A New Computationally Efficient Algorithm to solve Feature Selection for Functional Data Classification in High-dimensional Spaces

Tobia Boschi [1] **Francesca Bonin** [1] **Rodrigo Ordonez-Hurtado** [1] **Alessandra Pascale** [1] **Jonathan Epperlein** [1]

Abstract

This paper introduces a novel methodology for Feature Selection for Functional Classification, FSFC, that addresses the challenge of jointly performing feature selection and classification of functional data in scenarios with categorical responses and multivariate longitudinal features. FSFC tackles a newly defined optimization problem that integrates logistic loss and functional features to identify the most crucial variables for classification. To address the minimization procedure, we employ functional principal components and develop a new adaptive version of the Dual Augmented Lagrangian algorithm. The computational efficiency of FSFC enables handling high-dimensional scenarios where the number of features may considerably exceed the number of statistical units. Simulation experiments demonstrate that FSFC outperforms other machine learning and deep learning methods in computational time and classification accuracy. Furthermore, the FSFC feature selection capability can be leveraged to significantly reduce the problem's dimensionality and enhance the performances of other classification algorithms. The efficacy of FSFC is also demonstrated through a real data application, analyzing relationships between four chronic diseases and other health and demographic factors. FSFC source code is publicly available at https://github.com/IBM/funGCN.

1. Introduction

Many contemporary scientific domains that require the integration and interpretation of time-dependent variables have recently benefited from Functional Data Analysis (FDA) (Ramsay & Silverman, 2005; Horváth & Kokoszka, 2012), an active domain within statistics and Machine Learning (ML) that allows working with multivariate longitudinal measurements by estimating smooth curves from multiple observations over a continuous domain. This functional representation facilitates the comparison of different statistical units and variables throughout the entire temporal domain. FDA has demonstrated its effectiveness in many recent applications where the emergence of novel technologies, such as brain sensors, DNA sequencers, and wearable devices, has facilitated the measurement of a large number of longitudinal variables (Shcherbina et al., 2017; Cremona et al., 2019; Smuck et al., 2021). In these modern domain specific applications of FDA, two criticality emerge: *1)* high dimensionality, the number of features may significantly exceed the number of statistical units (Ovsyannikova et al., 2020; Paul et al., 2014), *2)* data scarcity (Brigato & Iocchi, 2021). The first issue may lead ML and Deep Learning (DL) models to overfit, the second may lead current feature selection models to sub-optimal efficiency.
Hence, there is a growing demand for new algorithms able to i) effectively reduce problem dimensionality and feature complexity, ii) deal with data scarcity (Zhu et al., 2023), and iii) preserve computational efficiency (Brigato & Iocchi, 2021).

To address these challenges in instances characterized by multivariate longitudinal variables and binary categorical responses, we introduce a new feature selection methodology called Feature Selection for Functional Classification (`FSFC`). The key innovative aspects of our work are the following:

- `FSFC` has a unique capability to concurrently solve feature selection and classification for multivariate longitudinal data across multiple statistical units.
- FSFC effectively reduces problem dimensionality by identifying the most relevant variables that allow two classes to be distinguished. This capability makes it particularly suitable for high-dimensional scenarios where the number of features exceeds the number of statistical units.
- We demonstrate the efficacy of `FSFC` both as a standalone tool and as a pre-processing step to enhance other ML and DL methods.

To achieve these objectives, we develop a new optimization

[1]IBM Research Europe, Dublin. Correspondence to: Tobia Boschi <tobia.boschi@ibm.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

problem that integrates logistic loss and multivariate functional features. Specifically, we employ Functional Principal Components (FPC) and introduce an innovative adaptive version of the Dual Augmented Lagrangian (DAL) algorithm. FPC are used to simplify the complexity of each feature while retaining its longitudinal representation. On the other hand, DAL employs the sparsity structure of the problem to reduce its dimensionality. This combination of FPC and DAL makes `FSFC` highly computationally efficient.

We validate our approach through simulations and a real data application. In the real data application, we examine the relationships between four chronic diseases and other health and socio-demographic factors, demonstrating `FSFC` adaptability to multi-modal data scenarios.

Related Work. Feature selection has been a fertile field of research and development since the 1970s and is now among the most renowned and widely utilized techniques in ML and DL to decrease the dimensionality of optimization problems, remove irrelevant and redundant features, improve efficiency in learning and predictive tasks, and enhance model interpretability (Blum & Langley, 1997; Dash & Liu, 1997). It has been applied to several tasks, including linear regression (Tibshirani, 1996; Zou & Hastie, 2005), and classification (Cai et al., 2018) to handle a vast number of instances and when dealing with high-dimensional data (Liu et al., 2002; Das, 2001). Additionally, feature selection has been extensively studied also in the context of linear functional regression (Chen et al., 2016; Parodi et al., 2018; Boschi et al., 2021). While there are numerous existing methods for classifying longitudinal observations, including *supporting vector machines (SVM)* (Müller et al., 1997), *long short-term memory (LSTM) networks* (Hochreiter & Schmidhuber, 1997), *shapelets* (Ye & Keogh, 2009), and *functional classification* (Leng & Müller, 2006; Fraiman et al., 2016), to the best of our knowledge, no study to date has enabled simultaneous feature selection and classification within the framework of multivariate functional data. Established techniques like *Variational AutoEncoders* (Kingma et al., 2019) and *Time Series Feature Extraction* (Barandas et al., 2020) can reduce the complexity of each longitudinal feature by identifying key attributes. Similarly, (Torrecilla & Suárez, 2016) and (Blanquero et al., 2019) focus on identifying salient attributes within individual sets of curves, such as time intervals or important patterns. However, these approaches are neither designed to reduce the total number of variables by detecting the most relevant ones, nor to work in scenarios where the number of variables exceeds the sample size. Finally, although novel studies have extended various DL architectures to multivariate time series, such as *transformers* (Jiang et al., 2022; Vijay et al., 2023; Zhou et al., 2024) and *graph convolutional networks* (Langbridge et al., 2023), they do not address problems involving multiple statistical units.

In the remainder of this paper, we first present our methodology in Section 2, and then we benchmark its performance against SVMs and LSTMs via simulations (Section 3) and a real-world application on the longitudinal SHARE dataset (Section 4). We draw our conclusions in Section 5.

2. Methods

2.1. Problem definition

`FSFC` introduces a novel optimization problem that can be described as follows. Let the number of observations and features be denoted by n and p, respectively. Let $\mathcal{T} = [a, b]$ be a closed bounded interval. Without loss of generality, we consider $\mathcal{T} = [0, 1]$. The categorical responses are represented by $Y_i \in \{-1, 1\}$, the functional features by $\mathcal{X}_{ij} \in \mathbb{L}^2(\mathcal{T})$, and the functional model coefficients by $\mathcal{B}_j \in \mathbb{L}^2(\mathcal{T})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. It is assumed, without loss of generality, that $\mathcal{X}_{ij}$ is standardized with a mean function of 0 and standard deviation of 1. Finally, for a generic function $f \in \mathbb{L}^2(\mathcal{T})$ the squared $\mathbb{L}^2$-norm is defined as $\|f\|^2_{\mathbb{L}^2} = \langle f, f\rangle_{\mathbb{L}^2} = \int_{\mathcal{T}} f^2(t)\mathrm{dt}$. The `FSFC` optimization problem is then formulated as follows:

$$\min_{\mathcal{B}_1,\ldots,\mathcal{B}_p} \left[\sum_{i=1}^{n} \log\left(1 + \exp\left(-Y_i \sum_{j=1}^{p} \int_{\mathcal{T}} \mathcal{B}_j(t)\mathcal{X}_{ij}(t)dt\right)\right) + \sum_{j=1}^{p} \omega_j \left(\lambda_1 \|\mathcal{B}_j\|_{\mathbb{L}^2} + \frac{\lambda_2}{2}\|\mathcal{B}_j\|^2_{\mathbb{L}^2}\right)\right] \tag{1}$$

The first term of the objective function (1) is the logistic loss with functional features and enables classification. The second term comprises two different penalties that induce sparsity to facilitate feature selection. The parameters λ_1 and λ_2 control the global importance of the penalties with respect to the logistic loss. The feature-specific weights ω_j extend the *adaptive* LASSO and Elastic Net (Zou, 2006; Zou & Zhang, 2009) to the functional classification settings and can improve feature selection by reducing the active set cardinality.

In order to solve (1), first, we approximate the functional variables via the Functional Principal Components (FPC) method (see, e.g. Horváth & Kokoszka, 2012), and subsequently, perform the minimization process using a Dual Augmented Lagrangian (DAL) algorithm. The DAL algorithm was introduced Tomioka & Sugiyama (2009); Li et al. (2018) to solve LASSO and Elastic-Net problems and has been extended by Boschi et al. (2023) to accommodate functional linear regression cases. DAL leverages the intrinsic sparsity of the problem and the sparsity induced by the Hessian matrix information, resulting in a substantial reduction in computational cost. In this work, we devise a novel adaptive version of DAL and develop the mathematical theory

underpinning the algorithm to incorporate logistic loss and functional features within the objective function.

2.2. Matrix representation

To address (1), we employ a matrix representation obtained by expressing functional variables as linear combinations of basis functions (Ramsay & Silverman, 2005; Kokoszka & Reimherr, 2017). Specifically, we represent the functional features using their first k FPCs. FPCs are a well-established technique to solve the functional regression (Reiss & Ogden, 2007) and classification (Preda et al., 2007; Wang et al., 2016) problems. We now show how to use FPCs to obtain a matrix representation of (1).

Consider $\mathcal{X}_j = [\mathcal{X}_{1j} \mid \ldots \mid \mathcal{X}_{nj}]^T$ representing the set of n curves of the j-th feature. Let $e^j = \left[e_1^j \mid \ldots \mid e_k^j\right]$ denote the matrix containing the first k FPCs of $\mathcal{X}_j$. Define $X = \left[X_{[1]} | \ldots | X_{[p]}\right] \in \mathbb{R}^{n\times pk}$ and $B = \left(B_1^T, \ldots, B_p^T\right)^T \in \mathbb{R}^{pk}$, where $X_{[j]} \in \mathbb{R}^{n\times k}$ and $B_j \in \mathbb{R}^k$ are the score matrix of $\mathcal{X}_j$ and the score vector of $\mathcal{B}_j$ with respect to the same basis system e^j, respectively, for $j = 1, \ldots, p$. Specifically, $X_{[j](is)} = \langle \mathcal{X}_{ji}, e_s^j\rangle_{\mathbb{L}^2}$ and $B_{j,s} = \langle \mathcal{B}_j, e_s^j\rangle_{\mathbb{L}^2}$. Then, we can approximate $\mathcal{X}_j(t) \approx X_{[j]}e^j(t)^T$, $\mathcal{B}_j(t) \approx e^j(t)B_j$, and $\int_{\mathcal{T}} \mathcal{B}_j(t)\mathcal{X}_{ij}(t)dt \approx X_{[j](i)} \left(\int e^j(t)^T e^j(t)dt\right) B_j$. Note that we use the notation $B_{j,s}$ and $X_{[j](i)}$ to indicate the s-th element of the vector B_j and i-row of the matrix $X_{[j]}$, respectively. Since e^j are orthonormal bases, then $\int e(t)^T e(t)dt = I_k$, i.e., the identity matrix of order k. Moreover, for a generic function $f \in \mathbb{L}^2([0,1])$, one has that $\|f\|_{\mathbb{L}^2}^2 = \sum_{i=1}^{\infty} \langle f, e_i\rangle_{\mathbb{L}^2}$ (Kokoszka & Reimherr, 2017). Thus, we can approximate the $\mathbb{L}^2$ functional norm and the standard l_2 vector norm, i.e., the Frobenius norm, denoted by $\|\cdot\|_2$. We can now express (1) in a matrix form as:

