Locally sparse function-on-function regression

Mauro Bernardi and Antonio Canale,

Dipartimento di Scienze Statistiche,

Università degli Studi di Padova, Padova, Italy,

and

Marco Stefanucci

Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche "Bruno de Finetti",

Università degli Studi di Trieste, Trieste, Italy

Abstract

In functional data analysis, functional linear regression has attracted significant attention recently. Herein, we consider the case where both the response and covariates are functions. There are two available approaches for addressing such a situation: concurrent and nonconcurrent functional models. In the former, the value of the functional response at a given domain point depends only on the value of the functional regressors evaluated at the same domain point, whereas, in the latter, the functional covariates evaluated at each point of their domain have a non-null effect on the response at any point of its domain. To balance these two extremes, we propose a locally sparse functional regression model in which the functional regression coefficient is allowed (but not forced) to be exactly zero for a subset of its domain. This is achieved using a suitable basis representation of the functional regression coefficient and exploiting an overlapping group-Lasso penalty for its estimation. We introduce efficient computational strategies based on majorization-minimization algorithms and discuss appealing theoretical properties regarding the model support and consistency of the proposed estimator. We further illustrate the empirical performance of the method through simulations and two applications related to human mortality and bidding the energy market.

arXiv:2109.13042v1 [stat.ME] 27 Sep 2021

1 Introduction

The undergoing technological advancement enables the collection and storage of high-resolution data that can be modelled as smooth functions (e.g., curves or surfaces). Functional data analysis (FDA) is a branch of statistics that models such data through suitable functional counterparts of successful methods and models developed for standard Euclidean data, such as clustering, regression, and classification (Ramsay and Silverman, 2005; Horváth and Kokoszka, 2012; Hsing and Eubank, 2015). Functional regression is one of the building blocks of FDA and has received remarkable attention in theory, methods, and applications (see Morris, 2015, for a recent review).

Herein, we focus on the general case of functional regression with functional responses. We developed a model and related methods for its implementation, representing a bridge between two successful alternative models. To set up the notation, we assume that for the generic i-th statistical unit ($i = 1, \ldots, n$), a functional response $y_i(s)$ with $s \in \mathcal{S}$ is available along with p functional covariates $x_{ij}(t)$ ($j = 1, \ldots, p$ and $t \in \mathcal{T}_j$) and $\mathcal{S}$ and $\mathcal{T}_j$ ($j = 1, \ldots, p$) subsets of $\mathbb{R}$, denoting the domains of the functional data y and x_j, respectively. The first available modelling strategy, namely, the concurrent functional linear model, assumes that the functional data are observed on the same domain, i.e., $\mathcal{T}_j = \mathcal{S}$ for $j = 1, \ldots, p$, and the relation between the response and the predictors is given as

$$y_i(s) = \alpha(s) + \sum_{j=1}^{p} x_{ij}(s)\psi_j(s) + e_i(s), \tag{1}$$

where $\alpha(s)$ is a functional intercept, $\psi_j(s)$ are functional regression coefficients, and $e_i(s)$ are functional zero-mean random errors. In the concurrent model, the covariates x_j influence $y(s)$ only through their values $x_j(s)$ at the domain point $s \in \mathcal{S}$. As a more general approach, the nonconcurrent functional linear model allows $y_i(s)$ to entirely depend on the functional regressors, and specifically,

$$y_i(s) = \alpha(s) + \sum_{j=1}^{p} \int x_{ij}(t)\psi_j(t, s)dt + e_i(s), \tag{2}$$

where $\psi_j(t, s)$ is the kernel function determining the impact of x_{ij} evaluated at domain point $t \in \mathcal{T}_j$ on $y_i(s)$. The nonconcurrent model is the default choice when the response variable and the covariates do not share the same domain or, even when sharing the same domain, the value of $y_i(s)$ can be assumed to depend on the functional regressors entirely and not just for their

values at $s \in \mathcal{S}$. The nonconcurrent model offers great flexibility, but the flexibility increases complexity, both in terms of interpretation and computation. An in-between solution, motivated by applications in which $\mathcal{T}_j = \mathcal{T} = \mathcal{S}$ is a time domain, is represented by the so-called historical functional linear model of Malfait and Ramsay (2003). This approach restricts the domain of integration of the integral in (2) to the set $\mathcal{T}(s) = \{t \in \mathcal{T} : t < s\}$ leading to

$$y_i(s) = \alpha(s) + \sum_{j=1}^{p} \int_{\mathcal{T}(s)} x_{ij}(t)\psi_j(t,s)dt + e_i(s). \tag{3}$$

Note that this model, for the general point $s \in \mathcal{S}$, can be interpreted as a nonconcurrent model up to point s. Despite providing an interesting intermediate solution, there are many situations in which the dependence through the interval $\mathcal{T}(s)$ is difficult to be justified.

Herein, we introduce a hybrid solution that combines the simplicity in terms of interpretation of the concurrent and historical models with the flexibility of the nonconcurrent model. Our goal is to introduce a noncconcurrent functional linear model that allows for local sparsity patterns. Specifically, we want that $\psi_j(t,s) = 0$ for $(t,s) \in \mathcal{F}_0$ with $\mathcal{F}_0$ being a suitable subset of $\mathcal{F}$, thus inducing locally sparse Hilbert-Schmidt operators ψ_j. In defining $\mathcal{F}_0$, we do not consider those regions where $\psi(t,s)$ is zero because the kernel changes its sign or where it is tangential to the plane 0. A formal definition of $\mathcal{F}_0$ used henceforth is thus

$$\mathcal{F}_0 = \{(t,s) \in \mathcal{F} : \psi(t,s) = 0, \exists \bar{N} \subseteq B_\epsilon((t,s)); \mu(\bar{N}) > 0 \text{ and } \psi(t',s') = 0, \forall (t',s') \in \bar{N}, \forall \epsilon > 0\},$$

where $B_\epsilon(t,s)$ is a ball of radius $\epsilon > 0$ of the point (t,s) and $\mu(\cdot)$ is the Lebesgue measure. Note that, although in (3) the region $\mathcal{F}_0$ is fixed *a priori* and set equal to $\mathcal{F}_0 = \{(t,s) \in \mathcal{F} : t \geq s\}$, we do not require any specific sparsity pattern, rather we learn it from the data, as discussed in the following sections.

Different approaches to sparsity and regularization have been used in regression models for functional data, but none of them can be readily adapted for our purpose since they are all defined in a simpler function on scalar regression setting, i.e., y_i are scalar responses. Lee and Park (2012), for example, after representing the functional regression coefficient with a splines basis expansion, introduced a Lasso-type penalty for basis coefficients. The proposed solution has the great benefit of regularizing the estimator of the functional regression coefficients and enjoys nice asymptotic properties. However, several zeroes in the basis function coefficient do

3

not map to a zero in the functional object represented by the basis, and thus, using this approach would not necessarily induce sparsity in the functional coefficients. James et al. (2009) proposed a model with an interpretable functional regression coefficient that can be exactly zero, flat, and different from zero, or linear in local areas of its univariate domain. This is achieved by inducing sparsity in the general D-th derivative of the functional regression coefficient using a constant basis expansion and an ℓ_1 penalization. On the surface, this solution seems intuitive and successful, but a zero on a general subregion of the regression function requires grid points that fall into the region to be simultaneously zero, which the plain ℓ_1 regularization does not warrant in general. In addition, in our setting, the regression coefficient is not a univariate curve but a bivariate surface, and the use derivatives in two dimensions is less straightforward to apply. Lin et al. (2017) proposed a smooth and locally sparse estimator of the coefficient function based on the combination of smoothing splines with a functional smoothly clipped absolute deviation (SCAD) penalty of Fan and Li (2001). Zhou et al. (2013) proposed a two-stage sparse estimator exploiting the properties of B-spline basis expansion, where an initial estimate is obtained using a Dantzig selector (Candes et al., 2007), and the spline coefficient is later refined using a group adaptation of the SCAD penalty. The only contribution dealing with a function-on-function situation is the recent manuscript by Centofanti et al. (2020) who estimated sparse functional coefficients through a functional version of the Lasso penalty.

Our proposed solution exploits a B-spline local property similarly to Zhou et al. (2013), but directly relies on minimizing a suitable objective function, as discussed in Section 2. This objective function includes an overlapping group-Lasso penalty (Jenatton et al., 2011) that ensures the desired sparsity in ψ_j. It is minimized using a fast and reliable numerical strategy proposed in Section 3. The properties of the induced estimator are discussed in Section 4. A detailed simulation study to evaluate the empirical performance of the proposed model compared to that of the state-of-the-art models is presented in Section 5. The results show that the proposed model outperforms other models in estimating both the region of sparsity and the value of the functional regression coefficient where it is different from zero. In Section 6, we specify the model in functional time-series settings and analyze two datasets related to human mortality and bidding in energy markets. Our analysis suggests that bridging between concurrent and nonconcurrent functional models results in better performance both in terms of goodness-of-fit and qualitative interpretation of the results. Section 7 concludes the paper.

4

2 Locally sparse functional model

Without loss of generality, we consider the case in which the functional data y_i are centered in zero, and a single functional covariate x_i is available for $i = 1, \ldots, n$, so that $\alpha(s) = 0$ for all $s \in \mathcal{S}$ and $p = 1$. We also assume that ψ is bounded and defined on a compact domain $\mathcal{F}$. Thus, (2) becomes

$$y_i(s) = \int x_i(t)\psi(t,s)dt + e_i(s). \tag{4}$$

The proposed locally sparse functional regression (LSFR) model relies on the introduction of a specific basis representation for the kernel ψ and the minimization of a suitable objective function. These are described in the next two sections.

2.1 Kernel basis representation

We employ the common FDA practice of representing the functional objects using of basis expansions. Specifically, we select two bases in L_2, e.g., $\{\theta_l(s), l = 1, \ldots, L\}$ and $\{\varphi_m(t), m = 1, \ldots, M\}$, where each θ_l is defined on $\mathcal{S}$, each φ_m is defined on $\mathcal{T}$, and the number of basis is L and M, respectively. Exploiting a tensor product expansion of these two, we represent the kernel ψ in (4) as

$$\psi(t,s) = \sum_{m=1}^{M}\sum_{l=1}^{L}\psi_{ml}\varphi_m(t)\theta_l(s),$$

alternatively, in matrix form as

$$\psi(t,s) = (\varphi_1(t), \ldots, \varphi_M(t))\begin{pmatrix} \psi_{11} & \cdots & \psi_{1L} \\ \vdots & \ddots & \vdots \\ \psi_{M1} & \cdots & \psi_{ML} \end{pmatrix}\begin{pmatrix} \theta_1(s) \\ \vdots \\ \theta_L(s) \end{pmatrix} = \boldsymbol{\varphi}(t)^T\boldsymbol{\Psi}\boldsymbol{\theta}(s). \tag{5}$$

where $\psi_{ml} \in \mathbb{R}$ for $l = 1, \ldots, L$ and $m = 1, \ldots, M$.

To achieve the desired sparsity property in $\psi(t,s)$, similarly to Zhou et al. (2013), first assume that the elements in equation (5) are B-splines (De Boor, 1978) of order d. A B-spline of order d is a piecewise polynomial function of degree $d-1$ and is defined by a set of knots, which represent the values of the domain where the polynomials meet. Based on (5), $\{\theta_l(s), l = 1, \ldots, L\}$ and $\{\varphi_m(t), m = 1, \ldots, M\}$ are B-splines of order d with $L - d$ and $M - d$ interior knots, respectively, and two external knots each.

