Locally sparse varying coefficient mixed model with application to longitudinal microbial differential abundance analysis

Simon Fontaine[1, *], Gen Li[2], and Ji Zhu[1]

[1]Department of Statistics, University of Michigan
[2]Department of Biostatistics, University of Michigan
[*]simfont@umich.edu

Abstract

Differential abundance (DA) analysis in microbiome studies has recently been used to uncover a plethora of associations between microbial composition and various health conditions. While current approaches to DA typically apply only to cross-sectional data, many studies feature a longitudinal design to better understand the underlying microbial dynamics. To perform DA on longitudinal microbial studies, we propose a novel varying coefficient mixed-effects model with local sparsity. The proposed method can identify time intervals of significant group differences while accounting for temporal dependence. Specifically, we exploit a penalized kernel-local polynomial smoothing approach for parameter estimation and extend local regression to include a random effect. Further, we obtain point-wise confidence intervals using bootstrapping to determine intervals of significant differences. Synthetic data experiments demonstrate the necessity of modelling dependence for precise estimation and support recovery. The application to a longitudinal study of mice oral microbiome undergoing cancer development with and without a mutation of interest reveals novel scientific insights.

Keywords: local regression, functional sparsity, differential abundance, longitudinal microbiome studies, semiparametric regression, kernel smoothing

1 Introduction

2 Locally sparse varying coefficient mixed model

2.1 Setting & Notation

We consider the following function-on-scalar regression problem. Let $i = 1, \dots, N$ denote the N sampling units (e.g., subjects). Let t_{ij}, $j = 1, \dots, n_i$, denote the sampling times for subject i and define $\mathbf{t}_i = (t_{i1}, \dots, t_{in_i})$; we do not assume any structure on the $\mathbf{t}_i$'s across subjects. The observed response for subject i at time t_{ij} is denoted $y_{ij} \in \mathbb{R}$ and we define $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})$ as the vector of responses for subject i. For each subject, we split covariates into two categories: time-varying effects $\mathbf{x}_i \in \mathbb{R}^{p_x}$ and constant effects $\mathbf{u}_i \in \mathbb{R}^{p_u}$.

In the present exposition, we assume the covariates for the varying coefficient terms to be constant through time, indicated by the absence of time index in $\mathbf{x}_i$, but our proposed model and implementation readily works for $\mathbf{x}_{ij} \in \mathbb{R}^{p_x}$ varying with time, provided it is observed at the same time points as the responses of subject i. In particular, we do not allow *asynchronous* covariates (see, e.g., Zhong et al., 2022.) A practical solution to using asyncrhonous covariates in our proposed method is to pre-process the curves using funcitonal PCA .

2.2 Varying coefficient mixed model

Our main goal is to study the relationship between covariates $\mathbf{x}_i$ and y_{ij}, while accounting for temporal dependence within subjects and other covariates. In particular, we are interested in identifying *if*, *when* and *how* y_{ij} changes with each entry in $\mathbf{x}_i$. To this end, we consider a *(semi)varying coefficient mixed model*:

$$y_{ij} \mid \theta_i(t_{ij}) \sim \mathcal{N}\left(\boldsymbol{\beta}(t_{ij})^\top \mathbf{x}_i + \boldsymbol{\alpha}^\top \mathbf{u}_i + \theta_i(t_{ij}), \sigma^2\right) \tag{1}$$

with (conditional) independence across i and j, where $\boldsymbol{\beta}(\cdot) : \mathbb{R} \to \mathbb{R}^{p_x}$ is the vector-valued function of time-varying coefficients, $\boldsymbol{\alpha} \in \mathbb{R}^{p_u}$ is the vector of constant effects, and where $\theta_i(\cdot)$ is a random process capturing the temporal dependence. In particular, we assume $\mathbb{E}\{\theta_i(t)\} \equiv 0$ with covariance kernel $\mathrm{Cov}(\theta_i(t), \theta_i(t')) = k_\theta(t, t')$ for some symmetric positive definite kernel k_θ. Define $\mathbf{K}_\theta(\mathbf{t})$ as the covariance matrix for a random effect evaluated at the time points in $\mathbf{t}$, that is, $[\mathbf{K}_\theta(\mathbf{t})]_{jj'} = k_\theta(t_j, t_{j'})$. Hence, marginally, $\mathbf{y}_i \sim \mathcal{N}(\mathbf{m}_i, \mathbf{V}_i)$ with

$$\mathbf{m}_i := \boldsymbol{\beta}(\mathbf{t}_i)^\top \mathbf{x}_i + \boldsymbol{\alpha}^\top \mathbf{u}_i \mathbf{1}_{n_i}, \qquad \mathbf{V}_i := \mathbf{K}_\theta(\mathbf{t}_i) + \sigma^2 \mathbf{I}_{n_i}, \tag{2}$$

where $\boldsymbol{\beta}(\mathbf{t}_i)$ is the $p_x \times n_i$ matrix with columns $\boldsymbol{\beta}(t_{ij})$. We further denote the inverse covariance $\mathbf{P}_i = \mathbf{V}_i^{-1}$. We thus find the marginal log-likelihood of subject i,

$$\ell_i(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) = -\frac{1}{2} \log\det(2\pi \mathbf{V}_i) - \frac{1}{2} [\mathbf{y}_i - \mathbf{m}_i]^\top \mathbf{P}_i [\mathbf{y}_i - \mathbf{m}_i], \tag{3}$$

where we omit the dependence on the variance parameters, σ^2 and K_θ. The log-likelihood across all subjects is given by $\ell(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) = \sum_{i=1}^N \ell_i(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha})$.

In our motivating example, we have two time-varying effects: the intercept $\beta_0(\cdot)$ and the group difference $\beta_1(\cdot)$, where $\mathbf{x}_i = (1, x_i)$ with $x_i \in \{0, 1\}$ being the group membership indicator. Differential abundance is therefore mainly interested in studying $\beta_1(\cdot)$: in particular, we want to identify the time points t, if any, where $\beta_1(t) \neq 0$, and estimate the direction and strength of the local difference.

2.3 Score equations

To motivate our GEE-based inference methodology, we start by investigating the score equations for the mean parameters. Let $\mathbf{r}_i = \mathbf{y}_i - \mathbf{m}_i$ denote the residual vector for subject i, which implicitly

depends on the mean parameters. Further denote $\mathbf{X}_i$ the $n_i \times p_x$ matrix with rows $\mathbf{x}_i(t_{ij})$; for fixed covariates, we simply have $\mathbf{X}_i = \mathbf{1}_{n_i}\mathbf{x}_i^\top$.

Consider computing the gradient with respect to $\boldsymbol{\beta}(t)$ for some t. Whenever $t_{ij} \neq t$, the mean m_{ij} does not depend on $\boldsymbol{\beta}(t)$, so we find

$$\nabla_{\boldsymbol{\beta}(t)}\ell(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) = -\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{D}_i(t)\mathbf{P}_i\mathbf{r}_i, \tag{4}$$

where $\mathbf{D}_i(t) = \text{diag}(\mathbb{1}[t = t_{ij}])$. Now, consider computing the gradient with respect to $\boldsymbol{\beta}$ by assuming that $\boldsymbol{\beta}(\cdot)$ is a constant function parameterized by $\boldsymbol{\beta}$, i.e., $\boldsymbol{\beta}(\cdot) \equiv \boldsymbol{\beta}$. We find

$$\nabla_{\boldsymbol{\beta}(t)}\ell(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) = -\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{P}_i\mathbf{r}_i = -\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{I}\mathbf{P}_i\mathbf{r}_i. \tag{5}$$

Looking at the difference between () and (), we see that the pointwise gradient weighs the precision-adjusted residuals $\mathbf{P}_i\mathbf{r}_i$ by $\mathbf{D}_i(t)$, while the constant gradient weighs them equally by $\mathbf{I}$. To obtain a nonconstant smooth estimate that borrows signal from neighboring time points, we utilize kernel smoothing, which interpolates between the pointwise estimator and the constant estimator by weighing the residuals using a kernel function $k_h(s) = k(s/h)/h$ depending on the distance from a time point of interest:

$$\nabla^{k_h}_{\boldsymbol{\beta}(t)}\ell(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) := -\sum_{i=1}^{N} \mathbf{X}_i^\top \,\text{diag}(\mathbf{k}_i(t))\mathbf{P}_i\mathbf{r}_i, \tag{6}$$

where $\mathbf{k}_i(t) = [k_h(t - t_{ij})]_{j=1}^{n_i}$. When $h \to 0$, $\text{diag}(\mathbf{k}_i(t))$ behaves as $\mathbf{D}_i(t)$, in which case we recover the pointwise estimator; when $h \to \infty$, $\text{diag}(\mathbf{k}_i(t))$ behaves as $\mathbf{I}$, in which case we recover the constant estimator. In practice, this is equivalent to weighing the design matrix: $\widetilde{\mathbf{X}}_i(t) := \text{diag}(\mathbf{k}_i(t))\mathbf{X}_i$.