$$\min_B \left[\sum_{i=1}^{n} \log\left(1 + \exp\left(-Y_i \cdot \left(X_{(i)}B\right)\right)\right) + \sum_{j=1}^{p} \omega_j \left(\lambda_1 \|B_j\|_2 + \frac{\lambda_2}{2}\|B_j\|_2^2\right)\right], \tag{2}$$

where $X_{(i)}$ indicates the i-th row of X. Mimicking the Group Elastic Net regularization (Zou & Hastie, 2005; Simon et al., 2013), Equation (2) combines two penalties: the first is non-differentiable, creating sparsity, while the second (Ridge-like) is differentiable, controlling multicollinearity and accelerating the convergence of the optimization algorithm. Once we obtain an estimate of the score vector $\hat{B}$, the coefficient curves can be recovered as $\hat{\mathcal{B}}_j(t) = e^j(t)\hat{B}_j(t)$, the class probability as $\hat{p}_i = 1/\left(1 + \exp\left(X_{(i)}\hat{B}\right)\right)$, and the categorical response as $\hat{Y}_i = 1$ if $\hat{p}_i > 0.5$ and -1 otherwise.

Selection of K. The selection of k determines the extent to which Equation (2) approximates Equation (1). FPCs have the significant advantage of being the most parsimonious orthonormal basis system: in most instances, a few components capture more than 90% of the curves' variability. As discussed in the next section, using a small k is critical for the efficiency of the DAL algorithm. The choice of k also accounts for the trade-off between computational efficiency and approximation accuracy. In the current version of `FSFC`, we require k to be identical across all features, i.e., one must use the same number of basis components for all features. In the Discussion section, we further detail the possibility of accommodating a distinct k for each feature.

2.3. Dual Augmented Lagrangian (DAL) Algorithm

The core idea underlying the DAL methodology involves minimizing the *Augmented Lagrangian* function associated with the *dual* problem. Prior to defining the dual problem, note that (2) can be expressed as

$$\min_B \left[\, h(XB) + \pi(B) \,\right], \tag{P}$$

where $h(XB) = \sum_{i=1}^{n} \log\left(1+\exp\left(-Y_i\cdot(X_{(i)}B)\right)\right)$ is the logistic loss function and $\pi(B) = \sum_{j=1}^{p} \omega_j(\lambda_1\|B_{[j]}\|_2 + \frac{\lambda_2}{2}\|B_{[j]}\|_2^2)$ is the adaptive Elastic Net-type penalty. Note that the penalty $\pi(B)$ is the same as that employed by Boschi et al. (2023) in the scalar-on-function regression scenario. Although their results can be utilized for penalty-related operators, a novel theoretical framework must be developed for the functional logistic loss. In particular, one needs to redefine a set of mathematical operators in broader dimensions while maintaining the DAL sparsity structure and efficiency.

Dual problem. A potential dual representation of the primal (P) can be derived from (Boyd & Vandenberghe, 2004; Tomioka et al., 2011) as:

$$\min_{V,Z} \left[\, h^*(V) + \pi^*(Z) \,\right] \quad s.t. \quad X^T V + Z = 0\,. \tag{D}$$

In this formulation, $V \in \mathbb{R}^n$ and $Z \in \mathbb{R}^{pk}$ are the dual variables. Following the notation introduced for B, we can express $Z = \left(Z_1^T, \ldots, Z_p^T\right)^T$, where $Z_j \in \mathbb{R}^k$ denotes the sub-vector of Z corresponding to the j-th feature. h^* and π^* denote the Fenchel-conjugate functions (Fenchel, 1949) of h and π, respectively. The computation of π^* is derived in (Boschi et al., 2023) as

$$\pi^*(Z) = \sum_{j=1}^{p} \pi^*(Z_j) = \sum_{j=1}^{p} (2\omega_j\lambda_2)^{-1} \left(\left[\,\|Z_j\|_2 - \omega_j\lambda_1\right]_+\right)^2, \tag{3}$$

where $[\ \cdot\ ]_+$ is the positive part operator: $[s]_+ = s$ if $s > 0$, and 0 otherwise. The form of h^* is given in the next Proposition (a proof can be found in the Appendix, Section A).

Algorithm 1 DAL Method

GOAL: minimize $\mathcal{L}_\sigma(V, Z, B)$. Start from the initial values V^0, Z^0, B^0, σ^0

`WHILE NOT CONVERGED:`

(1) Given B^s, find V^{s+1} and Z^{s+1} which approximately solve:

$$\left(V^{s+1}, Z^{s+1}\right) \approx \arg\min_{V,Z} \mathcal{L}_\sigma\left(V, Z \mid B^s\right) \tag{6}$$

> **Inner sub-problem:** to find (V^{s+1}, Z^{s+1}), update V and Z *independently*:
>
> `WHILE NOT CONVERGED`
>
> $V^{m+1} = \arg\min_V \mathcal{L}_\sigma(V \mid Z^m, B^s)$
> $\longrightarrow$ **Newton method** - see Proposition (2)
>
> $Z^{m+1} = \arg\min_Z \mathcal{L}_\sigma(Z, \mid V^{m+1}, B^s)$
> $\longrightarrow$ **closed-form** - see (7)

(2) Update the Lagrangian multiplier B and the parameter σ:

$$B^{s+1} = B^s - \sigma^s\left(X^T V^{s+1} + Z^{s+1}\right),\ \sigma^s \le \sigma^{s+1} < \infty$$

Proposition 2.1. *Considering h as in Equation (P), then the function h^* is defined for $|Y_i V_i| < 1$ as follows:*

$$h^*(V) = \sum_{i=1}^{n} (1 - |Y_i V_i|)\log(1 - |Y_i V_i|) + |Y_i V_i| \log(|Y_i V_i|)\ . \tag{4}$$

We are now ready to define the *Dual Augmented Lagrangian* function as

$$\begin{aligned}\mathcal{L}_\sigma(V, Z, B) &= h^*(V) + \pi^*(Z) \\ &- \sum_{j=1}^{p} \langle B_j, V^T X_j + Z_j\rangle + \frac{\sigma}{2}\sum_{j=1}^{p} \|V^T X_j + Z_j\|_2^2\ ,\end{aligned} \tag{5}$$

with $\sigma > 0$. Note that the Dual Augmented Lagrangian is defined by augmenting the dual objective function with a penalization term that accounts for the violation of the problem constraint. Our DAL implementation is outlined in Algorithm 1, as adapted from Boschi et al. (2023). The core part of the algorithm involves solving the inner sub-problem (6), which entails minimizing the Dual Augmented Lagrangian with respect to the dual variables V and Z. To obtain an approximate solution of (6), V and Z can be updated independently (Tomioka & Sugiyama, 2009). The primal variable B and the parameter σ are updated according to standard rules (Li et al., 2018). The choice of σ^0 is empirically studied in Tomioka et al. (2011): starting from very small values of σ results in an increased number of iterations required for convergence; however, if σ^0 is excessively large or increases too rapidly, the DAL method fails to converge to the optimal solution.

Update of Z. The minimization with respect to Z has a closed-form solution $\bar{Z}$, which is computed in Boschi et al. (2023) based on the *proximal operator* (Rockafellar, 1976b;a) of π, denoted as $\text{prox}_{\sigma\pi}$. Specifically, the following holds:

$$\begin{aligned}\bar{Z} &= \text{prox}_{\pi^*/\sigma}\left(B/\sigma - X^T\bar{V}\right) \\ &= B/\sigma - X^T\bar{V} - \text{prox}_{\sigma p}\left(B - \sigma X^T\bar{V}\right)/\sigma\ ,\end{aligned} \tag{7}$$

where $\text{prox}_{\sigma\pi}(B) = \left(\text{prox}_{\sigma\pi}(B_1), \ldots, \text{prox}_{\sigma\pi}(B_p)\right)^T$, and for each j

$$\text{prox}_{\sigma\pi}(B_j) = (1 + \sigma\omega_j\lambda_2)^{-1}\left[1 - \|B_j\|_2^{-1}\sigma\omega_j\lambda_1\right]_+ B_j\ . \tag{8}$$

Update of V. Given that the minimization with respect to V lacks a closed-form solution, we update V applying the *Newton Method* (Nocedal & Wright, 1999) to the function $\psi(V) := \mathcal{L}_\sigma(V \mid \bar{Z}, B)$. The Newton update is expressed by $V^{m+1} = V^m + sD$, where $D \in \mathbb{R}^n$ represents the descent direction and s the step-size. To obtain D, we solve the linear system

$$H_\psi(V)D = -\nabla\psi(V)\ , \tag{9}$$

where $H_\psi \in \mathbb{R}^{n\times n}$ and $\nabla\psi \in \mathbb{R}^n$ are the Hessian matrix and the gradient vector of ψ, respectively. To determine the step size, we implement the *line-search* strategy proposed by Li et al. (2018). Starting with $s = 1$, we iteratively reduce it by a factor of 0.5 until the condition

$$\psi\left(V^{m+1}\right) \le \psi(V^m) + \mu s \nabla\psi^T(V^m)D \tag{10}$$

is met, with $\mu \in (0, 0.5)$. Solving the linear system (9) is the most computationally demanding step in the DAL algorithm. However, the following proposition leverages the sparsity structure of H_ψ to significantly reduce the system dimension and computational complexity (a proof can be found in the Appendix, Section A).

