Modeling Longitudinal Data Using Matrix Completion

Łukasz Kidziński and Trevor Hastie

Department of Bioengineering, Stanford University, Stanford, CA; Department of Statistics, Stanford University, Stanford, CA

ABSTRACT
In clinical practice and biomedical research, measurements are often collected sparsely and irregularly in time, while the data acquisition is expensive and inconvenient. Examples include measurements of spine bone mineral density, cancer growth through mammography or biopsy, assessment of gait in patients with neurological disorders, or assessment of gait in patients with neurological disorders. Practitioners often need to infer the progression of diseases from such sparse observations. A classical tool for analyzing such data is a mixed-effect model where time is treated as both a fixed effect (population progression curve) and a random effect (individual variability). Alternatively, researchers use Gaussian processes or functional data analysis, assuming that observations are drawn from a certain distribution of processes. While these models are flexible, they rely on probabilistic assumptions, require very careful implementation, and tend to be slow in practice. In this study, we propose an alternative elementary framework for analyzing longitudinal data motivated by matrix completion. Our method yields estimates of progression curves by iterative application of the Singular Value Decomposition. Our framework covers multivariate longitudinal data, and regression and can be easily extended to other settings. As it relies on existing tools for matrix algebra, it is efficient and easy to implement. We apply our methods to understand trends of progression of motor impairment in children with Cerebral Palsy. Our model approximates individual progression curves and explains 30% of the variability. Low-rank representation of progression trends enables identification of different progression trends in subtypes of Cerebral Palsy.

1. Motivation
The key question in medical practice and clinical research is how diseases progress in individual patients. Accurate continuous monitoring of a patient’s condition could considerably improve prevention and treatment. Many medical tests, such as X-ray, magnetic resonance imaging (MRI), motion capture gait analysis, biopsy, or blood tests, cannot be taken routinely, since they are costly, harmful, or inconvenient. Therefore, practitioners and researchers need to reach out for statistical tools to analyze progression patterns, which then could be used as covariates for between-subject similarities, we can model individual patient’s life. By looking at individual processes and accounting for between-subject similarities, we can model individual progression curves despite observing only few measurements. Such modeling can yield personalized predictions, clustering of patients based on progression patterns, or latent representation of progression patterns, which then could be used as covariates in other predictive models.

For illustration, consider longitudinal measurements of the gait deviation index (GDI), which is a holistic measure of motor impairment in children. GDI is measured using advanced motion capture hardware and software (Figure 1). Due to high costs, such measurements are taken only a few times in a patient’s life. By looking at individual processes and accounting for between-subject similarities, we can model individual progressions despite observing only few measurements. Such modeling can yield personalized predictions, clustering of patients based on progression patterns, or latent representation of progression patterns, which then could be used as covariates in other predictive models.

These kinds of data have been commonly analyzed using linear mixed models, where we treat time as a random effect, and the nonlinearity of this effect is imposed by choice of the functional basis (Zeger, Liang, and Albert 1988; Verbeke 1997; McCulloch and Neuhaus 2001). When data is very sparse, additional constraints on the covariance structure of trajectories are inferred using cross-validation or information criteria (Rice and Wu 2001; Bigelow and Dunson 2009). To further constrain the dimensionality of the problem, practitioners model the covariance as a low-rank matrix (Berkey and Kent 1983; Besse and Ramsay 1986; James, Hastie, and Sugar 2000; Hall, Müller, and Wang 2006; Yao and Lee 2006; Greven et al. 2011; Yan, Lin, and Huang 2017). Many models have been developed for incorporating additional covariates (Song, Davidian, and Tsiatis 2002; Liu and Huang 2009; Rizopoulos et al. 2014). While these methods are used in practice, they tend to be slow and require careful implementation.