Let $\mathbf{U}_i$ be the $n_i \times p_u$ matrix with rows $\mathbf{u}_i(t_{ij})$. The gradient with respect to $\boldsymbol{\alpha}$ is given by

$$\nabla_{\boldsymbol{\alpha}}\ell(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) = -\sum_{i=1}^{N} \mathbf{U}_i^\top \mathbf{P}_i\mathbf{r}_i.$$

2.4 Hessian

The point-wise Hessian is given by

$$\nabla^2_{\boldsymbol{\beta}(t)}\ell(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) = \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{D}_i(t)\mathbf{P}_i\mathbf{D}_i(t)\mathbf{X}_i, \tag{7}$$

and the constant Hessian is given by

$$\nabla^2_{\boldsymbol{\beta}(t)}\ell(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) = \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{I}\mathbf{P}_i\mathbf{I}\mathbf{X}_i. \tag{8}$$

This suggests the Hessian under the kernel approximation

$$\nabla^{2,k_h}_{\boldsymbol{\beta}(t)}\ell(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) := \sum_{i=1}^{N} \mathbf{X}_i^\top \,\text{diag}(\mathbf{k}_i(t))\mathbf{P}_i \,\text{diag}(\mathbf{k}_i(t))\mathbf{X}_i. \tag{9}$$

The Hessian with respect to $\boldsymbol{\alpha}$ is given by

$$\nabla^2_{\boldsymbol{\alpha}}\ell(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) = \sum_{i=1}^{N} \mathbf{U}_i^\top \mathbf{P}_i\mathbf{U}_i.$$

The cross terms are given by

$$\nabla^{2,k_h}_{\boldsymbol{\beta}(t),\boldsymbol{\alpha}} \ell(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) := \sum_{i=1}^{N} \mathbf{X}_i^\top \operatorname{diag}(\mathbf{k}_i(t)) \mathbf{P}_i \mathbf{U}_i. \tag{10}$$

2.5 Estimating the covariance parameters

Recall that $\mathbf{P}_i = \mathbf{V}_i^{-1}$ where $\mathbf{V}_i = \mathbf{K}_\theta(\mathbf{t}_i) + \sigma^2 \mathbf{I}_{n_i}$. To obtain the most efficient estimator of the mean parameters, we need $\mathbf{V}_i$ to accurately capture the dependence structure in the residuals. For a regular design, i.e, $\mathbf{t}_i \equiv \mathbf{t}$ for some $\mathbf{t}$, we could simply estimate $\mathbf{V} \approx \frac{1}{N} \sum_{i=1}^{N} \mathbf{r}_i \mathbf{r}_i^\top$.

To avoid overfitting, we could alternatively consider a smoothed version of the residuals. Wang et al. (2022) propose a two-step estimator where the residuals are first obtained under an independent model assumption. Then, the residuals are smoothed using a local linear kernel smoother whose empirical covariance is then used as the working covariance in the final mean parameter estimator.

For irregular designs, more care is required as empirical covariance are not applicable. A first approach would be to directly estimate the covariance function $k_\theta(t, s)$ from which we can reconstruct $\mathbf{K}_\theta$ by evaluations.

2.5.1 Parametrically

As observed by (...), efficiency gains can be obtained even though the w

With our motivating example in mind, where only a few time points are sampled, a non-parametric estimation of the covariance function might be overkill. Instead, we propose to specify a working parametric model whose covariance function is determined by a few parameters. Some notable examples include the compound symmetry structure, equivalently a random intercept model, with covariance function

$$k_\theta(t, s; r_\theta) = \sigma^2 r_\theta,$$

and the AR(1) model, with covariance function

$$k_\theta(t, s; r_\theta, \rho) = \sigma^2 r_\theta \rho^{|t-s|},$$

where r_θ denotes the variance ratio with the noise variance σ^2 and ρ controls the long-range dependency.

To estimate the variance parameters $\boldsymbol{\tau}$, we proceed iteratively: we estimate the mean parameters, extract the residuals and estimate the variance parameters by minimizing the profile likelihood

$$\ell(\boldsymbol{\tau}) := -\frac{1}{2} \sum_{i=1}^{N} \log\det(2\pi \mathbf{V}_i) + \mathbf{r}_i^\top \mathbf{P}_i \mathbf{r}_i,$$

where $\mathbf{P}_i$ and $\mathbf{V}_i$ implicitly depend on the variance parameters. By parameterising the covariance function as a multiple of σ^2, we can write $\mathbf{V}_i = \sigma^2 \mathbf{C}_i$, where $\mathbf{C}_i = \frac{1}{\sigma^2} \mathbf{K}_\theta(\boldsymbol{\tau}) + \mathbf{I}$. Then, the estimate for σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{N} \mathbf{r}_i^\top \mathbf{C}_i \mathbf{r}_i$$

where $n = \sum_{i=1}^{N} n_i$ is the total number of observations.

2.5.2 Parametrically with proxy matrix

In a variable selection context within mixed models, Fan and Li (2012) propose to use a fixed proxy matrix during estimation of the fixed effect. Suppose we can write $\theta_i(\mathbf{t}_i) = \mathbf{Z}_i \boldsymbol{\gamma}_i$ for some random

effect design matrix $\mathbf{Z}_i \in \mathbb{R}^{n_i \times q}$ and some random effect vector $\boldsymbol{\gamma}_i \sim \mathcal{N}_q(\mathbf{0}, G)$. Then, the marginal covariance of $\mathbf{y}_i$ can be written as

$$\mathbf{V}_i = \mathbf{Z}_i G \mathbf{Z}_i^\top + \sigma^2 \mathbf{I}.$$

Instead of estimating G, Fan and Li (2012) propose to use a fixed proxy matrix M in the place of G, leading to the (rescaled) precision matrix

$$\mathbf{P}_i \approx [\mathbf{Z}_i M \mathbf{Z}_i^\top + \mathbf{I}]^{-1}.$$

Assuming some regularity conditions, they show that $M = \log(n)\mathbf{I}$ is sufficient for asymptotic model selection consistency.

As an example, consider a random intercept design, namely, $\mathbf{Z}_i = \mathbf{1}_{n_i}$ with $\boldsymbol{\gamma}_i \sim \mathcal{N}(0, \sigma^2 r_\theta)$. Then,

$$\mathbf{P}_i = \sigma^{-2}[\mathbf{I} + r_\theta \mathbf{1}\mathbf{1}^\top]^{-1} = \sigma^{-2}\left[\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n_i + r_\theta^{-1}}\right].$$

We see that whenever n_i and r_θ are moderately large, the denominator does not change much with the value of r_θ so estimating the best value for r_θ will not be that impactful for mean parameter estimation.

2.6 Extension to generalized linear models

Using a link function for the mean, e.g.,

$$\mathbb{E}\{y_{ij}\} = m_{ij} = g^{-1}(\boldsymbol{\beta}(t_{ij})^\top \mathbf{x}_i + \boldsymbol{\alpha}^\top \mathbf{u}_i),$$

with a mean-variance relationship $\nu(\mu)$, we find the following marginal variance for $\mathbf{y}_i$:

$$\mathbf{V}_i = \phi \mathbf{S}_i^{1/2}[\mathbf{K}_\theta(\mathbf{t}_i) + \mathbf{1}]\mathbf{S}_i^{1/2},$$

where $\mathbf{S}_i = \text{diag}(\nu(m_{ij}), i = 1, \ldots, n_i)$, ϕ is the dispersion parameter.

Estimation proceeds by computing $\mathbf{S}_i$ using the current mean values and taking a gradient step for mean parameters. The main difference will be the inclusion of the derivative of g^{-1} in the gradient:

$$\nabla^{k_h}_{\boldsymbol{\beta}(t)} \ell_Q(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) := -\sum_{i=1}^{N} \mathbf{X}_i^\top \text{diag}(\mathbf{g}(\mathbf{m}_i)) \text{diag}(\mathbf{k}_i(t)) \mathbf{S}_i^{-1/2} \mathbf{P}_i \mathbf{S}_i^{-1/2} \mathbf{r}_i, \tag{11}$$

where $\mathbf{g}(\mathbf{m}_i)$ has entries $\frac{\mathrm{d}}{\mathrm{d}\eta} g^{-1}(\eta)|_{\eta=\eta_{ij}}$, and where $\mathbf{P}_i$ is the inverse of the (rescaled) working covariance matrix.