Proposition 2.2. *Let $T = B - \sigma X^T V$, $T_j = B_j - X_{[j]}^T V$, $\mathcal{J} = \left\{j\ :\ \|T_j\|_2 \ge \sigma\omega_j\lambda_1\right\}$, and $r = |\mathcal{J}|$ be the cardinality of $\mathcal{J}$. Next, let $X_\mathcal{J} \in \mathbb{R}^{n\times rk}$ be the sub-matrix of X restricted to the blocks X_j, $j \in \mathcal{J}$. Define the square $k \times k$ matrix $P_{[j]}$ as*

$$P_{[j]} = (1 + \sigma\omega_j\lambda_2)^{-1}\left(\left(1 - \frac{\sigma\omega_j\lambda_1}{\|T_j\|_2}\right)I_k + \frac{\sigma\omega_j\lambda_1}{\|T_j\|_2^3}T_j T_j^T\right)\ .$$

Finally, let $Q_\mathcal{J} \in \mathbb{R}^{rk\times rk}$ be the block-diagonal matrix formed by the blocks $P_{[j]}$, $j \in \mathcal{J}$. Then:

(i) $\psi(V) = h^*(V)$
$$+ \frac{1}{2\sigma}\sum_{j=1}^{p}\left((1 + \sigma\omega_j\lambda_2)\left\|\text{prox}_{\sigma\pi}\left(T_j\right)\right\|_2^2 - \|B_j\|_2^2\right)\ ,$$

(ii) $\nabla\psi(V) = \nabla h^*(V) - X\,\text{prox}_{\sigma\pi}(T)\ ,$

(iii) $H_\psi(V) = H_{h^*}(V) + \sigma X_\mathcal{J} Q_\mathcal{J} X_\mathcal{J}^T\ ,$

where $\nabla h^*(V) \in \mathbb{R}^n$ and $H_{h^*}(V) \in \mathbb{R}^{n\times n}$ are the gradient vector and the Hessian matrix of h^*, respectively. Specifically, when $|Y_iV_i| < 1$, each element i of $\nabla h^*(V)$ is equal to

$$Y_i \log\left((1-|Y_iV_i|)^{-1}|Y_iV_i|\right),$$

and $H_{h^*}(V)$ is a diagonal matrix with elements given by

$$\left(|Y_iV_i|\,(1-|Y_iV_i|)\right)^{-1}.$$

Computational efficiency. As stated in Proposition 2, the DAL algorithm takes advantage of the sparsity information embedded within the Hessian matrix, selecting a subset $\mathcal{J}$ of r active features, with r decreasing at each iteration. By selecting an r-feature subset, the overall computational expense of solving the linear system (9) is reduced from $\mathcal{O}\left(n(n^2+npk+p^2k^2)\right)$ to $\mathcal{O}\left(n(n^2+nrk+r^2k^2)\right)$ with $r < p$. In sparse scenarios, where the number of active features is low, r might be significantly smaller than n, further reducing the computational burden through the application of the *Sherman-Morrison-Woodbury* formula (Van Loan & Golub, 1983): $\left(H_{h^*}+\sigma X_{\mathcal{J}}Q_{\mathcal{J}}X_{\mathcal{J}}^T\right)^{-1} = H_{h^*}^{-1} - H_{h^*}^{-1}X_{\mathcal{J}}\left((\sigma Q_{\mathcal{J}})^{-1} + X_{\mathcal{J}}^T H_{h^*}^{-1} X_{\mathcal{J}}\right)^{-1} X_{\mathcal{J}}^T H_{h^*}^{-1}$. This equivalence enables the factorization of an $rk \times rk$ matrix, resulting in a total cost of $\mathcal{O}\left(rk(k^2+nrk+r^2k^2+n^2)\right)$ since H_{h^*} is diagonal and computing its inverse is straightforward. Remarkably, the computational burden is not dependent on the total number of features p, but solely on the number of active features r. Provided that sparsity is maintained (i.e., r remains small), the number of features can grow without impacting the efficiency of the linear system resolution. However, it is important to acknowledge that a larger p increases the cost of the principal components computation during the matrix representation stage, since the FPC scores have to be derived for each feature. Note that the computational cost depends on k^3. By maintaining a small value for k, which describes the number of bases used to represent the functional features, the efficiency of the method is significantly enhanced.

Convergence criteria. DAL's convergence properties have been studied in both classical (Tomioka & Sugiyama, 2009; Li et al., 2018) and functional (Boschi et al., 2023) contexts. In order to guarantee the DAL super-linear convergence rate, we must implement consistent stopping-criteria to evaluate the convergence of both the inner sub-problem and the overall algorithm. With respect to the inner sub-problem, Z and V are updated iteratively until the following condition is satisfied (Tomioka et al., 2011):

$$\|\nabla\psi(V^s)\|_2 \le 2\sigma^{1/2}\sum_{j=1}^{p}\|(X^TV^{s+1}+Z^{s+1})_j\|_2\,.$$

For the overall algorithm's convergence, we monitor one of the *Karush-Kuhn-Tucker* (KKT) conditions associated with the dual problem (D), namely, $X^TV = -Z$. This condition is exclusively satisfied by the optimal solutions of (D) (Boyd & Vandenberghe, 2004). Then, the algorithm is halted when the standardized residual of the KKT equation is less than a specified tolerance tol, as follows:

$$\left(1+\|V\|_2+\sum_{j=1}^{p}\|Z_j\|_2\right)^{-1}\sum_{j=1}^{p}\left\|(X^TV+Z)_j\right\|_2 < \text{tol}\,. \quad (11)$$

2.4. Model selection and adaptive implementation

To assess the solution of (2) with varying values of the penalty parameter λ_1, we implement a path search mechanism. We perform the search for λ_1 using the formula $\lambda_1 = c_\lambda\lambda_{max}$, with c_λ belonging to a grid of 100 values evenly spaced on a logarithmic scale from 1 to 0.01, and $\lambda_{max} = 0.5\max_j\|(X^TY)_{(j)}/\omega_j\|$. For $c_\lambda = 1$ (i.e., $\lambda_1 = \lambda_{max}$), 0 active features are selected, and as λ_1 decreases for further values of c_λ, the number of features selected by the solution increases. At this stage, the weights w_j's in (2) are all equal to 1. We select λ_1 along the path that minimizes a *5-fold cross-validation* classification accuracy score, denoting the optimal value of λ_1 and the corresponding solution as $\tilde{\lambda}_1$ and $\tilde{B}$, respectively. The *adaptive solution* is then computed starting from $\tilde{B}$. We set $\omega_j = \text{sd}_B/\|\tilde{B}_j\|_2$, where sd_B is the standard deviation of $\left(\|\tilde{B}_1\|_2,\ldots,\|\tilde{B}_r\|_2\right)$, and we execute a single DAL minimization considering $\lambda_1 = \tilde{\lambda}_1$. The adaptive feature-specific weights w_j's allow for imposing a greater penalty on the coefficient curves that, despite a small norm, have not been screened out by the unweighted minimization, thus promoting their removal from the active set.

3. Simulation Study

Settings. In this section, we evaluate the performance of `FSFC` using synthetic data. We benchmark `FSFC` against two representative techniques for longitudinal data classification: a `kernlab R package` (Karatzoglou et al., 2023) implementation of `SVM`, and a `TensorFlow` (Abadi et al., 2015) implementation of `LSTM`. Furthermore, we apply `SVM` and `LSTM` on a reduced streamlined problem that considers solely the active features identified by `FSFC` and their FPC representation – essentially, employing the `FSFC` output as input. Notably, `SVM` and `LSTM` do not have any selection or problem-reduction capabilities. These approaches are called `r-SVM` and `r-LSTM`, respectively. The hyper-parameters used for all the methods are discussed and detailed in the Appendix, Section B.

We consider two distinct scenarios: one with $n = 300$ and $p = 800$, and another with $n = 600$ and $p = 2000$. We denote the number of active features (i.e., non-zero regression coefficient curves) as p_0. For each scenario, we examine four different levels of sparsity by setting $p_0 =$

Table 1. Simulation results. Average CPU processing time, measured in seconds, for 50 instances in each scenario. All the computations were executed on a MacBookPro 2021 with an M1 Max processor and 32GB of RAM.

	p_0	LSTM	SVM	FSFC	rLSTM	rSVM
$n = 300$	2	140.61	16.17	1.44	5.56	0.01
$p = 800$	5	141.25	16.19	1.72	5.61	0.01
	10	139.93	16.17	1.84	5.66	0.02
	20	140.05	16.12	2.11	5.72	0.04
$n = 600$	2	355.66	144.18	5.44	9.52	0.01
$p = 2000$	5	354.16	142.14	7.09	9.65	0.02
	10	348.51	141.96	7.58	9.86	0.05
	20	349.58	142.82	8.16	10.11	0.13

$2, 5, 10, 20$. The synthetic data are generated as described in Appendix Section C.

We assess FSFC in terms of *selection* performance, while all the methods are evaluated with respect to *classification accuracy* and *computational efficiency*. To assess selection performance, we compute the *recall* and *precision* scores. Classification accuracy is computed as the proportion of observation correctly classified, in both the *training* and the *test* sets, where the latter is generated independently from the former. We set the size of the test set $n_{test} = n/3$, with n being the size of the training set.

Results. The simulation study results are presented in Figure 1. The displayed boxplots are derived from the distribution of the various scores obtained across 50 replications of each scenario. In terms of FSFC feature selection performance, the recall score consistently surpasses the precision score, maintaining an average above 80%, except for the scenario with $n = 600, p = 2000$, and $p_0 = 20$. This suggests that while FSFC might occasionally select non-relevant features, it is proficient in identifying the active ones. Both metrics show a decline as the number of active features grows, indicating challenges in more complex scenarios.

Except for LSTM, all models exhibit high classification accuracy in the training set. FSFC outperforms all competitors in the test set, with an average accuracy larger than 85% when $p_0 = 2, 5, 10$. As the proportion of active features increases, i.e., in scenarios characterized by larger p_0 and smaller p, distinguishing non-relevant information becomes more challenging also for FSFC. Notably, r-SVM and r-LSTM test accuracy outperforms the standard SVM and LSTM (especially in the scenario with a larger p): employing FSFC for a preliminary feature screening and problem dimension reduction results in a significant performance enhancement, potentially mitigating overfitting.

The average CPU time reported in Table 1 demonstrates that FSFC is significantly more efficient than SVM and LSTM. This difference is even larger when p increases. As expected, FSFC's computational cost does not depend on p, but only on the number of active features p_0. By exploiting the properties of DAL optimization, FSFC can solve a problem with a large number of features ($p = 2000$) in under 10 seconds. Given that a 5-fold cross-validation is implemented on a grid of 100 values, this implies that solving DAL takes, on average, less than 0.2 seconds per repetition. Moreover, leveraging FSFC as a preliminary reduction step significantly trims the CPU time for both r-SVM and r-LSTM, given the much smaller feature set they operate upon.

Finally, in Appendix Table E1, we report the accuracy and the CPU time related to a hybrid approach that combines Feature Extraction (Barandas et al., 2020) with LSTM and SVM. This approach does not reduce the number of features but simplifies the longitudinal variables by extracting salient attributes before performing the classification algorithms. Contrary to rLSTM and rSVM, results indicate that Feature Extraction does not enhance the accuracy of LSTM and SVM while significantly increasing the computation time.

Additional analyses. In addition to the above *classification accuracy* analysis, we have included further analyses in the Appendix that also consider the *average AUC score* as a new evaluation metric. Our findings demonstrate that both metrics consistently align.

In Appendix Table E2, we include *random forest* (Breiman, 2001) and *multilayer perceptron networks* (Popescu et al., 2009) in the comparison. These approaches do not naturally account for temporal dependencies, and consequently the longitudinal data need to be flattened into single-dimension vectors, which results in the loss of time-dependent information. FSFC still outperforms these competitors. Notably, leveraging FPC coefficients and FSFC-selected features as input significantly increases their performance (as observed for LSTM and SVM).

Appendix Table E3 showcases how varying the hyperparameters k and λ_2 – set to $(1 - \alpha)\lambda_1$ – do not impact FSFC selection and classification performances. As discussed in Subsection 2.4 and in Appendix Section B, the most critical FSFC tuning parameter is λ_1, which is linked to the Lasso-penalty and creates sparsity, determining the number of features in the model. The ridge penalty λ_2 influences convergence speed and addresses feature collinearity without notably affecting feature selection.