Suitable zero patterns in the B-spline basis coefficients of $\boldsymbol{\Psi}$ induce sparsity of $\psi(t,s)$. Let $\tau_1 < \cdots < \tau_m < \cdots < \tau_{M-d+2}$ and $\sigma_1 < \cdots < \sigma_l < \cdots < \sigma_{L-d+2}$ denote the knots defining the tensor product splines in (5), with $\tau_m \in \mathcal{T}$ and $\sigma_l \in \mathcal{S}$. For $m = 1, \ldots, M - d + 1$ and $l = 1, \ldots, L - d + 1$ let $\mathcal{F}_{m,l} \in \mathcal{F}$ be the rectangular subset of $\mathcal{F}$ defined as

$$\mathcal{F}_{m,l} = (\tau_m, \tau_{m+1}) \times (\sigma_l, \sigma_{l+1}). \tag{6}$$

Hence, to obtain $\psi(t,s) = 0$ for each $(t,s) \in \mathcal{F}_{m,l}$, it is sufficient that all the coefficients $\psi_{m',l'}$ with $m' = m, \ldots, m + d - 1$ and $l' = l, \ldots, l + d - 1$ are jointly zero. In general $\psi(t,s)$ equals zero in the region identified by two pairs of consecutive knots if the related $d \times d$ block of coefficients of $\boldsymbol{\Psi}$ is entirely set to zero. This suggests that $\boldsymbol{\Psi}$ should be suitably partitioned in several blocks of dimensions $d \times d$ on which a joint sparsity penalty is induced. This is discussed further in the next section.

2.2 Sparsity-inducing norm

Consistent with the above discussion and with successful approaches in statistics and machine learning, we minimize an objective function having the following form:

$$\frac{1}{2} \sum_{i=1}^{n} \int \left(y_i(s) - \int x_i(t)\psi(t,s)dt \right)^2 ds + \lambda\Omega(\boldsymbol{\Psi}), \tag{7}$$

where the first summand is a goodness-of-fit index while the second a suitable penalty.

Studies on statistics and machine learning have proposed numerous approaches for the sparsity-inducing Ω, including Lasso (Tibshirani, 1996; Efron et al., 2004). In our setting, however, Lasso would yield sparsity by treating each parameter individually, regardless of its position in $\boldsymbol{\Psi}$, which is against our desiderata. The first extension of Lasso involving the concept of groups of coefficients is the group-Lasso reported by Yuan and Lin (2006). This approach considers a partition of all the coefficients into a certain number of (disjoint) subsets and eventually allows some of these groups to shrink to zero.

On the surface, the group-Lasso approach is a promising solution to our problem, but we discuss this in more detail in what follows considering the graphical representation of a possible grouping, coherent with the group-Lasso definition, reported in panel (a) of Figure 1. Setting all coefficients belonging to one of the tiles of panel (a) to zero will force $\psi(t,s) = 0$ for (t,s) in a
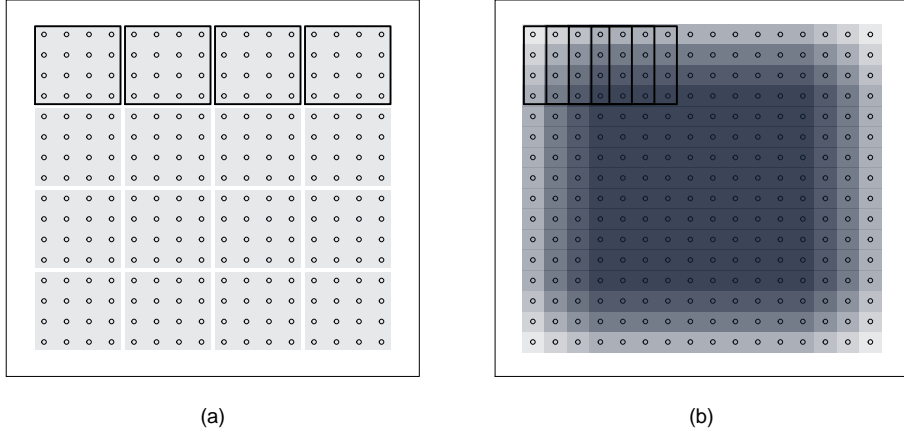
6

Figure 1: Possible coverings of the matrix Ψ ($L = M = 16$) with blocks of size $d \times d$ with $d = 4$: disjoint covering (a) and overlap covering (b). Darker color denotes the superimposition of different blocks. The first four blocks are represented with thick black borders in each subfigure.

specific set $\mathcal{F}_{m,l}$. In fact, this kind of solution does not provide enough model flexibility in many respects. First, the construction requires L and M to be multiple of d. Second, it is impossible to induce $\psi(t,s) = 0$ for $(t,s) \in \mathcal{F}_{m,l}$ and $\psi(t',s') \neq 0$ for (t',s') belonging to neighboring sets of $\mathcal{F}_{m,l}$ for any m and l not equal to $k_1 d + 1$ and $k_2 d + 1$, respectively, where k_1 and k_2 are nonnegative integers. Third, $\psi(t,s) = 0$ for (t,s) belonging to the union of two contiguous rectangles $\mathcal{F}_{m,l} \cup \mathcal{F}_{m+1,l}$ if and only if two consecutive and disjoint blocks of coefficients are jointly zero, leading to $\psi(t,s) = 0$ for each (t,s) belonging to the superset $\bigcup_{j=0}^{d+1} \mathcal{F}_{m+j,l}$. Figure 6 in the Appendix shows two examples of these key limitations of the disjoint grouping of coefficients.

To overcome these issues, instead of a disjoint partition, we define an overlapping sequence of blocks of size $d \times d$. Specifically, we introduce the block index $b = 1, \ldots, B$ with $B = (M - d + 1) \times (L - d + 1)$ denoting the total number of blocks. Notably, there is a block for each set $\mathcal{F}_{m,l}$. The generic b-th block contains the coefficients in the set

$$\big\{ \psi_{ml} \;:\; m = m^*, \ldots, m^* + d - 1,$$
$$l = l^*, \ldots, l^* + d - 1, m^* = \left\lceil \frac{b}{L - d + 1} \right\rceil, l^* = b \bmod (L - d + 1) \big\}, \qquad (8)$$

where $A \bmod B$ represents the remainder of the division A/B. A graphical representation of this construction is shown in panel (b) of Figure 1. This overlapping group structure allows $\mathcal{F}_0$ to be the union of any set $\mathcal{F}_{m,l}$ in (6) by moving a block of minimum size, depending on d,

7

with steps of size one, and thus, allowing for greater flexibility in the definition of the subset $\mathcal{F}_0$, with respect to the disjoint grouping of panel (a). A precise characterization is formalized in Proposition 1, in the next section.

The above construction suggests specifying a penalty Ω for overlapping groups of coefficients, which has attracted significant interest in the last decade. For example, Zhao et al. (2009) focused on overlapping and nested groups of coefficients motivated by modeling hierarchical relationships between predictors. More general overlapping group-Lasso penalties have been proposed by Jacob et al. (2009) and Jenatton et al. (2011), which define suitable norms for inducing a penalty that can model specific patterns for the *support* of the vector of coefficients—the latter being the set of coefficients different from zero. The difference between these two approaches is that Jenatton et al. (2011) introduced a penalty inducing supports that arise as intersections of a subset of suitable groups, whereas Jacob et al. (2009) introduced a penalty that induces supports that are rather the unions of a subset of the groups. Lim and Hastie (2015) exploited the construction of Jacob et al. (2009) to learn the main and pairwise interaction terms of categorical covariates in linear and logistic regression models, imposing through the overlapping group structure hierarchical regularization with a similar motivation as Zhao et al. (2009) but also proposing an efficient computational strategy.

In classical regression, the interest lies in the support of the vector of regression coefficients. Consistent with our motivations, instead, we focus on the sparsity structure of the matrix of coefficients $\boldsymbol{\Psi}$ rather than its support. Hence, we specify (7) as

$$\frac{1}{2} \sum_{i=1}^n \int \left(y_i(s) - \int x_i(t)\psi(t,s)dt \right)^2 ds + \lambda \sum_{b=1}^{B+1} ||c_b \odot \boldsymbol{\psi}||_2, \tag{9}$$

where $\lambda > 0$ is a fixed penalization term, and Ω in (7) specifies in the sum of $B+1$ Euclidean norms $||c_b \odot \boldsymbol{\psi}||_2$, where $\boldsymbol{\psi} = \text{vec}(\boldsymbol{\Psi})$, and $\odot$ represents the Hadamard product. The index b denotes the block of coefficients in $\boldsymbol{\Psi}$, with the first B blocks being consistent with (8) and the last block containing all coefficients in $\boldsymbol{\Psi}$. Vectors of size ML, denoted by c_b, contain known constants that equally balance the penalization of the coefficients in $\boldsymbol{\Psi}$. This *balancing* is needed to account for the fact that the parameters close to the boundaries of the matrix $\boldsymbol{\Psi}$ appear in fewer groups than central ones, as shown by the color scaling of panel (b) of Figure 1. Specifically, the generic vector c_b is defined as $c_b = \text{vec}(\mathbf{S}_b \odot \mathbf{C})$, where $\mathbf{S}_b$ is the $M \times L$ selection matrix with

general entry $s_{ml}^{(b)}$, defined as 1 if the parameter ψ_{ml} belongs to group b and 0 otherwise and $\mathbf{C}$ is the matrix with general element c_{ml} defined as $c_{ml} = \left(\sum_{b=1}^{B+1} s_{ml}^{(b)}\right)^{-1}$. Note that this penalty constitutes a special case of the norm defined by Jenatton et al. (2011).

3 Computation

Before describing an efficient computational strategy for our LSFR model, we introduce the empirical counterparts of the quantities described in the previous section assuming to observe a sample of response curves y_i with $i = 1, \ldots, n$ on a common grid of G points, i.e. $y_i = (y_i(s_1), \ldots, y_i(s_G))^T$. Let also x_i be the related functional covariate observed on a possibly different but— common across i—grid of points, that for simplicity and without loss of generality, we assume of length G. Let $\mathbf{X}$ be the $n \times G$ matrix with x_i in the rows. Let $\mathbf{\Phi}$ and $\mathbf{\Theta}$ be the $M \times G$ and $L \times G$ matrices defined as

$$\mathbf{\Phi} = \begin{pmatrix} \varphi_1(t_1) & \cdots & \varphi_1(t_G) \\ \vdots & & \vdots \\ \varphi_m(t_1) & \cdots & \varphi_m(t_G) \\ \vdots & & \vdots \\ \varphi_M(t_1) & \cdots & \varphi_M(t_G) \end{pmatrix}, \qquad \mathbf{\Theta} = \begin{pmatrix} \theta_1(s_1) & \ldots & \theta_1(s_G) \\ \vdots & & \vdots \\ \theta_l(s_1) & \cdots & \theta_l(s_G) \\ \vdots & & \vdots \\ \theta_L(s_1) & \cdots & \theta_L(s_G) \end{pmatrix}.$$

Let $\mathbf{Y}$ and $\mathbf{E}$ be the $n \times G$ matrices obtained as $\mathbf{Y} = (y_1, \ldots, y_n)^T$ and $\mathbf{E} = (e_1, \ldots, e_n)^T$, with $e_i = (e_i(s_1), \ldots, e_i(s_G))^T$. Model (4) can be equivalently written in matrix form as

$$\mathbf{Y} = \mathbf{X}\mathbf{\Phi}^T\mathbf{\Psi}\mathbf{\Theta} + \mathbf{E}.$$

Applying the vectorization operator on each side of the equality above, we have

$$\mathbf{y} = \text{vec}(\mathbf{X}\mathbf{\Phi}^T\mathbf{\Psi}\mathbf{\Theta}) + \text{vec}(\mathbf{E}) = (\mathbf{\Theta}^T \otimes \mathbf{X}\mathbf{\Phi}^T)\text{vec}(\mathbf{\Psi}) + \text{vec}(\mathbf{E}) = \mathbf{Z}\boldsymbol{\psi} + \mathbf{e},$$

where $\mathbf{y} = \text{vec}(\mathbf{Y})$, $\boldsymbol{\psi} = \text{vec}(\mathbf{\Psi})$ is the vector of coefficients of dimension LM, $\mathbf{e} = \text{vec}(\mathbf{E})$, and $\mathbf{Z} = \mathbf{\Theta}^T \otimes \mathbf{X}\mathbf{\Phi}^T$ is the design matrix of dimension $nG \times LM$. Therefore, for a given tuning

parameter $\lambda > 0$, the optimization problem becomes

$$\widehat{\boldsymbol{\psi}}_\lambda = \arg \min_{\boldsymbol{\psi}} \ell(\boldsymbol{\psi}), \quad \ell(\boldsymbol{\psi}) \ = \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\psi}\|_2^2 + \lambda \sum_{b=1}^{B+1} \|\mathbf{D}_b\boldsymbol{\psi}\|_2, \tag{10}$$

where $\mathbf{D}_b = \mathrm{diag}(c_b)$ is a diagonal matrix whose elements correspond to the elements of the vector c_b defined in the previous section. The following result is about the uniqueness of the solution of the optimization problem in equation (10) thus leading to the uniqueness of the estimator $\widehat{\psi}_\lambda$. Its proof is reported in the Appendix.

