}) \approx \mathbf{V}_p \left\{ \sum_{g=1}^G \text{Var} \mathbf{u}_g(\theta^*) \right\} \mathbf{V}_p, \quad (7)$$

so $\mathbf{V}_p$ is the bread and only the meat $\sum_g \text{Var} \mathbf{u}_g$ depends on the within-curve dependence. Under the working model, $\text{Var}(\mathbf{z}_g) = \mathbf{I}$ and the meat is $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$, which returns $\mathbf{V}_e$. The unpenalized bread $(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}$ linearizes a different estimator. Scores and penalty are both on the scale of the summed log-likelihood; a construction that averages the scores over the N observations must use the penalty $\sum_j \lambda_j \mathbf{S}_j / (N\phi)$, or the bread and the shrinkage are mis-scaled by the factor N (and by ϕ if it is dropped).

By Equation 6 the curve scores at the estimate do not sum to zero: $\sum_g \tilde{\mathbf{X}}_g^\top \mathbf{z}_g = \hat{\mathbf{S}}\hat{\boldsymbol{\theta}}$. Unlike the unpenalized cluster sandwich (Liang and Zeger 1986), whose curve scores at the estimate sum to zero, the meat $\sum_g \mathbf{U}_g \mathbf{U}_g^\top$ below is not centred; Proposition A.1(b) gives the consequence.

A.2 The leverage adjustment

The CL2 covariance of the main text is

$$\hat{\mathbf{V}}_{\text{CL2}} = \frac{G}{G-1} \mathbf{V}_p \left(\sum_{g=1}^G \mathbf{U}_g \mathbf{U}_g^\top \right) \mathbf{V}_p + (\mathbf{V}_p - \mathbf{V}_e), \quad \mathbf{U}_g = \tilde{\mathbf{X}}_g^\top \mathbf{R}_g^{-1/2} \mathbf{z}_g, \quad (8)$$

with $\mathbf{R}_g = [(\mathbf{I} - \tilde{\mathbf{H}})(\mathbf{I} - \tilde{\mathbf{H}})^\top]_{gg}$. Since $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} = \mathbf{V}_p^{-1} - \mathbf{S}$, $(\mathbf{I} - \tilde{\mathbf{H}})^2 = \mathbf{I} - \tilde{\mathbf{X}}(2\mathbf{V}_p - \mathbf{V}_p \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{V}_p) \tilde{\mathbf{X}}^\top$ and

$$\mathbf{R}_g = \mathbf{I} - \tilde{\mathbf{X}}_g (2\mathbf{V}_p - \mathbf{V}_p \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{V}_p) \tilde{\mathbf{X}}_g^\top = \mathbf{I} - \tilde{\mathbf{H}}_{gg} - \tilde{\mathbf{X}}_g \mathbf{V}_p \mathbf{S} \mathbf{V}_p \tilde{\mathbf{X}}_g^\top. \quad (9)$$

$\mathbf{R}_g$ needs only $\tilde{\mathbf{X}}_g$ and one $p \times p$ matrix. It is computed in the column space of $\tilde{\mathbf{X}}_g$ in $\mathbb{R}^{D_g}$ (dimension at most $\min(D_g, p)$), because $\mathbf{R}_g$ equals $\mathbf{I}$ on its orthogonal complement and $\tilde{\mathbf{X}}_g^\top$ annihilates that complement.

Relation to CR2. The bias-reduced linearization of Bell and McCaffrey (2002) chooses $\mathbf{A}_g$ with $\mathbf{A}_g[(\mathbf{I} - \mathbf{H})\Phi(\mathbf{I} - \mathbf{H})^\top]_{gg} \mathbf{A}_g = \Phi_g$ for a working covariance Φ (McCaffrey and Bell 2006; Pustejovsky and Tipton 2018). With $\Phi = \mathbf{I}$ this is $\mathbf{A}_g = \mathbf{R}_g^{-1/2}$. For a projection, $\mathbf{R}_g = \mathbf{I} - \mathbf{H}_{gg}$ and the usual CR2 results. The penalized hat is not idempotent: the last term of Equation 9 is positive semidefinite, so $\mathbf{R}_g \preceq \mathbf{I} - \tilde{\mathbf{H}}_{gg}$, and the approximation $\mathbf{A}_g^s = (\mathbf{I} - \tilde{\mathbf{H}}_{gg})^{-1/2}$, which keeps the full block $\tilde{\mathbf{H}}_{gg}$ but drops the last term of Equation 9 (the *shortcut*), under-corrects: $\mathbf{A}_g^s \mathbf{R}_g \mathbf{A}_g^s \preceq \mathbf{I}$. The dropped term is small when the per-curve leverage is small.

Expectation of the meat. Take the Gaussian working model with fixed λ and fixed design: $\mathbf{W}^{1/2} \mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ with $\text{Var}(\boldsymbol{\varepsilon}) = \mathbf{I}$. Then $\mathbf{z} = (\mathbf{I} - \tilde{\mathbf{H}})(\tilde{\mathbf{X}}\boldsymbol{\theta} + \boldsymbol{\varepsilon})$, and $(\mathbf{I} - \tilde{\mathbf{H}})\tilde{\mathbf{X}} = \tilde{\mathbf{X}}\mathbf{V}_p\mathbf{S}$ gives

$$\mathbb{E}(\mathbf{z}_g) = \mathbf{m}_g = \tilde{\mathbf{X}}_g \mathbf{V}_p \mathbf{S} \boldsymbol{\theta}, \quad \mathbb{E}(\mathbf{z}_g \mathbf{z}_g^\top) = \mathbf{R}_g + \mathbf{m}_g \mathbf{m}_g^\top. \quad (10)$$

The penalized residuals have nonzero mean unless $\mathbf{S}\boldsymbol{\theta} = \mathbf{0}$. With $\mathbf{R}_g$ invertible and the floor inactive,

$$\mathbb{E}\left(\sum_g \mathbf{U}_g \mathbf{U}_g^\top\right) = \sum_g \tilde{\mathbf{X}}_g^\top \tilde{\mathbf{X}}_g + \sum_g \tilde{\mathbf{X}}_g^\top \mathbf{R}_g^{-1/2} \mathbf{m}_g \mathbf{m}_g^\top \mathbf{R}_g^{-1/2} \tilde{\mathbf{X}}_g. \quad (11)$$

The first term is the working-model sampling meat, $\text{Var}(\tilde{\mathbf{X}}^\top \boldsymbol{\varepsilon})$, so $\mathbf{V}_p(\sum_g \mathbf{U}_g \mathbf{U}_g^\top) \mathbf{V}_p$ is unbiased for $\mathbf{V}_e$ when the truth lies in the null space of the penalty and biased upwards by a positive semidefinite term otherwise. The factor $G/(G-1)$ adds a further upward bias. Under within-curve dependence, $\text{Var}(\boldsymbol{\varepsilon}_g) = \Sigma_g \neq \mathbf{I}$, and the adjustment removes the residual shrinkage only approximately; it does not whiten correlated residuals.

Floor and factor. The eigenvalues of $\mathbf{R}_g$ are floored at 10^{-6} before the inverse square root, which caps the inflation of any residual direction at a factor 10^3 . The floor binds only for curves whose data nearly determine a coefficient direction on their own; how often it bound in the studies was not recorded. CR2 carries no factor $G/(G-1)$; we keep it. It inflates frequentist CL2 standard errors by 1.3% at $G = 40$ and 0.5% at $G = 100$; in the Bayesian form it multiplies only the sandwich part, so these are upper bounds. CR1 is Equation 8 with $\mathbf{R}_g = \mathbf{I}$ and no further degrees-of-freedom factor.