[S] Old stuff from here

2.7 Localized model

Suppose we are interested in estimating $\boldsymbol{\beta}(t)$ at a specific value of t. A common approach is *local polynomial regression* (Fan, 1993) where we assume that $\boldsymbol{\beta}(\cdot)$ is well approximated by a polynomial function around t. A special case is the *locally constant approximation* where we assume $\boldsymbol{\beta}(t) = \mathbf{b} \in \mathbb{R}^{p_x}$. To estimate $\mathbf{b}$, we weight observation according to their distance from t using a kernel function k: in the simple case of no repeated measurements ($n_i \equiv 1$), no random effect $\theta_i(\cdot)$, and no constant effects $\boldsymbol{\alpha}$ we find

$$\hat{\mathbf{b}} = \arg\min_{\mathbf{b}} \sum_{i=1}^{N} k(t_i, t) \left[y_i - \mathbf{b}^\top \mathbf{x}_i\right]^2. \tag{12}$$

Now, extending local regression to repeated measurements without random effects and including constant effects, we find, assuming $\boldsymbol{\alpha}$ known for now,

$$\hat{\mathbf{b}} = \arg\min_{\mathbf{b}} \sum_{i=1}^{N} \sum_{j=1}^{n_i} k(t_{ij}, t) \left[y_i j - \mathbf{b}^\top \mathbf{x}_i - \boldsymbol{\alpha}^\top \mathbf{u}_i \right]^2. \tag{13}$$

The main difficulty arises when we wish to introduce the random effect. First, we rewrite (13) as the sum of squared weighted residuals:

$$\sum_{i=1}^{N} \sum_{j=1}^{n_i} k(t_{ij}, t) \left[y_i j - \mathbf{b}^\top \mathbf{x}_i - \boldsymbol{\alpha}^\top \mathbf{u}_i \right]^2 = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left[w_{ij}(t)(y_i j - \mathbf{b}^\top \mathbf{x}_i - \boldsymbol{\alpha}^\top \mathbf{u}_i) \right]^2 \tag{14}$$

$$= \sum_{i=1}^{N} \left\| \operatorname{diag}(\mathbf{w}_i(t))(\mathbf{y}_i - \tilde{\mathbf{m}}_i) \right\|^2 \tag{15}$$

where $\tilde{\mathbf{m}}_i := \mathbf{b}^\top \mathbf{x}_i + \boldsymbol{\alpha}^\top \mathbf{u}_i$ is the vector of localized fitted values, and where $\mathbf{w}_i(t)$ is the vector of length n_i with entries $w_{ij}(t) = \sqrt{k(t_{ij}, t)}$. Hence, we find that local regression implicitly assumes a covariance of $\operatorname{diag}(\mathbf{w}_i(t))^{-2}$ among $\mathbf{y}_i$. Since the covariance matrix is diagonal, this objective function does not contain any within-subject correlation, but weighs down observation away from the time point of interest t. Our proposed approach consists of replacing the precision $\operatorname{diag}(\mathbf{w}_i(t))^2$ by a non-diagonal precision matrix which both captures local within-subject correlation and weighs observation according to their distance from t.

In particular, we propose to consider a localized version of the random effect $\theta_i(\cdot)$. We utilize the same local constant approximation to approximate $\theta_i(\cdot) \approx \tilde{\theta}_i$ around t. Then, since the random effect is constant across time, it behaves as a random intercept. In that case, the localized model for subject i becomes $\mathbf{y}_i \sim \mathcal{N}(\tilde{\mathbf{m}}_i, \tilde{\mathbf{V}}_i)$ where $\tilde{\mathbf{V}}_i := \sigma_\theta^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^\top + \sigma^2 \mathbf{I}_{n_i}$, for some prior variance σ_θ^2. Let's further define the localized precision matrix $\tilde{\mathbf{P}}_i = \tilde{\mathbf{V}}_i^{-1}$. We then find the unweighted localized marginal squared loss, $\sum_{i=1}^{N} (\mathbf{y}_i - \tilde{\mathbf{m}}_i)^\top \tilde{\mathbf{P}}_i (\mathbf{y}_i - \tilde{\mathbf{m}}_i)$. We finally introduce the kernel weighting by replacing the residuals by their weighted analogs:

$$\sum_{i=1}^{N} \left[\operatorname{diag}(\mathbf{w}_i(t))(\mathbf{y}_i - \tilde{\mathbf{m}}_i) \right]^\top \tilde{\mathbf{P}}_i \left[\operatorname{diag}(\mathbf{w}_i(t))(\mathbf{y}_i - \tilde{\mathbf{m}}_i) \right] \tag{16}$$

$$= \sum_{i=1}^{N} (\mathbf{y}_i - \tilde{\mathbf{m}}_i)^\top \tilde{\mathbf{P}}_i(t) (\mathbf{y}_i - \tilde{\mathbf{m}}_i) \tag{17}$$

where $\mathbf{W}_i(t) = \mathbf{w}_i(t)\mathbf{w}_i(t)^\top$ is the $n_i \times n_i$ matrix with entries given by $\sqrt{k(t_{ij}, t) k(t_{ij'}, t)}$, and where $\tilde{\mathbf{P}}_i(t) := [\mathbf{W}_i(t) \odot \tilde{\mathbf{P}}_i]$. When we assume no within-subject correlation, we naturally recover (13) up to a scaling factor: in such case, $\tilde{\mathbf{P}}_i = \sigma^2 \mathbf{I}_{n_i}$ so that $\tilde{\mathbf{P}}_i(t) = \sigma^2 \operatorname{diag}(\mathbf{w}_i(t))^2$. Lin and Carroll (2001, see also Wang, 2003) propose a similar objective function for semiparametric longitudinal regression.

We note that the choice of localized correlation structure—here, a random intercept—is of minor importance. Indeed, since it always appears masked by the weight matrix $\mathbf{W}_i(t)$, only the behavior close to the diagonal matters, and there is generally little difference between relevant correlation structures. In fact, the choice of kernel function k and bandwidth h will be of larger importance since it can completely mask long-range correlation. We choose a random intercept structure for its simplicity and agreement with the local constant approximation approach used throughout.

The localized residual sum of squares naturally extends to a localized Gaussian likelihood:

$$\tilde{\ell}_i(t) = \frac{1}{2} \log \det(\tilde{\mathbf{P}}_i(t)/2\pi) - \frac{1}{2} (\mathbf{y}_i - \tilde{\mathbf{m}}_i)^\top \tilde{\mathbf{P}}_i(t) (\mathbf{y}_i - \tilde{\mathbf{m}}_i).$$

However, the weighted localized precision $\tilde{\mathbf{P}}_i(t)$ is typically singular, preventing the computation of the log determinant. Recalling that $\mathbf{W}_i(t) = \mathbf{w}_i(t)\mathbf{w}_i(t)^\top$, we find that for t_{ij} far from t, $k_h(t_{ij}, t)$ can

be very small, if not 0 for finite-support kernels. Then, entire rows and columns of $\widetilde{\mathbf{P}}_i(t)$ will be close to 0. Instead, we compute the log determinant as the sum of the logarithm of the non-zero eigenvalues of $\widetilde{\mathbf{P}}_i(t)$.

Inspecting the localized model more closely reveals an alternative narrative. The kernel choice k and scale h and the variance parameters $\sigma^2, \sigma_\theta^2$ were introduced for different reasons and at different steps. However, these objects ultimately define the localized precision matrix $\widehat{\mathbf{P}}_i(t)$, and nothing else. In particular, writing

$$\widehat{\mathbf{P}}_i(t) = \operatorname{diag}(\mathbf{w}_i(t))\widetilde{\mathbf{P}}_i \operatorname{diag}(\mathbf{w}_i(t))$$

shows that $k, h, \sigma^2, \sigma_\theta^2$ implicitly parameterize $\widetilde{\mathbf{P}}_i(t)$. Reparametrizing $\rho_\theta = \sigma_\theta^2/\sigma^2$ enables us to disentangle scale from correlation:

$$\widetilde{\mathbf{P}}_i = \sigma^{-2}\left[\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n_i + \rho_\theta^{-1}}\right].$$

Then, σ^2 controls the overall scale of $\widetilde{\mathbf{P}}_i(t)$, k and h control how quickly $\widetilde{\mathbf{P}}_i(t)$ vanishes to 0 away from t, and ρ_θ controls the ratio between the diagonal and off-diagonal terms.

2.8 Joint criterion & quasi likelihood

To estimate $\boldsymbol{\beta}(\cdot)$ at a collection of T timepoints $\mathcal{T} \subset \mathbf{R}$, we simply sum the localized objectives:

$$\underset{\mathbf{B},\boldsymbol{\alpha}}{\text{minimize}} \quad \sum_{t\in\mathcal{T}}\sum_{i=1}^{N}(\mathbf{y}_i - \tilde{\mathbf{m}}_i(t))^\top[\mathbf{W}_i(t)\odot\widetilde{\mathbf{P}}_i](\mathbf{y}_i - \tilde{\mathbf{m}}_i(t))$$

where $\tilde{\mathbf{m}}_i(t) = \mathbf{b}^{(t)\top}\mathbf{x}_i + \boldsymbol{\alpha}^\top\mathbf{u}_i$ and where $\mathbf{B} = [\mathbf{b}^{(1)}, \dots \mathbf{b}^{(T)}] \in \mathbb{R}^{p_x \times T}$. We note that the only quantities that are shared across $t \in \mathcal{T}$ are the constant effects $\boldsymbol{\alpha}$ (apart from the variance parameters $\sigma^2, \sigma_\theta^2$.) In fact, when there are no constant effects, the objective is completely separable across $t \in \mathcal{T}$.