Theorem 1. *Under the representation* (5), *for any* $\lambda > 0$, *the minimization of* (10) *with respect to the coefficients* ψ_{ml} *for* $l = 1, \ldots, L$ *and* $m = 1, \ldots, M$ *leads to a unique solution* $\widehat{\psi}_\lambda$.

In practice, however, the non-separability of the penalty function when groups overlap, makes the optimization problem in equation (10) not straightforward. The non-separability of the overlap group-Lasso penalty function prevents the application of standard coordinate descent algorithms that cycle through the parameters and updates them either individually (as for the Lasso) or by groups (as for the group-Lasso), (see, e.g. Wu and Lange, 2008; Bach et al., 2012; Huang et al., 2012; Yang and Zou, 2015). We propose to map the optimization in equation (10) to an optimization of a fully convex and differentiable function by leveraging the Majorization-Minimization (MM, hereafter) principle firstly introduced by Ortega and Rheinboldt (1970) and developed by Hunter and Lange (2004); Lange (2010, 2016). The MM approach is a general prescription for constructing optimization algorithms that operates by creating a surrogate function that minorizes (or majorizes) the objective function. The surrogate function is then maximized (or minimized) in place of the original function. Various existing approaches in statistics and machine learning can be interpreted from the majorization-minorization point of view, including the EM algorithm (Neal and Hinton, 1999; Wu and Lange, 2010) or boosting and some variants of the variational Bayes methods (Wainwright and Jordan, 2008). Here the MM approach is employed for the purpose of delivering a quadratic function that majorizes the convex objective function in (10). The MM algorithm has been introduced within the context of ℓ_1 and ℓ_2 penalized regressions by Wu and Lange (2008). The authors consider both the Lasso and group-Lasso penalty and provide cyclic coordinate descent algorithms for both problems. Coordinate descent algorithms are simple, fast and stable and they usually do not require the inversion of large matrices (Hunter and Lange, 2004). However, for large dimensional models with high level of

sparsity and separable penalty, the coordinate-wise gradient provide all the relevant information to update parameters and most of the parameters are never updated from their starting value of zero. Nevertheless, when the level of sparsity is either unknown or moderately low, coordinate-wise updates no longer represent the best strategy. This is exactly the framework here considered. Therefore, leveraging the supporting hyperplane paradigm (see Lange, 2016) and the Sherman–Morrison–Woodbury identity we deliver an efficient MM algorithm for our overlap group-Lasso penalty that jointly updates the non-zero regression coefficients at any iteration. Observe also that, because of the analytical form of the overlap penalty function (see Huang et al., 2012), efficient block-wise updates as in Qin et al. (2013) cannot be considered here. Our MM algorithm instead serves the purpose of delivering an efficient solution to the otpimization problem in equation (10) without imposing any specific group conformation and finds application even in more structured group-type penalties, (e.g. Bach et al., 2012; Jenatton et al., 2011).

Any MM algorithm iterates between two steps: (i) at iteration k, a majorizer function $\mathcal{Q}(\boldsymbol{\psi}|\widehat{\boldsymbol{\psi}}^{(k-1)})$ is obtained conditioning on $\widehat{\boldsymbol{\psi}}^{(k-1)}$; (ii) the value for $\widehat{\boldsymbol{\psi}}^{(k)}$ is obtained minimizing $\mathcal{Q}(\boldsymbol{\psi}|\widehat{\boldsymbol{\psi}}^{(k-1)})$. As a further benefit of exploiting the MM majorization principle, the MM iterations $\widehat{\boldsymbol{\psi}}^{(k)}$, for $k = 1, 2, \ldots$ possesses the decent properties driving the target function downhill.

Several paradigms can be exploited to derive a valid majorizer of the overlap group penalty term in (10), (see, e.g. Lange, 2016, for an exhaustive discussion). Here, we leverage the dominating hyperplane principle that introduces an upper bound for the strictly concave function $\sqrt{x}$, i.e. $\sqrt{x} \leq \sqrt{x^{(k)}} + (2\sqrt{x^{(k)}})^{-1}(x - x^{(k)})$ for any $x, x^{(k)} \in \mathbb{R}^+$. This manoeuvres separate parameters and reduce the surrogate to a sum of linear terms and squared Euclidean norms. Consistently with this, we introduce the following surrogate function that majorizes (10) at $\widehat{\boldsymbol{\psi}}^{(k)}$

$$\mathcal{Q}(\boldsymbol{\psi}|\widehat{\boldsymbol{\psi}}^{(k)}) = \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\psi}\|_2^2 + \lambda \sum_{b=1}^{B+1}\left(\|\mathbf{D}_b\widehat{\boldsymbol{\psi}}^{(k)}\|_2 + \frac{\|\mathbf{D}_b\boldsymbol{\psi}\|_2^2 - \|\mathbf{D}_b\widehat{\boldsymbol{\psi}}^{(k)}\|_2^2}{2\|\mathbf{D}_b\widehat{\boldsymbol{\psi}}^{(k)}\|_2}\right), \qquad (11)$$

where $\widehat{\boldsymbol{\psi}}^{(k)}$ is the value of the parameter $\boldsymbol{\psi}$ at the k-th iteration of the MM algorithm. The important consequence of this result is that solution of problem (10) can be obtained through the iterative minimization of (11), where (11) is convex and differentiable. A compact form for

the minimization problem of the surrogate function of (11) is

$$\widehat{\psi}_\lambda^{(k+1)} = \arg\min_{\psi} \mathcal{Q}(\psi|\widehat{\psi}^{(k)}) \tag{12}$$

$$\mathcal{Q}(\psi|\widehat{\psi}^k) = \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\psi\|_2^2 + \widehat{d}_0^{(k)} + \lambda \sum_{b=1}^{B+1} \widehat{d}_b^{(k)}\|\mathbf{D}_b\psi\|_2^2, \tag{13}$$

where $\widehat{d}_0^{(k)}$ and $\widehat{d}_b^{(k)}$ are constants that depend on the k-th iteration:

$$\widehat{d}_0^{(k)} = \sum_{b=1}^{B+1}\left(\|\mathbf{D}_b\widehat{\psi}^{(k)}\|_2 - \frac{\|\mathbf{D}_b\widehat{\psi}^{(k)}\|_2^2}{2\|\mathbf{D}_b\widehat{\psi}^{(k)}\|_2}\right), \quad \text{and} \quad \widehat{d}_b^{(k)} = \frac{1}{2\|\mathbf{D}_b\widehat{\psi}^{(k)}\|_2}. \tag{14}$$

The explicit solution of the minimization problem in equations (12)-(13) at the k-th iteration is

$$\widehat{\psi}^{(k+1)} = (\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{H}^{(k)})^{-1}\mathbf{Z}^T\mathbf{y}, \tag{15}$$

where $\mathbf{H}^{(k)} = \left(\sum_{b=1}^{B+1}\widehat{d}_b^{(k)}\mathbf{D}_b^T\mathbf{D}_b\right)^{1/2}$. The MM algorithm is described in Algorithm 1, reported in the Appendix. An important consideration in using a particular algorithm is the amount of work the computer is required to carry out in running it which is measured in terms of the number of floating-point operations (FLOPS) that are needed. The computational cost of Algorithm 1 is provided in the Appendix, at Proposition 2.

3.1 Efficient MM for the generalised ridge inversion

The MM parameters update in equation (15) suffers from two major drawbacks. First, it requires the inversion of a potentially large-dimensional ridge-type design matrix and, second, it does not exclude the pathological case where one or more of the denominators of the weights in (14) are exactly zero. The usual solution for the latter problem consists to perturb $\widehat{d}_b^{(k)}$ by adding a small $\epsilon > 0$ to the ℓ_2-norm, (see Hunter and Lange, 2000). In what follows, instead, we rely on a different solution. Specifically, let $\mathbf{H}^{(k)}$ be a $LM \times LM$ diagonal matrix and define the $LM \times LM$ symmetric and positive definite matrix $\mathbf{A}_\lambda^{(k)} = \mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{H}^{(k)} \in \mathbb{S}_{++}^{LM}$. The computation of the MM update in equation (15) requires the inversion of the $LM \times LM$ symmetric full matrix $\mathbf{A}_\lambda^{(k)}$ at any iteration which takes on the order of $\mathcal{O}((LM)^3)$ arithmetic operations (Golub and Van Loan, 2013). Since the matrix $\mathbf{H}^{(k)}$ changes at any iteration performing the QR decomposition of $\mathbf{A}_\lambda^{(k)}$ becomes prohibitive even for moderately large values of LM.

Given the diagonal structure of the matrix $\mathbf{H}^{(k)}$, using the Sherman-Morrison-Woodbury matrix identity reduces the problem of inverting $\mathbf{A}_\lambda^{(k)}$ for a fixed λ to the simpler problem of computing

$$\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{H}^{(k)}\right)^{-1} = \frac{1}{\lambda}\left(\mathbf{H}^{(k)}\right)^{-1} - \frac{1}{\lambda^2}\left(\mathbf{H}^{(k)}\right)^{-1}\mathbf{Z}^T\mathbf{B}_\lambda^{(k)}\mathbf{Z}\left(\mathbf{H}^{(k)}\right)^{-1}, \qquad (16)$$

where the full matrix $\mathbf{B}_\lambda^{(k)} = \lambda\left(\lambda\mathbf{I}_{nG} + \mathbf{J}^{(k)}\right)^{-1} \in \mathbb{S}_{++}^{nG}$ with $\mathbf{J}^{(k)} = \mathbf{Z}\left(\mathbf{H}^{(k)}\right)^{-1}\mathbf{Z}^T \in \mathbb{S}_{++}^{nG}$ can be computed without loss of generality through the spectral decomposition of $\mathbf{J}^{(k)}$. Specifically, let $\mathbf{U}^{(k)}\mathbf{\Lambda}^{(k)}(\mathbf{U}^{(k)})^T = \mathbf{J}^{(k)}$ be such a decomposition, then $\mathbf{B}_\lambda^{(k)} = \lambda\mathbf{U}^{(k)}\left(\lambda\mathbf{I}_{nG} + \mathbf{\Lambda}^{(k)}\right)^{-1}(\mathbf{U}^{(k)})^T$. Therefore, the MM update only requires the spectral decomposition of the symmetric matrix $\mathbf{J}^{(k)}$ to be computed at any iteration. However, as pointed by the following remark, as a byproduct of our procedure, we obtain the indirect solution to the problem of zeros in the surrogate penalty function.