A.3 Bayesian and frequentist forms

The frequentist form omits $\mathbf{V}_p - \mathbf{V}_e = \mathbf{V}_p \mathbf{S} \mathbf{V}_p$. For Gaussian responses with identity link and fixed λ , the bias of $\hat{\theta}$ is $-\mathbf{V}_p \mathbf{S} \theta$. Under the smoothing prior $\theta \sim N(0, \mathbf{S}^-)$, $E(\mathbf{V}_p \mathbf{S} \theta \theta^\top \mathbf{S} \mathbf{V}_p) = \mathbf{V}_p \mathbf{S} \mathbf{V}_p$: the allowance is the prior-averaged squared bias, and $\mathbf{V}_p$ the prior-averaged mean squared error matrix (Wahba 1983; Nychka 1988; Marra and Wood 2012). For a fixed truth and any contrast $\mathbf{a}$, by the Cauchy–Schwarz inequality,

$$(\mathbf{a}^\top \mathbf{V}_p \mathbf{S} \theta)^2 \leq (\mathbf{a}^\top \mathbf{V}_p \mathbf{S} \mathbf{V}_p \mathbf{a}) (\theta^\top \mathbf{S} \theta), \quad \theta^\top \mathbf{S} \theta = \sum_j (\lambda_j / \phi) \theta^\top \mathbf{S}_j \theta. \quad (12)$$

The allowance bounds the squared bias of every contrast if $\theta^\top \mathbf{S} \theta \leq 1$; for larger values the bound permits, but does not imply, a bias above the allowance. Where the data identify θ ($\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$ positive definite), the squared bias scales as λ^2 and the allowance as λ for small λ . Undersmoothing, as REML produces under within-curve dependence, therefore makes the allowance large relative to the bias but small relative to the dependence-inflated variance, and the two forms nearly coincide. For a truth that is rough at the selected λ ($\theta^\top \mathbf{S} \theta$ large), or after oversmoothing, the bias can exceed the allowance, most likely where the truth is most curved. Under independent errors with a smooth truth the allowance exceeds the bias and the Bayesian form overcovers.

A.4 Extended families

The CL2 construction needs a factorization $\mathbf{s}_{gd} = z_{gd} \tilde{\mathbf{x}}_{gd}$ in which $\tilde{\mathbf{x}}_{gd}$ carries the Fisher weight of the fit, so that $\text{tr} \tilde{\mathbf{H}}$ equals the effective degrees of freedom.

Gaussian location–scale (mgcv’s gaulss). The two linear predictors η_1 (mean) and η_2 (inverse standard deviation τ) have block-diagonal Fisher information with weights $W_1 = \tau^2 (d\mu/d\eta_1)^2$ and $W_2 = 2(d\tau/d\eta_2)^2/\tau^2$. Each observation contributes two pseudo-rows, both assigned to its curve,

$$\tilde{\mathbf{x}}_k = W_k^{1/2} \mathbf{x}_k, \quad z_k = \frac{\partial \ell / \partial \eta_k}{W_k^{1/2}}, \quad k = 1, 2, \quad (13)$$

with $\partial \ell / \partial \eta_1 = \tau^2 (y - \mu) d\mu/d\eta_1$ and $\partial \ell / \partial \eta_2 = (1/\tau - \tau(y - \mu)^2) d\tau/d\eta_2$; $\mathbf{x}_k$ is the design row of predictor k , zero in the other block.

Scaled t (scat). With degrees of freedom ν and scale σ at their estimates, the exact location score and Fisher weight are

$$\frac{\partial \ell}{\partial \mu} = \frac{(\nu + 1)(y - \mu)}{\nu \sigma^2 + (y - \mu)^2}, \quad W = \left(\frac{d\mu}{d\eta} \right)^2 \frac{\nu + 1}{(\nu + 3)\sigma^2}, \quad (14)$$

and $z = (\partial \ell / \partial \mu)(d\mu/d\eta)/W^{1/2}$. The working-residual score Equation 5 would treat the t response as Gaussian and give no downweighting of outliers.

All other extended families of mgcv (negative binomial, Tweedie, beta, ordered categorical) use the working-residual score Equation 5 with the fitted variance function. For the negative binomial, with its shape parameter at the estimate, this is the exact score for the mean; for the beta family it is an approximation. The simulations use CL2 for the beta and negative-binomial responses. Families that define their own sandwich (multinomial) have no cluster-robust covariance.

A.5 Asymptotic validity for fixed smoothing parameters

Proposition A.1. Assume (i) the curves (y_g, X_g) , $g = 1, \dots, G$, are independent and identically distributed with $D_g \leq \bar{D} < \infty$, with arbitrary dependence within a curve; (ii) the basis, hence p , is fixed and the link is canonical; (iii) $\lambda = \lambda_G$ is non-random with $S_G/G \rightarrow S_\infty$; (iv) the weighted design rows are uniformly bounded, the curve scores have bounded moments of order $2 + \delta$ and dominated derivatives near the target, and $J = E(\tilde{X}_1^\top \tilde{X}_1) + S_\infty$ is positive definite at $\theta_\infty^* = \lim_G \theta_G^*$, where θ_G^* solves

$$E \mathbf{u}_1(\theta) = (S_G/G) \theta. \quad (15)$$

Then (a) $\sqrt{G}(\hat{\theta} - \theta_G^*) \rightarrow N(\mathbf{0}, J^{-1} M J^{-1})$ with $M = \text{Var} \mathbf{u}_1(\theta_\infty^*)$. (b) With $\mu = E \mathbf{u}_1(\theta_\infty^*) = S_\infty \theta_\infty^*$, the variance part of Equation 8 (and of CR1) satisfies

$$G \frac{G}{G-1} \mathbf{v}_p \left(\sum_g \mathbf{U}_g \mathbf{U}_g^\top \right) \mathbf{v}_p \rightarrow_p J^{-1} (M + \mu \mu^\top) J^{-1}, \quad G(\mathbf{V}_p - \mathbf{V}_e) \rightarrow J^{-1} S_\infty J^{-1}, \quad (16)$$

so the variance part is consistent for the covariance in (a) if $S_\infty \theta_\infty^* = \mathbf{0}$ and conservative otherwise; the Bayesian form adds $J^{-1} S_\infty J^{-1}$. (c) If the mean model is correct and $\sqrt{G}(\theta_G^* - \theta) \rightarrow \mathbf{0}$ (which forces $S_\infty \theta = \mathbf{0}$ even if $S_\infty \neq \mathbf{0}$), the frequentist CL2 interval for any fixed contrast is asymptotically valid for the true θ , and the Bayesian one is valid and conservative, asymptotically exact if $S_\infty = \mathbf{0}$.

For Gaussian responses with identity link, the finite-sample target given the design is $E(\hat{\theta}) = (\mathbf{I} - \mathbf{V}_p \mathbf{S})\theta$, of which Equation 15 is the population version; for GLMs the target solves the expected penalized score equation.

Proof. $\hat{\theta}$ solves an M-estimating equation with independent curve contributions and a deterministic penalty offset whose expectation vanishes at θ_G^* ; (a) follows from the standard argument for such equations (Liang and Zeger 1986; Hansen and Lee 2019). For (b), bounded curve sizes and fixed p make the leverages $O_p(G^{-1})$, so CL2 and CR1 meats agree to first order, and the meat converges to the second moment of a score with mean μ . For (c), a correct mean model gives curve scores with mean zero whatever the dependence within the curve. $\square$

The proposition does not cover smoothing parameters selected from the data, nor a number of coefficients comparable to or larger than G ; both hold in our applications and simulations.

A.6 Covariance components

The comparison uses the 18 main factorial settings at the default signal level ($D = 61$, $G = 40$ or 100 , 200 replicates per setting) with common random numbers. Except for the fixed- λ interval, every covariance is computed for the same REML fit of each replicate; these fits reproduce the stored fits of the main study to numerical precision. The intervals are the model-based interval ($\mathbf{V}_e$); CR1, CL2 with the $\mathbf{I} - \tilde{\mathbf{H}}_{gg}$ approximation, and CL2, all in the Bayesian form of Equation 8; CL2 in the frequentist form; CL2 with the critical value $t_{G-1, 0.975}$ in place of $z_{0.975}$; and CL2 after a refit of the same data with the smoothing parameters fixed, per penalty, at the geometric mean of the setting's REML estimates. The fixed- λ refit removes the replicate-to-replicate variability of the smoothing parameters and also moves them to their typical level. Table 6, Table 7 and Table 9 give the coverage, Table 8 the paired differences.