In this work, we pose the problem of trajectory prediction as a matrix completion problem, and we solve it using sparse matrix factorization techniques (Rennie and Srebro 2005; Candès and Recht 2009; Fithian and Mazumder 2018). The objective of the classical matrix completion problem is to predict elements of a sparsely observed matrix using its known elements while minimizing a specific criterion, commonly Mean Squared Error (MSE). The motivating example is the “Netflix Prize” competition (Bennett and Lanning 2007), where teams were tasked to predict unknown movie ratings using a sparse set of observed ratings. We can represent these data as a matrix of N users and M movies, with a subset of known elements measured on a fixed scale, for example, 1–5.
One popular approach to approximate the observed matrix is to estimate its low-rank decomposition (Srebro, Alon, and Jaakkola 2005). In the low-rank representation $W A'$, columns of $A$ spanning the space of movies can be loosely associated with some implicit (latent) characteristics such as taste, style, or genre, and each rater is represented as a weighted sum of their preferred characteristics, that is, a row in the matrix of latent variables $W$ (see Figure 2).

We can use the same idea to predict sparsely sampled curves as long as we introduce an additional smoothing step. The low-dimensional latent structure now corresponds to progression patterns, and the trajectory of each individual can be represented as a weighted sum of these “principal” patterns. In Figure 2, the patterns are given by $A' B'$, while the individual weights are encoded in $W$.

The key advantage of this approach is that the implementation only relies on the iterative application of matrix multiplication and singular value decomposition. We can leverage existing literature and packages for these operations, making the implementation straightforward and very efficient.

To the best of our knowledge, this is the first application of matrix completion for filling in entire curves to sparsely observed data. Nevertheless, matrix completion techniques have been previously used in this context for estimating the covariance operator Descary and Panaretos (2016), which can subsequently be used for fitting curves.

The article is organized as follows. In Section 2, we introduce the notation of the functional representation of the problem and we discuss optimization techniques. Next, in Section 3, we introduce the matrix representation of the problem and show how it maps to the functional representation. We develop our main results and conclude with links to other results. In Section 4, we extend our framework to the multivariate case. We show example uses of our methods in a simulation study (Section 5) and in a data study (Section 6). We conclude with a discussion section (Section 7).

2. Background and Related Work

To focus our attention and aid further discussion, we start by introducing notation and methodology for the univariate case. Let $N$ denote the number of subjects. For each individual $i \in \{1,2,\ldots,N\}$, we take measurements at $n_i$ irregularly sampled time-points $t_i = [t_{i,1},t_{i,2},\ldots,t_{i,n_i}]$. We assume that $t_{i,1} < t_{i,2} < \cdots < t_{i,n_i} < t_{\text{max}}$ for each $i$ and some $t_{\text{min}}, t_{\text{max}} \in \mathbb{R}$. In this work, we ignore the stochastic nature of sampling timepoints and we assume that points $t_i$ are fixed for each individual. Vectors $y_i = [y_{i,1},\ldots,y_{i,n_i}]$ denote observations corresponding to $t_i$ for each individual $i$.

To model individual trajectories given pairs $(t,y)$ practitioners map observations into a low-dimensional space that represents progression patterns. Conceptually, a small distance between individuals in the latent space reflects similar progression patterns.

In this section, we discuss state-of-the-art approaches to estimating this low-dimensional latent embedding. We classify them into three broad categories: the direct approach, mixed-effect models, and low-rank approximations.

2.1. Direct Approach

If the set of observed values for each individual is dense, elementary interpolation using a continuous basis can be sufficient for approximating the entire trajectory. Let $\{b_i : i \in \mathbb{N}\}$ be
a basis of $L^2([t_{\min}, t_{\max}])$, that is the space of functions $f : [t_{\min}, t_{\max}] \rightarrow \mathbb{R}$ such that the integral of $f^2$ on $[t_{\min}, t_{\max}]$ is finite, such as splines or polynomials. In practice, we truncate the basis to a finite set of the first $K \in \mathbb{N}$ basis elements. Let $\mathbf{b}(t) = [b_1(t), b_2(t), \ldots, b_K(t)]'$ be a vector of $K$ basis elements evaluated at a timepoint $t \in (t_{\min}, t_{\max})$. Throughout this article we use the word basis to refer to some truncated basis of $K$ elements. The choice of the basis is relevant and should be done carefully. We refer the reader to the vast literature on functional data analysis for recommendations on this topic (Ramsey 2006). Throughout this article, we assume the basis is fixed upfront.