While the localized models do not directly translate to a model of the responses $\mathbf{y}_i$, the form of (18) does suggest a model induced by the choice of time points $\mathcal{T}$. Indeed, the squared loss is nothing but a quadratic form in $\mathbf{y}_i$. In particular, define the localized residuals $\tilde{\mathbf{r}}_i(t) = \mathbf{y}_i - \tilde{\mathbf{m}}_i(t)$, and consider the sum of localized squared errors,

$$\sum_{t\in\mathcal{T}}\tilde{\mathbf{r}}_i(t)^\top\widetilde{\mathbf{P}}_i(t)\tilde{\mathbf{r}}_i(t) = \mathbf{y}_i^\top\left[\sum_{t\in\mathcal{T}}\widetilde{\mathbf{P}}_i(t)\right]\mathbf{y}_i - 2\mathbf{y}_i^\top\left[\sum_{t\in\mathcal{T}}\widetilde{\mathbf{P}}_i(t)\tilde{\mathbf{m}}_i(t)\right] + \sum_{t\in\mathcal{T}}\tilde{\mathbf{m}}_i(t)^\top\widetilde{\mathbf{P}}_i(t)\tilde{\mathbf{m}}_i(t) \tag{18}$$

which suggests the induced model $\mathbf{y}_i \sim \mathcal{N}(\tilde{\mathbf{m}}_{i,\mathcal{T}}, \widetilde{\mathbf{V}}_{i,\mathcal{T}})$ where

$$\widetilde{\mathbf{V}}_{i,\mathcal{T}} := \widetilde{\mathbf{P}}_{i,\mathcal{T}}^{-1} = \left[\sum_{t\in\mathcal{T}}\widetilde{\mathbf{P}}_i(t)\right]^{-1}, \qquad \tilde{\mathbf{m}}_{i,\mathcal{T}} := \widetilde{\mathbf{V}}_{i,\mathcal{T}}\left[\sum_{t\in\mathcal{T}}\widetilde{\mathbf{P}}_i(t)\tilde{\mathbf{m}}_i(t)\right]. \tag{19}$$

In particular, this enables us to define an induced Gaussian marginal log-likelihood:

$$\ell_{i,\mathcal{T}}(\mathbf{B}, \boldsymbol{\alpha}, \sigma^2, \sigma_\theta^2) := \frac{1}{2}\log\det(\widetilde{\mathbf{P}}_{i,\mathcal{T}}/2\pi) - \frac{1}{2}\tilde{\mathbf{r}}_{i,\mathcal{T}}^\top\widetilde{\mathbf{P}}_{i,\mathcal{T}}\tilde{\mathbf{r}}_{i,\mathcal{T}} \tag{20}$$

where $\tilde{\mathbf{r}}_{i,\mathcal{T}} = \mathbf{y}_i - \tilde{\mathbf{m}}_{i,\mathcal{T}}$ are the global residuals. While this induced model share the same functional form in $\mathbf{y}_i$ as the sum of localized residual sums of squares, it does not share the same gradients with respect to the regression coefficients. Indeed, the gradient is computed using different residuals:

$$\nabla_{\mathbf{b}^{(t)}}\tilde{\mathbf{r}}_{i,\mathcal{T}}^\top\widetilde{\mathbf{P}}_{i,\mathcal{T}}\tilde{\mathbf{r}}_{i,\mathcal{T}} = -2\mathbf{X}_i^\top\widetilde{\mathbf{P}}_i(t)\tilde{\mathbf{r}}_{i,\mathcal{T}} \qquad \nabla_{\mathbf{b}^{(t)}}\sum_{t\in\mathcal{T}}\tilde{\mathbf{r}}_i(t)^\top\widetilde{\mathbf{P}}_i(t)\tilde{\mathbf{r}}_i(t) = -2\mathbf{X}_i\widetilde{\mathbf{P}}_i(t)\tilde{\mathbf{r}}_i(t).$$

Indeed, the term independent of $\mathbf{y}_i$ involving mean parameters differ:

$$\begin{aligned}\tilde{\mathbf{r}}_{i,\mathcal{T}}^\top \widetilde{\mathbf{P}}_{i,\mathcal{T}} \tilde{\mathbf{r}}_{i,\mathcal{T}} - \sum_{t\in\mathcal{T}} \tilde{\mathbf{r}}_i(t)^\top \widetilde{\mathbf{P}}_i(t) \tilde{\mathbf{r}}_i(t) &= \tilde{\mathbf{m}}_{i,\mathcal{T}}^\top \widetilde{\mathbf{P}}_{i,\mathcal{T}} \tilde{\mathbf{m}}_{i,\mathcal{T}} - \sum_{t\in\mathcal{T}} \tilde{\mathbf{m}}_i(t)^\top \widetilde{\mathbf{P}}_i(t) \tilde{\mathbf{m}}_i(t) \\ &= \sum_{t\in\mathcal{T}} \left[\tilde{\mathbf{m}}_{i,\mathcal{T}} - \tilde{\mathbf{m}}_i(t)\right]^\top \widetilde{\mathbf{P}}_i(t) \tilde{\mathbf{m}}_i(t)\end{aligned}$$

The induced covariance matrix $\widetilde{\mathbf{V}}_{i,\mathcal{T}}$ can be interpreted as an approximation to the underlying variance $\mathbf{V}_i$ emerging from the covariance kernel k_θ evaluated over all pairs in $\mathbf{t}_i$. However, it is unclear whether any choice of $k, h, \sigma^2, \rho_\theta, \mathcal{T}$ can generally be mapped back to a particular choice of kernel function k_θ, but we can make a few observations. The trivial case of independence in the local model (either from $\rho_\theta = 0$, $h = 0$ or the independence kernel $k_h(t, t') = \mathbb{1}[t = t']$) leads to a diagonal global precision $\widetilde{\mathbf{P}}_{i,\mathcal{T}}$ and a diagonal global variance $\widetilde{\mathbf{V}}_{i,\mathcal{T}}$. At the other extreme, when the weight vector is flat (typically for h large, or a uniform kernel with h large enough), we find that $\widetilde{\mathbf{P}}_{i,\mathcal{T}} \propto \mathbf{P}_i$ and thus $\widetilde{\mathbf{V}}_{i,\mathcal{T}} \propto \mathbf{V}_i$. In practice, we often get something in between: for regularly space sampling points $\mathbf{t}_i$ and for common choices of kernel functions, the total weight matrix $\mathbf{W}_{i,\mathcal{T}}$ will resemble a Toeplitz matrix with entries decreasing away from the diagonal. Then, taking the Hadamard product with $\mathbf{P}_i$ yields another Toeplitz-like matrix with small negative off-diagonal entries, again vanishing away from the diagonal. Its inverse again resembles a Toeplitz matrix with positive off-diagonal entries; the various parameters control if and how fast the entries vanishes to zero away from the diagonal.

2.9 Local & global sparsity

A convenient feature of local regression is that the value of $\boldsymbol{\beta}(t)$ at a specific time t is directly parameterized by $\mathbf{b}^{(t)}$. This is in contrast to spline basis expansion where the value of $\boldsymbol{\beta}(t)$ is a linear combination of basis functions active at time t. Hence, to find $\beta_j(t) = 0$, we only need $b_j^{(t)} = 0$, compared to requiring a consecutive set of spline weights to be zero. This allows us to use a simple sparsity-inducing penalty on the entries of $\mathbf{B}$, compared to overlapping group penalties used in spline methods (Wang and Kai, 2015; Wang et al., 2022).

We thus propose to encourage local sparsity by adding a Lasso penalty (Tibshirani, 1996) on $\mathbf{B}$ to the objective, namely, $\lambda \sum_{j=1}^{p_x} \sum_{t\in\mathcal{T}} \omega_j^{(t)} |b_j^{(t)}|$, where $\omega_j^{(t)}$ are weights and where $\lambda \geqslant 0$ is the regularization parameter. For example, when $\boldsymbol{\beta}(\cdot)$ consists of two components—an intercept and a group difference—, we are only interested in encouraging zeros in the second component so we would set $\omega_0^{(t)} = 0$.