Under dependence the leverage adjustment adds 3.7 pp of coverage over CR1 at $G = 40$ and 1.2 pp at $G = 100$; the full block adds at most 0.4 pp in any setting over the $\mathbf{I} - \tilde{\mathbf{H}}_{gg}$ approximation (under dependence 0.1–0.4 pp), in line with $\mathbf{A}_g^s \mathbf{R}_g \mathbf{A}_g^s \leq \mathbf{I}$. The allowance $\mathbf{V}_p - \mathbf{V}_e$ matters under independent errors (6.0 pp at

Table 6: Covariance components, $\beta(s, t)$: grid-averaged coverage of nominal 95% pointwise intervals by family, error process and number of curves G , main factorial settings at the default signal level, 200 replicates. Model: model-based ($\mathbf{V}_c$); CR1, CL2 $I - H$ (CL2 with the $\mathbf{I} - \tilde{\mathbf{H}}_{gg}$ approximation) and CL2 in the Bayesian form; CL2 freq.: frequentist form; CL2 t_{G-1} : CL2 with t critical value; CL2 fixed λ : CL2 after a refit with smoothing parameters fixed at the setting's geometric-mean REML values.

family	error	G	model	CR1	CL2 $I - H$	CL2	CL2 freq.	CL2 t_{G-1}	CL2 fixed λ
Gaussian	independent	40	0.992	0.983	0.988	0.989	0.930	0.991	0.989
		100	0.989	0.983	0.985	0.986	0.940	0.987	0.986
	rough stationary	40	0.638	0.889	0.924	0.927	0.915	0.935	0.939
		100	0.672	0.929	0.940	0.940	0.931	0.944	0.949
	smooth	40	0.615	0.886	0.923	0.926	0.915	0.934	0.940
		100	0.657	0.931	0.943	0.944	0.935	0.947	0.951
Poisson	independent	40	0.993	0.984	0.988	0.989	0.928	0.991	0.990
		100	0.989	0.983	0.985	0.986	0.937	0.987	0.986
	rough stationary	40	0.684	0.887	0.924	0.927	0.911	0.935	0.942
		100	0.698	0.928	0.941	0.942	0.930	0.945	0.950
	smooth	40	0.658	0.886	0.925	0.928	0.914	0.936	0.943
		100	0.682	0.931	0.944	0.945	0.935	0.948	0.953
binary	independent	40	0.984	0.969	0.972	0.973	0.869	0.976	0.990
		100	0.994	0.989	0.990	0.990	0.917	0.991	0.993
	rough stationary	40	0.800	0.906	0.934	0.937	0.907	0.945	0.958
		100	0.819	0.941	0.950	0.951	0.929	0.954	0.961
	smooth	40	0.749	0.889	0.922	0.926	0.901	0.934	0.947
		100	0.779	0.941	0.951	0.952	0.933	0.954	0.964

Table 7: Covariance components, mean $E(Y | X)$: grid-averaged coverage of nominal 95% pointwise intervals, intervals and settings as for $\beta(s, t)$.

family	error	G	model	CR1	CL2 $I - H$	CL2	CL2 freq.	CL2 t_{G-1}	CL2 fixed λ
Gaussian	independent	40	0.968	0.943	0.956	0.958	0.926	0.964	0.960
		100	0.967	0.954	0.960	0.961	0.939	0.963	0.962
	rough stationary	40	0.624	0.895	0.928	0.931	0.925	0.939	0.937
		100	0.646	0.930	0.941	0.942	0.937	0.945	0.946
	smooth	40	0.606	0.895	0.929	0.931	0.926	0.939	0.939
		100	0.634	0.932	0.943	0.944	0.940	0.947	0.948
Poisson	independent	40	0.968	0.942	0.956	0.958	0.922	0.964	0.960
		100	0.966	0.953	0.959	0.960	0.935	0.962	0.961
	rough stationary	40	0.657	0.893	0.927	0.930	0.922	0.938	0.939
		100	0.664	0.929	0.942	0.943	0.937	0.946	0.947
	smooth	40	0.642	0.896	0.930	0.934	0.926	0.941	0.942
		100	0.653	0.932	0.944	0.945	0.940	0.948	0.950
binary	independent	40	0.952	0.911	0.923	0.925	0.857	0.933	0.943
		100	0.963	0.944	0.949	0.949	0.898	0.952	0.960
	rough stationary	40	0.747	0.894	0.922	0.927	0.909	0.935	0.946
		100	0.752	0.929	0.938	0.940	0.927	0.943	0.952
	smooth	40	0.704	0.884	0.917	0.921	0.906	0.929	0.941
		100	0.722	0.929	0.939	0.941	0.930	0.944	0.952

Table 8: Covariance components: paired coverage differences in percentage points, averaged over families and over the mean and $\beta(s, t)$, by error process and number of curves G ; in brackets the range of the per-setting paired Monte Carlo SEs (each from the replicate-level differences of one setting and estimand). Each difference compares two intervals on the same replicates; unless stated, both are CL2 in the Bayesian form with z critical value. The SE of the averages cannot be recovered from the retained per-setting summaries, because the settings share replicates.

contrast	independent		dependent	
	$G = 40$	$G = 100$	$G = 40$	$G = 100$
CL2 minus CR1	+1.0 (0.03–0.05)	+0.4 (0.01–0.03)	+3.7 (0.09–0.19)	+1.2 (0.04–0.07)
CL2 minus CL2 with $I - \tilde{H}_{gg}$	+0.1 (0.01–0.01)	+0.1 (0.00–0.01)	+0.3 (0.01–0.02)	+0.1 (0.01–0.01)
Bayesian minus frequentist, CL2	+6.0 (0.09–0.45)	+4.4 (0.06–0.32)	+1.4 (0.02–0.12)	+1.0 (0.02–0.10)
Bayesian minus frequentist, CR1	+8.1 (0.10–0.50)	+5.2 (0.06–0.35)	+2.1 (0.02–0.14)	+1.3 (0.02–0.10)
t_{G-1} minus z	+0.4 (0.02–0.03)	+0.2 (0.01–0.01)	+0.8 (0.02–0.04)	+0.3 (0.01–0.02)
fixed minus reselected λ , CL2	+0.7 (0.01–0.48)	+0.3 (0.01–0.11)	+1.4 (0.08–0.32)	+0.8 (0.05–0.16)
fixed minus reselected λ , model-based	-0.1 (0.01–0.40)	-0.1 (0.01–0.08)	-0.3 (0.17–0.52)	-0.6 (0.10–0.43)

Table 9: Covariance components, intercept $\alpha(t)$ and coefficient function $\gamma(t)$: grid-averaged coverage of nominal 95% pointwise intervals, settings as for $\beta(s, t)$; model-based ($\mathbf{V}_c$), CR1 and CL2 (Bayesian form), and CL2 after a refit with smoothing parameters fixed at the setting’s geometric-mean REML values.

family	error	G	$\alpha(t)$				$\gamma(t)$			
			model	CR1	CL2	CL2 fixed λ	model	CR1	CL2	CL2 fixed λ
Gaussian	independent	40	0.956	0.932	0.947	0.950	0.959	0.927	0.951	0.953
		100	0.956	0.946	0.956	0.958	0.960	0.942	0.950	0.951
	rough stationary	40	0.632	0.909	0.940	0.945	0.653	0.906	0.938	0.943
		100	0.659	0.940	0.951	0.952	0.663	0.925	0.938	0.940
	smooth	40	0.614	0.906	0.937	0.941	0.621	0.910	0.946	0.950
		100	0.641	0.941	0.950	0.951	0.655	0.938	0.948	0.951
Poisson	independent	40	0.960	0.942	0.953	0.957	0.955	0.924	0.948	0.950
		100	0.954	0.945	0.951	0.953	0.963	0.945	0.955	0.957
	rough stationary	40	0.655	0.910	0.935	0.940	0.667	0.900	0.937	0.945
		100	0.668	0.943	0.950	0.951	0.677	0.921	0.937	0.939
	smooth	40	0.641	0.907	0.933	0.938	0.647	0.903	0.943	0.948
		100	0.665	0.940	0.948	0.951	0.662	0.933	0.945	0.948
binary	independent	40	0.903	0.845	0.859	0.867	0.931	0.872	0.890	0.929
		100	0.943	0.919	0.924	0.939	0.945	0.914	0.923	0.949
	rough stationary	40	0.712	0.870	0.904	0.929	0.724	0.880	0.920	0.951
		100	0.749	0.919	0.926	0.942	0.740	0.912	0.924	0.946
	smooth	40	0.668	0.855	0.885	0.925	0.691	0.874	0.922	0.950
		100	0.702	0.910	0.920	0.936	0.730	0.925	0.939	0.953