To find an individual trajectory, for each subject $i \in \{1, \ldots, N\}$, we might use least squares and estimate a set of coefficients $\mathbf{w}_i \in \mathbb{R}^K$, minimizing squared Euclidean distance to observed points

$$\arg\min_{\mathbf{w}_i} \sum_{j=1}^{n_i} |y_{ij} - \mathbf{w}_i'b(t_{ij})|^2. \quad (1)$$

This direct approach has two main drawbacks. First, it ignores correlations within the curves, which could potentially improve the fit. Second, if the number of observations $n_i$ for an individual $i$ is smaller or equal to the size of the basis $K$, we can fit a curve with no error leading to overfitting and an unreliable estimator of the variance.

To remedy the first issue, it is common to estimate the covariance operator, compute principal functional components across all individuals and represent curves in the space spanned by the first few of them. Such representation, is referred to as a Karhunen-Loève expansion (Watanabe 1965; Kosambi 2016), has become a foundation of many functional data analysis workflows (Laird 1988; Ramsay and Dalzell 1991; Besse, Cardot, and Ferraty 1997; Cnaan, Laird, and Slasor 1997; Yao, Müller, and Wang 2005; Horváth and Kokoszka 2012).

A basic idea to remedy the second issue is to estimate both the basis coefficients and the covariance structure simultaneously. Linear mixed-effect models provide a convenient solution to this problem.

### 2.2. Linear Mixed-Effect Models

A common approach to modeling longitudinal observation $(t_i, y_i)$ is to assume that data come from a linear mixed-effect model (LMM) (Zeger, Liang, and Albert 1988; Verbeke 1997). We operate in a functionspace with a basis $\mathbf{b}(t)$ of $K$ elements and we assume there exists a fixed effect $\mu(t) = m'(\mathbf{b}(t))$, where $m = [m_1, \ldots, m_K]'$ for $m_k \in \mathbb{R}$ for $1 \leq k \leq K$. We model the individual random effect as a vector of basis coefficients. In the simplest form, we assume

$$\mathbf{w}_i \sim \mathcal{N}(0, \Sigma), \quad (2)$$

where $\Sigma$ is a $K \times K$ covariance matrix. We model individual observations as

$$y_i | \mathbf{w}_i \sim \mathcal{N}(\mu_i + B_i \mathbf{w}_i, \sigma^2 I_{n_i}), \quad (3)$$

where $\mu_i = [\mu(t_{i,1}), \mu(t_{i,2}), \ldots, \mu(t_{i,n_i})]'$, $\sigma$ is the standard deviation of observation error and $B_i = [\mathbf{b}(t_{i,1}), \ldots, \mathbf{b}(t_{i,n_i})]'$ is the basis evaluated at timepoints defined in $t_i$. Estimation of the model parameters is typically accomplished by the expectation-maximization (EM) algorithm (Laird and Ware 1982). For predicting coefficients $\mathbf{w}_i$ one can use the best unbiased linear predictor (BLUP) (Henderson 1950; Robinson 1991).

Since the LMM estimates the covariance structure and the individual fit simultaneously, it reduces the problem of overfitting of $\mathbf{w}_i$, present in the direct approach. However, this model is only applicable if we observe a relatively large number of observations per subject since we attempt to estimate $K$ coefficients for every subject.

To model trajectories from a small number of observations, practitioners further constrain the covariance structure. If we knew the functions which contribute the most to the random effect, we could fit an LMM in a smaller space spanned by these functions. We explore possibilities to learn the basis from the data using low-rank approximations of the covariance matrix.

### 2.3. Low-Rank Approximations

There are multiple ways to constrain the covariance structure. We can use cross-validation or information criteria to choose the best basis, the number of elements, or the positions of spline knots (Rice and Wu 2001; Bigelow and Dunson 2009). Alternatively, we can place a prior on the covariance matrix (MacLehose and Dunson 2009).