Appendix Table E4 presents FSFC performance across three varying noise level scenarios for $k = 5, 10$. Appendix Figure E1 depicts a sample of 10 curves for each noise scenario. For k=5, the average explanation of curve variability is 95%, 70%, and 50%, respectively. For k=10, these figures rise to 99%, 95%, and 80%. In less smooth scenarios, FPCs are less effective, and more components are required to reconstruct the curves accurately. Despite a slight performance drop in

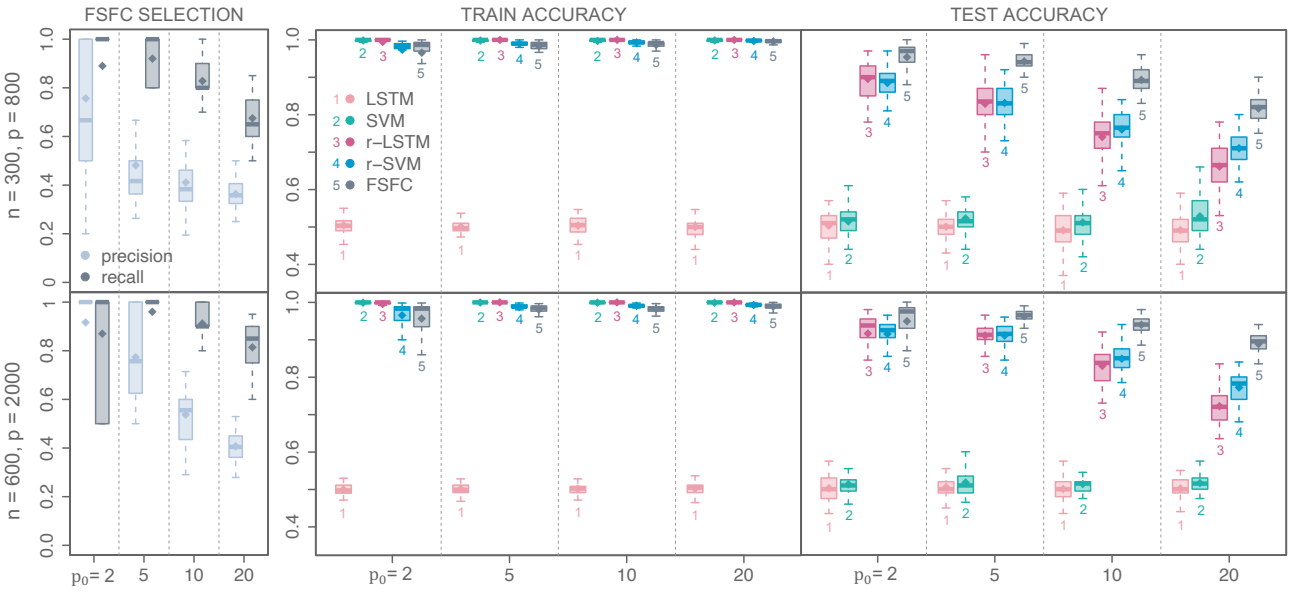

Figure 1. Simulation results. Boxplots generated from the distribution obtained across 50 replications of each scenario, with gray diamonds and horizontal lines indicating means and medians of the distributions, respectively. Selection performances (precision and recall) are computed just for `FSFC`, while classification accuracy in the training/test set is reported for all the examined algorithms (`LSTM`, `SVM`, `r-LSTM`, `r-SVM`, `FSFC`). The rows illustrate two distinct scenarios ($n = 300$, $p = 800$, and $n = 600$, $p = 2000$). In each scenario, we investigate $p_0 = 2, 5, 10, 20$ (x-axes).

noisier settings, `FSFC` still achieves high accuracy. Notably, classification metrics and recall values remain stable across varying k. Even if the first five components capture less variability, they still effectively identify crucial features and perform classification accurately.

4. SHARE Application

The Survey of Health, Ageing and Retirement in Europe - SHARE (Alcser et al., 2005) is a research infrastructure that aims to investigate the effects of health, social, economic, and environmental policies on the life course of European citizens (Börsch-Supan et al., 2013; Bergmann et al., 2017; Börsch-Supan, 2020). SHARE is a longitudinal study, where the same subjects are followed over multiple years. Specifically, eight surveys or "waves" were conducted from 2004 to 2020 (see Appendix Figure E2 for the SHARE timeline). We focus on the 1518 subjects who participated in at least seven out of the eight waves, ensuring a sufficient number of measurements for reliable curve estimation. We investigate a subset of $p = 36$ variables from the EasySHARE dataset (Gruber et al., 2014), a preprocessed version of the SHARE data. In our application, we aim to study the relationship between 4 chronic diseases – *i.e., diabetes, myocardial infarction, high cholesterol*, and *hypertension* – and various physical and mental health, socio-demographic, and healthcare factors (a comprehensive list of the examined variables can be found in Appendix Table E5). While some of them are described by values that change over time (e.g., CASP index, max grip) and are suitable for a functional representation, others are scalar (e.g., education years) or categorical (e.g., gender) and do not evolve over different waves. We smoothed time-varying variables using cubic *B-splines* with knots at each survey date and roughness penalty on the curve's second derivative (Ramsay & Silverman, 2005). For each curve, the smoothing parameter was selected by minimizing the average generalized cross-validation error (Craven & Wahba, 1978). Although the surveys' dates and the amount of measurements may vary across subjects, the functional representation provides a natural imputation for missing values and facilitates the comparison of different statistical units across the entire temporal domain. Scalar and categorical variables were treated as constant functions over the time domain, facilitating their inclusion in the analysis without compromising feature selection consistency and showing `FSFC`'s proficiency in dealing with multi-modal data scenarios. This adjustment results in `FSFC` estimating functional coefficients for these features rather than real numbers. These coefficients will be constant functions across the domain where the function's value aligns with the conventional logistic coefficient. We indicate the models for the selected categorical response variables as m_{d} (diabetes), m_{inf} (myocardial infarction), m_{c} (high cholesterol), and m_{hyp} (hypertension).

We explore two different experiment settings, detailed in

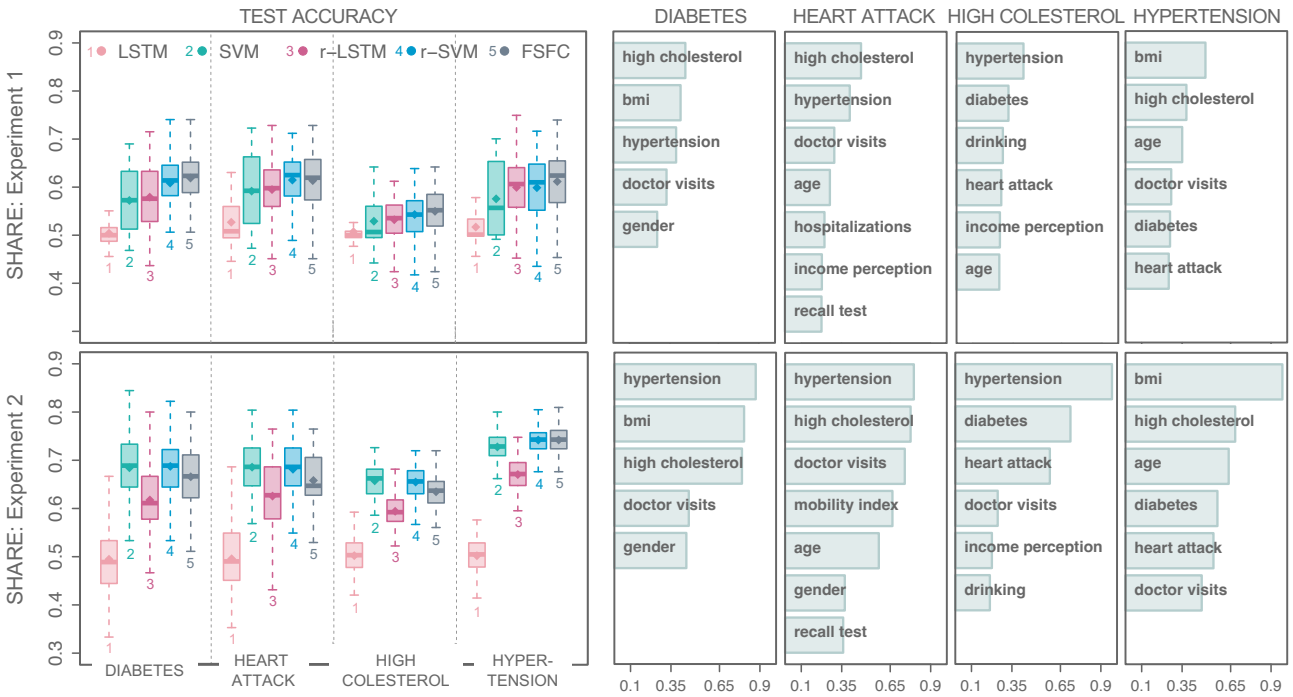

Figure 2. Experiment 1 (upper panel) and Experiment 2 (lower panel) SHARE results. The test set classification accuracy boxplots (on the left) are generated from 100 replications. The dots and the horizontal lines indicate the means and medians of the distributions, respectively. On the right, features selected by FSFC for more than 80 out of 100 replications are displayed for each response variable. The bar plots illustrate the average ratio of λ_{max} at which each feature entered the active set. The higher the ratio, the earlier the feature is included in the model during the λ path search.

Appendix Section D. In *Experiment 1*, we limit our analysis to a subset of 20 subjects to replicate a scenario where $p > n$. In *Experiment 2*, we consider the complete data. In this case, $n > p$. For each response, in both experiments, all the methods were executed for 100 replications using the hyper-parameters detailed in the Appendix, Section B.

Classification results. Figure 2 presents classification and selection results. In *Experiment 1*, FSFC achieves the best median accuracy and boosts the performance of SVM and LSTM. In *Experiment 2*, SVM and FSFC achieve superior and comparable outcomes. In this setting, with $n > p$, applying FSFC as a pre-processing step improves the effectiveness of LSTM and does not significantly impact SVM. Applying FSFC as a pre-processing step improves the effectiveness of LSTM – which already avoids model overfitting. Nonetheless, FSFC still has the unique capability of identifying crucial features and producing insightful and meaningful outcomes.

Appendix Table E6 presents the accuracy results from integrating Feature Extraction with LSTM and SVM. In situations where the number of features is comparable to the sample size, Feature Extraction significantly boosts the performance of SVM, especially in the Experiment 2 settings. However, this approach does not identify the key features related to the binary response and does not preserve their longitudinal nature.

Feature selection results. It is worth noticing that the most frequently selected features by the four models are consistent across the two experiments, showing FSFC feature selection efficiency even when the number of samples is small. Moreover, Appendix Table E7 displays the top five selected features for each disease when pre-processing the SHARE dataset with three different smoothing levels. Notably, the selected features for each disease are remarkably consistent across the three scenarios, underlying FSFC's robustness in selection across varying smoothness and noise levels.