$G = 40$), much less under dependence (1.4 pp). With reselected smoothing parameters, the grid average of CL2’s root-mean-square standard error divided by the empirical standard deviation of $\hat{\beta}$ under dependence is 0.92 at $G = 40$ and 0.97 at $G = 100$; with fixed smoothing parameters it is 1.05 and 1.06, and CL2 coverage of β rises from 0.929 to 0.945 at $G = 40$. Fixing the smoothing parameters, which Equation 8 conditions on, at their typical level closes most of the remaining shortfall at $G = 40$; the t_{G-1} critical value recovers 0.8 pp of it. Fixed smoothing parameters also raise CL2 coverage of $\alpha(t)$ and $\gamma(t)$ under dependence (Table 9): for binary responses with smooth errors at $G = 40$, CL2 coverage of $\alpha(t)$ rises from 0.885 to 0.925 and that of $\gamma(t)$ from 0.922 to 0.950. Under independent errors the binary $\alpha(t)$ at $G = 40$ stays below nominal with fixed smoothing parameters (0.867).

A.7 Satterthwaite degrees of freedom

For a contrast $\mathbf{a}^\top \theta$ the sandwich part of Equation 8 is a quadratic form in the whitened residuals, $\mathbf{a}^\top \mathbf{V}_p (\sum_g \mathbf{U}_g \mathbf{U}_g^\top) \mathbf{V}_p \mathbf{a} = \sum_g (\mathbf{q}_g^\top \mathbf{z}_g)^2$ with $\mathbf{q}_g = \mathbf{R}_g^{-1/2} \tilde{\mathbf{X}}_g \mathbf{V}_p \mathbf{a}$ (times $G/(G-1)$). Under the Gaussian working model with fixed smoothing parameters, $\mathbf{z} = (\mathbf{I} - \tilde{\mathbf{H}})\boldsymbol{\varepsilon}$ with independent errors. The quadratic form then has mean $\text{tr } \Gamma$ and variance $2 \text{tr}(\Gamma^2)$, where the $G \times G$ matrix Γ has entries $\Gamma_{gh} = \mathbf{q}_g^\top [(\mathbf{I} - \tilde{\mathbf{H}})(\mathbf{I} - \tilde{\mathbf{H}})^\top]_{gh} \mathbf{q}_h$. Matching a scaled χ^2_ν gives the Satterthwaite degrees of freedom

$$\nu(\mathbf{a}) = \frac{(\text{tr } \Gamma)^2}{\text{tr}(\Gamma^2)} \quad (17)$$

(Bell and McCaffrey 2002; Pustejovsky and Tipton 2018), computed for every evaluation point; the interval uses $t_{v,0.975}$ in place of $z_{0.975}$. The Satterthwaite intervals evaluated here compute Equation 17 for the sandwich part only. A variant matches the moments of the full Bayesian CL2 variance, whose term $\mathbf{a}^\top(\mathbf{V}_p - \mathbf{V}_e)\mathbf{a}$ is fixed, which increases v . Using only the diagonal of Γ gives practically the same v .

The Satterthwaite intervals use the REML fits and CL2 covariances of Appendix A.6 (the 18 main factorial settings at the default signal level, 200 replicates); Section 4.1, Section 5.2 and Supplement S1.2, S1.7 and S2.1 give them with 20 curves, in the extensions and on the plasmode data. The median per-point degrees of freedom are 12–30 with $G = 40$ curves and 30–87 with $G = 100$, between 0.30–0.88 times $G - 1$. Under dependent errors, Satterthwaite intervals cover the mean, β and γ at 0.95–0.96 with 40 curves (normal critical value: 0.92–0.94) and at 0.94–0.96 with 100 (0.93–0.95), at 1.10–1.12 and 1.04–1.05 times the width (next table). The binary intercept is not repaired: 0.91–0.93 under dependence and 0.87–0.93 under independent errors. For this term the Satterthwaite degrees of freedom are large (the intervals widen by only 1.01–1.05 times), so the reference distribution cannot make up for a standard error that is too small under independent errors or for the additional shift of the studentized error under dependence (Section 4.1, Supplement S1.4). Under independent errors Satterthwaite intervals overcover β further (0.98–0.99). The variant whose degrees of freedom include $\mathbf{V}_p - \mathbf{V}_e$ lies between t_{G-1} and the plain Satterthwaite reference.

Table 10: Critical values for REML + CL2: grid-averaged coverage (5% quantile of pointwise coverage) and width relative to the normal critical value, by dependence, number of curves G , family and estimand; means over the error processes of the main factorial at the default signal level, 200 replicates. Satterthwaite: per-point degrees of freedom for the sandwich part; incl. $\mathbf{V}_p - \mathbf{V}_e$: degrees of freedom for the full Bayesian CL2 variance.