Furthermore, we may be interested in encouraging $\beta_j(\cdot) \equiv 0$ altogether to identify if the jth covariate has any effect on the response. This suggests to add a group lasso penalty on the whole vector $\mathbf{b}_j = (b_j^{(1)}, \ldots, b_j^{(T)})$, leading to the sparse group Lasso penalty (Friedman et al., 2010; Simon et al., 2013):

$$\mathcal{P}_{\lambda,\alpha}(\mathbf{B}; \Omega) := \lambda \sum_{j=1}^{p_x} \left[(1-\alpha)\sqrt{T}\omega_j \|\mathbf{b}_j\|_2 + \alpha \sum_{t\in\mathcal{T}} \omega_j^{(t)} |b_j^{(t)}| \right] \tag{21}$$

where $\alpha \in [0, 1]$ is a tuning parameter balancing between encouraging global sparsity ($\alpha = 0$) and local sparsity ($\alpha = 1$), where ω_j are weights for the group lasso penalty, and where Ω contains all weights.

Our final objective is therefore given by

$$\mathcal{L}_\mathcal{T}(\mathbf{B}, \boldsymbol{\alpha}, \sigma^2, \sigma_\theta^2) = -\sum_{i=1}^{N} \ell_{i,\mathcal{T}}(\mathbf{B}, \boldsymbol{\alpha}, \sigma^2, \sigma_\theta^2) + \mathcal{P}_{\lambda,\alpha}(\mathbf{B}; \Omega). \tag{22}$$

The Lasso and group Lasso penalty are famously known for their estimation bias due to the shrinkage applied to non-zero values. To alleviate this issue, we propose to use adaptive penalties (Zou, 2006; Nardi and Rinaldo, 2008; Poignard, 2020), where the weights are set to $\omega_j = \|\hat{\mathbf{b}}_j^{\text{mle}}\|_2^{-\gamma}$ and $\omega_j^{(t)} = |\hat{b}_j^{(t),\text{mle}}|^{-\gamma}$ for some $\gamma > 0$, where $\widehat{\mathbf{B}}^{\text{mle}}$ contains the coefficients estimated without penalty.

3 Inference

3.1 Generalized Estimating Equations

For a model of the structure

$$y_{ij} = \beta(t_{ij}) + \varepsilon_{ij}, \qquad \boldsymbol{\varepsilon}_i \sim (\mathbf{0}, \Sigma),$$

Wang (2003, and other closely related papers) proposed the *seemingly unrelated kernel estimator* based on GEEs. Consider a kernel function $k_h(s) = k(s/h)/h$ and a q-th order polynomial kernel estimator. Then, $\beta^{(\ell)}(t)$ is updated to $\beta^{(\ell+1)}(t) = \alpha_0$ by solving

$$0 = \sum_{i=1}^{N} \sum_{j=1}^{n_i} k_h(t_{ij} - t)\mathbf{B}_{ij}(t)\mathbf{P}_i[\mathbf{y}_i - \boldsymbol{\mu}_{i(j)}] \tag{23}$$

for $\alpha_0, \ldots, \alpha_k$, where $\mathbf{P}_i$ is a working inverse precision matrix, where

$$\boldsymbol{\mu}_{i(j)}(t) = \begin{bmatrix} \beta^{(\ell)}(t_{i1}) \\ \vdots \\ \beta^{(\ell)}(t_{ij-1}) \\ \sum_{k=0}^{q} \alpha_k (t_{ij} - t)^k \\ \beta^{(\ell)}(t_{ij+1}) \\ \vdots \\ \beta^{(\ell)}(t_{in_i}) \end{bmatrix}$$

and where $\mathbf{B}_{ij}(t)$ is the design matrix for the q-th order polynomial, restricted only to the j-th observation (essentially the gradient of $\boldsymbol{\mu}_{i(j)}(t)$ wrt $\boldsymbol{\alpha}$.)

The main differences with the current approach are:

- The residuals for $j' \neq j$ are the true current residuals, not the ones implied by the local model.
- The weights $k_h(t_{ij} - t)$ are only applied once to the full expression rather than as observation weights.

The special case of $q = 0$ yields the GEE equation

$$0 = \sum_{i=1}^{N} \sum_{j=1}^{n_i} k_h(t_{ij} - t)\mathbf{e}_j\mathbf{P}_i[\mathbf{y}_i - \boldsymbol{\mu}_{i(j)}] \tag{24}$$

with

$$\boldsymbol{\mu}_{i(j)}(t) = \begin{bmatrix} \beta^{(\ell)}(t_{i1}) \\ \vdots \\ \beta^{(\ell)}(t_{ij-1}) \\ \alpha_0 \\ \beta^{(\ell)}(t_{ij+1}) \\ \vdots \\ \beta^{(\ell)}(t_{in_i}) \end{bmatrix}$$

In a GD scheme, we can view this as using the gradient

$$-\sum_{i=1}^{n} \mathbf{k}_i(t)^\top \mathbf{P}_i \mathbf{r}_i$$

where $\mathbf{r}_i$ is the vector of current residuals and where $\mathbf{k}_i(t)$ has entries $k_h(t_{ij} - t)$. The natural extension to a more general design $\mathbf{X}_i$ would be

$$-\sum_{i=1}^{n} \mathbf{X}_i^\top \operatorname{diag}(\mathbf{k}_i(t))\mathbf{P}_i\mathbf{r}_i$$

The main differences with the gradient below are

- using the true residuals $\mathbf{r}_i$ rather than the local residuals $\tilde{\mathbf{r}}_i(t)$
- using the matrix $\operatorname{diag}(\mathbf{k}_i(t))\mathbf{P}_i$ instead of $\widetilde{\mathbf{P}}_i(t) = \operatorname{diag}(\mathbf{k}_i(t))^{1/2}\mathbf{P}_i \operatorname{diag}(\mathbf{k}_i(t))^{1/2}$

3.2 Estimation of regression coefficients

For fixed variance parameters, $\mathcal{L}_\mathcal{T}$ is nothing but a sparse group Lasso penalized least squares objective in $\mathbf{B}$ and $\boldsymbol{\alpha}$. A natural approach for such problems is to use proximal gradient descent, which only requires gradient computation, step sizes and proximal evaluations. The gradients are given by

$$\nabla_{\boldsymbol{\alpha}}\mathcal{L}_\mathcal{T} = -\sum_{i=1}^{N}\sum_{t\in\mathcal{T}} \mathbf{U}_i^\top \widetilde{\mathbf{P}}_i(t)\tilde{\mathbf{r}}_i(t) \qquad \nabla_{\mathbf{b}^{(t)}}\mathcal{L}_\mathcal{T} = -\sum_{i=1}^{N} \mathbf{X}_i^\top \widetilde{\mathbf{P}}_i(t)\tilde{\mathbf{r}}_i(t) \tag{25}$$

where $\mathbf{U}_i = \mathbf{1}_{n_i}\mathbf{u}_i^\top$ and $\mathbf{X}_i = \mathbf{1}_{n_i}\mathbf{x}_i^\top$. Since the gradients are on different scales—$\nabla_{\boldsymbol{\alpha}}\mathcal{L}_\mathcal{T}$ should be around T times larger—, we utilize different step sizes for the two blocks of coefficients. To choose appropriate step lengths, we produce a diagonal upper bound on the Hessian matrix of $(\boldsymbol{\alpha}, \mathbf{B})$ as follows. First, we obtain upper bounds for the Hessian of $\boldsymbol{\alpha}$ and $\mathbf{B}$ separately, given by $L_{\boldsymbol{\alpha}}\mathbf{I}_{p_u}$ and $L_{\mathbf{B}}\mathbf{I}_{Tp_x}$ where $L_{\boldsymbol{\alpha}}$ and $L_{\mathbf{B}}$ are the largest eigenvalues of the respective Hessians. Then, we find the smallest c such that $\widetilde{L}_{\boldsymbol{\alpha}} = cL_{\boldsymbol{\alpha}}$ and $\widetilde{L}_{\mathbf{B}} = cL_{\mathbf{B}}$ produce an upper bound on the joint Hessian. The step-sizes are then chosen to be $1/\widetilde{L}_{\boldsymbol{\alpha}}$ and $1/\widetilde{L}_{\mathbf{B}}$, respectively. Since we do not penalize $\boldsymbol{\alpha}$, its update is simply $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \frac{1}{\widetilde{L}_{\boldsymbol{\alpha}}}\nabla_{\boldsymbol{\alpha}}\mathcal{L}_\mathcal{T}$. The update for $\mathbf{b}_j$ is given by the usual proximal gradient updates, namely,

$$\mathbf{b}_j \leftarrow \operatorname{prox}_{\mathcal{P}_{\lambda,\alpha}(\cdot;\Omega)/\widetilde{L}_{\mathbf{B}}}\left(\mathbf{b}_j^*\right), \qquad \mathbf{b}_j^* := \mathbf{b}_j - \frac{1}{\widetilde{L}_{\mathbf{B}}}\nabla_{\mathbf{b}_j}\mathcal{L}_\mathcal{T}. \tag{26}$$