Remark 1. *As iterations proceeds, it may happen that some of the weights $\|\mathbf{D}_b\widehat{\boldsymbol{\psi}}^{(k)}\|_2$ associated to a sequence of zeros on the vector $\widehat{\boldsymbol{\psi}}^{(k)}$ larger than d^2 becomes closer and closer to zero. Leveraging the Sherman-Morrison-Woodbury matrix identity prevents the weights to explode.*

We can exploit the fact that $\widehat{\boldsymbol{\psi}}^{(k)}$ in equation (15) at some iteration $k \geq k_0$ may become zero, i.e. $\|\mathbf{D}_b\widehat{\boldsymbol{\psi}}^{(k)}\|_2 = 0$, to provide a fast and efficient solution to the problem of finding the spectral decomposition of $\mathbf{J}^{(k)}$ for nG moderately large. Let $\mathbf{Z}_0 \in \mathbb{R}^{nG \times p_0}$ and $\mathbf{Z}_1 \in \mathbb{R}^{nG \times (LM-p_0)}$ with $1 < p_0 < LM$ be a disjoint partition of the space spanned by $\mathbf{Z}$ such that $\mathbf{Z} = (\mathbf{Z}_0 \quad \mathbf{Z}_1)$, and let $\mathbf{H}_0^{(k)} \in \mathbb{R}^{p_0 \times p_0}$ and $\mathbf{H}_1^{(k)} \in \mathbb{R}^{(LM-p_0) \times (LM-p_0)}$ be the corresponding partition of the diagonal matrix $\mathbf{H}^{(k)}$, then

$$\mathbf{Z}\left(\mathbf{H}^{(k)}\right)^{-1}\mathbf{Z}^T = \mathbf{Z}_0\left(\mathbf{H}_0^{(k)}\right)^{-1}\mathbf{Z}_0^T + \mathbf{Z}_1\left(\mathbf{H}_1^{(k)}\right)^{-1}\mathbf{Z}_1^T.$$

Now, assume that at the k-th iteration, $(\mathbf{H}_0^{(k)})^{-1} = \mathrm{diag}\{0, \ldots, 0\}$, then $\mathbf{Z}(\mathbf{H}^{(k)})^{-1}\mathbf{Z}^T = \mathbf{Z}_1(\mathbf{H}_1^{(k)})^{-1}\mathbf{Z}_1^T$, which only requires the spectral decomposition of the matrix $\mathbf{Z}_1(\mathbf{H}_1^{(k)})^{-1}\mathbf{Z}_1^T$. Additional computational considerations and results are reported in the Appendix.

4   Theoretical properties

In this section, we provide appealing theoretical properties for our LSFR model and for the related estimator $\widehat{\psi}_\lambda(t, s)$ of $\psi(t, s)$ arising from (5) and (10). All the proofs are reported in the Appendix. We first describe the sparsity patterns that we can induce through our LSFR

construction in the following proposition.

Proposition 1. *Let $F = \{\mathcal{F}_{m,l}; m = 1, \ldots, M - d + 1; l = 1, \ldots, L - d + 1\}$ be the collection of all the sets described in (6), $\mathcal{P}(F)$ the power set of F, and $\mathcal{F}_0(\psi)$ be the subset of $\mathcal{F}$ in which $\psi(t, s) = 0$. Then $\widehat{\psi}_\lambda \in S$ with $S \subseteq L_2$ and $S = \{\psi \in L_2 : \mathcal{F}_0(\psi) \in \mathcal{P}(F)\}$.*

The following two results show that the introduced model structure is sufficiently flexible to cover the two extreme situations represented by the concurrent and nonconcurrent models.

Remark 2. *Assume $L = M$. Let $\mathcal{F}_0^{concurrent} = \{(t, s) \in \mathcal{F} : t \neq s\}$ be the subset of $\mathcal{F}$ where $\psi(t, s) = 0$ in the concurrent model. Let $\bar{\mathcal{F}}_0 \in \mathcal{P}(F)$ the superset of $\mathcal{F}_0^{concurrent}$ defined as*

$$\bar{\mathcal{F}}_0 = \arg \min_{\mathcal{F}_0 \in \mathcal{P}(F)} \mu(\mathcal{F}_0 \backslash \mathcal{F}_0^{concurrent}).$$

Then $\mu(\bar{\mathcal{F}}_0 \backslash \mathcal{F}_0^{concurrent}) = O(1/M)$.

Remark 3. *If $\psi^*(t, s) \neq 0$ for each $(t, s) \in \mathcal{F}$ then $\psi^* \in S$.*

Remark 2 implies that for a sufficiently high number of knots in the tensor product spline expansion, the sparsity structure of a concurrent Hilbert-Schmidt operator can be well approximated by the proposed formulation. A similar property holds also for the historical model (Malfait and Ramsay, 2003). At the same time, Remark 3 trivially states that nonconcurrent model is included in our class, implying that relatively small values of the penalty parameter λ lead to no sparsity of ψ.

The following two results determine interesting consistency properties, namely, that the method correctly identifies the region where the true kernel is null and vice versa. Specifically Theorem 2 assumes that the true kernel lies in the vector space generated by the tensor product $V_{\theta \otimes \varphi}$ and Theorem 3 relaxes this assumption. Both theorems hold under the correct model specification and when $e_i(s)$ are uncorrelated Gaussian errors with variance $\sigma^2 > 0$. Both results are adaptations of the consistency results of Jenatton et al. (2011).

Before stating the results, we define the following quantities based on the assumption that there exists a true kernel ψ_T. In Theorem 2 we assume that ψ_T belongs to the vector space $V_{\varphi \otimes \theta}$ and thus that there exist a unique matrix of coefficients $\boldsymbol{\Psi}_T$ representing it. Furthermore, let $\mathcal{B} \subset \{1, \ldots, B, B + 1\}$ be the set of indices of the true non-zero groups and $\mathcal{I} \subset \{(m, l) : m = 1, \ldots, M; l = 1, \ldots, L\}$ the set of the indices of the non-zero coefficients $\psi_{m,l}$. Let $\mathbf{Z}_1$ the matrix

containing the columns of $\mathbf{Z}$ associated to the true non zero coefficients and $\mathbf{Z}_0$ its complement. Define the norm

$$\Omega_{\mathcal{B}}^c(u) = \sum_{b \in \mathcal{B}} ||c_b^{(\mathcal{B}^C)} \odot u||_2,$$

which is the modification of the penalising norm of (9) considering the sum over the inactive groups. Let further $\left(\Omega_{\mathcal{B}}^c\right)^*$ be its dual norm. Finally let $\mathbf{r}_1$ be the vector containing, for $(m, l) \in \mathcal{I}$, the elements

$$\psi_{m,l} c_{m,l}^2 \sum_{b \in \mathcal{B}} ||c_b \odot \boldsymbol{\psi}||_2^{-1}.$$

Theorem 2. *Let ψ_T be the true HS operator with $\psi_T \in V_{\varphi \otimes \theta}$ where $V_{\varphi \otimes \theta}$ is the vector space defined by the tensor product of the two B-splines basis θ and φ both with fixed dimensions L and M, respectively. Let $\mathcal{F}_{0T}$ be the induced true sparsity set, i.e. $\psi_T(t, s) = 0$ for each $(t, s) \in \mathcal{F}_{0T}$. Define the event E_0 as*

$$E_0 = \left\{ \int_{\mathcal{F}_{0T}} \left[\widehat{\psi}_\lambda(t, s) \right]^2 dt ds = 0, \quad \left(\inf_{\mathcal{F}_1 \subset \mathcal{F}_{0T}^C} \int_{\mathcal{F}_1} \left[\widehat{\psi}_\lambda(t, s) \right]^2 dt ds \right) > 0 \right\}.$$

For $\lambda > 0$, $nG > LM$, $\lambda \to 0$, $\lambda \sqrt{n} \to \infty$, $\left(\Omega_{\mathcal{B}}^c\right)^* \left(\mathbf{Z}_0^T \mathbf{Z}_1 (\mathbf{Z}_1^T \mathbf{Z}_1)^{-1} \mathbf{r}_1\right) < 1$, then $\lim_{n \to \infty} \mathbb{P}(E_0) = 1$.

The next theorem avoids the strict assumption of $\psi_T \in V_{\varphi \otimes \theta}$. The main idea is to introduce suitable approximations of ψ_T and the induced $\mathcal{F}_{0T}$ and study consistency of the estimator to those approximations. Specifically, we let $\bar{\mathcal{F}}_{0T}$ be the subset of $\mathcal{F}_{0T}$ of maximum Lebesgue measure belonging to the power set $\mathcal{P}(F)$ and define ψ_T^P as

$$\psi_T^P = \arg \min_{\psi \in V_{\varphi \otimes \theta}; \mathcal{F}_{0\psi} = \bar{\mathcal{F}}_{0T}} \int_{\mathcal{F}} \left(\psi(t, s) - \psi_T(t, s) \right)^2 dt ds.$$

In what follows $\mathcal{B}, \mathcal{I}, \mathbf{Z}_1, \mathbf{Z}_2$ and the related norm $\Omega_{\mathcal{B}}^c$ and its dual, are defined with respect to the unique $\boldsymbol{\Psi}_T^P$ associated to ψ_T^P.

Theorem 3. *Let ψ_T be the true HS operator and $\mathcal{F}_{0T}$ the true induced sparsity set, i.e. $\psi_T(t, s) = 0$ for each $(t, s) \in \mathcal{F}_{0T}$. Define the event E_k as*

$$E_k = \left\{ \int_{\mathcal{F}_{0T}} \left[\widehat{\psi}_\lambda(t, s) \right]^2 dt ds \leq \frac{k}{LM}, \quad \left(\inf_{\mathcal{F}_1 \subset \mathcal{F}_{0T}^C} \int_{\mathcal{F}_1} \left[\widehat{\psi}_\lambda(t, s) \right]^2 dt ds \right) > 0 \right\},$$

with k a positive constant. Let $D < LM$ be the number of columns of $\mathbf{Z}_1$. For $\tau > 0$, assume $\left(\Omega_{\mathcal{B}}^c\right)^* \left(\mathbf{Z}_0^T \mathbf{Z}_1 (\mathbf{Z}_1^T \mathbf{Z}_1)^{-1} \mathbf{r}_1\right) < 1 - \tau$, with

$$\frac{\mu(\mathcal{F}_{0T}^C)}{\mu(\mathcal{F})} \leq \frac{(nG-1)(d+1)^2}{(L-1)(M-1)}. \tag{17}$$

If $\tau\lambda\sqrt{n} \geq \sigma C_3$ and $\lambda\sqrt{D} \leq C_4$, then the probability of the event E_k is lower bounded by

$$1 - \left(\exp\left\{-\frac{n\lambda^2\tau^2 C_1}{2\sigma^2}\right\} + 2D \exp\left\{-\frac{n\lambda^2\tau^2 C_2}{2D\sigma^2}\right\}\right),$$

where C_1, C_2, C_3, and C_4 are positive constants.

5 Simulation study

In this section we assess the empirical performance of our LSFR model by means of simulations. We generated a sample of a functional covariate $x_1(t), \ldots, x_n(t)$ from cubic B-splines basis with 15 evenly spaced knots between 0 and 1. These functional data are assumed to be observed on an equispaced grid of 100 points and the sample size n is $n \in \{50, 150\}$. Conditionally on these functional covariates we generated functional responses as $y_i(s) = \int \psi_k(t, s)x_i(t)dt + e_i(s)$ under four different levels of sparsity of the operator ψ_k and two different signal-to-noise ratios.

The first and the second scenario focuses on a quasi–concurrent and historical relations, respectively. The third scenario is associated to a local dependence between the dependent variable and the functional covariate with the region $\mathcal{F}_0$ being the union of two non–intersecting rectangles. Finally, last scenario assumes no sparsity. Graphical representations of the kernels ψ_k are showed in Figure 2.We vary the signal-to-noise ratio SNR $\in \{2, 4\}$ with

$$\text{SNR} = \frac{1}{n}\sum_{i=1}^{n}\left\{\int\left[\int \psi(t, s)x_i(t)dt\right]^2 ds \bigg/ \int e_i^2(s)ds\right\}.$$

We choose $d = 4$ and $L = M = 20$ leading to 400 coefficients in $\mathbf{\Psi}$. For each combination of ψ_k, n, and SNR we simulated $R = 100$ replicated datasets, leading to $4 \times 2 \times 2 \times 100 = 1600$ independent datasets.