dependence	G	family	estimand	z	t(G-1)	Satterthwaite	Satterthwaite incl. $\mathbf{V}_p - \mathbf{V}_e$
independent	40	Gaussian	E(Y X)	0.958 (0.93) 1.00	0.964 (0.94) 1.03	0.973 (0.95) 1.10	0.968 (0.94) 1.06
independent	40	Gaussian	$\alpha(t)$	0.947 (0.91) 1.00	0.953 (0.92) 1.03	0.956 (0.92) 1.05	0.954 (0.92) 1.03
independent	40	Gaussian	$\beta(s,t)$	0.989 (0.97) 1.00	0.991 (0.97) 1.03	0.995 (0.98) 1.10	0.991 (0.97) 1.03
independent	40	Gaussian	$\gamma(t)$	0.951 (0.92) 1.00	0.959 (0.93) 1.03	0.969 (0.94) 1.11	0.963 (0.94) 1.07
independent	40	Poisson	E(Y X)	0.958 (0.93) 1.00	0.964 (0.94) 1.03	0.974 (0.94) 1.12	0.967 (0.94) 1.07
independent	40	Poisson	$\alpha(t)$	0.953 (0.93) 1.00	0.958 (0.93) 1.03	0.961 (0.94) 1.05	0.957 (0.93) 1.03
independent	40	Poisson	$\beta(s,t)$	0.989 (0.96) 1.00	0.991 (0.97) 1.03	0.995 (0.98) 1.11	0.991 (0.97) 1.03
independent	40	Poisson	$\gamma(t)$	0.948 (0.92) 1.00	0.953 (0.92) 1.03	0.967 (0.94) 1.12	0.961 (0.93) 1.07
independent	40	binary	E(Y X)	0.925 (0.82) 1.00	0.933 (0.84) 1.03	0.946 (0.86) 1.10	0.935 (0.84) 1.05
independent	40	binary	$\alpha(t)$	0.859 (0.67) 1.00	0.868 (0.68) 1.03	0.872 (0.69) 1.04	0.865 (0.68) 1.02
independent	40	binary	$\beta(s,t)$	0.973 (0.92) 1.00	0.976 (0.92) 1.03	0.982 (0.94) 1.10	0.975 (0.92) 1.02
independent	40	binary	$\gamma(t)$	0.890 (0.79) 1.00	0.900 (0.80) 1.03	0.918 (0.83) 1.11	0.904 (0.81) 1.05
independent	100	Gaussian	E(Y X)	0.961 (0.93) 1.00	0.963 (0.94) 1.01	0.967 (0.94) 1.04	0.965 (0.94) 1.02
independent	100	Gaussian	$\alpha(t)$	0.956 (0.93) 1.00	0.959 (0.94) 1.01	0.959 (0.94) 1.01	0.959 (0.94) 1.01
independent	100	Gaussian	$\beta(s,t)$	0.986 (0.97) 1.00	0.987 (0.97) 1.01	0.989 (0.97) 1.04	0.987 (0.97) 1.01
independent	100	Gaussian	$\gamma(t)$	0.950 (0.92) 1.00	0.953 (0.93) 1.01	0.959 (0.93) 1.04	0.957 (0.93) 1.03
independent	100	Poisson	E(Y X)	0.960 (0.93) 1.00	0.962 (0.94) 1.01	0.967 (0.94) 1.04	0.965 (0.94) 1.03
independent	100	Poisson	$\alpha(t)$	0.951 (0.92) 1.00	0.953 (0.92) 1.01	0.954 (0.92) 1.01	0.953 (0.92) 1.01
independent	100	Poisson	$\beta(s,t)$	0.986 (0.96) 1.00	0.987 (0.97) 1.01	0.989 (0.97) 1.04	0.987 (0.97) 1.01
independent	100	Poisson	$\gamma(t)$	0.955 (0.92) 1.00	0.958 (0.93) 1.01	0.965 (0.93) 1.05	0.962 (0.93) 1.03
independent	100	binary	E(Y X)	0.949 (0.89) 1.00	0.952 (0.89) 1.01	0.957 (0.90) 1.04	0.953 (0.89) 1.02
independent	100	binary	$\alpha(t)$	0.924 (0.78) 1.00	0.926 (0.78) 1.01	0.927 (0.78) 1.01	0.926 (0.78) 1.01
independent	100	binary	$\beta(s,t)$	0.990 (0.96) 1.00	0.991 (0.96) 1.01	0.993 (0.97) 1.04	0.991 (0.96) 1.01
independent	100	binary	$\gamma(t)$	0.923 (0.86) 1.00	0.927 (0.86) 1.01	0.932 (0.88) 1.04	0.928 (0.86) 1.02
dependent	40	Gaussian	E(Y X)	0.931 (0.90) 1.00	0.939 (0.91) 1.03	0.953 (0.93) 1.10	0.951 (0.93) 1.09
dependent	40	Gaussian	$\alpha(t)$	0.938 (0.91) 1.00	0.947 (0.92) 1.03	0.952 (0.93) 1.05	0.952 (0.93) 1.05
dependent	40	Gaussian	$\beta(s,t)$	0.926 (0.90) 1.00	0.934 (0.90) 1.03	0.950 (0.92) 1.10	0.946 (0.92) 1.09
dependent	40	Gaussian	$\gamma(t)$	0.942 (0.92) 1.00	0.948 (0.93) 1.03	0.961 (0.94) 1.11	0.959 (0.94) 1.10
dependent	40	Poisson	E(Y X)	0.932 (0.90) 1.00	0.940 (0.91) 1.03	0.954 (0.93) 1.12	0.951 (0.93) 1.11
dependent	40	Poisson	$\alpha(t)$	0.934 (0.91) 1.00	0.942 (0.92) 1.03	0.945 (0.92) 1.05	0.944 (0.92) 1.05
dependent	40	Poisson	$\beta(s,t)$	0.928 (0.90) 1.00	0.936 (0.91) 1.03	0.953 (0.93) 1.11	0.948 (0.92) 1.09
dependent	40	Poisson	$\gamma(t)$	0.940 (0.91) 1.00	0.947 (0.92) 1.03	0.963 (0.94) 1.12	0.961 (0.94) 1.11
dependent	40	binary	E(Y X)	0.924 (0.89) 1.00	0.932 (0.90) 1.03	0.946 (0.92) 1.10	0.942 (0.91) 1.08
dependent	40	binary	$\alpha(t)$	0.894 (0.79) 1.00	0.904 (0.81) 1.03	0.909 (0.81) 1.05	0.907 (0.81) 1.04
dependent	40	binary	$\beta(s,t)$	0.932 (0.90) 1.00	0.939 (0.91) 1.03	0.954 (0.93) 1.11	0.948 (0.92) 1.08
dependent	40	binary	$\gamma(t)$	0.921 (0.89) 1.00	0.931 (0.90) 1.03	0.945 (0.92) 1.11	0.942 (0.92) 1.09
dependent	100	Gaussian	E(Y X)	0.943 (0.91) 1.00	0.946 (0.92) 1.01	0.951 (0.92) 1.04	0.951 (0.92) 1.03
dependent	100	Gaussian	$\alpha(t)$	0.951 (0.93) 1.00	0.953 (0.94) 1.01	0.953 (0.94) 1.02	0.953 (0.94) 1.01
dependent	100	Gaussian	$\beta(s,t)$	0.942 (0.92) 1.00	0.945 (0.92) 1.01	0.951 (0.92) 1.04	0.950 (0.92) 1.03
dependent	100	Gaussian	$\gamma(t)$	0.943 (0.92) 1.00	0.947 (0.92) 1.01	0.953 (0.93) 1.04	0.952 (0.93) 1.04

Table 10: Critical values for REML + CL2: grid-averaged coverage (5% quantile of pointwise coverage) and width relative to the normal critical value, by dependence, number of curves G , family and estimand; means over the error processes of the main factorial at the default signal level, 200 replicates. Satterthwaite: per-point degrees of freedom for the sandwich part; incl. Vp - Ve: degrees of freedom for the full Bayesian CL2 variance. (*continued*)

dependence	G	family	estimand	z	t(G-1)	Satterthwaite	Satterthwaite incl. Vp - Ve
dependent	100	Poisson	$E(Y X)$	0.944 (0.92) 1.00	0.947 (0.92) 1.01	0.953 (0.93) 1.05	0.952 (0.93) 1.04
dependent	100	Poisson	$\alpha(t)$	0.949 (0.93) 1.00	0.952 (0.93) 1.01	0.952 (0.93) 1.01	0.952 (0.93) 1.01
dependent	100	Poisson	$\beta(s,t)$	0.943 (0.92) 1.00	0.946 (0.92) 1.01	0.953 (0.93) 1.04	0.952 (0.93) 1.04
dependent	100	Poisson	$\gamma(t)$	0.941 (0.91) 1.00	0.943 (0.92) 1.01	0.951 (0.93) 1.05	0.951 (0.93) 1.04
dependent	100	binary	$E(Y X)$	0.940 (0.91) 1.00	0.943 (0.92) 1.01	0.949 (0.92) 1.04	0.947 (0.92) 1.03
dependent	100	binary	$\alpha(t)$	0.923 (0.87) 1.00	0.927 (0.87) 1.01	0.928 (0.87) 1.01	0.927 (0.87) 1.01
dependent	100	binary	$\beta(s,t)$	0.951 (0.93) 1.00	0.954 (0.94) 1.01	0.960 (0.94) 1.04	0.957 (0.94) 1.03
dependent	100	binary	$\gamma(t)$	0.931 (0.90) 1.00	0.935 (0.91) 1.01	0.941 (0.92) 1.04	0.939 (0.91) 1.03

B Set-ups of the two studies

B.1 Simulation design

The data follow model Equation 1 with $i = 1, \dots, G$ curves. Responses are generated on a grid of 241 points in $t \in [0, 1]$ and subsampled to $D = 61$ (default) or 241 points; the functional covariate is observed on 51 points in $s \in [0, 1]$: $X_i(s) = \xi_{i0} + \sum_{j=1}^{29} \xi_{ij} \sqrt{2} \cos(j\pi s)$ with independent $\xi_{ij} \sim N(0, (j+1)^{-2})$. Its centred spectrum (first principal component 62% of the variance, four components 88%, rank 24 at 99.5%) matches the corpus-callosum profiles of the baseline DTI data in refund (67%, 86%, 26). The low-rank variant uses variances $(j+1)^{-3}$ (rank 9). The scalar covariate is $z_i \sim N(0, 1)$. The smooth-effect model replaces the functional term by $f(x_i, t)$ with $x_i \sim U(0, 1)$.

