Locally Sparse Varying Coefficient Mixed Model with Application to Longitudinal Microbial Differential Abundance Analysis

Simon Fontaine[1, *], Nisha J D'Silva[2], Marcell Costa de Medeiros[2], Ji Zhu[1], and Gen Li[3]

[1]Department of Statistics, University of Michigan
[2]Department of Periodontics and Oral Medicine, University of Michigan
[3]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 $y_{ij} = y_i(t_{ij}) \in \mathbb{R}$ and we define $\mathbf{y}_i = y_i(\mathbf{t}_i) = (y_{i1}, \dots, y_{in_i})$ 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, for a method that does.) A practical solution to using asynchronous covariates in our proposed method is to pre-process the curves using functional PCA .

2.2 Varying coefficient mixed model

Our main goal is to study the relationship between covariates $\mathbf{x}_i$ and the functional response $y_i(\cdot)$, 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 $\mathrm{Cov}(\theta_i(t), \theta_i(t')) = \sigma^2 k_\theta(t, t')$ for some symmetric positive definite kernel k_θ. Define $\mathbf{K}_\theta(\mathbf{t})$ as the (unscaled) covariance matrix for a random process 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

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

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 $\mathbf{D}_i(t) = \operatorname{diag}(\mathbb{1}[t = t_{ij}])$. 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

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

Looking at the difference between (4) and (5), 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:

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

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

The previous gradient can be seen as a special case of the *seemingly unrelated kernel estimator* proposed by Wang (2003). Indeed, a solution to $\nabla^{k_h}_{\boldsymbol{\beta}(t)}\ell(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) = \mathbf{0}$ is a solution to the corresponding estimating equation.

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

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

2.3.1 Time points of interest

The score equation is defined for a specific choice of t. Then, for every time point of interest $\mathcal{T} = \{t^{(1)}, \ldots, t^{(S)}\}$ (e.g., all observed sampling times), we impose

$$\nabla^{k_h}_{\boldsymbol{\beta}(t)}\ell(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) = \mathbf{0}, \qquad t \in \mathcal{T}.$$

Then, we collect all evaluations of $\boldsymbol{\beta}(\cdot)$ of interest into the $p_x \times S$ matrix $\mathbf{B}$ with columns $\boldsymbol{\beta}(t^{(s)}) \in \mathbb{R}^{p_x}$ and with rows $\mathbf{b}_j = \beta_j(\mathcal{T}) \in \mathbb{R}^S$.

2.3.2 Hessian

To obtain a starting value for the step-size we define a kernel-weighted Hessian matrix whose eigenvalues can provide an approximate Lipschitz constant.

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

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

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

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

2.4 Local and 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):

$$\mathcal{P}_{\lambda,\alpha}(\mathbf{B};\Omega) := \lambda\sum_{j=1}^{p_x}\left[(1-\alpha)\sqrt{T}\omega_j\|\mathbf{b}_j\|_2 + \alpha\sum_{t\in\mathcal{T}}\omega_j^{(t)}|b_j^{(t)}|\right] \tag{11}$$

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

$$\mathcal{L}(\mathbf{B},\boldsymbol{\alpha}) = -\sum_{i=1}^{N}\ell_{i,}(\mathbf{B},\boldsymbol{\alpha}) + \mathcal{P}_{\lambda,\alpha}(\mathbf{B};\Omega). \tag{12}$$

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.

2.4.1 Proximal updates

The update for $\mathbf{b}_j$ is given by the usual proximal gradient updates, namely,

$$\mathbf{b}_j \leftarrow \text{prox}_{\eta \mathcal{P}_{\lambda,\alpha}(\cdot;\Omega)} \left(\mathbf{b}_j^*\right), \qquad \mathbf{b}_j^* := \mathbf{b}_j - \eta \nabla_{\mathbf{b}_j} \mathcal{L}_{\mathcal{T}}. \tag{13}$$

for some stepsize η.