Comparing FSFC selection results with medical literature, one can find evidence of the model's accuracy in identifying important connections between features. The four examined chronic diseases exhibit strong interconnectivity. Both m_{d} and m_{inf} identify hypertension and high cholesterol as crucial features, while m_{c} and m_{hyp} incorporate the remaining three investigated diseases in the model. Substantial medical literature supports the association between these diseases, including the following studies: Kearney

et al. (2005); Collaboration et al. (2010); Cosentino et al. (2020); Baigent et al. (2010). *Gender* has been identified as significant in diabetes (Kautzky-Willer et al., 2016) and myocardial infarction (i.e. hearth attack) (Vaccarino et al., 1999; Bairey Merz et al., 2006); *age* is associated with myocardial infarction (White et al., 1996; Avezum et al., 2005) and hypertension (Franklin et al., 1997; Vasan et al., 2002); *bmi* is related to diabetes (Chan et al., 1994; Mokdad et al., 2003) and hypertension (Chobanian et al., 2003; Gelber et al., 2007); and the level of *mobility* and *recall test score* are factors relevant to myocardial infarction (Mora et al., 2007; Haring et al., 2013). Model m_c also selects *drinking* behavior and quality of life factors such as *income perception*, findings supported in Rimm et al. (1999); Hare et al. (2014). Lastly, the number of doctor visits is selected as a relevant factor by all models, aligning with multiple studies that document the significant impact of chronic diseases on healthcare utilization (Hoffman et al., 1996; Lehnert et al., 2011).

5. Conclusions

In this paper, we present an innovative method, `FSFC`, that jointly performs feature selection and classification of multivariate functional data. We utilize the properties of Functional Principal Components and implement a novel variant of the DAL algorithm that leverages the sparsity structure of the dual Hessian information to significantly reduce the problem's dimensionality. `FSFC`'s computational efficiency enables (i) handling high-dimensional scenarios where the number of features may far exceed the number of statistical units, (ii) performing an exhaustive search through the algorithm's hyperparameters.

A simulation study demonstrates that `FSFC` outperforms `SVM` and `LSTM` in terms of classification accuracy and computational time when the number of features is much larger than the sample size. Moreover, `FSFC`'s unique feature selection capability reduces problem dimensionality and enhances the efficacy of the competing methods. By applying `FSFC` to data from the SHARE study, we identify well-documented relationships between four chronic diseases in the literature. Furthermore, `FSFC` uncovers other critical health and socio-demographic factors that play a significant role in differentiating between the healthy and affected groups, as supported by numerous research studies.

Future work and limitations. A detailed exploration of `FSFC`'s performance on non-smooth data, backed by theoretical insights, is a promising avenue for future research. This would significantly enhance our understanding of `FSFC`'s reliability in various data conditions. For instance, `FSFC`'s performance tends to decline in unbalanced datasets. To mitigate this issue, we recommend to consider balancing techniques, such as oversampling the under-represented class or undersampling the over-represented class.

Moreover, while the SHARE application showcases `FSFC`'s capability with longitudinal, scalar, and categorical data, modern applications often involve diverse data types like text and images. Future implementations of `FSFC` should consider developing specific embeddings for each data modality, allowing for a unified representation. Another limitation of the current `FSFC` implementation concerns the parameter k, i.e., the number of principal components used to approximate the functional features. Currently, k must be the same for all features. Adapting `FSFC` to accommodate varying dimensions for different modalities would broaden its applicability and facilitate the integration of heterogeneous data types.

Finally, we plan to incorporate functional responses in the model. Instead of a static category, the response would describe a class-belonging probability that evolves over time, and it could also consider more than two different classes, similar to multinomial regression (Kwak & Clayton-Matthews, 2002).

Impact Statement

While the primary goal of this paper is to advance the field of Machine Learning, FSFC has the potential to improve analysis processes in sectors that rely on complex, high-dimensional data, such as healthcare, environmental science, and economics. For instance, in healthcare, FSFC can help identify critical health factors from longitudinal patient data, leading to more accurate diagnoses and personalized treatment plans. As FSFC is applied in real-world scenarios, for a positive societal impact it is essential to ensure data privacy, prevent biases, and promote transparency and interpretability of the results.

Acknowledgements

The authors received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 945449. The authors also thank their partners in IMEC for helping access the SHARE data.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan,

V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Alcser, K. H., Benson, G., BÃ¶rsch-Supan, A., Brugiavini, A., Christelis, D., Croda, E., Das, M., de Luca, G., Harkness, J., Hesselius, P., Jappelli, T., JÃ¼rges, H., Kalwij, A., Kemperman, M.-L., Klevmarken, A., Lipps, O., Paccagnella, O., Padula, M., Perrachi, F., Rainato, R., van Soest, A., Swensson, B., Vis, C., Weber, G., and Weerman, B. *The Survey of Health, Aging, and Retirement in Europe: Methodology*. Mannheim Research Institute for the Economics of Aging (MEA), 2005.

Avezum, A., Makdisse, M., Spencer, F., Gore, J. M., Fox, K. A., Montalescot, G., Eagle, K. A., White, K., Mehta, R. H., Knobel, E., et al. Impact of age on management and outcome of acute coronary syndrome: observations from the global registry of acute coronary events (grace). *American heart journal*, 149(1):67–73, 2005.

Baigent, C., Blackwell, L., Emberson, J., Holland, L., Reith, C., Bhala, N., Peto, R., Barnes, E., Keech, A., Simes, J., et al. Efficacy and safety of more intensive lowering of ldl cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials. *Lancet (London, England)*, 376(9753):1670–1681, 2010.

Bairey Merz, C. N., Shaw, L. J., Reis, S. E., Bittner, V., Kelsey, S. F., Olson, M., Johnson, B. D., Pepine, C. J., Mankad, S., Sharaf, B. L., et al. Insights from the nhlbi sponsored women's ischemia syndrome evaluation (wise) study: Part ii: gender differences in presentation, diagnosis, and outcome with regard to gender-based pathophysiology of atherosclerosis and macrovascular and microvascular coronary disease. *Journal of the American College of Cardiology*, 47(3S):S21–S29, 2006.

Barandas, M., Folgado, D., Fernandes, L., Santos, S., Abreu, M., Bota, P., Liu, H., Schultz, T., and Gamboa, H. Tsfel: Time series feature extraction library. *SoftwareX*, 11: 100456, 2020.

Bergmann, M., Kneip, T., De Luca, G., and Scherpenzeel, A. Survey participation in the survey of health, ageing and retirement in europe (share), wave 1-6. *Munich: Munich Center for the Economics of Aging*, 2017.

Blanquero, R., Carrizosa, E., Jiménez-Cordero, A., and Martín-Barragán, B. Variable selection in classification for multivariate functional data. *Information Sciences*, 481:445–462, 2019.

Blum, A. L. and Langley, P. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1-2):245–271, 1997.

Börsch-Supan, A. Survey of health, ageing and retirement in europe (share) wave 5. *Release version*, 7(0), 2020.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., and Zuber, S. Data resource profile: the survey of health, ageing and retirement in europe (share). *International journal of epidemiology*, 42(4):992–1001, 2013.

Boschi, T., Reimherr, M., and Chiaromonte, F. A highly-efficient group elastic net algorithm with an application to function-on-scalar regression. *Advances in Neural Information Processing Systems*, 34, 2021.

Boschi, T., Testa, L., Chiaromonte, F., and Reimherr, M. Fasten: an efficient adaptive method for feature selection and estimation in high-dimensional functional regressions. *arXiv preprint arXiv:2303.14801*, 2023.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Breiman, L. Random forests. *Machine learning*, 45:5–32, 2001.

Brigato, L. and Iocchi, L. A close look at deep learning with small data. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2490–2497, 2021. doi: 10.1109/ICPR48806.2021.9412492.

Cai, J., Luo, J., Wang, S., and Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2018.

Chan, J. M., Rimm, E. B., Colditz, G. A., Stampfer, M. J., and Willett, W. C. Obesity, fat distribution, and weight gain as risk factors for clinical diabetes in men. *Diabetes care*, 17(9):961–969, 1994.

Chen, Y., Goldsmith, J., and Ogden, R. T. Variable selection in function-on-scalar regression. *Stat*, 5(1):88–101, 2016.

Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo Jr, J. L., Jones, D. W., Materson, B. J., Oparil, S., Wright Jr, J. T., et al. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. *hypertension*, 42(6):1206–1252, 2003.

Collaboration, E. R. F. et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *The lancet*, 375(9733):2215–2222, 2010.

Cosentino, F., Grant, P. J., Aboyans, V., Bailey, C. J., Ceriello, A., Delgado, V., Federici, M., Filippatos, G., Grobbee, D. E., Hansen, T. B., et al. 2019 esc guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the easd: The task force for

diabetes, pre-diabetes, and cardiovascular diseases of the european society of cardiology (esc) and the european association for the study of diabetes (easd). *European heart journal*, 41(2):255–323, 2020.

Craven, P. and Wahba, G. Smoothing noisy data with spline functions. *Numerische mathematik*, 31(4):377–403, 1978.

Cremona, M. A., Xu, H., Makova, K. D., Reimherr, M., Chiaromonte, F., and Madrigal, P. Functional data analysis for computational biology. *Bioinformatics (Oxford, England)*, 35(17):3211, 2019.

Cressie, N. and Huang, H.-C. Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448):1330–1339, 1999.

Das, S. Filters, wrappers and a boosting-based hybrid for feature selection. In *Icml*, volume 1, pp. 74–81, 2001.

Dash, M. and Liu, H. Feature selection for classification. *Intelligent data analysis*, 1(1-4):131–156, 1997.

Fenchel, W. On conjugate convex functions. *Canadian Journal of Mathematics*, 1(1):73–77, 1949.

Fraiman, R., Gimenez, Y., and Svarc, M. Feature selection for functional data. *Journal of Multivariate Analysis*, 146: 191–208, 2016.

Franklin, S. S., Gustin IV, W., Wong, N. D., Larson, M. G., Weber, M. A., Kannel, W. B., and Levy, D. Hemodynamic patterns of age-related changes in blood pressure: the framingham heart study. *Circulation*, 96(1):308–315, 1997.

Gelber, R. P., Gaziano, J. M., Manson, J. E., Buring, J. E., and Sesso, H. D. A prospective study of body mass index and the risk of developing hypertension in men. *American journal of hypertension*, 20(4):370–377, 2007.

Gruber, S., Hunkler, C., and Stuck, S. Generating easyshare: guidelines, structure, content and programming. Technical report, SHARE Working Paper Series 17-2014. Munich, 2014.

Hare, D. L., Toukhsati, S. R., Johansson, P., and Jaarsma, T. Depression and cardiovascular disease: a clinical review. *European heart journal*, 35(21):1365–1372, 2014.

Haring, B., Leng, X., Robinson, J., Johnson, K. C., Jackson, R. D., Beyth, R., Wactawski-Wende, J., von Ballmoos, M. W., Goveas, J. S., Kuller, L. H., et al. Cardiovascular disease and cognitive decline in postmenopausal women: Results from the w omen's h ealth i nitiative m emory s tudy. *Journal of the American Heart Association*, 2(6): e000369, 2013.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hoffman, C., Rice, D., and Sung, H.-Y. Persons with chronic conditions: their prevalence and costs. *Jama*, 276(18): 1473–1479, 1996.

Horváth, L. and Kokoszka, P. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

Jiang, H., Liu, L., and Lian, C. Multi-modal fusion transformer for multivariate time series classification. In *2022 14th International Conference on Advanced Computational Intelligence (ICACI)*, pp. 284–288. IEEE, 2022.