Truths are fixed functions projected by least squares onto the fitted bases, so they are exactly representable: $\beta(s, t) = \cos(\pi s) \sin(\pi t) + (s - 1/2) \cos(2\pi t)$, $f(x, t) = 2(x - 1/2) \cos(\pi t) + t \sin(\pi x)$, $\gamma(t) = \sin(2\pi t)$. The rough truth adds $0.6 \sin(3\pi s) \sin(3\pi t)$ to β and f and $0.5 \sin(5\pi t)$ to γ . Truths are scaled so that the non-intercept signal $S = \int \text{Var}_{X,z} \{\eta(t) - \alpha(t)\} dt$ is 1 (80% from the main term, 20% from γ) and multiplied by $\sqrt{S}$ per family. The intercept is a Gaussian bump plus a constant that sets the average response level (mean count 3, prevalence 0.2).

Fitted bases: cubic B-splines with first-order difference penalties (Eilers and Marx 1996; Wood 2017) (mgcv's ps basis with $m = c(2, 1)$), so constants along each direction are unpenalized; $k = 12$ along t for the intercept and γ ; an 8×10 tensor product for β ; 8×12 for f . The larger bases halve and quarter the knot spacing (about 100, 260 and 790 coefficients). Within-curve errors are a zero-mean process with correlation kernel R , scaled for Gaussian responses and used as the latent layer of a Gaussian copula for the other families, $Y_{ij} = F_{ij}^{-1}\{\Phi(Z_{ij})\}$, so each marginal is exactly the fitted family and the working model is wrong only about the dependence. Kernels: independent; rough stationary $0.8 \exp(-|s - t|/0.089) + 0.2 \mathbb{1}\{s = t\}$; smooth: the residual covariance of a REML fit to the DTI data (24 eigenfunctions plus 6% white noise, as a correlation kernel); exact AR(1) $\exp(-|s - t|/0.071)$; sign-changing $0.8 \exp(-|s - t|/0.3) \cos(2\pi|s - t|/0.4) + 0.2 \mathbb{1}\{s = t\}$; heteroskedastic versions of the smooth kernel with the DTI variance profile along t , a subject scale $\exp(0.5z_i - 0.25)$, or both. Misregistration: $Y_i(t_j) = \eta_i(h_i(t_j)) + \varepsilon_i(h_i(t_j))$ with $h_i(t) = t + a_i \sin(\pi t)/\pi$, $a_i \sim N(0, 0.314^2)$ truncated at $|a_i| \leq 0.9$, independent errors. Gaussian misregistration settings score the warp-averaged coefficient functions, GLM settings only the exact observed-clock mean.

Signal: $R^2 \in \{0.2, 0.5, 0.8\}$ for Gaussian responses; $\sqrt{S} \in \{0.5, 1\}$ on the link scale for Poisson and binary responses. Extended families at the default signal: scaled t (4 df) as Gaussian at $R^2 = 0.5$; beta ($\phi = 20$,

mean 0.3) and negative binomial (size 5, mean 3) at $\sqrt{S} = 0.5$. The 20-curve block uses the first 20 curves of each replicate of the main factorial, so it shares its draws with the 40- and 100-curve settings. The zero-surface block sets $\beta \equiv 0$ in the smooth-truth setting at the default signal, for Gaussian and binary responses with independent and smooth errors and 40 and 100 curves; γ , the intercept bump and the Gaussian error variance are those of the non-zero setting, and the intercept constant of binary responses is reset to keep the prevalence at 0.2. All random draws depend only on the replicate index, so contrasts between settings and intervals are paired. Each setting has 200 replicates; none failed. Concurvity of the fitted terms (mgcv's worst case) is at most 0.25. The next table lists the blocks of settings: the main factorial (family $\times$ error process $\times$ G $\times$ signal) and extensions that each change one factor.

Table 11: Blocks of the simulation design. Error processes: indep. independent, rough rough stationary, smooth from the DTI residuals, AR(1) exact AR(1), sign-chg. sign-changing, var(t)/var(z)/var(t,z) heteroskedastic versions of smooth, misreg. misregistration. Truth: smooth, rough or zero. Signal: low, default, high. Bases: default, $2\times$ or $4\times$ finer knots.

block	settings	families	errors	G	signal	truth	D	model	covariate	basis
main factorial	42	Gaussian, Poisson, binary	indep., rough, smooth	40, 100	low, default, high	smooth	61	functional	rich	default
main factorial, G = 20	21	Gaussian, Poisson, binary	indep., rough, smooth	20	low, default, high	smooth	61	functional	rich	default
misregistration	12	Gaussian, Poisson, binary	misreg.	40, 100	high	smooth, rough	61	functional	rich	default
dense grid	9	Gaussian, Poisson, binary	indep., rough, smooth	100	default	smooth	241	functional	rich	default
rough truth	14	Gaussian, Poisson, binary	indep., smooth	100	low, default, high	rough	61	functional	rich	default
zero surface	8	Gaussian, binary	indep., smooth	40, 100	default	zero	61	functional	rich	default
smooth effect	9	Gaussian, Poisson, binary	indep., rough, smooth	100	default	smooth	61	smooth effect	rich	default
extended families	8	scaled t, beta, negative binomial, negative binomial (fitted as Poisson)	indep., smooth	100	default	smooth	61	functional	rich	default
exact AR(1)	1	Gaussian	AR(1)	100	default	smooth	61	functional	rich	default
sign-changing	3	Gaussian, Poisson, binary	sign-chg.	100	default	smooth	61	functional	rich	default
misreg. (AR(1) fit)	1	Gaussian	misreg.	100	high	rough	61	functional	rich	default
heteroskedastic	6	Gaussian	var(t), var(z), var(t,z)	40, 100	default	smooth	61	functional	rich	default
low-rank covariate	6	Gaussian, Poisson, binary	indep., smooth	100	default	smooth	61	functional	low-rank	default
larger bases	12	Gaussian, Poisson, binary	indep., smooth	100	default	smooth	61	functional	rich	$2\times$ finer knots, $4\times$ finer knots

Fits per replicate: REML and curve-blocked NCV (mgcv gam); pointwise (non-blocked) NCV in the Gaussian settings at $G = 100$ with independent, smooth and sign-changing errors; and the AR(1) working model in the Gaussian settings at $G = 100$ with dependent errors. The AR(1) model is fitted with mgcv's `bam()` by fast REML with AR(1) errors that restart at the first grid point of each curve. Its correlation ρ is profiled over the grid $\{0, 0.1, \dots, 0.9, 0.95, 0.99\}$ by the fast-REML criterion, which includes the log-determinant of the whitening transform; the best grid value is then refined by a one-dimensional search between its two neighbouring grid values. Its intervals are model-based from that fit. The plasmode study uses the same fit. NCV fits flagged as not converged: 78% of scaled- t fits, 1.0% of beta fits, fewer than 0.1% otherwise. For the first 5 flagged scaled- t fits, refits from the REML smoothing parameters and from ten times them reach the same criterion within $8e-05$ and coefficients within $6e-04$: for these fits the optimum is flat, the fit stable. The check covers these five fits only.

B.2 Plasmode datasets

A plasmode simulation (Franklin et al. 2014) keeps the real covariates and real error structure of a dataset and replaces only the true regression function by a known one.

ECG strain (independent curves; myocardial-strain study, first beat per subject; non-public; aggregates only). 78 curves; response: global longitudinal strain, 4-chamber view (%), one cardiac cycle normalised to $[0, 1]$, native grid about 52 frames per beat (29–83) (fitted: interpolated to 61 points); functional covariate: simultaneously recorded ECG of the same beat (digitised pixel intensity, shape usable, amplitude up to scale), 51 points; scalar covariate: age (standardised). Preprocessing: cycle time normalised to the R–R interval; subjects with at least two beats (the second beat supplies the beat-to-beat difference curves). REML fit: R^2 0.47, surface EDF 59.6 (NCV fit: 44.0); residual design effect 19.9, lag-one correlation 0.96, 15% negative correlations, 4 principal components for 90% of the residual variance, covariate rank 15.