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,

$$[\mathbf{b}_j^{**}]^{(t)} = \text{prox}_{\eta\lambda\alpha\omega_j^{(t)}|\cdot|}([\mathbf{b}_j^*]^{(t)}) = \text{sgn}([\mathbf{b}_j^*]^{(t)}) \left(|[\mathbf{b}_j^*]^{(t)}| - \eta\lambda\alpha\omega_j^{(t)}\right)_+, \tag{14}$$

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

$$\mathbf{b}_j \leftarrow \text{prox}_{\eta\lambda(1-\alpha)\sqrt{T}\omega_j\|\cdot\|_2}(\mathbf{b}_j^{**}) = \left(1 - \frac{\eta\lambda(1-\alpha)\sqrt{T}\omega_j}{\|\mathbf{b}_j^{**}\|_2}\right)_+ \mathbf{b}_j^{**}. \tag{15}$$

2.4.2 Solution path

To select the regularization parameter λ, we first find the smallest value such that we find a completely sparse solution. We can find by setting $\mathbf{B} = \mathbf{0}$ in checking what conditions the gradients must satisfy in order for the proximal update to remain $\mathbf{0}$. We find $\mathbf{b}_j^* = \eta \nabla_{\mathbf{b}_j}^{k_h} \mathcal{L}$ so that the Lasso penalty shrinks back to 0 provided

$$|\nabla_{\mathbf{b}_j^{(t)}}^{k_h} \mathcal{L}| \leqslant \lambda\alpha\omega_j^{(t)}.$$

Similarly, the group Lasso penalty proximal will remain null provided

$$\|\nabla_{\mathbf{b}_j}^{k_h} \mathcal{L}\|_2 \leqslant \lambda(1-\alpha)\sqrt{T}\omega_j.$$

This leads to the condition

$$\lambda \geqslant \max_j \left[\max_t \frac{1}{\alpha\omega_j^{(t)}} |\nabla_{\mathbf{b}_j^{(t)}}^{k_h} \mathcal{L}|\right] \vee \frac{1}{(1-\alpha)\sqrt{T}\omega_j} \|\nabla_{\mathbf{b}_j}^{k_h} \mathcal{L}\|_2$$

Of course, anytime we would divide by zero, we simply omit the corresponding bound.

2.5 Estimating the covariance parameters

Recall that $\mathbf{P}_i = \mathbf{V}_i^{-1}$ where $\mathbf{V}_i = \sigma^2 \left(\mathbf{K}_\theta(\mathbf{t}_i) + \mathbf{I}_{n_i}\right)$. 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 covariances 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 Fan et al. (2007), efficiency gains can be obtained even though the working covariance does not exactly match the true covariance. In particular, even a rough optimization of the working covariance can lead to near-optimal estimation efficiency.

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

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

and the AR(1) model, with covariance function

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

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 parameterizing the covariance function as a multiple of σ^2, we can write $\mathbf{V}_i = \sigma^2 \mathbf{C}_i$, where $\mathbf{C}_i = \mathbf{K}_\theta(\boldsymbol{\tau}) + \mathbf{I}$. Then, the estimate for σ^2 is given by

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

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

To update the parameters of $\mathbf{C}_i$, we perform Newton-Raphson steps. Directly, the first derivative may be computed using the following expressions

$$\begin{aligned}
\partial \ell(\boldsymbol{\tau}) &= -\frac{1}{2} \sum_{i=1}^{N} \partial \log\det \mathbf{C}_i + \sigma^{-2} \partial [\mathbf{r}_i^\top \mathbf{C}_i^{-1} \mathbf{r}_i] \\
\partial \log\det \mathbf{C}_i &= \operatorname{Tr}\left(\mathbf{C}_i^{-1} \partial \mathbf{C}_i\right) \\
\partial [\mathbf{r}_i^\top \mathbf{C}_i^{-1} \mathbf{r}_i] &= -\mathbf{r}_i^\top \left(\mathbf{C}_i^{-1} \partial \mathbf{C}_i \mathbf{C}_i^{-1}\right) \mathbf{r}_i,
\end{aligned}$$

where $\partial \mathbf{C}_i$ depends on the parameterization. For example, in the compound symmetry structure, namely $\mathbf{C}_i = \mathbf{I} + r_\theta \mathbf{1}\mathbf{1}^\top$, we find $\partial \mathbf{C}_i = \mathbf{1}\mathbf{1}^\top$. For the second derivative, we have