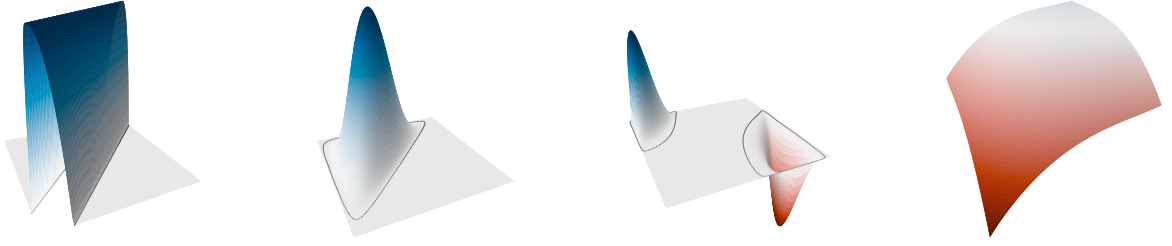We compare our LSFR with the results obtained minimizing (7) with Ridge, Lasso, and

Figure 2: Three dimensional representations of ψ_k with, from left to right, $k = 1, 2, 3, 4$.

Elastic-net penalties (Hastie et al., 2009). Classical group-Lasso penalty has not been fitted due to its limitations discussed in Section 2.1. We expect the Ridge approach to be the best competitor in the last scenario and in general if estimation is evaluated in the non-sparse regions of the true kernels. On the other side we expect Lasso to be the best competitor in detecting the sparse regions. The Elastic-net minimization and our LSFR solution are expected to take the best of both Lasso and Ridge.

All methods depend on the choice of the tuning parameter λ. We run each model on a fine grid of λ and choose as final estimate for each method, the one minimizing the prediction error for the general estimator $\breve{\psi}_\lambda$ of ψ, i.e.

$$\breve{\psi} = \arg\min_\lambda \sum_{i=1}^{200} \left(y_i(s) - \int \breve{\psi}_\lambda(t,s) x_i(t) dt \right)^2,$$

where the sum is defined on external validation set of size 200. The range of possible λ depends on the specific methods and for LFSR we refer to (20) in Appendix. For all competing methods we rely on their implementation in the package `glmnet` (Friedman et al., 2010).

The final estimates are evaluated in both subsets of $\mathcal{F}$ where the true function is zero or not, on a third independent test set of size 1,000. Specifically, we define two partial integrated square error measures as

$$ISE_0(\breve{\psi}) = \int_{\mathcal{F}_0} \left(\breve{\psi}(t,s) - \psi(t,s) \right)^2 dt ds, \quad ISE_1(\breve{\psi}) = \int_{\mathcal{F}_0^C} \left(\breve{\psi}(t,s) - \psi(t,s) \right)^2 dt ds, \quad (18)$$

representing the error in not estimating the sparsity of the true kernel and the estimation error when the true kernel is smooth, respectively. As a global measure, we use the integrated squared

17

error on the whole $\mathcal{F}$, which can be obtained as a suitable convex combination of the previous measures, i.e.

$$ISE(\breve{\psi}) = ISE_0(\breve{\psi})\mu(\mathcal{F}_0) + ISE_1(\breve{\psi})\mu(\mathcal{F}_1). \tag{19}$$

5.1 Results

Table 1 reports ISE_0 and ISE_1 for four different settings. As default setting, we consider $n = 50$, SNR $= 4$ and $M = L = 20$. As second setting, we decrease the SNR $= 2$ while keeping fixed $n = 50$ and $M = L = 20$. In the third, we assess the effect of changing the resolution of the kernels' basis representations letting $M = L = 40$ with $n = 50$ and SNR $= 4$. Finally, we assess the effect of a bigger sample size letting $n = 150$ while keeping the default SNR $= 4$ and $M = L = 20$. For all these settings, the performance of each method is evaluated for all the four kernels.

The proposed LSFR model exhibits good performance consistently in all settings and considering both the evaluation metrics. Specifically, our approach achieves a uniformly lower ISE_0 both with respect to Ridge and Elastic-net in all settings while presenting similar performance with respect to Lasso. In fact the results in terms of ISE_0 are broadly comparable for our LSFR and the Lasso for the estimation of the first and third kernel, while our LSFR is even slightly better in estimating the second kernel. Looking at ISE_1, as expected, the performance of the Lasso and of the Ridge and Elastic-net are flipped. The proposed LSFR model, however, consistently outperforms all competing methods also in terms of ISE_1.

Different values of SNR, n, or M and L have different impacts on the results. Lowering the signal-to-noise ratio or increasing the sample size reduces —resp. increases— the errors almost proportionally among different estimation procedures. On the contrary, changing the number of basis from 20 to 40 has a relevant impact on the results. Particularly, a dramatic increase of ISE_1 is observed for Lasso, in all scenarios. This can be explained by the fact that using a larger number of bases is equivalent to reduce the support of each element of the basis. Consistently with this, Lasso penalizes coefficients that are much more correlated, yielding to

Table 1: Monte Carlo means and standard deviations ($\times 10^5$) of the ISE_0 and ISE_1 between the estimated coefficients and the true coefficients ψ_k, $k = 1, 2, 3, 4$ over 100 replicates. Rows labelled S1 report the results for the default simulation settings, i.e. $n = 50$, $SNR = 4$, $L = M = 20$; Rows labelled S2 report the results for $SNR = 2$ and default remaining settings; Rows labelled S3 report the results for $L = M = 40$ and default remaining settings; Rows labelled S4 report the results for $n = 150$ and default remaining settings.

|	LSFR		Lasso		Ridge		Enet	
	ISE_0	ISE_1	ISE_0	ISE_1	ISE_0	ISE_1	ISE_0	ISE_1
ψ_1								
S1	0.18 (0.04)	1.07 (0.20)	0.18 (0.06)	2.76 (0.44)	0.66 (0.11)	1.12 (0.22)	0.26 (0.06)	1.34 (0.23)
S2	0.28 (0.08)	1.63 (0.34)	0.31 (0.12)	4.08 (0.67)	0.98 (0.17)	1.74 (0.35)	0.42 (0.12)	2.03 (0.40)
S3	0.14 (0.05)	1.66 (0.36)	0.37 (0.13)	16.24 (2.06)	1.02 (0.20)	1.48 (0.31)	0.39 (0.11)	2.08 (0.40)
S4	0.11 (0.02)	0.58 (0.09)	0.08 (0.03)	1.45 (0.24)	0.34 (0.06)	0.56 (0.09)	0.14 (0.03)	0.71 (0.11)
ψ_2								
S1	0.04 (0.03)	0.51 (0.14)	0.12 (0.07)	1.80 (0.37)	0.52 (0.14)	0.75 (0.18)	0.16 (0.08)	0.68 (0.17)
S2	0.08 (0.05)	0.82 (0.22)	0.23 (0.13)	2.42 (0.47)	0.86 (0.21)	1.21 (0.26)	0.30 (0.13)	1.09 (0.25)
S3	0.08 (0.05)	0.74 (0.21)	0.32 (0.16)	9.98 (2.75)	0.96 (0.25)	1.16 (0.26)	0.39 (0.14)	1.04 (0.23)
S4	0.02 (0.01)	0.23 (0.05)	0.04 (0.02)	1.12 (0.22)	0.26 (0.07)	0.33 (0.07)	0.06 (0.03)	0.29 (0.06)
ψ_3								
S1	0.26 (0.06)	0.58 (0.11)	0.25 (0.09)	1.01 (0.19)	0.86 (0.19)	1.10 (0.16)	0.32 (0.09)	0.69 (0.11)
S2	0.32 (0.09)	0.82 (0.20)	0.32 (0.15)	1.29 (0.27)	1.14 (0.30)	1.57 (0.27)	0.40 (0.15)	0.96 (0.20)
S3	0.25 (0.07)	0.62 (0.14)	0.47 (0.14)	7.17 (1.45)	1.18 (0.29)	1.40 (0.20)	0.41 (0.12)	0.91 (0.17)
S4	0.22 (0.04)	0.36 (0.06)	0.21 (0.05)	0.68 (0.13)	0.56 (0.11)	0.61 (0.09)	0.27 (0.07)	0.46 (0.08)
ψ_4								
S1	–	0.76 (0.11)	–	2.13 (0.26)	–	0.93 (0.12)	–	1.08 (0.13)
S2	–	1.07 (0.16)	–	2.62 (0.26)	–	1.30 (0.17)	–	1.48 (0.17)
S3	–	1.04 (0.18)	–	5.87 (0.81)	–	1.10 (0.18)	–	1.34 (0.20)
S4	–	0.42 (0.06)	–	1.09 (0.17)	–	0.53 (0.06)	–	0.61 (0.08)
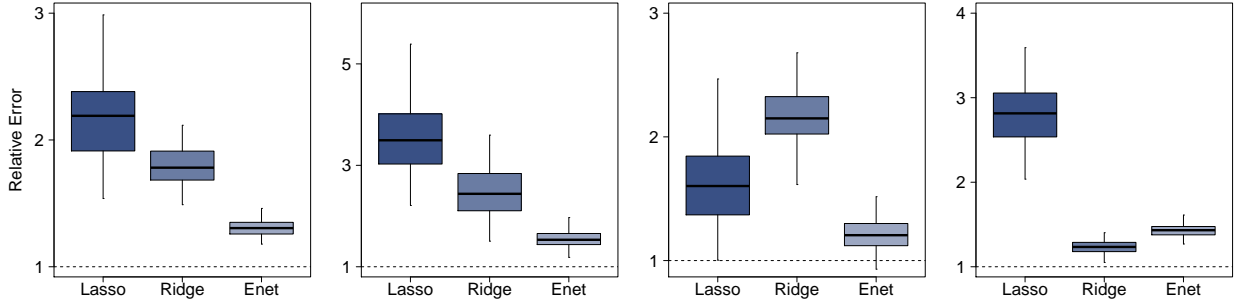
Figure 3: Relative ISE for the simulation study under the default setting and for the four scenarios.

an erratic sequence of estimated coefficients. Conversely, the larger number of bases of the third setting does not negatively affect the remaining methods including our LSFR model, but instead it allows for a potentially better identification of small sparse regions impossible to detect with a lower number of bases.

Figure 3 reports the boxplots of the global relative efficiency of the three competing methods, defined as the ratio between the three different ISE and the ISE achieved by our LSFR model in the default setting. All boxplots are above one, witnessing that the proposed method uniformly attains a lower ISE. Exception made for the third scenario, Lasso is the worst method, while Elastic-net is the most flexible, as expected. Similar results can be noticed for the other simulation settings.

6 Applications

6.1 Swedish Mortality

We apply the proposed LSFR model to the well-known Swedish Mortality dataset, available from the Human Mortality Database and considered one of the most reliable dataset on long-term longitudinal mortality. We focus on the analysis of the log-hazard rate functions of the Swedish female population between the years 1751 and 1894. The goal of our analysis is to model the log-hazard function y_i at a specific calendar year i by using the log-hazard function at previous year y_{i-1}. The log-hazard rate for year i and age s is computed as the logarithm of the ratio
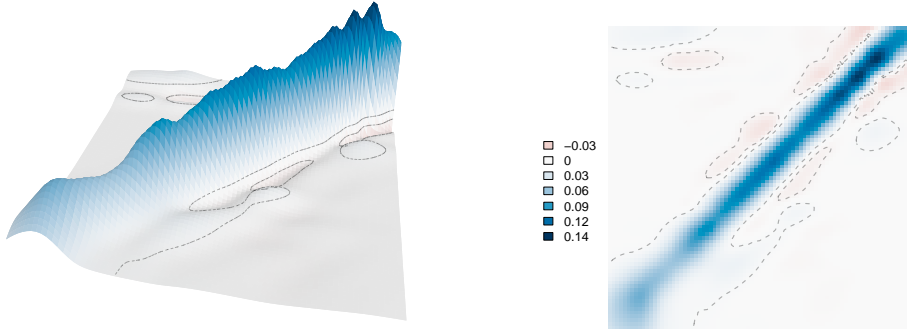
Figure 4: Estimated regression surface for Swedish Mortality data. Left: perspective plot. Right: heatmap.

between women born on year i who died at age s and women born on year i still alive at age s. An underlying autoregressive linear relation between y_{i-1} and y_i is assumed and specifically

$$y_i(s) = \int y_{i-1}(t)\psi(t,s)dt + e_i(s).$$

The estimated kernel can be interpreted as the influence of the log-hazard rate at year $i-1$ and age t on the log-hazard rate at year i and age s. Existing studies (Ramsay et al., 2009; Chiou and Muller, 2009) show that the hazard function at year i and age s is mainly influenced by the hazard function at the previous year $i-1$ at age $t = s-1$, resembling a quasi–concurrent relation. However, none of these studies reports the total absence of relation when t and s are far away and the corresponding estimated surfaces exhibit nonvanishing fluctuations even near the boundaries of their bivariate domains.