Karatzoglou, A., Smola, A., and Hornik, K. *kernlab: Kernel-Based Machine Learning Lab*, 2023. URL https://CRAN.R-project.org/package=kernlab. R package version 0.9-32.

Kautzky-Willer, A., Harreiter, J., and Pacini, G. Sex and gender differences in risk, pathophysiology and complications of type 2 diabetes mellitus. *Endocrine reviews*, 37(3):278–316, 2016.

Kearney, P. M., Whelton, M., Reynolds, K., Muntner, P., Whelton, P. K., and He, J. Global burden of hypertension: analysis of worldwide data. *The lancet*, 365(9455):217–223, 2005.

Kingma, D. P., Welling, M., et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Kokoszka, P. and Reimherr, M. *Introduction to functional data analysis*. CRC Press, 2017.

Kwak, C. and Clayton-Matthews, A. Multinomial logistic regression. *Nursing research*, 51(6):404–410, 2002.

Langbridge, A., O'Donncha, F., Ba, A., Lorenzi, F., Lohse, C., and Ploennigs, J. Causal temporal graph convolutional neural networks (ctgcn). *arXiv preprint arXiv:2303.09634*, 2023.

Lehnert, T., Heider, D., Leicht, H., Heinrich, S., Corrieri, S., Luppa, M., Riedel-Heller, S., and König, H.-H. Health care utilization and costs of elderly persons with multiple chronic conditions. *Medical Care Research and Review*, 68(4):387–420, 2011.

Leng, X. and Müller, H.-G. Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, 22(1):68–76, 2006.

Li, X., Sun, D., and Toh, K.-C. A highly efficient semismooth newton augmented lagrangian method for solving lasso problems. *SIAM Journal on Optimization*, 28(1): 433–458, 2018.

Liu, H., Hussain, F., Tan, C. L., and Dash, M. Discretization: An enabling technique. *Data mining and knowledge discovery*, 6:393–423, 2002.

Mokdad, A. H., Ford, E. S., Bowman, B. A., Dietz, W. H., Vinicor, F., Bales, V. S., and Marks, J. S. Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *Jama*, 289(1):76–79, 2003.

Mora, S., Cook, N., Buring, J. E., Ridker, P. M., and Lee, I.-M. Physical activity and reduced risk of cardiovascular events: potential mediating mechanisms. *Circulation*, 116(19):2110–2118, 2007.

Müller, K.-R., Smola, A. J., Rätsch, G., Schölkopf, B., Kohlmorgen, J., and Vapnik, V. Predicting time series with support vector machines. In *International conference on artificial neural networks*, pp. 999–1004. Springer, 1997.

Nocedal, J. and Wright, S. J. *Numerical optimization*. Springer, 1999.

Ovsyannikova, I. G., Haralambieva, I. H., Crooke, S. N., Poland, G. A., and Kennedy, R. B. The role of host genetics in the immune response to sars-cov-2 and covid-19 susceptibility and severity. *Immunological reviews*, 296(1):205–219, 2020.

Parodi, A., Reimherr, M., et al. Simultaneous variable selection and smoothing for high-dimensional function-on-scalar regression. *Electronic Journal of Statistics*, 12 (2):4602–4639, 2018.

Paul, I. M., Williams, J. S., Anzman-Frasca, S., Beiler, J. S., Makova, K. D., Marini, M. E., Hess, L. B., Rzucidlo, S. E., Verdiglione, N., Mindell, J. A., et al. The intervention nurses start infants growing on healthy trajectories (insight) study. *BMC pediatrics*, 14(1):184, 2014.

Popescu, M.-C., Balas, V. E., Perescu-Popescu, L., and Mastorakis, N. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7):579–588, 2009.

Preda, C., Saporta, G., and Lévéder, C. Pls classification of functional data. *Computational Statistics*, 22(2):223–235, 2007.

Ramsay, J. O. and Silverman, B. W. *Functional data analysis*. Springer, 2 edition, 2005.

Reiss, P. T. and Ogden, R. T. Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102(479): 984–996, 2007.

Rimm, E. B., Williams, P., Fosher, K., Criqui, M., and Stampfer, M. J. Moderate alcohol intake and lower risk of coronary heart disease: meta-analysis of effects on lipids and haemostatic factors. *Bmj*, 319(7224):1523–1528, 1999.

Rockafellar, R. T. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976a.

Rockafellar, R. T. Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research*, 1(2):97–116, 1976b.

Shcherbina, A., Mattsson, C. M., Waggott, D., Salisbury, H., Christle, J. W., Hastie, T., Wheeler, M. T., and Ashley, E. A. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of personalized medicine*, 7(2):3, 2017.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

Smuck, M., Odonkor, C. A., Wilt, J. K., Schmidt, N., and Swiernik, M. A. The emerging clinical role of wearables: factors for successful implementation in healthcare. *NPJ Digital Medicine*, 4(1):45, 2021.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Tomioka, R. and Sugiyama, M. Dual-augmented lagrangian method for efficient sparse reconstruction. *IEEE Signal Processing Letters*, 16(12):1067–1070, 2009.

Tomioka, R., Suzuki, T., and Sugiyama, M. Super-linear convergence of dual augmented lagrangian algorithm for sparsity regularized estimation. *Journal of Machine Learning Research*, 12(5), 2011.

Torrecilla, J. L. and Suárez, A. Feature selection in functional data classification with recursive maxima hunting. *Advances in Neural Information Processing Systems*, 29, 2016.

Vaccarino, V., Parsons, L., Every, N. R., Barron, H. V., and Krumholz, H. M. Sex-based differences in early mortality after myocardial infarction. *New England journal of medicine*, 341(4):217–225, 1999.

Van Loan, C. F. and Golub, G. H. *Matrix computations*. Johns Hopkins University Press, 1983.

Vasan, R. S., Beiser, A., Seshadri, S., Larson, M. G., Kannel, W. B., D'Agostino, R. B., and Levy, D. Residual lifetime risk for developing hypertension in middle-aged women

and men: The framingham heart study. *Jama*, 287(8): 1003–1010, 2002.

Vijay, E., Jati, A., Nguyen, N., Sinthong, G., and Kalagnanam, J. Tsmixer: lightweight mlp-mixer model for multivariate time series forecasting. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2023.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. Functional data analysis. *Annual Review of Statistics and its application*, 3:257–295, 2016.

White, H. D., Barbash, G. I., Califf, R. M., Simes, R. J., Granger, C. B., Weaver, W. D., Kleiman, N. S., Aylward, P. E., Gore, J. M., Vahanian, A., et al. Age and outcome with contemporary thrombolytic therapy: results from the gusto-i trial. *Circulation*, 94(8):1826–1833, 1996.

Ye, L. and Keogh, E. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 947–956, 2009.

Zhou, T., Niu, P., Sun, L., Jin, R., et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36, 2024.

Zhu, Y., Li, W., and Li, T. A hybrid artificial immune optimization for high-dimensional feature selection. *Know.-Based Syst.*, 260(C), jan 2023. ISSN 0950-7051. doi: 10.1016/j.knosys.2022.110111. URL https://doi.org/10.1016/j.knosys.2022.110111.

Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Zou, H. and Zhang, H. H. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37 (4):1733, 2009.

Appendix

A. Proofs

A.1. Proof of Proposition 1

Note that $h(XB)$ can be expressed as a separable sum, i.e.: $h(XB) = \sum_{i=1}^{n} h(X_{(i)}B)$, with $h(X_{(i)}B) = \log\big(1+\exp\big(-Y_i \cdot (X_{(i)}B)\big)\big)$, where $Y_i \in \{-1, 1\}$. Hence, we have (Boyd & Vandenberghe, 2004):

$$h^*(V) = \sum_{i=1}^{n} h^*(V_i)\ .$$

By definition, $h^*(V_i) = g\big(\tilde{b}\big)$, where $g(b) = bV_i - h(b)$ and $\tilde{b}_i = \max_b g(b)$.

<u>*Case* $Y_i = -1$, $V_i \in (0, 1)$</u>

We have $g(b) = bV_i - \log\big(1 + \mathrm{e}^b\big)$. To find $\tilde{b}$, we derive $g(b)$ and set the derivative equal to 0, obtaining:

$$V_i = \big(1 + \mathrm{e}^{\tilde{b}}\big)^{-1}\mathrm{e}^{\tilde{b}}\ . \tag{12}$$

To solve for $\tilde{b}$, we take the logarithm of both side and note $\log\big(1 + \mathrm{e}^{\tilde{b}}\big) = h\big(\tilde{b}\big)$. Therefore,

$$\tilde{b} = \log(V_i) + h\big(\tilde{b}\big).$$

To find an explicit form for $h\big(\tilde{b}\big)$, we manipulate (12):

$$(12) \quad \Leftrightarrow \quad 1 - V_i = 1 - \big(1 + \mathrm{e}^{\tilde{b}}\big)^{-1}\mathrm{e}^{\tilde{b}} \quad \Leftrightarrow \quad \log(1 - V_i) = -\log\big(1 + \mathrm{e}^{\tilde{b}}\big)\ .$$

The last equality gives us $h\big(\tilde{b}\big) = -\log(1 - V_i)$. Then, we can compute:

$$h^*(V_i) = g\big(\tilde{b}\big) = V_i\big(\log(V_i) - \log(1 - V_i)\big) + \log(1 - V_i) = V_i \log(V_i) + (1 - V_i)\log(1 - V_i)\ .$$

<u>*Case* $Y_i = 1$, $V_i \in (-1, 0)$</u>

We have $g(b) = bV_i - \log\left(1 + \mathrm{e}^b\right)$. Following the same steps of the previous case, we obtain:

$$h^*(V_i) = -V_i \log(-V_i) + (1 + V_i)\log(1 + V_i)\ .$$

Considering $|Y_1, V_1| < 1$ and combining the two cases, we finally obtain the desired result:

$$h^*(V_i) = (1 - |Y_iV_i|)\log(1 - |Y_iV_i|) + |Y_iV_i|\log(|Y_iV_i|)\ .$$

A.2. Proof of Proposition 2

A significant part of this proof relies on the results of Boschi et al. (2023). The authors prove a similar proposition in the case of a scalar-on-function linear regression scenario where the same penalty π is considered.