AF trial (independent curves; atrial-fibrillation trial, patients randomised to digoxin or a beta-blocker; non-public; aggregates only). 118 curves; response: left-atrial strain, 4-chamber view, at day 30 (%), one cycle on [0, 1], native grid 30 points per cycle (fitted: used as is); functional covariate: the same strain curve at day 0, 51 points; scalar covariate: randomised treatment (digoxin = 1). Preprocessing: none beyond the cycle normalisation; a baseline-adjusted functional ANCOVA. REML fit: R^2 0.19, surface EDF 33.2 (NCV fit: 25.3); residual design effect 10.9, lag-one correlation 0.89, 20% negative correlations, 3 principal components for 90% of the residual variance, covariate rank 8.

running (independent curves; Fukuchi, Fukuchi & Duarte 2017 (PeerJ 5:e3298); one row per person, first condition; public (tidyfundata, MIT; original data public)). 90 curves; response: sagittal knee moment over normalised stance, native grid 101 points (fitted: interpolated to 61 points); functional covariate: sagittal knee angle over the same stance, 51 points; scalar covariate: body mass (standardised). Preprocessing: one row per person (post-intervention rows dropped); speed and backpack conditions differ between persons and stay in the residuals. REML fit: R^2 0.46, surface EDF 53.7 (NCV fit: 54.8); residual design effect 33.3, lag-one correlation 0.99, 12% negative correlations, 2 principal components for 90% of the residual variance, covariate rank 7.

DTI (independent curves; refund::DTI, baseline visit (Goldsmith et al. 2011): multiple-sclerosis patients and controls; public (refund, GPL ≥ 2)). 92 curves; response: fractional anisotropy along the right corticospinal tract, native grid 55 points (fitted: used as is); functional covariate: fractional anisotropy along the corpus callosum, 51 points; scalar covariate: case status (MS = 1). Preprocessing: complete cases of both tracts. REML fit: R^2 0.21, surface EDF 53.8 (NCV fit: 35.1); residual design effect 9.9, lag-one correlation 0.81, 19% negative correlations, 16 principal components for 90% of the residual variance, covariate rank 24.

gait (independent curves; Van Crielinge et al. 2023 (Sci Data 10:852; figshare 24192489), able-bodied adults walking barefoot at preferred speed; public (CC0)). 138 curves; response: sagittal knee moment over one stride, one mean stride per subject, native grid 1001 points (0.1% of the stride) (fitted: every 10th point (101)); functional covariate: sagittal knee angle over the stride, 51 points; scalar covariate: age (standardised; from the subject-characteristics file). Preprocessing: the authors' stride normalisation and averaging over strides (within-subject stride variability is averaged out). REML fit: R^2 0.40, surface EDF 68.7 (NCV fit: 60.8); residual design effect 37.0, lag-one correlation 0.98, 33% negative correlations, 6 principal components for 90% of the residual variance, covariate rank 8.

ECG 8-lead (independent curves; roahd::mfD_healthy and mfD_LBBB (Ieva et al.), 50 healthy and 50 left-bundle-branch-block subjects; public (roahd, GPL-3)). 100 curves; response: ECG lead V5 (μ V) over one beat, native grid 1024 points at 1 kHz (fitted: interpolated to 101 points); functional covariate: ECG lead I, 51 points; scalar covariate: LBBB indicator. Preprocessing: registered and smoothed by the package authors. REML fit: R^2 0.21, surface EDF 39.4 (NCV fit: 11.2); residual design effect 6.2, lag-one correlation 0.94, 32% negative correlations, 4 principal components for 90% of the residual variance, covariate rank 15.

ocean (dependent curves; Hawaii Ocean Time-series profiles (FRegSigCom::ocean), monthly casts; public (FRegSigCom, GPL ≥ 2 ; HOT data public)). 116 curves; response: dissolved oxygen over depth, native grid 101 points (fitted: used as is); functional covariate: temperature over depth, 51 points; scalar covariate: cast order (standardised). Preprocessing: none; casts kept in time order. REML fit: R^2 0.36, surface EDF 59.2 (NCV fit: 49.8); residual design effect 39.7, lag-one correlation 0.99, 22% negative correlations, 5 principal components for 90% of the residual variance, covariate rank 6.

weather (dependent curves; AEMET open data via fda.usc::aemet, 73 Spanish stations, daily means 1980–2009; public (fda.usc, GPL-2)). 73 curves; response: log precipitation over the year, native grid 365 days (fitted: every 5th day (73)); functional covariate: daily mean temperature over the year, 51 points; scalar covariate: altitude (standardised). Preprocessing: none beyond the day thinning. REML fit: R^2 0.41, surface

EDF 29.5 (NCV fit: 38.6); residual design effect 29.1, lag-one correlation 0.50, 1% negative correlations, 19 principal components for 90% of the residual variance, covariate rank 4.

electricity (dependent curves; Adelaide electricity demand (fds::mondaydemand, fds::mondaytempairport; Magnano, Boland & Hyndman 2008), every 5th Monday 1997–2007; public (fds, GPL-3)). 102 curves; response: half-hourly electricity demand over the day, native grid 48 half-hours (fitted: used as is); functional covariate: airport temperature over the same day, 51 points; scalar covariate: week index (standardised; absorbs the linear trend). Preprocessing: every 5th Monday (thins, not removes, the week-to-week dependence; the annual cycle stays). REML fit: R^2 0.31, surface EDF 33.2 (NCV fit: 10.0); residual design effect 35.8, lag-one correlation 0.98, 0% negative correlations, 2 principal components for 90% of the residual variance, covariate rank 11.

The ECG strain beats were recorded with 29 to 83 frames and interpolated to 61 grid points; interpolation adds smoothness within a curve, so part of the residual dependence of that dataset, and of its design effect, comes from the interpolation.

Six datasets have independent source curves. The other three have source curves that are serially (ocean, electricity) or spatially (weather) dependent units. The residual *design effect* of a dataset is $\mathbf{1}^\top \Sigma \mathbf{1} / \text{tr} \Sigma$ for the covariance Σ of its REML residual curves: the factor by which the variance of a grid mean exceeds its working-independence value.

Table 12: Fitted models: bases (intercept, γ ; surface 8×10), R^2 of the REML fit, surface and total EDF of the three fits used as truths (REML / NCV / midpoint), and whether the midpoint fit lies between the two for every term.

dataset	intercept basis k	gamma(t) basis k	surface basis (s $\times$ t)	R^2	surface EDF REML / NCV / midpoint	total EDF REML / NCV / midpoint	midpoint interior
ECG strain	12	12	8×10	0.469	59.6 / 44.0 / 52.9	77.1 / 54.1 / 63.4	FALSE
AF trial	12	12	8×10	0.192	33.2 / 25.3 / 29.5	44.7 / 32.8 / 38.9	TRUE
running	12	12	8×10	0.458	53.7 / 54.8 / 60.4	68.0 / 65.9 / 73.1	FALSE
DTI	12	12	8×10	0.210	53.8 / 35.1 / 47.0	68.5 / 47.9 / 60.7	TRUE
gait	12	12	8×10	0.397	68.7 / 60.8 / 65.3	89.0 / 80.1 / 85.2	FALSE
ECG 8-lead	40	30	8×10	0.211	39.4 / 11.2 / 23.9	103.2 / 77.0 / 88.9	TRUE
ocean	12	12	8×10	0.362	59.2 / 49.8 / 60.3	74.3 / 58.7 / 72.1	FALSE
weather	12	12	8×10	0.407	29.5 / 38.6 / 41.9	45.4 / 60.8 / 61.5	FALSE
electricity	12	12	8×10	0.312	33.2 / 10.0 / 23.9	45.8 / 22.0 / 35.7	FALSE

Table 13: Residual dependence and covariate diagnostics (REML residual curves): design effect DE and DE/D, residual correlation at one grid step and at 10% and 25% of the domain, share of negative correlations and the most negative one, max/min SD along t , principal components for 90% of the residual variance, centred covariate rank, share of response variance removed by registration.