The hierarchical structure of the sparse group Lasso penalty enables its proximal to be computed by composing the two proximal operators emerging from the two penalties. First, the Lasso proximal is the soft-thresholding operator,

$$[\mathbf{b}_j^{**}]^{(t)} = \operatorname{prox}_{\lambda\alpha\omega_j^{(t)}|\cdot|/\widetilde{L}_{\mathbf{B}}}([\mathbf{b}_j^*]^{(t)}) = \operatorname{sgn}([\mathbf{b}_j^*]^{(t)})\left(|[\mathbf{b}_j^*]^{(t)}| - \lambda\alpha\omega_j^{(t)}/\widetilde{L}_{\mathbf{B}}\right)_+, \tag{27}$$

and, second, the group Lasso penalty proximal is given by

$$\mathbf{b}_j \leftarrow \operatorname{prox}_{\lambda(1-\alpha)\sqrt{T}\omega_j\|\cdot\|_2/\widetilde{L}_{\mathbf{B}}}(\mathbf{b}_j^{**}) = \left(1 - \frac{\lambda(1-\alpha)\sqrt{T}\omega_j}{\widetilde{L}_{\mathbf{B}}\|\mathbf{b}_j^{**}\|_2}\right)_+ \mathbf{b}_j^{**}. \tag{28}$$

3.3 Optimization of localized covariance parameters

We are not interested in modeling or estimating the overall variance $\mathbf{V}_i$ of an observation $\mathbf{y}_i$, but as noted by Fan et al. (2007), there is some utility in estimating the parameters of the working covariance $\widetilde{\mathbf{V}}_i$ to improve the efficiency of the regression parameter estimates. In particular, we do not need a perfect match between the working correlation and the true correlation to obtain efficiency gain. Fan et al. (2007) propose to minimize the determinant of the covariance of estimate for the non-varying

regression effect $\boldsymbol{\alpha}$. We cannot proceed as such since our target quantity is the varying coefficient term $\boldsymbol{\beta}(\cdot)$ and the sparsity-inducing penalty prevents us from computing a covariance.

We estimate the variance parameters σ^2 and σ_θ^2 by directly maximizing the induced marginal log-likelihood (20). We note that the random intercept variance σ_θ^2 only occurs in the ratio σ^2/σ_θ^2; this suggests the reparameterization into the ratio with the noise variance $\rho_\theta = \sigma_\theta^2/\sigma^2$. This amounts to a localized variance of $\widetilde{\mathbf{V}}_i = \sigma^2 \left(\rho_\theta \mathbf{1}_{n_i}\mathbf{1}_{n_i}^\top + \mathbf{I}_{n_i}\right)$ and a localized precision of $\widetilde{\mathbf{P}}_i = \sigma^{-2}\left[\mathbf{I}_{n_i} - \mathbf{1}_{n_i}\mathbf{1}_{n_i}^\top/(n_i + \rho_\theta^{-1})\right]$.

Under this parameterization, the MLE for the noise variance σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{N} \tilde{\mathbf{r}}_{i,\mathcal{T}}^\top (\sigma^2 \widetilde{\mathbf{P}}_{i,\mathcal{T}}) \tilde{\mathbf{r}}_{i,\mathcal{T}},$$

where $\sigma^2\widetilde{\mathbf{P}}_{i,\mathcal{T}}$ is free of σ^2. We note that this estimate will be on the wrong scale. Indeed, suppose we rescale the kernel weights by some constant c, then this estimate will also be rescaled by the same quantity. However, for our purpose, we only need an estimate of σ^2 to evaluate the likelihood; the value of σ^2 itself is of little use otherwise. Then, any rescaling of the kernel will exactly cancel out with the rescaling of the MLE, so this is a non-issue.

The estimation of the random effect variance ratio ρ_θ does not have an analytical MLE. We therefore resort to Newton-Raphson updates, whose details are included in the Supplementary material. Again, the actual value of ρ_θ is of little interest apart from producing appropriate localized precision matrices.

Overall, we estimate all the model parameters in nested loops. The outer loop monitors convergence over the update of both groups of parameters; two inner loops update respectively mean parameters $(\mathbf{B}, \boldsymbol{\alpha})$ and variance parameters (σ^2, ρ_θ) until convergence while holding the other fixed. Mean parameter estimate do not tend to change much with changes in ρ_θ, so to outer loop typically converge in a few iterations and iterations beyond the first generally have short inner loops.

3.4 Estimation using proxy precision matrices

Alternatively, we can avoid the estimation of the variance parameters provided we have a good approximation for the precision matrices $\widetilde{\mathbf{P}}_i(t)$. First, we observe that the $-\mathbf{1}_{n_i}\mathbf{1}_{n_i}^\top/(n_i + \rho_\theta^{-1})$ term in $\widetilde{\mathbf{P}}_i$ does not change significantly with ρ_θ, suggesting that a rough estimate should be good enough for estimating the regression coefficients. This is essentially the approach suggested in Fan and Li (2012) where the parameterized covariance matrix emerging from a mixed model is replaced by a fixed proxy matrix during estimation. This avoids a significant computational burden as there is no need to recompute the precision matrices every time we update ρ_θ and we skip the estimation of ρ_θ altogether. The loss in accuracy will be largest when either n_i or ρ_θ are small, i.e., when the random effect importance vanishes.

Applied to our case, the proposed proxy matrix in Fan and Li (2012) amounts to fixing $\rho_\theta = \log(n)$. Typical microbial studies contain a few hundreds samples, meaning that $\log(n)$ will be around 5–6, so that ρ_θ^{-1} becomes negligible compared to n_i. When estimating variance parameters, we use this proxy matrix method as the initialization.

We observe some similarity with the spline method proposed by Wang et al. (2022) where a first step estimates a common precision matrix $\mathbf{P} \equiv \mathbf{P}_i$ using the empirical covariance of the residuals in an OLS model without sparsity, random effects nor smoothing; that estimate $\mathbf{P}$ is then used in the regularized problem. An important requirement for this approach is that all subjects are measured at a common set of time points, which we do not require here.

3.5 Tuning parameter selection

Our proposed method contains multiple parameters to be determined or optimized.

First, the time points at which to estimate the localized models, $\mathcal{T}$, need to be specified before estimation. When the number of sample time points is relatively small, we suggest using those directly. Otherwise, if observations are sample at irregular times or with high frequency, we suggest using a

suitably fine grid over the observed domain. The main principle guiding this choice is about the desired output: what are the time points of interest, and what is a relevant timescale to identify differential intervals?

Second, the smoothing is mostly controlled by the choice of the kernel function k and the kernel scale h. We use the squared exponential kernel $k(t, t') = \exp(-|t - t'|^2)$ by default, but finite-support kernels can be useful, especially in implementations exploiting sparsity. Our target application has few samples per subjects, so large $\mathbf{P}_i$ is not an issue, but other applications with large n_i may benefit from exact zeros. The choice of h can be done manually by the user informed by the sampling domain and frequency; in our implementation, we rescale time to the unit interval by default. Alternatively, we propose two information criteria described below as well as cross-validation to select h empirically.

Third, the sparsity-inducing penalty contains three main parameters. The regularization strength $\lambda > 0$ will be selected similarly to h using information criteria or cross-validation. The global-local weight $\alpha \in [0, 1]$ should be user-selected informed by the expected and desired sparsity. The adaptive strength $\gamma \geqslant 0$ is best set to be around $1/2$ or 1.

3.6 Information criteria

To select the kernel scale h and the regularization parameter λ, we propose an AIC and a BIC. The main difficulty resides in the ambiguous number of model parameters and sample size. To address these issues, we follow our general approach of localized models, inspired from other ICs used in kernel regression.

The naïve AIC would add $2\|\mathbf{B}\|_0$ to the log-likelihood; the naïve BIC would rather add $\|\mathbf{B}\|_0 \log(n)$. However, these penalties do not depend on h, so it is unclear how helpful they would be in selecting h. Many sparse kernel regression method utilizes a BIC penalty of the form $\|\mathbf{B}\|_0 \log(nh)/nh$ (Wang and Xia, 2009), which implicitly assumes an effective sample size of nh. We have found this penalty not very effective for selection purposes, especially with respect to h, so we propose a penalty using better estimates of degrees of freedom and effective sample size.

For the localized model at time $t \in \mathcal{T}$, not all n samples are used. In fact, we can use the weights $w_{ij}(t) = \sqrt{k_h(t_{ij}, t)}$ to estimate the effective sample size. The maximum contribution towards the likelihood a sample can have is when $t_{ij} = t$, in which case, its weight will be $k_h(0)$. This suggests the estimated sample size

$$\hat{n}_h(t) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{k_h(t_{ij}, t)}{k_h(0)}.$$

It turns out that nh is a crude estimate (or special case) of $\hat{n}_h(t)$. Consider uniformly distributed sampling times $t_{ij} \sim \text{Uniform}[0, 1]$ with a uniform kernel of radius h. Then, there will be, on average, $2nh$ samples falling within h of t (for t far enough from the boundary); the uniform kernel has a scaling factor of $1/2$, so we recover nh. The proposed estimate $\hat{n}_h(t)$ better captures the effective number of samples used to estimate the model at time t than nh. For large h, all n samples will have $k_h(t_{ij}, t) \approx k_h(0)$ and $\hat{n}_h(t)$ is capped by n from the scaling by the maximal value $k_h(0)$, while nh can exceed n. For very small $h \to 0$, only the t_{ij} exactly equal to t will contribute: if there are $n(t)$ such points, then $\hat{n}(t) = n(t)$, while $nh \to 0$.