We implement the proposed locally sparse estimator by minimizing the objective (10) with $d = 4$, $M = L = 20$ basis functions on each dimension and we select the optimal value of λ by means of cross–validation. Perspective and contour plots of the estimated kernel are depicted in Figure 4.

Our estimate shows a marked positive diagonal confirming the positive influence on the log-
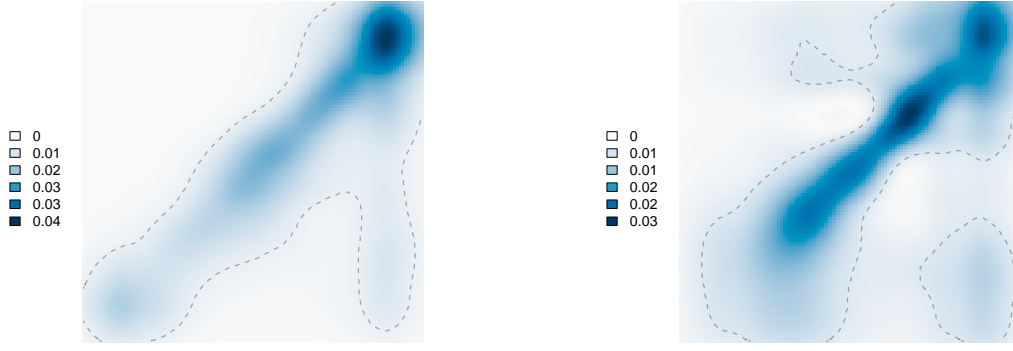
Figure 5: Estimated regression surfaces for demand (left) and supply (right) curves in the Italian natural gas dataset.

hazard rate at age s of the previous year's curve evaluated on a neighborhood of s. At the same time, the flat zero regions outside the diagonal suggest that there is no influence of the curves evaluated at distant ages. Our estimate is more regular than previous approaches and its qualitative interpretation sharper and easier. Refer, for example, to Figure 10.11 of Ramsay et al. (2009).

This result witnesses the practical relevance of adopting the proposed approach. Indeed, the resulting estimates, while being reminiscent of a concurrent model—inheriting its ease of interpretation—it gives further insights and improves the fit, representing the desired intermediate solution between the concurrent and nonconcurrent models.

6.2   Italian Gas Market

As second illustrative example, we focus on the Italian natural gas data first analyzed by Canale and Vantini (2016) by means of a concurrent functional time series regression fitted under a Ridge-like penalty. Data consists on 375 pairs of demand and supply curves related to the Italian gas trading platform in 2012. Details on how the curves are constructed from the original bids can be found in Canale and Vantini (2016). The domain of the demand and supply curves represent the quantity and the price of the natural gas, respectively. The demand (resp. supply) curves are subject to equality and inequality constraints at the two edges of the domain and are

monotonically increasing (resp. decreasing). Consistently with the so-called transform/back-transform method (Ramsay and Silverman, 2005) we map the actual monotone and bounded curves using the logH transformation introduced by Canale and Vantini (2016) and fit a LSFR autoregressive model to the transformed data similarly to the previous section. Specifically, let z_i be the demand (resp. supply) curve at day i. We let

$$y_i(s) = 1 - \exp\left\{-\int_0^s \exp\{z_i(u)\}du\right\},$$

and then model transformed demand (resp. supply) curve y_i at day i with a first-order autoregressive linear model as done in Section 6.1. We set $d = 4$ and $M = L = 20$ basis functions and select λ via cross-validation, in the previous application. Two models for the demand and supply series are fitted separately.

Figure 5 reports the contour plots of the estimated kernels. Similarly to Section 6.1, both estimates present marked positive diagonals while the influence of the curve at time $i-1$ on the curve at time i vanishes if we move far away from the diagonal. The qualitative interpretation is clear: the price of a given quantity for day i mainly depends on the price of the $i-1$ demand curve in a neighborhood of the same quantity.

Also in this practical case we stress the usefulness of the proposed approach and its ability to lie between the concurrent and nonconcurrent approaches.

7   Discussion

Motivated by the current gap between the concurrent and nonconcurrent models, we introduced a model for function-on-function regression which allows for local sparsity patterns, exploiting some of the properties of B-spline basis and a specifically tailored overlap group-Lasso penalization. The proposed MM algorithm directly tackles the minimization arising from such a group-Lasso penalty, but it is also amenable to generalizations beyond the specific group structure motivated by our model and, furthermore, beyond the FDA context. The empirical assessment through

simulation and applications to real datasets, provides evidence of improved estimation accuracy with respect to standard competitors and straightforward and effortless interpretability of the results. The latter is of substantial interest when dealing with infinite dimensional objects as in FDA.

Despite we focused on the case of a single functional covariate, the extension of our modelling framework to p covariates is straightforward. At the same time, our approach can be specified also in the simpler case when the y_i's are scalar responses.

References

Bach, F., Jenatton, R., Mairal, J. and Obozinski, G. (2012) Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, **4**, 1–106.

Canale, A. and Vantini, S. (2016) Constrained functional time series: Applications to the italian gas market. *International Journal of Forecasting*, **32**, 1340–1351.

Candes, E., Tao, T. et al. (2007) The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, **35**, 2313–2351.

Centofanti, F., Fontana, M., Lepore, A. and Vantini, S. (2020) Smooth lasso estimator for the function-on-function linear regression model. *arXiv*, 1–31.

Chiou, J.-M. and Muller, H.-G. (2009) Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *Journal of the American Statistical Association*, **104**, 572–585.

De Boor, C. (1978) *A practical guide to splines*, vol. 27. Springer-Verlag New York.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. et al. (2004) Least angle regression. *The Annals of statistics*, **32**, 407–499.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**, 1348–1360.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.

Golub, G. and Uhlig, F. (2009) The QR algorithm: 50 years later its genesis by John Francis and Vera Kublanovskaya and subsequent developments. *IMA J. Numer. Anal.*, **29**, 467–485.

Golub, G. H. and Van Loan, C. F. (2013) *Matrix computations.* Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edn.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* Springer Series in Statistics. Springer New York.

Horváth, L. and Kokoszka, P. (2012) *Inference for Functional Data with Applications.* Springer Series in Statistics. Springer New York.

Hsing, T. and Eubank, R. (2015) *Theoretical foundations of functional data analysis, with an introduction to linear operators.* John Wiley & Sons.

Huang, J., Breheny, P. and Ma, S. (2012) A selective review of group selection in high-dimensional models. *Statist. Sci.*, **27**, 481–499.

Hunter, D. R. and Lange, K. (2000) Quantile regression via an MM algorithm. *J. Comput. Graph. Statist.*, **9**, 60–77.

— (2004) A tutorial on MM algorithms. *Amer. Statist.*, **58**, 30–37.

Jacob, L., Obozinski, G. and Vert, J.-P. (2009) Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, 433–440.

James, G. M., Wang, J., Zhu, J. et al. (2009) Functional linear regression that's interpretable. *The Annals of Statistics*, **37**, 2083–2108.

Jenatton, R., Audibert, J.-Y. and Bach, F. (2011) Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*, **12**, 2777–2824.

Lange, K. (2010) *Numerical analysis for statisticians*. Springer, New York.

— (2016) *MM optimization algorithms.* Society for Industrial and Applied Mathematics, Philadelphia, PA.

Lee, E. R. and Park, B. U. (2012) Sparse estimation in functional linear regression. *Journal of Multivariate Analysis*, **105**, 1–17.

Lim, M. and Hastie, T. (2015) Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, **24**, 627–654.

Lin, Z., Cao, J., Wang, L. and Wang, H. (2017) Locally sparse estimator for functional linear regression models. *Journal of Computational and Graphical Statistics*, **26**, 306–318.

Malfait, N. and Ramsay, J. O. (2003) The historical functional linear model. *Canadian Journal of Statistics*, **31**, 115–128.

McLachlan, G. J. and Krishnan, T. (2008) *The EM algorithm and extensions.* Wiley Series in Probability and Statistics. Wiley-Interscience, John Wiley & Sons, Hoboken, NJ, second edn.

Morris, J. S. (2015) Functional regression. *Annual Review of Statistics and Its Application*, **2**, 321–359.

Neal, R. M. and Hinton, G. E. (1999) *A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants*, 355–368. Cambridge, MA, USA: MIT Press.

Ortega, J. M. and Rheinboldt, W. C. (1970) *Iterative Solution of Nonlinear Equations in Several Variables.* USA: Society for Industrial and Applied Mathematics.

Polak, E. (1987) On the mathematical foundations of nondifferentiable optimization in engineering design. *SIAM Rev.*, **29**, 21–89.

Qin, Z., Scheinberg, K. and Goldfarb, D. (2013) Efficient block-coordinate descent algorithms for the group Lasso. *Math. Program. Comput.*, **5**, 143–169.

Ramsay, J. O., Hooker, G. and Graves, S. (2009) *Functional data analysis with R and Matlab.* New York: Springer.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional data analysis.* New York: Springer.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267–288.

Trefethen, L. N. and Bau, III, D. (1997) *Numerical linear algebra.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Wainwright, M. J. and Jordan, M. I. (2008) Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, **1**, 1–305.

Watkins, D. S. (2011) Francis's algorithm. *Amer. Math. Monthly*, **118**, 387–403.

Wu, T. T. and Lange, K. (2008) Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.*, **2**, 224–244.

— (2010) The MM alternative to EM. *Statist. Sci.*, **25**, 492–505.

Yang, Y. and Zou, H. (2015) A fast unified algorithm for solving group-lasso penalize learning problems. *Stat. Comput.*, **25**, 1129–1141.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 49–67.

Zhao, P., Rocha, G., Yu, B. et al. (2009) The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, **37**, 3468–3497.

Zhou, J., Wang, N.-Y. and Wang, N. (2013) Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica*, **23**, 25.

Appendix

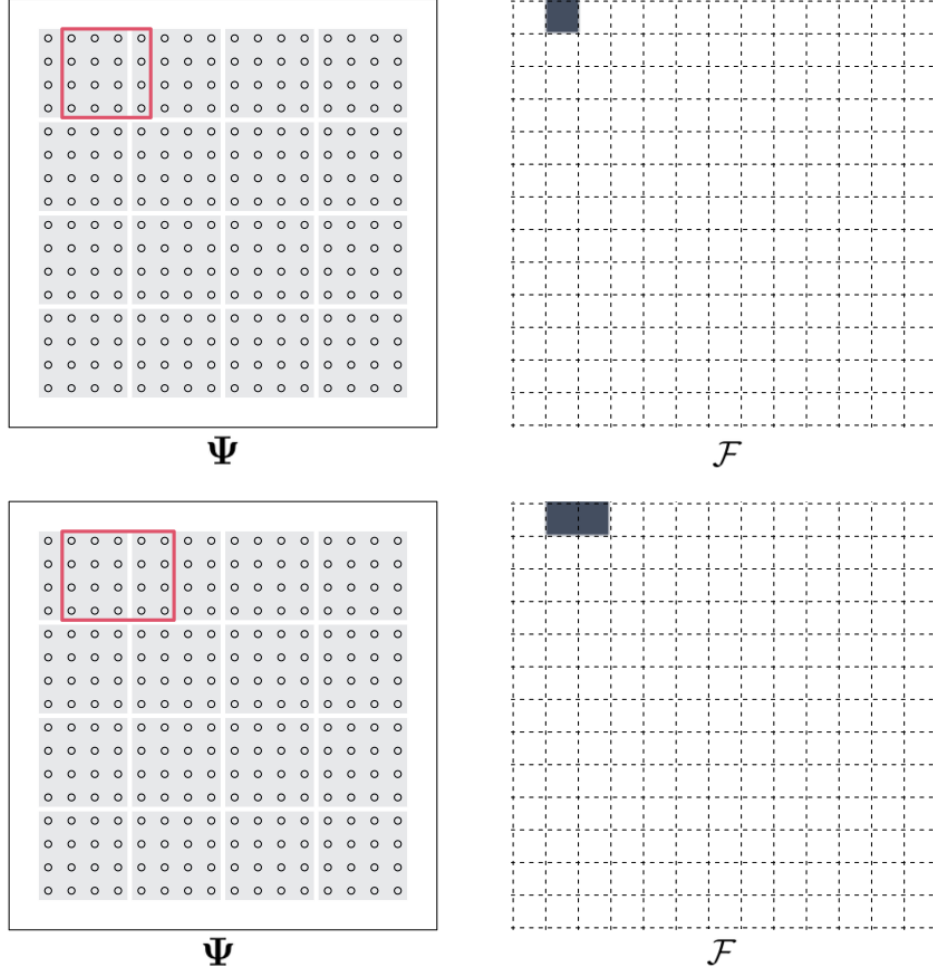## Examples of group-Lasso not allowed sparsity patterns