(i) We need to compute $\psi(V) = \mathcal{L}_\sigma(V \mid \bar{Z}, B)$. Plugging $\bar{Z}$ in Equation (5) , we get:

$$\psi(V) = h^*(V) + \pi^*(\bar{Z}) + \frac{1}{2\sigma}\sum_{j=1}^{p}\left\|\operatorname{prox}_{\sigma\pi}\left(T_j\right)\right\|_2^2 - \frac{1}{2\sigma}\sum_{j=1}^{p}\|B_j\|_2^2\ . \tag{13}$$

Next, from (Boschi et al., 2023), we know:

$$\pi^*(\bar{Z}_j) = (\omega_j\lambda_2/2)\left((1 + \sigma\omega_j\lambda_2)^{-1}\left(\|T_j\|_2 - \sigma\omega_j\lambda_1\right)\right)^2\ .$$

Furthermore, starting from the definition of $\mathrm{prox}_{\sigma\pi}$ given in Equation (8) , we have:

$$\begin{aligned}\|\mathrm{prox}_{\sigma\pi}(T_j)\|_2^2 &= \left((1+\sigma\omega_j\lambda_2)^{-1}\left(1-\|T_j\|_2^{-1}\sigma\omega_j\lambda_1\right)\|T_j\|_2\right)^2 \\ &= \left((1+\sigma\omega_j\lambda_2)^{-1}\left(\|T_j\|_2-\sigma\omega_j\lambda_1\right)\right)^2 .\end{aligned}$$

Therefore,

$$\pi^*(\bar{Z}_j) = (\omega_j\lambda_2/2)\left\|\mathrm{prox}_{\sigma\pi}\left(T_{[j]}\right)\right\|_2^2,$$

which leads to $\pi^*(\bar{Z}) = \sum_{j=1}^{p}(\omega_j\lambda_2/2)\|\mathrm{prox}_{\sigma\pi}(T_j)\|_2^2$. If we plug this expression of $\pi^*(\bar{Z})$ in (13), we obtain the desired result.

(ii) Again, from (Boschi et al., 2023), we know:

$$\nabla_V\left(\frac{1}{2\sigma}\sum_{j=1}^{p}(1+\sigma\omega_j\lambda_2)\left\|\,\mathrm{prox}_{\sigma\pi}\left(T_j\right)\right\|_2^2\right) = -X\,\mathrm{prox}_{\sigma\pi}(T)\,.$$

To complete the proof, it is sufficient to note that the derivative of $h^*(V_i) = (1-|Y_iV_i|)\log(1-|Y_iV_i|)+|Y_iV_i|\log(|Y_iV_i|)$ is:

$$\nabla h_i^*(V) = Y_i\log\left((1-|Y_iV_i|)^{-1}|Y_iV_i|\right)\,.$$

(iii) From (Boschi et al., 2023) we have :

$$\nabla_V\big(-X\,\mathrm{prox}_{\sigma\pi}(T)\big) = \sigma X_{\mathcal{J}}Q_{\mathcal{J}}X_{\mathcal{J}}^T\,.$$

Note that $\nabla_i h^*(V)$ in (ii) depends only on V_i, which implies that all the off-diagonal elements of H_{h^*} are 0. The i-th diagonal element of H_{h^*} is then computed as follows:

$$H_{h^*(ii)} = \nabla_V\Big(Y_i\log\left((1-|Y_iV_i|)^{-1}|Y_iV_i|\right)\Big) = \big(|Y_iV_i|\,(1-|Y_iV_i|)\big)^{-1}\,.$$

Combining the two equations, we complete the proof.

B. Algorithms' hyper-parameters

For both `SVM` and `LSTM`, the inputs are the functional features evaluated on a grid of 100 uniformly spaced time points for the simulations scenarios, and 192 uniformly spaced time points (one for each month) for the SHARE application.

The `SVM` method utilizes the default parameters from the `kernlab R package`, with `kernel = rbfdot`.

The `LSTM` architecture consists of two layers with 50 units each, activated by the `relu` function. This is followed by a dense layer that outputs class probabilities and is activated by the `softmax` function. Training is conducted over 200 epochs, utilizing batches of 32 and the `adam` optimizer.

In all scenarios, `FSFC` performs a path search on a grid of 100 different λ_1 values. Specifically, $\lambda_1 = c_\lambda\lambda_{max}$ and $\lambda_2 = (1-\alpha)\lambda_1$, where $\lambda_{max} = 0.5\max_j\|(X^TY)_{(j)}/\omega_j\|$, c_λ belongs to a grid of 100 values spaced on a logarithmic scale from 1 to 0.01, and α is set to 0.2. After selecting the model that minimizes a 5-fold cross-validation criterion, the adaptive procedure is implemented as detailed in the main text. For each instance, `FSFC` starts from an initial value of $\sigma^0 = 0.1c_\lambda/\lambda_{max}$ increased by a factor of $\max(\min(5, 1+10c_\lambda), 1.1)$ at each iteration. Choosing σ^0 based on λ_{max} and c_λ is a common practice in the DAL optimization literature (Tomioka et al., 2011). The tolerance in Equation (11), employed for determining algorithm convergence, is set to 10^{-4}, and the parameter μ in the line-search procedure in Equation (10) is set to 0.2. Finally, we set $k = 5$. Across all examined scenarios and for every feature, five functional principal components are enough to capture more than 95% of the curves' variability, allowing for a good approximation of the features while maintaining manageable problem dimensions.

C. Synthetic data generation

Following other work on functional regression (Parodi et al., 2018), we draw each functional feature $\mathcal{X}_j$ and each non-zero regression coefficient curve $\mathcal{B}_j$ from a zero-mean Gaussian process with a Matern covariance function (Cressie & Huang, 1999) of the form

$$C(t,s) = \frac{\eta^2}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{2\nu}}{l} |t-s| \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{l} |t-s| \right),$$

where K_ν is a modified Bessel function. We employ point-wise variance $\eta^2 = 1$, range $l = 0.25$, and smoothness parameter $\nu = 3.5$ for both the $\mathcal{X}$'s and the $\mathcal{B}$'s. Before running `FSFC`, we standardize each feature individually as $(\mathcal{X}_j - \text{mean}(t))/\text{sd}(t)$, where $\text{mean}(t)$ and $\text{sd}(t)$ are the point-wise average and standard deviation of all instances computed at t. Appendix Figure E2 shows some instances of $\mathcal{X}$ and $\mathcal{B}$ for one specific scenario. The categorical response Y is generated according to the logistic regression procedure: each Y_i is drawn from a Bernoulli distribution on $\{-1, 1\}$ with probability $\hat{p}_i = 1/\left(1 + \exp\left(X_{(i)}\hat{B}\right)\right)$.

D. SHARE experiment settings

In *Experiment 2*, the cardinalities of the "affected" and "healthy" subject groups, denoted as n_a and n_h, are as follows: $n_a = 89, 102, 313, 419$ and $n_h = 1235, 1160, 747, 577$ for diabetes, myocardial infarction, high cholesterol, and hypertension, respectively. To prevent unbalanced scenarios, for each replication we randomly select n_a subjects from the healthy pool, resulting in a total of $n = 2n_a$ observations evenly balanced between the two classes. The observations are then divided into training and test sets, with $n_{test} = n/4$.

In *Experiment 1*, each replication uses a training set of only 20 random subjects. All the other observations are part of the testing set.

E. Additional figures and tables

Table E1. Simulation results for the integration of Feature Extraction with LSTM (`feLSTM`) and SVM (`feSVM`). For each curve, the following 12 attributes have been extracted: *mean, variance, median, maximum, minimum, skewness, kurtosis, zero crossing rate, spectral entropy, spectral kurtosis, area under the curve,* and *autocorrelation*. The table reports the average classification test accuracy and CPU processing time (in seconds) for 50 instances in each scenario.

		Accuracy test		**CPU time (s)**	
	p_0	`feLSTM`	`feSVM`	`feLSTM`	`feSVM`
$n = 300$	2	0.535	0.495	67.7	52.6
$p = 800$	5	0.532	0.504	67.8	52.9
	10	0.538	0.502	67.7	52.6
	20	0.539	0.514	67.8	52.6
$n = 600$	2	0.526	0.504	298.1	271.2
$p = 2000$	5	0.526	0.509	298.1	273.5
	10	0.531	0.504	298.2	271.2
	20	0.530	0.505	297.9	271.3

Table E2. Additional metrics and competitors. The table reports the averages of two classification metrics: the average test accuracy, and the test AUC average score. Additionally, we investigate two new competitors: *Random Forests* (RF) and Multilayer Perceptron networks (MLP). The experiment also includes the approaches that leverage FSFC reduced-spaced output as a pre-processing step, denoted as r−RF and r−MLP, respectively. The analysis focuses on the scenario with $n = 300$, $p = 800$, and $p_0 = 2, 5, 10, 20$ with hyperparameters detailed in Appendix Section B.

	$p_0 = 2$		$p_0 = 5$		$p_0 = 10$		$p_0 = 20$	
	accuracy	AUC	accuracy	AUC	accuracy	AUC	accuracy	AUC
FSFC	0.958	0.993	0.949	0.991	0.889	0.961	0.814	0.906
SVM	0.540	0.549	0.554	0.601	0.558	0.592	0.548	0.534
r−SVM	0.899	0.971	0.856	0.943	0.790	0.887	0.750	0.841
LSTM	0.491	0.491	0.499	0.504	0.488	0.489	0.498	0.506
r−LSTM	0.893	0.947	0.850	0.922	0.727	0.806	0.669	0.724
RF	0.828	0.914	0.741	0.833	0.666	0.732	0.614	0.658
r−RF	0.860	0.943	0.806	0.897	0.722	0.800	0.664	0.724
MLP	0.566	0.590	0.577	0.603	0.580	0.606	0.576	0.615
r−MLP	0.901	0.971	0.845	0.933	0.748	0.838	0.693	0.767

Table E3. Varying α and k. The table presents the average FSFC performance over 20 replications in terms of classification (accuracy and AUC) and selection (recall and precision) metrics for various values of α (0.2, 0.5, 0.8) and k (3, 5, 10), with $\alpha = 0.2$ and $k = 5$ being the settings investigated in the main text. The analysis focuses on the scenario with $n = 300$, $p = 800$, and $p_0 = 2, 5, 10, 20$.

Different α, k = 5

	$p_0 = 2$		$p_0 = 5$		$p_0 = 10$		$p_0 = 20$	
	accuracy	AUC	accuracy	AUC	accuracy	AUC	accuracy	AUC
α=0.2	0.958	0.993	0.949	0.991	0.889	0.961	0.814	0.906
α=0.5	0.956	0.976	0.954	0.993	0.895	0.969	0.819	0.910
α=0.8	0.942	0.952	0.949	0.982	0.896	0.969	0.820	0.909

	recall	prec	recall	prec	recall	prec	recall	prec
α=0.2	0.875	0.858	0.910	0.507	0.845	0.431	0.702	0.365
α=0.5	0.844	1.00	0.911	0.799	0.842	0.604	0.700	0.439
α=0.8	0.773	1.00	0.911	0.876	0.847	0.651	0.690	0.454

Different k, α = 0.2

	$p_0 = 2$		$p_0 = 5$		$p_0 = 10$		$p_0 = 20$	
	accuracy	AUC	accuracy	AUC	accuracy	AUC	accuracy	AUC
k=3	0.938	0.986	0.948	0.989	0.890	0.962	0.814	0.908
k=5	0.958	0.993	0.949	0.991	0.889	0.961	0.814	0.906
k=10	0.958	0.991	0.946	0.991	0.892	0.963	0.818	0.907

	recall	prec	recall	prec	recall	prec	recall	prec
k=3	0.850	0.752	0.910	0.416	0.845	0.370	0.718	0.326
k=5	0.875	0.858	0.910	0.507	0.845	0.431	0.702	0.365
k=10	0.921	0.886	0.910	0.623	0.845	0.478	0.690	0.417

Table E4. Varying noise levels. The table shows the average FSFC performance over 20 replications in terms of classification (accuracy and AUC) and selection (recall and precision) metrics for $k = 5, 10$. Data generation follows the procedure outlined in Appendix Section C but with range parameter $l = 0.1, 0.05$ and smoothness parameter $\nu = 2.5, 1.5$, resulting in rougher and noisier curves. Appendix Figure E1 depicts a sample of 10 curves for the first feature under each scenario. The analysis focuses on the scenario with $n = 300$, $p = 800$, and $p_0 = 2, 5, 10, 20$ with hyperparameters detailed in Appendix B.