The number of non-zero entries in $\mathbf{B}$, namely $\|\mathbf{B}\|_0$, is not an appropriate estimate of the degrees of freedom. Indeed, suppose we have a large h inducing estimated functions that are essentially constant but non-zero. Then, $\|\mathbf{b}_j\|_0 = T$, while there is really only a single degree of freedom. Kernel smoothing methods usually use $k_h(0)|\mathcal{D}|$, where $\mathcal{D} \subset \mathbb{R}$ is the time domain, as the degree of freedom. This quantity naturally scales with h since $k_h(0) = k(0)/h$. We propose to replace $\mathcal{D}$ by $\widehat{\mathcal{D}}_{h,\lambda}$ defined by the domain estimated to be non-zero with tuning parameters h, λ. In practice, we only estimate $\widehat{\mathcal{D}}_{h,\lambda}$ at a finite set of points $\mathcal{T}$, so we estimate $|\widehat{\mathcal{D}}_{h,\lambda}| \approx |\mathcal{D}| \|\mathbf{b}_j\|_0 / T$, that is, the original domain length $|\mathcal{D}|$ multiplied by the proportion of time points estimated to be non-zero $\|\mathbf{b}_j\|_0/T$. If $\mathcal{T}$ is irregularly-spaced over $\mathcal{D}$,

one may improve on this approximation by taking the unevenness into account: for instance, the term $|\mathcal{D}|/T$ can be interpreted as if the time t model is responsible for a fraction $1/T$ of the domain.

We finally define the AIC and BIC penalties as follows:

$$\mathrm{AIC}(h,\lambda) = -2\sum_{i=1}^{N} \ell_{i,\mathcal{T}} + 2\frac{k(0)|\mathcal{D}|\|\mathbf{B}\|_0}{hT} \tag{29}$$

$$\mathrm{BIC}(h,\lambda) = -2\sum_{i=1}^{N} \ell_{i,\mathcal{T}} + \sum_{t\in\mathcal{T}} \frac{k(0)|\mathcal{D}|\|\mathbf{b}^{(t)}\|_0}{hT} \log \hat{n}_h(t) \tag{30}$$

In particular, both criteria penalize complexity induced by rougher curves (small h) and denser estimates (large $\|\mathbf{B}\|_0$). We note that the BIC is not too different from the simpler $\|\mathbf{B}\|_0 \log(nh)/nh$. Indeed, with T often chosen as large as n (all observed time points) in kernel smoothing (Fan, 1993) and replacing $n_h(t) \leftarrow nh$, we recover the usual kernel BIC penalty (Wang and Xia, 2009), up to the scaling factor $k(0)|\mathcal{D}|$.

We also note that $k(0)|\mathcal{D}/hT$ can be understood as a rescaling of the number of parameter. For small h, that scale can exceed 1, leading to an estimated degree of freedom exceeding the practical number of parameters. At the opposite end, when h gets larger, this quantity will tend to 0, even though there is still one effective parameter in the model that can be understood as the time average. Hence, we finally suggest to clamp this scale between $1/T$ and 1: $k(0)|\mathcal{D}/hT \leftarrow (k(0)|\mathcal{D}/hT \wedge 1) \vee 1/T$.

3.7 Estimation of random effects

A natural estimate for the random effect $\theta_i(t_{ij})$ is the posterior mean. We can see reinterpret the localized model as the marginalization of the following hierarchical model:

$$\begin{aligned}
\tilde{\theta}_i &\sim \mathcal{N}(0, \sigma_\theta^2) \\
y_{ij} \mid \tilde{\theta}_i &\sim \mathcal{N}(\mathbf{b}^{(t)\top}\mathbf{x}_i + \mathbf{a}^\top \mathbf{u}_i + w_{ij}^{-1}(t)\tilde{\theta}_i, \sigma^2 w_{ij}^{-2}(t)).
\end{aligned}$$

Indeed,

$$\begin{aligned}
\mathbb{E}\{y_{ij}\} &= \mathbf{b}^{(t)\top}\mathbf{x}_i + \mathbf{a}^\top \mathbf{u}_i \\
\mathrm{Cov}(y_{ij}, y_{ik}) &= \mathbb{E}\{\mathrm{Cov}(y_{ij}, y_{ik} \mid \tilde{\theta}_i)\} + \mathrm{Cov}(\mathbb{E}\{y_{ij} \mid \tilde{\theta}_i\}, \mathbb{E}\{y_{ik} \mid \tilde{\theta}_i\}) \\
&= \mathbb{1}[j=k] w_{ij}^{-2}(t)\sigma^2 + w_{ij}^{-1}(t) w_{ik}^{-1}(t)\sigma_\theta^2 \\
&= w_{ij}^{-1}(t)[\mathbb{1}[j=k]\sigma^2 + \sigma_\theta^2] w_{ik}^{-1}(t).
\end{aligned}$$

Then, the posterior mean of $\tilde{\theta}_i$ can be shown to be given by

$$\mathbb{E}\{\tilde{\theta}_i(t) \mid \mathbf{y}_i\} = \rho_\theta \mathbf{1}^\top [\sigma^2 \widetilde{\mathbf{P}}_i] \,\mathrm{diag}(\mathbf{w}_i(t)) \tilde{\mathbf{r}}_i(t)$$

This allows us to consider adjusted residuals (I don't know what is the common name for these?):

$$\tilde{\mathbf{r}}_i(t) - \tilde{\theta}_i(t)\mathbf{1},$$

from which we can get an estimate of the noise variance by taking the average over all observations. Perhaps, an estimate of σ_θ^2 would be given by the empirical variance of the $\tilde{\theta}_i(t)$ (possibly restricted at the observed times). This also provides a natural quantity to be used in an information criterion:

$$n \log \hat{\sigma}^2 + \text{penalty},$$

which probably has some profile likelihood interpretation since the other terms will be constant.

This is not good. When h is small, the estimated RE is the residuals so we get a 0 estimate. Indeed, looking at the model, when $w_{ij} = 0$, we essentially drop the observation; for h small, we drop all but one observation. Then, the model is unidentifiable since the RE can capture all the error.

3.8 Implementation details

grid search: h first, then max lambda and warm start. mention -1/5.
solution path: largest lambda
warm start
tried acceleration (FISTA)

3.9 Bootstrap simultaneous confidence intervals

4 Numerical experiments

5 Application

6 Discussion

References

Jianqing Fan. Local Linear Regression Smoothers and Their Minimax Efficiencies. *The Annals of Statistics*, 21(1):196–216, 1993. ISSN 0090-5364. URL https://www.jstor.org/stable/3035587.

Jianqing Fan, Tao Huang, and Runze Li. Analysis of Longitudinal Data With Semiparametric Estimation of Covariance Function. *Journal of the American Statistical Association*, 102(478):632–641, June 2007. ISSN 0162-1459. doi: 10.1198/016214507000000095. URL https://doi.org/10.1198/016214507000000095.

Yingying Fan and Runze Li. Variable Selection in Linear Mixed Effect Models. *Annals of statistics*, 40(4):2043–2068, August 2012. ISSN 0090-5364. doi: 10.1214/12-AOS1028. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4026175/.

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso, January 2010. URL http://arxiv.org/abs/1001.0736. arXiv:1001.0736 [math, stat].

Xihong Lin and Raymond J Carroll. Semiparametric Regression for Clustered Data Using Generalized Estimating Equations. *Journal of the American Statistical Association*, 96(455):1045–1056, September 2001. ISSN 0162-1459. doi: 10.1198/016214501753208708. URL https://doi.org/10.1198/016214501753208708.

Yuval Nardi and Alessandro Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2(none):605–633, January 2008. ISSN 1935-7524, 1935-7524. doi: 10.1214/08-EJS200. URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-2/issue-none/On-the-asymptotic-properties-of-the-group-lasso-estimator-for/10.1214/08-EJS200.full.

Benjamin Poignard. Asymptotic theory of the adaptive Sparse Group Lasso. *Annals of the Institute of Statistical Mathematics*, 72(1):297–328, February 2020. ISSN 1572-9052. doi: 10.1007/s10463-018-0692-7. URL https://doi.org/10.1007/s10463-018-0692-7.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, April 2013. ISSN 1061-8600. doi: 10.1080/10618600.2012.681250. URL https://doi.org/10.1080/10618600.2012.681250.

Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x.

Hansheng Wang and Yingcun Xia. Shrinkage Estimation of the Varying Coefficient Model. *Journal of the American Statistical Association*, 104(486):747–757, June 2009. ISSN 0162-1459. doi: 10.1198/jasa.2009.0138. URL https://doi.org/10.1198/jasa.2009.0138.

Haonan Wang and Bo Kai. Functional Sparsity: Global Versus Local. *Statistica Sinica*, 25(4):1337–1354, 2015. ISSN 1017-0405. URL https://www.jstor.org/stable/24721236.

Naisyin Wang. Marginal Nonparametric Kernel Regression Accounting for Within-Subject Correlation. *Biometrika*, 90(1):43–52, 2003. ISSN 0006-3444. URL https://www.jstor.org/stable/30042018.

Zhengjia Wang, John Magnotti, Michael S. Beauchamp, and Meng Li. Functional group bridge for simultaneous regression and support estimation. *Biometrics*, 2022. ISSN 1541-0420. doi: 10.1111/biom.13684. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13684.

Rou Zhong, Chunming Zhang, and Jingxiao Zhang. Locally sparse estimator of generalized varying coefficient model for asynchronous longitudinal data, June 2022. URL http://arxiv.org/abs/2206.04315. arXiv:2206.04315 [stat].

Hui Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429, December 2006. ISSN 0162-1459. doi: 10.1198/016214506000000735. URL https://doi.org/10.1198/016214506000000735.

A Calculations

A.1 Variance parameter estimation

To estimate the localized random intercept variance parameter $\rho_\theta = \sigma_\theta^2/\sigma^2$, we proceed using Newton-Raphson updates. The first derivative of the induced log-likelihood with respect to ρ_θ is given by:

$$\frac{\partial}{\partial \rho_\theta}\mathcal{L}_\mathcal{T} = \sum_{i=1}^{N}\frac{\partial}{\partial \rho_\theta}\ell_{i,\mathcal{T}}(\mathbf{B},\boldsymbol{\alpha},\sigma,\rho_\theta) = \frac{1}{2}\sum_{i=1}^{N}\left[\frac{\partial}{\partial \rho_\theta}\log\det(\widetilde{\mathbf{P}}_{i,\mathcal{T}}) - \frac{1}{\sigma^2}\frac{\partial}{\partial \rho_\theta}\tilde{\mathbf{r}}_{i,\mathcal{T}}^\top(\sigma^2\widetilde{\mathbf{P}}_{i,\mathcal{T}})\tilde{\mathbf{r}}_{i,\mathcal{T}}\right].$$

Then,

$$\frac{\partial}{\partial \rho_\theta}\log\det(\widetilde{\mathbf{P}}_{i,\mathcal{T}}) = \operatorname{Tr}\left(\widetilde{\mathbf{P}}_{i,\mathcal{T}}^{-1}\frac{\partial}{\partial \rho_\theta}\widetilde{\mathbf{P}}_{i,\mathcal{T}}\right),$$

where

$$\frac{\partial}{\partial \rho_\theta}\widetilde{\mathbf{P}}_{i,\mathcal{T}} = \mathbf{W}_{i,\mathcal{T}} \odot \frac{\partial}{\partial \rho_\theta}\widetilde{\mathbf{P}}_i$$

with

$$\frac{\partial}{\partial \rho_\theta}\widetilde{\mathbf{P}}_i = \sigma^{-2}\frac{\partial}{\partial \rho_\theta}\left[\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n_i + \rho_\theta^{-1}}\right] = -\frac{\sigma^{-2}}{(n_i\rho_\theta+1)^2}\mathbf{1}\mathbf{1}^\top,$$

which implies

$$\begin{aligned}
\frac{\partial}{\partial \rho_\theta}\widetilde{\mathbf{P}}_i(t) &= -\frac{\sigma^{-2}}{(n_i\rho_\theta+1)^2}\mathbf{W}_i(t)\\
\frac{\partial}{\partial \rho_\theta}\widetilde{\mathbf{P}}_{i,\mathcal{T}} &= -\frac{\sigma^{-2}}{(n_i\rho_\theta+1)^2}\mathbf{W}_{i,\mathcal{T}}\\
\frac{\partial}{\partial \rho_\theta}\log\det(\widetilde{\mathbf{P}}_{i,\mathcal{T}}) &= -\frac{\sigma^{-2}}{(n_i\rho_\theta+1)^2}\operatorname{Tr}\left(\widetilde{\mathbf{P}}_{i,\mathcal{T}}^{-1}\mathbf{W}_{i,\mathcal{T}}\right).
\end{aligned}$$

We thus find

$$\frac{\partial}{\partial \rho_\theta}\mathcal{L}_\mathcal{T} = \frac{1}{2}\sum_{i=1}^{N}\left[\frac{-1}{(n_i\rho_\theta+1)^2}\operatorname{Tr}\left([\sigma^2\widetilde{\mathbf{P}}_{i,\mathcal{T}}]^{-1}\mathbf{W}_{i,\mathcal{T}}\right) + \frac{1}{\sigma^2(n_i\rho_\theta+1)^2}\tilde{\mathbf{r}}_{i,\mathcal{T}}^\top\mathbf{W}_{i,\mathcal{T}}\tilde{\mathbf{r}}_{i,\mathcal{T}}\right]$$

For the second derivative, we first note that

$$\begin{aligned}
\frac{\partial}{\partial \rho_\theta}\frac{1}{(n_i\rho_\theta+1)^2} &= \frac{-2n_i}{(n_i\rho_\theta+1)^3}\\
\frac{\partial}{\partial \rho_\theta}\operatorname{Tr}\left([\sigma^2\widetilde{\mathbf{P}}_{i,\mathcal{T}}]^{-1}\mathbf{W}_{i,\mathcal{T}}\right) &= \operatorname{Tr}\left(\frac{\partial}{\partial \rho_\theta}[\sigma^2\widetilde{\mathbf{P}}_{i,\mathcal{T}}]^{-1}\mathbf{W}_{i,\mathcal{T}}\right)\\
\frac{\partial}{\partial \rho_\theta}[\sigma^2\widetilde{\mathbf{P}}_{i,\mathcal{T}}]^{-1} &= -[\sigma^2\widetilde{\mathbf{P}}_{i,\mathcal{T}}]^{-1}\frac{\partial}{\partial \rho_\theta}[\sigma^2\widetilde{\mathbf{P}}_{i,\mathcal{T}}][\sigma^2\widetilde{\mathbf{P}}_{i,\mathcal{T}}]^{-1}\\
&= \frac{1}{(n_i\rho_\theta+1)^2}[\sigma^2\widetilde{\mathbf{P}}_{i,\mathcal{T}}]^{-1}\mathbf{W}_{i,\mathcal{T}}[\sigma^2\widetilde{\mathbf{P}}_{i,\mathcal{T}}]^{-1}.
\end{aligned}$$

Finally,

$$\begin{aligned}\frac{\partial^2}{\partial \rho_\theta^2}\mathcal{L}_\mathcal{T} = &\sum_{i=1}^{N} \frac{n_i}{(n_i\rho_\theta + 1)^3} \operatorname{Tr}\left([\sigma^2\widetilde{\mathbf{P}}_{i,\mathcal{T}}]^{-1}\mathbf{W}_{i,\mathcal{T}}\right) \\ &- \sum_{i=1}^{N} \frac{1}{2(n_i\rho_\theta + 1)^4} \operatorname{Tr}\left([\sigma^2\widetilde{\mathbf{P}}_{i,\mathcal{T}}]^{-1}\mathbf{W}_{i,\mathcal{T}}[\sigma^2\widetilde{\mathbf{P}}_{i,\mathcal{T}}]^{-1}\mathbf{W}_{i,\mathcal{T}}\right) \\ &- \sum_{i=1}^{N} \frac{n_i\tilde{\mathbf{r}}_{i,\mathcal{T}}^\top\mathbf{W}_{i,\mathcal{T}}\tilde{\mathbf{r}}_{i,\mathcal{T}}}{\sigma^2(n_i\rho_\theta + 1)^3}.\end{aligned}$$

Finally, the N-R update is given by

$$\rho_\theta \leftarrow \rho_\theta - \left(\frac{\partial^2\mathcal{L}_\mathcal{T}}{\partial\rho_\theta^2}\right)^{-1}\left(\frac{\partial\mathcal{L}_\mathcal{T}}{\partial\rho_\theta}\right)$$

We note that the global marginal likelihood may not be convex in ρ_θ, so N-R updates may be inappropriate. We propose a simple heuristic to escape local minima. If the second derivative is positive, the update for ρ_θ is taken to be doubling or halving, following the sign of the first derivative.

A.2 Hessian calculations & Lipschitz constant

A.3 Solution path initialization