Figure 6: On the left, the matrices of coefficients $\mathbf{\Psi}$ partitioned in $d \times d$ ($d = 4$) blocks (grey background) coherent with a group-Lasso approach. On the right, the set $\mathcal{F}$ partitioned by the B-splines knots (dashed lines). Each square corresponds to a set $\mathcal{F}_{m,l}$. In blue we depict two sparsity patterns $\mathcal{F}_0$ that are not allowed under the group-Lasso definition, consistent with Section 2.2. Specifically in the upper plot $\psi(t,s) = 0$ for $(t,s) \in \mathcal{F}_{1,2}$ and no other neighboring sets. In the lower plot $\psi(t,s) = 0$ for $(t,s) \in \mathcal{F}_{1,2} \cup \mathcal{F}_{1,3}$. To get these sparsity patterns, all the coefficient inside the red contours in the matrix $\mathbf{\Psi}$ (left plot) need to be set to zero. The red tiles are not consistent with the disjoint grouping (in grey). Notably, the overlapping group structure that we adopt allows for these situations.

Additional computational results

In this Section we provide additional results concerning the MM algorithm introduced in Section 3. We first introduce the MM algorithm 1, then we provide the exact computational costs and we conclude the section by considering the problem of finding the minimum value of λ for which all the coefficients are equal to zero.

Algorithm 1: MM algorithm for penalized functional regression

---

Set $k = 0$, initialise the parameters $\widehat{\boldsymbol{\psi}}^{(0)}$ and evaluate $\ell\left(\widehat{\boldsymbol{\psi}}^{(0)}\right)$ as in equation (10), compute $\mathbf{Z}^T\mathbf{y}$, $\mathbf{y}^T\mathbf{y}$ and $\mathbf{D}_b^2 = \mathbf{D}_b^T\mathbf{D}_b$ for $b = 1, 2, \ldots, B+1$, and set $\ell\left(\widehat{\boldsymbol{\psi}}^{(-1)}\right) = 0$;

while $\left|\ell\left(\widehat{\boldsymbol{\psi}}^{(k)}\right) - \ell\left(\widehat{\boldsymbol{\psi}}^{(k-1)}\right)\right| > \epsilon$ **do**

    update $\widehat{\boldsymbol{\psi}}^{(k+1)}$, as in equation (15);

    update $\widehat{d}_b^{(k)}$ as in equation (14) and compute $\mathbf{H}^{(k)} = \left(\sum_{b=1}^{B+1} \widehat{d}_b^{(k)}\mathbf{D}_b^2\right)^{1/2}$;

    evaluate $\ell\left(\widehat{\boldsymbol{\psi}}^{(k+1)}\right)$ as in equation (10);

    set $k = k + 1$;

end

Proposition 2. *The exact computational cost in terms of floating point operations (see, e.g. Golub and Van Loan, 2013) of performing Algorithm 1 is* $13(nG)^3 + LM\big[2(nG)^2 + 3nG + 7B + 15\big] + (LM)^2 + 7nG + 6$ *which is of order* $\mathcal{O}\big((nG)^3\big)$.

Proof. The exact computational cost is

- $(LM+1)(2nG-1)$ to compute $\mathbf{Z}^T\mathbf{y}$ and $2LM + nG(2LM-1) + 2nG + 3LM(B+1) + 2$ to compute $\ell(\boldsymbol{\psi}^{(-1)})$ and $2LM - 1$ to compute $\mathbf{D}_b^2$, in the initialization step, totalling $4nGLM + 6LM + 3BLM + 2nG$;

- $3LM + 2$ to compute $\widehat{d}_n^{(k)}$ and $LM(B+1) + 1$ to compute $\mathbf{H}^{(k)}$, totalling $(B+4)LM + 3$;

- $13(nG)^3$ to compute the spectral decomposition of the matrix $\mathbf{J}^{(k)}$ leveraging the Francis' algorithm (Watkins, 2011; Golub and Uhlig, 2009), (see Trefethen and Bau, 1997);

- $4nG + (LM)^2$ to compute $(\mathbf{B}_\lambda^{(k)})^{1/2}$;

- $(2nG-1)nGLM + (LM)^2 + 3LM + 2$ to compute $\widehat{\boldsymbol{\psi}}^{(k+1)}$, totalling $(2(nG)^2 + nG + 3)LM + (LM)^2 + 3$;

- $2LM + nG(2LM - 1) + 2nG + 3LM(B + 1) + 2$ to compute $\ell(\boldsymbol{\psi}^{(k+1)})$, totalling $(2 + 2nG)LM + nG + 3LM(B + 1)$,

totalling $13(nG)^3 + LM\big[2(nG)^2 + 3nG + 7B + 15\big] + (LM)^2 + 7nG + 6.$ □

It is worth noting that the leading element of the computational cost is cubic in nG which is assumed to be less than LM. Moreover, leveraging the Shermann-Morrison-Woodbury formula has the further benefit that, as the MM algorithm iterates, the columns of the design matrix $\mathbf{Z}$ that correspond to groups of coefficients that have been previously set to zero are discarded.

The next proposition provides the minimum value of the tuning parameter λ for which all the coefficients in the minimization problem in Equation (10) are equal to zero.

Proposition 3. *For the overlap group-Lasso problem in equation* (10) *the smallest λ at which all coefficients $\widehat{\boldsymbol{\psi}}_{\lambda_{\max}}$ are equal to zero is*

$$\widehat{\lambda}_{\max} = \max_j \left\{ \frac{|\mathbf{Z}_j^T \mathbf{y}|}{\widetilde{\boldsymbol{D}}_{jj}} \right\}, \tag{20}$$

where $\mathbf{Z}_j$ denotes the j-th column vector of the matrix $\mathbf{Z}$ and $\widetilde{\boldsymbol{D}}_{jj}$ denotes the j-th diagonal element of the matrix $\widetilde{\mathbf{D}} = \sum_{b=1}^{B+1} \mathbf{D}_b^2$.

Proof. The objective function of the overlap group-Lasso in equation (10) can be decomposed into the sum of three terms:

$$\ell(\boldsymbol{\psi}) = \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\psi}\|_2^2 + \lambda \sum_{b=1}^{B+1} \sqrt{\boldsymbol{\psi}^T \mathbf{D}_b^2 \boldsymbol{\psi}} = \mathrm{RSS}(\boldsymbol{\psi}) + \mathcal{P}_\lambda^{\mathrm{OG}}(\boldsymbol{\psi}), \tag{21}$$

having subdifferential

$$\partial \ell(\boldsymbol{\psi}) = \left\{ -\mathbf{Z}^T(\mathbf{y} - \mathbf{Z}\boldsymbol{\psi}) + \mathcal{P}_\lambda^{\mathrm{OG}}(\boldsymbol{\psi}) \right\}, \tag{22}$$

where $\partial\mathcal{P}_\lambda^{\mathrm{OG}}(\boldsymbol{\psi}) = \partial\big(\lambda\sum_{b=1}^{B+1}\sqrt{\boldsymbol{\psi}^T\mathbf{D}_b^2\boldsymbol{\psi}}\big)$ is the subdifferential of the OG penalty term

$$\partial\mathcal{P}_\lambda^{\mathrm{OG}}(\boldsymbol{\psi}) = \partial\big(\lambda\sum_{b=1}^{B+1}\sqrt{\boldsymbol{\psi}^T\mathbf{D}_b^2\boldsymbol{\psi}}\big)$$

$$= \bigg\{\mathbf{s}\in\mathbb{R}^{LM} : \mathbf{s} = \sum_{b=1}^{B+1}\frac{\mathbf{D}_b^2\boldsymbol{\psi}}{\sqrt{\boldsymbol{\psi}^T\mathbf{D}_b^2\boldsymbol{\psi}}},\ \text{if }\|\boldsymbol{\psi}\|_2 > 0,\ \text{and }\|\boldsymbol{\psi}\|_2 \leq \lambda\sum_{b=1}^{B+1}\mathbf{D}_b^2,\ \text{otherwise}\bigg\}.$$

(23)

Since the objective function is convex the first order optimality condition $\partial\ell(\boldsymbol{\psi}) = 0$ and Theorem 1 ensure the existence of a unique $\widehat{\mathbf{s}}\in\partial\mathcal{P}_\lambda^{\mathrm{OG}}(\boldsymbol{\psi})$ fulfilling the condition $-\mathbf{Z}^T(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\psi}}) + \widehat{\mathbf{s}} = 0$. Plugging $\widehat{\boldsymbol{\psi}} = 0$ we get $\mathbf{Z}^T\mathbf{y} = \lambda\widehat{\mathbf{s}}$ from which we obtain the equivalent condition $\big(\sum_{b=1}^{B+1}\mathbf{D}_b^2\big)^{-1}\mathbf{Z}^T\mathbf{y} = \lambda\widehat{\mathbf{z}}$ and $\widehat{\mathbf{z}}$ is the dual variable satisfying $\widehat{z}_j = \mathrm{sign}(\widehat{\psi}_j)$ if $\widehat{\psi}_j \neq 0$ and $\widehat{z}_j \in [-1, 1]$ if $\widehat{\psi}_j = 0$ for $j = 1, \ldots, LM$. Therefore $\big\|\big(\sum_{b=1}^{B+1}\mathbf{D}_b^2\big)^{-1}\mathbf{Z}^T\mathbf{y}\big\|_\infty = \lambda\|\widehat{\mathbf{z}}\|_\infty$, which completes the proof. $\qquad\square$

Algorithm convergence properties

The following propositions state the local and global convergence properties of the MM algorithm. In what follows, let $\ell(\boldsymbol{\psi})$ be the objective function to be minimized and $\mathcal{Q}(\boldsymbol{\psi}|\widehat{\boldsymbol{\psi}}^{(k)})$ be the majorizing function at the current k-iteration, where $\widehat{\boldsymbol{\psi}}^{(k)}$ is the estimate of the parameter $\boldsymbol{\psi}$ at the k-iteration. Moreover, let $\widehat{\boldsymbol{\psi}}^{(k+1)} = M(\widehat{\boldsymbol{\psi}}^{(k)})$ denote the mimimizer of $\mathcal{Q}(\boldsymbol{\psi}|\widehat{\boldsymbol{\psi}}^{(k)})$ at iteration k. Following Lange (2010) we provide conditions for local and global convergence of the MM algorithm for proving the local and global convergence of the MM sequence delivered by the MM algorithm introduced in Section 3.

The following proposition (adapted from Lange, 2010, Proposition 15.3.2) provides the conditions for local convergence of any MM sequence.