$\boldsymbol{l = 0.1, \nu = 2.5}$

	$p_0 = 2$		$p_0 = 5$		$p_0 = 10$		$p_0 = 20$	
	accuracy	AUC	accuracy	AUC	accuracy	AUC	accuracy	AUC
k=5	0.948	0.991	0.928	0.986	0.872	0.953	0.767	0.863
k=10	0.956	0.992	0.927	0.983	0.876	0.950	0.749	0.848

	recall	prec	recall	prec	recall	prec	recall	prec
k=5	0.925	0.596	0.920	0.436	0.850	0.358	0.656	0.341
k=10	0.950	0.787	0.905	0.578	0.826	0.492	0.622	0.388

$\boldsymbol{l = 0.05, \nu = 1.5}$

	$p_0 = 2$		$p_0 = 5$		$p_0 = 10$		$p_0 = 20$	
	accuracy	AUC	accuracy	AUC	accuracy	AUC	accuracy	AUC
k=5	0.906	0.973	0.898	0.965	0.825	0.913	0.748	0.837
k=10	0.923	0.980	0.897	0.970	0.821	0.920	0.739	0.825

	recall	prec	recall	prec	recall	prec	recall	prec
k=5	0.950	0.551	0.940	0.364	0.826	0.305	0.606	0.319
k=10	0.925	0.803	0.916	0.485	0.788	0.403	0.560	0.339

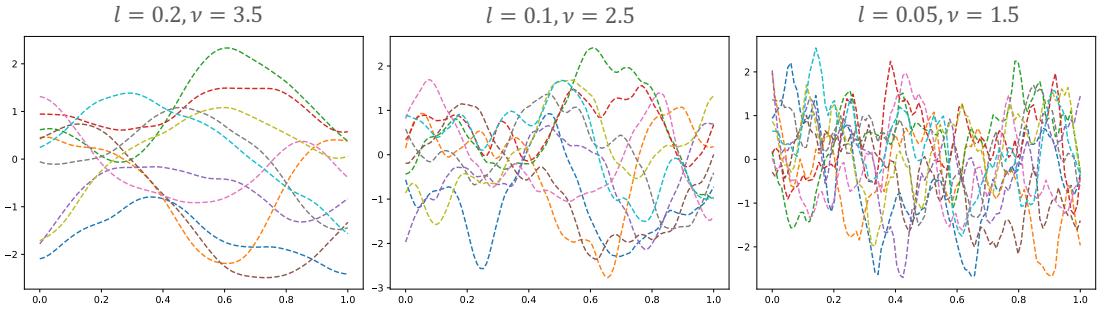

Figure E1. Varying noise levels. The figure depicts a sample of 10 curves for the first feature across three distinct noise scenarios, differentiated by range parameter $l = 0.2, 0.1, 0.05$ and smoothness parameter $\nu = 3.5, 2.5, 1.5$. The main text experiments use the parameters $l = 0.2$ and $\nu = 3.5$.

Table E5. List of the variables analyzed within the SHARE application. The letter adjacent to the variable name denotes whether it is *longitudinal* (l), *scalar* (s), or *categorical* (c). The letter (a) denotes a scalar variable obtained taking an average across the waves where the values were available. For more detailed information, the reader should consult the SHARE project website: https://share-eric.eu/

Variable	**Short description**
CASP (l)	Quality of life index
doctor visits (l)	Number of doctor visits within the past year
recall test (l)	Number of words recalled in the first trial
bmi (l)	Body mass index
adlwa (l)	Activities of daily living index
adla (l)	Sum of five daily activities
lgmuscle (l)	Large muscle index
mobilityind (l)	Mobility index
grossmotor (l)	Grossmotor skills index
finemotor (l)	Finemotor skills index
income perception (l)	Household able to make ends meet
eurod (l)	Depression index with EURO-D scale
income pct (l)	Household income percentiles
heart attack (l)	1 if the subject ever had the disease, 0 o.w.
high cholesterol (l)	1 if the subject ever had the disease, 0 o.w.
stroke (l)	1 if the subject ever had the disease, 0 o.w.
diabetes (l)	1 if the subject ever had the disease, 0 o.w.
COPD (l)	1 if the subject ever had the disease, 0 o.w.
cancer (l)	1 if the subject ever had the disease, 0 o.w.
ulcer (l)	1 if the subject ever had the disease, 0 o.w.
parkinson (l)	1 if the subject ever had the disease, 0 o.w.
cataracts (l)	1 if the subject ever had the disease, 0 o.w.
hip fracture (l)	1 if the subject ever had the disease, 0 o.w.
age (l)	Age of the subject
education years (s)	Years of education
number of children (a)	Number of children that are still alive
numeracy test 1 (a)	Mathematical performance
numeracy test 2 (a)	Mathematical performance
drinking behavior (a)	Times a patient drunk in the last 6 months
hospitalization days (a)	Days spent at the hospital in the last 6 months
number of hospitalization (a)	Number of hospitalizations in the last 6 months
gender (c)	Female or male
vaccinated (c)	Being vaccinated during childhood
ever smoked (c)	Ever smoked daily

Table E6. SHARE results for the integration of Feature Extraction with LSTM (feLSTM) and SVM (feSVM). For each curve, the extracted attributes are the ones listed in Table E1. The table reports the average classification test accuracy for 100 instances for each response.

		feLSTM	feSVM
Experiment 1	Diabetes	0.603	0.731
	Heart Attack	0.612	0.732
	High Colesterol	0.556	0.659
	Hypertension	0.614	0.703
Experiment 2	Diabetes	0.652	0.979
	Heart Attack	0.632	0.974
	High Colesterol	0.596	0.985
	Hypertension	0.675	0.987

Table E7. Varying smoothness in SHARE data analysis. The table replicates the SHARE analysis from the main text, exploring three distinct smoothness levels: unsmoothed data ($s = 0$), data smoothed according to a GCV-criterion optimal level ($s = s^{\text{opt}}$) utilized in the main analysis, and smoothing parameter $s = 10^4$, exceeding s^{opt} by at least an order of magnitude. For each disease and smoothing level, we present the top five variables based on the highest average λ_{max}-ratio (indicated in parentheses) over 100 replications. The higher the ratio, the earlier the feature is included in the model during the λ path search.

DIABETES

$s = 0$	$s = s^{\text{opt}}$	$s = 10^4$
hypertension (0.87)	hypertension (0.87)	hypertension (0.86)
bmi (0.79)	bmi (0.80)	bmi (0.80)
high cholesterol (0.79)	high cholesterol (0.79)	high cholesterol (0.79)
doctor visits (0.45)	doctor visits (0.46)	doctor visits (0.52)
gender (0.44)	gender (0.45)	gender (0.45)

HEART ATTACK

$s = 0$	$s = s^{\text{opt}}$	$s = 10^4$
hypertension (0.80)	hypertension (0.80)	mobility index (0.80)
high cholesterol (0.76)	high cholesterol (0.78)	hypertension (0.76)
mobility index (0.74)	mobility index (0.75)	high cholesterol (0.75)
doctor visits (0.66)	doctor visits (0.64)	doctor visits (0.71)
age (0.58)	age (0.59)	age (0.55)

HIGH CHOLESTEROL

$s = 0$	$s = s^{\text{opt}}$	$s = 10^4$
hypertension (0.97)	hypertension (0.97)	hypertension (0.97)
diabetes (0.71)	diabetes (0.71)	diabetes (0.71)
heart attack (0.58)	heart attack (0.58)	heart attack (0.58)
lgmuscle (0.36)	lgmuscle (0.39)	lgmuscle (0.43)
casp (0.32)	age (0.45)	casp (0.34)

HYPERTENSION

$s = 0$	$s = s^{\text{opt}}$	$s = 10^4$
bmi (0.97)	bmi (0.97)	bmi (0.97)
high cholesterol (0.68)	high cholesterol (0.68)	high cholesterol (0.66)
age (0.64)	age (0.64)	age (0.62)
diabetes (0.59)	diabetes (0.57)	doctor visits (0.56)
heart attack (0.54)	heart attack (0.54)	diabetes (0.55)

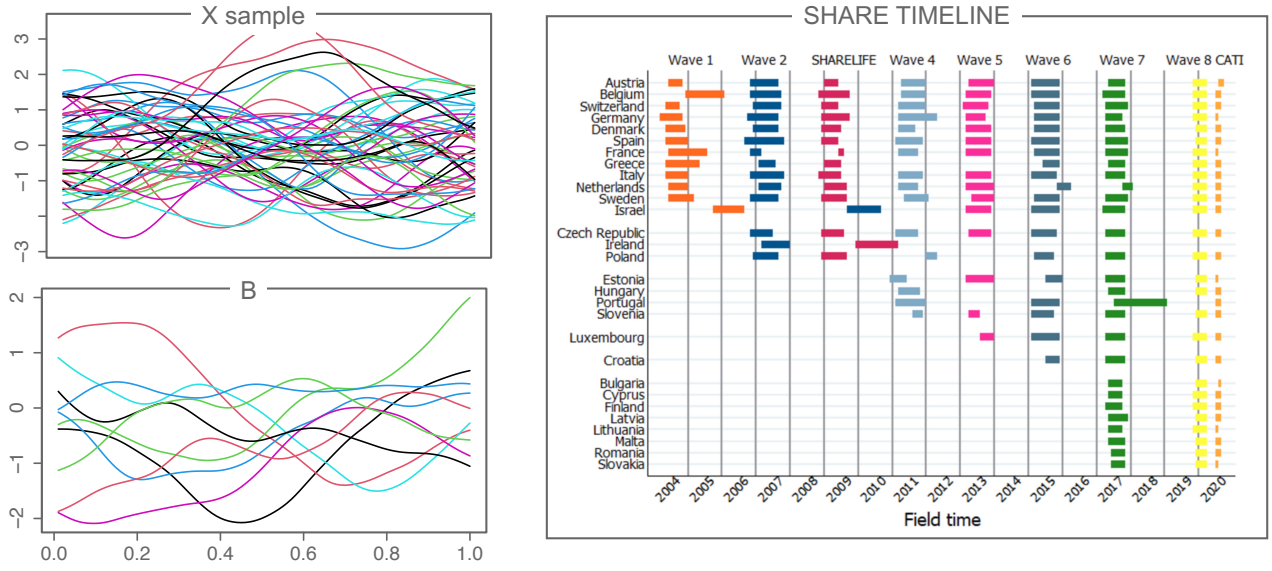

Figure E2. On the left, a sample of 50 curves from the first feature of the design matrix $\mathcal{X}$ (top) and the 10 non-zero $\mathcal{B}$ coefficients (bottom) are displayed for the given scenario with $n = 300$, $p = 800$, $p_0 = 10$. On the right, the SHARE project timeline is depicted, which has been sourced from the SHARE website: https://share-eric.eu/data/data-documentation/waves-overview/