dataset	DE	DE/D	lag 1 step	lag 10%	lag 25%	neg. corr. share	min r	sd ratio	FPC90	rank X	phase share
ECG strain	19.88	0.326	0.962	0.620	0.208	0.146	-0.351	24.53	4	15	0.63
AF trial	10.93	0.364	0.889	0.684	0.359	0.195	-0.269	29.78	3	8	0.43
running	33.31	0.546	0.994	0.871	0.594	0.119	-0.095	4.63	2	7	0.30
DTI	9.91	0.180	0.811	0.233	0.106	0.193	-0.340	2.31	16	24	0.22
gait	36.96	0.366	0.978	0.483	0.258	0.327	-0.535	6.15	6	8	0.03
ECG 8-lead	6.19	0.061	0.938	0.363	0.078	0.324	-0.681	42.26	4	15	0.33
ocean	39.74	0.393	0.987	0.738	0.432	0.217	-0.268	2.86	5	6	0.00
weather	29.12	0.399	0.503	0.489	0.393	0.006	-0.101	2.22	19	4	0.06
electricity	35.77	0.745	0.981	0.894	0.822	0.000	0.308	3.07	2	11	0.00

Table 14: Datasets with dependent source curves: block construction for the block flips and block-wise NCV; residual-score correlation within and between blocks.

dataset	block rule	blocks	block size	alt. blocks	score corr. within	score corr. between	between pairs
ocean	3 consecutive monthly casts (cost check; the autocorrelation rule gives single casts)	39	2–3		0.045	-0.100	38
weather	complete-linkage clusters of station distance cut at 250 km (mismatched partition: 150 km)	16	1–8	25	0.548	0.356	276
electricity	2 consecutive frame curves (10 weeks)	51	2		0.214	0.011	50

B.3 Construction of the plasmode data

Each dataset’s model, $Y_i(t) = \alpha(t) + \int X_i(s)\beta(s, t) ds + z_i\gamma(t) + \varepsilon_i(t)$ (Gaussian), is fitted by REML and by curve-blocked NCV. The fitted means of three fits serve as truths: the REML fit (*REML-fitted truth*), the NCV fit (*NCV-fitted truth*) and a refit with every smoothing parameter at the geometric mean of the two fits’ values (*midpoint truth*). Replicate data are these fitted means plus each subject’s own residual curve from the REML or the NCV fit, multiplied by an independent random sign per curve. Sign flipping keeps each residual curve’s within-curve covariance and makes the curves independent across subjects. For the ECG strain data, heartbeat difference curves $(Y_{i1} - Y_{i2})/\sqrt{2}$ of two beats per subject are a further error source that no fit shaped, on their own scale or rescaled to the average variance of the REML residuals. Each combination of dataset, truth and residual source is run with all subjects and with 40 randomly chosen subjects (200 replicates each). The three datasets with dependent source curves add block flips (one sign per block of neighbouring curves) and block-wise NCV and CL2. The mean is evaluated for a fixed set of 50 subjects’ covariates of the dataset, at the native grid points. 130 combinations in all.

Table 15: Choice of the fixed set of 50 subjects for the mean: the rule used, and the balance check of the first 50 subjects against the remaining subjects (one-sided test at 5%, permutation p-value, standardized mean differences of the scalar covariate and the first three covariate scores). Where this check failed, a seeded random sample of 50 subjects was used; the balance of that sample was not recorded.

dataset	G	cohort	one-sided at 5%	permutation p	SMD z	SMD PC1	SMD PC2	SMD PC3
ECG strain	78	first 50 rows	FALSE	0.302	0.313	0.021	-0.180	-0.402
AF trial	118	seeded sample of 50	TRUE	0.000	1.788	0.046	-0.231	0.317
running	90	seeded sample of 50	TRUE	0.000	-0.331	-0.398	0.916	-0.008
DTI	92	seeded sample of 50	TRUE	0.000	-1.149	-0.408	-0.258	-0.036
gait	138	seeded sample of 50	TRUE	0.000	1.668	0.118	0.299	0.377
ECG 8-lead	100	seeded sample of 50	TRUE	0.000	-1.990	-1.344	0.235	0.274
ocean	116	seeded sample of 50	TRUE	0.000	-1.725	-0.255	-0.346	0.109
weather	73	first 50 rows	FALSE	0.411	-0.173	0.130	0.225	0.378
electricity	102	seeded sample of 50	TRUE	0.000	-1.724	0.029	-0.091	-0.297

C Computing time

Times are wall-clock seconds on one core. They come from four sources: (A) a benchmark on one local core with Gaussian data, $G = 100$ curves, $D = 61$ grid points and the default bases (90 replicates over the dependent error processes of the simulation); (B) the comparator runs on a computing cluster, Gaussian main-factorial settings and plasmode datasets with independent source curves, $G = 40$ to 138; (C) the smoothing-parameter-term runs on the main-factorial settings at the default signal ($G = 40$ and 100); (D) the same terms on the plasmode data. Table 16 gives the median, the quartiles, the 90% quantile and the maximum per method.

Table 16: Computing time in seconds per fit or per covariance, by method and data source (A: local benchmark, Gaussian, $G = 100$, $D = 61$; B: comparator runs, Gaussian synthetic settings and plasmode datasets with independent source curves; C: main-factorial settings; D: plasmode data). n: number of timed fits or covariances. CL2 and the smoothing-parameter terms are computed on top of an existing fit.

method	data	n	median	25%	75%	90%	max
REML fit	A	90	2.02	1.85	2.66	4.04	5.1
curve-blocked NCV fit	A	90	8.56	7.18	10.60	13.60	17.1
AR(1) working model (profiled)	A	90	39.00	35.10	42.20	51.50	58.7
CL2 covariance (REML fit)	A	90	0.95	0.92	1.13	1.27	1.6
REML fit	B	3600	2.39	1.42	3.92	7.37	36.7
GLS, FPCA residual covariance	B	3600	3.24	1.83	4.31	8.23	37.3
GLS, raw residual covariance	B	3600	3.32	1.87	4.58	9.92	47.1
pcre fit	B	3600	17.70	7.90	87.90	241.00	1340.0
curve bootstrap (199 refits)	B	1798	507.00	315.00	801.00	1610.00	6640.0
CL2 on a GLS fit	B	7200	0.12	0.07	0.18	0.32	1.1
CL2 on a pcre fit	B	1799	1.40	0.73	2.80	4.30	11.0
smoothing-parameter term, REML	C	3600	1.16	0.49	1.35	1.53	3.8
smoothing-parameter term, REML	D	26000	1.17	0.74	1.68	3.48	7.2
smoothing-parameter term, NCV	C	1200	2.10	1.12	3.70	3.95	5.8
smoothing-parameter term, NCV	D	13000	2.48	1.36	3.66	8.61	38.1
smoothing-parameter term, NCV (Poisson, binary)	C	2400	5.25	2.29	11.60	12.20	13.9

A curve-blocked NCV fit costs about 4.2 times a REML fit; the profiled AR(1) model about 19 times, because it refits the model at each value of ρ . The CL2 covariance of a REML fit adds a median 0.95 s, its Satterthwaite degrees of freedom a median 1.8 s per replicate for all four estimands in the main factorial, 2.1 s in the other synthetic settings and 1.7 s on the plasmode data (cluster nodes). GLS needs about the time of a REML fit plus the estimation of the residual covariance; pcre's time grows quickly with the number of curves and grid points (median 18 s, 90% quantile 241 s, maximum 1342 s). CL2 on these fits adds a median of 0.12 s (GLS) and 1.4 s (pcre). The curve bootstrap with 199 refits takes a median of 507 s and up to 6645 s per data set. The smoothing-parameter uncertainty terms (curve scores, Hessian and infinitesimal jackknife) add a median of 1.2 s to a REML fit and 2.1 s to a Gaussian NCV fit. The one-step bias correction of the NCV estimate needs one linear solve with $p \times p$ matrices and was not timed separately.