Proposition 4. *Since the majorizing function is strictly convex, if the Hessian of the surrogate function $d^{20}\mathcal{Q}(\boldsymbol{\psi}^\infty|\boldsymbol{\psi}^\infty)$ is invertible, then the proposed MM algorithm is locally attracted to a local minimum $\boldsymbol{\psi}^\infty$ at a linear rate equal to the spectral radius $\rho(dM(\boldsymbol{\psi}^\infty))$ of $dM(\boldsymbol{\psi}^\infty) = \mathbf{I}_{LM} - \big(d^{20}\mathcal{Q}(\boldsymbol{\psi}^\infty|\boldsymbol{\psi}^\infty)\big)^{-1}d^2\ell(\boldsymbol{\psi}^\infty)$, where $\mathbf{I}_{LM}$ denotes the identity matrix of dimension LM.*

Proof. The mapping function in equation (15) is differentiable and the surrogate function is strictly convex. Moreover, the second derivative of $\mathcal{Q}(\boldsymbol{\psi}^\infty|\boldsymbol{\psi}^\infty)$ is invertible for any $\lambda > 0$. By Proposition 15.3.1 in Lange (2010) it is then sufficient to show that all the eigenvalues of the differential $dM(\boldsymbol{\psi}^\infty)$ lie in the half-open interval $[0,1)$. All the eigenvalues of the $dM(\boldsymbol{\psi}^\infty)$ can be determined as stationary points of the Rayleight quotient $\mathcal{R}_v(\widehat{\boldsymbol{\psi}}^{(\infty)}) = \frac{v^T(d^{20}\mathcal{Q}(\boldsymbol{\psi}^\infty|\boldsymbol{\psi}^\infty) - d^2\ell(\boldsymbol{\psi}^\infty))v}{v^T(d^{20}\mathcal{Q}(\boldsymbol{\psi}^\infty|\boldsymbol{\psi}^\infty)v} = 1 - \frac{v^T d^2\ell(\boldsymbol{\psi}^\infty)v}{v^T(d^{20}\mathcal{Q}(\boldsymbol{\psi}^\infty|\boldsymbol{\psi}^\infty)v}$. Moreover, strict convexity of both $\ell(\boldsymbol{\psi})$ and $\mathcal{Q}(\boldsymbol{\psi}|\widehat{\boldsymbol{\psi}}^{(k)})$ implies that $d^2\ell(\boldsymbol{\psi}^\infty)$ and $d^{20}\mathcal{Q}(\boldsymbol{\psi}^\infty|\boldsymbol{\psi}^\infty)$ are positive definite (see Polak, 1987) and therefore $\mathcal{R}_v(\widehat{\boldsymbol{\psi}}^{(\infty)}) < 1$ for any non null vector of length one. Positive semi-definiteness of $d^{20}\mathcal{Q}(\boldsymbol{\psi}^\infty|\boldsymbol{\psi}^\infty) - d^2\ell(\boldsymbol{\psi}^\infty)$ also implies $\mathcal{R}_v(\widehat{\boldsymbol{\psi}}^{(\infty)}) \geq 0$. $\qquad\square$

The convergence rate is usually used to characterize the convergence behavior of an iterative algorithm. It is well known that the convergence rate of an MM algorithm is, in general, linear. If $\{\widehat{\boldsymbol{\psi}}^{(k)}\}$ converges to some optimal point $\boldsymbol{\psi}^\infty$ of $\ell(\boldsymbol{\psi})$ and $M(\boldsymbol{\psi})$ is continuous, then $\boldsymbol{\psi}^\infty$ is a fixed point and $\boldsymbol{\psi}^\infty = M(\boldsymbol{\psi}^\infty)$. By Taylor expansion, $\widehat{\boldsymbol{\psi}}^{(k+1)} - \boldsymbol{\psi}^\infty = dM(\boldsymbol{\psi}^\infty)(\widehat{\boldsymbol{\psi}}^{(k)} - \boldsymbol{\psi}^\infty)$ where $dM(\boldsymbol{\psi}^\infty)$ is the matrix rate of convergence. The spectral radius of $dM(\boldsymbol{\psi}^\infty)$ is usually defined as the local convergence rate of the sequence (see, e.g. Lange, 2010; McLachlan and Krishnan, 2008). The following proposition provides conditions for global convergence of the MM algorithm.

Proposition 5. *If $\ell(\boldsymbol{\psi})$ is coercive, the subset $\{\boldsymbol{\psi} \in \boldsymbol{\Theta} \ni \ell(\boldsymbol{\psi}) \leq \ell(\widehat{\boldsymbol{\psi}}^{(k)})\}$ of parameter domain $\boldsymbol{\Theta}$ is compact and all stationary points of $\ell(\boldsymbol{\psi})$ are isolated. The majorizing function $\mathcal{Q}(\boldsymbol{\psi}|\widehat{\boldsymbol{\psi}}^{(k)})$ is strictly convex and differentiable in both $\boldsymbol{\psi}$ and $\widehat{\boldsymbol{\psi}}^{(k)}$, then the MM sequence $\{\widehat{\boldsymbol{\psi}}^{(k)}\}$ convergence to the stationary point of $\ell(\boldsymbol{\psi})$. Moreover, since $\ell(\boldsymbol{\psi})$ is strictly convex, then the limiting point of $\{\widehat{\boldsymbol{\psi}}^{(k)}\}$ is the minimum.*

Proof. For any fixed $\lambda > 0$, the objective function $\ell(\boldsymbol{\psi})$ is convex with one bounded local minimizer (see Proposition 1), which implies that $\ell(\boldsymbol{\psi})$ is coercive, (see Lange, 2010, 2016). The convexity of $\ell(\boldsymbol{\psi})$ further implies that $\ell(\boldsymbol{\psi})$ is Lipschitz continuous on each compact subset of $\mathbb{R}^{LM}$ (i.e. locally Lipschitz continuous). It follows that the gradient of $\ell(\boldsymbol{\psi})$ exists for almost all $\boldsymbol{\psi}$. Moreover, by the Liapunov theorem (Lange, 2010, Proposition 15.4.1) the set $\mathcal{D}$ of clustering

points generated by the sequence $\{\widehat{\boldsymbol{\psi}}^{(k)}\}$, for $k = 0, 1, \ldots$, is contained in the set $\mathcal{S}$ of stationary points of $\ell(\boldsymbol{\psi})$. Then, according to (Lange, 2010, Proposition 8.2.1), $\mathcal{D}$ is a connected set since any closed subset of a compact set is also compact. The condition that all stationary points of $\ell(\boldsymbol{\psi})$ are isolated easily implies that the number of stationary points in the compact set $\{\boldsymbol{\psi} \in \boldsymbol{\Psi} \ni \ell(\boldsymbol{\psi}) \leq \ell(\widehat{\boldsymbol{\psi}}^{(k)})\}$ can only be finite and since the cluster set $\mathcal{D}$ is a connected subset of a finite set $\mathcal{S}$, $\mathcal{D}$ reduces to a singleton. $\qquad\square$

Proofs of Section 4

Remark 2.

$$\mu(\bar{\mathcal{F}}_0 \backslash \tilde{\mathcal{F}}_0^{\text{concurrent}}) = \mu(\bar{\mathcal{F}}_0) = (M - d + 2)\frac{c}{(M - d + 2)^2} = c(M - d + 2)^{-1}.$$

$\qquad\square$

Theorem 1. Using Proposition 1 of Jenatton et al. (2011) and noting that the B-th element of the second sum in (10) is the norm of all the coefficients ψ_{ml}, the minimization in (10) leads to a unique set of coefficients $\widehat{\psi}_{ml}$. Since d and the knots defining the bases $\{\varphi_m(t), m = 1, 2, \ldots, M\}$ and $\{\theta_l(s), l = 1, 2, \ldots, L\}$ are fixed each B-spline basis represents a vector space of piecewise polynomials. Since their tensor product also generates a vector space, the elements of the tensor product basis are also linearly independent and so the coefficients ψ_{ml} of any given kernel $\psi(t, s)$ in this vector space are uniquely determined. Thus, the estimator $\widehat{\psi}_\lambda$ obtained by the minimization of (10) is unique. $\qquad\square$

Theorem 2. We first note that if $\psi_T \in V_{\varphi \otimes \theta}$, there exists a unique true $\boldsymbol{\Psi}_T$ and thus the event E_0 is equivalent to consistently estimate the non-zero patterns in $\boldsymbol{\Psi}$. To show this, it is sufficient to rely on Theorem 6 of Jenatton et al. (2011). The rest of the proof consists in noting that the conditions required by Theorem 6 of Jenatton et al. (2011) are satisfied under our set of assumptions. $\qquad\square$

Theorem 3. We study the consistency of $\widehat{\psi}_\lambda(t,s)$ with respect to $\psi_T^P(t,s)$ exploiting its representation through the vector space $V_{\varphi \otimes \theta}$ and using the results of Theorem 7 of Jenatton et al. (2011). Eventually the consistency of $\widehat{\psi}_\lambda(t,s)$ with respect to the true $\psi_T(t,s)$ is obtained letting the number of B-splines basis to grow with n since if both L and M are increasing we have that the event E_k approaches E_0 and $\int[\psi_T^P(t,s) - \psi_T(t,s)]^2 dt ds \to 0$. Note that for any $\mathcal{F}_1 \subset \mathcal{F}_{0T}^C$ we also have $\mathcal{F}_1 \subset \bar{\mathcal{F}}_{0T}^C$ since $\bar{\mathcal{F}}_{0T}$ is a subset of $\mathcal{F}_{0T}$. Thus the correct estimation of $\mathbf{\Psi}_T^P$ automatically leads to

$$\left(\inf_{\mathcal{F}_1 \subset \mathcal{F}_{0T}^C} \int_{\mathcal{F}_1} \left[\widehat{\psi}_\lambda(t,s) \right]^2 dt ds \right) > 0.$$

Despite correctly estimating $\mathbf{\Psi}_T^P$, however, it may happen that $\widehat{\psi}_\lambda(t,s) \neq 0$ for $(t,s) \in \mathcal{F}_{0T}$ for subregions of a finite number of sets $\mathcal{F}_{ml}$ in (6) of Lebesgue measure $(LM)^{-1}$. Since $\widehat{\psi}_\lambda(t,s)$ is bounded we can conclude that a correct estimation of the non zero patterns in $\mathbf{\Psi}_T^P$ would also lead to

$$\int_{\mathcal{F}_{0T}} \left[\widehat{\psi}_\lambda(t,s) \right]^2 dt ds \leq \frac{k}{LM}.$$

Hence, similarly to Theorem 2 we focus on the consistent estimation of the non-zero patterns in $\mathbf{\Psi}_T^P$. To this end, we exploit Theorem 7 of Jenatton et al. (2011) and show how the conditions required therein translate to our settings. We first sightly rewrite the minimization in (10) as

$$\widehat{\psi}_\lambda = \arg\min_{\psi} \left\{ \frac{1}{2} ||\mathbf{y} - \tilde{\mathbf{Z}}\psi||_2^2 + \lambda \sum_{b=1}^{B+1} ||\tilde{c}_b \odot \psi||_2 \right\} \odot \mathrm{V}_{\mathbf{Z}}^{-1}, \tag{24}$$

where $\tilde{\mathbf{Z}}$ is obtained dividing each column of $\mathbf{Z}$ by its standard deviation. Consistently with this operation we are intrinsically rescaling ψ and thus the final estimator is multiplied through the Hadamart product for the inverse of $\mathrm{V}_{\mathbf{Z}}$ which is the vector containing the standard deviations of $\mathbf{Z}$. To maintain the penalization consistent with our settings, the weights inside the norm are also redefined as $\tilde{c}_b = c_b \odot \mathrm{V}_{\mathbf{Z}}^{-1}$. With this operation, the minimization is conducted on a matrix $\tilde{\mathbf{Z}}$ such that $\tilde{\mathbf{Z}}^T \tilde{\mathbf{Z}}$ has unit diagonal. In addition, let κ be the minimum eigenvalue of the matrix $\tilde{\mathbf{Z}}_1^T \tilde{\mathbf{Z}}_1$. From condition (17) the number of nonzero coefficients in $\mathbf{\Psi}_T^P$ is strictly less than nG

and then $\tilde{\mathbf{Z}}_1^T \tilde{\mathbf{Z}}_1$ is positive definite and $\kappa > 0$.

Hence, defining the constants $C_1, C_2, C_3,$ and C_4 consistently with Jenatton et al. (2011) all the requirements of their Theorem 7 are satisfied. Hence the proof. $\qquad\square$