$$\begin{aligned}
\partial \ell(\boldsymbol{\tau}) &= -\frac{1}{2} \sum_{i=1}^{N} \partial^2 \log\det \mathbf{C}_i + \sigma^{-2} \partial^2 [\mathbf{r}_i^\top \mathbf{C}_i^{-1} \mathbf{r}_i] \\
\partial^2 \log\det \mathbf{C}_i &= \operatorname{Tr}\left(-\mathbf{C}_i^{-1} \partial \mathbf{C}_i \mathbf{C}_i^{-1} \partial \mathbf{C}_i + \mathbf{C}_i^{-1} \partial^2 \mathbf{C}_i\right) \\
\partial^2 [\mathbf{r}_i^\top \mathbf{C}_i^{-1} \mathbf{r}_i] &= -\mathbf{r}_i^\top \left(-2\mathbf{C}_i^{-1} \partial \mathbf{C}_i \mathbf{C}_i^{-1} \partial \mathbf{C}_i \mathbf{C}_i^{-1} + \mathbf{C}_i^{-1} \partial^2 \mathbf{C}_i \mathbf{C}_i^{-1}\right) \mathbf{r}_i
\end{aligned}$$

In the compound symmetry structure, $\partial^2 \mathbf{C}_i = \mathbf{0}$.

2.6 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.

To select the kernel scale h and the regularization parameter λ, we propose an AIC, a BIC and an EBIC. 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.

2.6.1 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 $k_h(t_{ij}, t)$ 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 $\hat{n}_h(t) \to n(t)$, while $nh \to 0$.

2.6.2 Effective degrees of freedom

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 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: $k(0)|\mathcal{D}|/hT \leftarrow (k(0)|\mathcal{D}|/hT \wedge 1) \vee 1/T$.

2.6.3 Information criterion

We finally define the ICs as follows:

$$\mathrm{AIC}(h, \lambda) = -2 \sum_{i=1}^{N} \ell + 2 \frac{k(0)|\mathcal{D}|\|\mathbf{B}\|_0}{hT} \tag{16}$$

$$\mathrm{BIC}(h, \lambda) = -2 \sum_{i=1}^{N} \ell + \sum_{t \in \mathcal{T}} \frac{k(0)|\mathcal{D}|\|\mathbf{b}^{(t)}\|_0}{hT} \log \hat{n}_h(t) \tag{17}$$

$$\mathrm{EBIC}(h, \lambda, \nu) = -2 \sum_{i=1}^{N} \ell + \sum_{t \in \mathcal{T}} \frac{k(0)|\mathcal{D}|\|\mathbf{b}^{(t)}\|_0}{hT} \log \hat{n}_h(t) + \nu \frac{k(0)|\mathcal{D}|\|\mathbf{B}\|_0}{hT} \log(p_{\max}) \tag{18}$$

where $p_{\max}$ is the maximal number of parameters, namely Sp_x and where $\nu \geqslant 0$ (typically $\nu \in \{0.5, 1\}$.) Empirical studies suggest that $\mathrm{EBIC}(h, \lambda, 0.5)$ provides the most accurate support recovery.

2.7 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 $\mathbf{y}_i$:

$$\mathbf{V}_i = \phi \mathbf{S}_i^{1/2} \mathbf{C}_i \mathbf{S}_i^{1/2},$$

where $\mathbf{S}_i = \mathrm{diag}(\nu(m_{ij}), i = 1, \ldots, n_i)$, 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:

$$\nabla^{k_h}_{\boldsymbol{\beta}(t)} \ell_Q(\boldsymbol{\beta}(\cdot), \boldsymbol{\alpha}) := -\sum_{i=1}^{N} \mathbf{X}_i^\top \mathrm{diag}(\mathbf{k}_i(t)) \mathrm{diag}(\mathbf{g}(\mathbf{m}_i)) \mathbf{S}_i^{-1/2} \mathbf{C}_i^{-1} \mathbf{S}_i^{-1/2} \mathbf{r}_i, \tag{19}$$

where $\mathbf{g}(\mathbf{m}_i)$ has entries $\frac{\mathrm{d}}{\mathrm{d}\eta} g^{-1}(\eta)|_{\eta=\eta_{ij}}$.

3 Numerical experiments

4 Application

5 Discussion

6 Aknowledgments

Nisha J D'Silva's research was supported by the National Institutes of Health grant R35DE027551. Gen Li's research was partially supported by the National Institutes of Health grant R03DE031296.

We are grateful to Prof. Grace Chen (UM Hematology and Oncology) and Prof. Tom Schmidt (UM Microbiology and Immunology, UM Evolutionary Biology) and his graduate students for data collection, sequencing and preparation, and for their useful comments.

References

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.

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].

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.