Another solution is to restrict the latent space to $q < K$ dimensions and learn from the data a mapping $A \in \mathbb{R}^{K \times q}$ between the latent space and the basis. In the simplest scenario with Gaussian noise, observations can be modeled as

$$y_i | \mathbf{w}_i \sim \mathcal{N}(\mu_i + B_i A \mathbf{w}_i, \sigma^2 I_{n_i}), \quad (4)$$

following the notation from (3).

James, Hastie, and Sugar (2000) propose an EM algorithm for finding model parameters and predicting latent variables $\mathbf{w}_i \in \mathbb{R}^q$ in (4). In the expectation stage, they compute the conditional mean of $\mathbf{w}_i$ given $y_i$ and the model parameters, while in the maximization stage, with $\mathbf{w}_i$ assumed observed, they maximize the likelihood with respect to $[\mu, A, \sigma]$. The likelihood, given $\mathbf{w}_i$, takes the form

$$\prod_{i=1}^{N} \frac{1}{(2\pi)^{n_i/2} \sigma^{n_i/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2} - \frac{1}{2} \mathbf{w}_i' \Sigma^{-1} \mathbf{w}_i\right\}.$$

Another approach to estimating parameters of (4) is to optimize over $\mathbf{w}_i$ and marginalize $A$ (Lawrence 2004). This approach allows modifying the distance measure in the latent space, using the kernel trick (Schulam and Arora 2016).

Estimation of the covariance structure of processes is central to the estimation of individual trajectories. Descary and Panaretos (2016) propose a method where the estimate of the covariance matrix is obtained through matrix completion.

Methods based on low-rank approximations are widely adopted and applied in practice (Berke and Kent 1993; Besse and Ramsay 1986; Hall, Müller, and Wang 2006; Yao and Lee 2006; Greven et al. 2011; Yan, Lin, and Huang 2017). However, due to their probabilistic formulation and reliance on the distribution assumptions, these models usually need...
to be carefully fine-tuned for specific situations and existing implementations tend to be slow. This shortcoming motivates us to develop an elementary optimization framework, using existing, extensively studied, and well-optimized tools for matrix algebra.

3. Modeling Sparse Longitudinal Processes Using Matrix Completion

We first introduce the methodology for univariate sparsely-sampled processes. The direct method, mixed-effect models, and low-rank approximations described in Section 2 have their counterparts in the matrix completion setting. We discuss these analogies in Sections 3.2 and 3.3. Next, in Section 4, we show that the simple representation of the problem allows for extension to multivariate sparsely-sampled processes and a regression setting.

3.1. Notation

For each individual \( i \in \{1, 2, \ldots, N\} \) we observe \( n_i \) measurements \( \{\hat{y}_{i1}, \ldots, \hat{y}_{in_i}\} \) at time-points \( \{t_{i1}, t_{i2}, \ldots, t_{in_i}\} \). Unlike in the prior work introduced in Section 2 operating on entire curves, here we discretize time to \( T \) time-points \( G = \{t_1, t_2, \ldots, t_T\} \), where \( t_{\min} = t_1, t_{\max} = t_T \) and \( T \) is arbitrarily large. Each individual \( i \) is expressed as a partially observed vector \( r_i \in \mathbb{R}^T \). For each time-point \( t_{ij} \) for \( 1 \leq j \leq n_i \) we find a corresponding grid-point \( g_j = \arg\min_{k \leq T} |t_j - t_{ij}| \). We assign \( y_{ij} = \hat{y}_{ij} \). Let \( O_i = \{g_j : 1 \leq j \leq n_i\} \) be a set of grid indices corresponding to observed grid points for an individual \( i \). All elements outside \( O_i \), that is \( \{y_{ij} : j \notin O_i\} \) are considered missing.

For \( T \) sufficiently large, our observations can be uniquely represented as a \( N \times T \) matrix \( Y \) with missing values. We denote the set of all observed elements by pairs of indices as \( \Omega = \{(i,j) : i \in \{1, 2, \ldots, N\}, j \notin O_i\} \). Let \( Y_i \) be the projection onto observed indices, that is \( P_{\Omega_i}(Y) = \breve{Y}_i \), such that \( \breve{Y}_{ij} = Y_{ij} \) for \( (i,j) \in \Omega \) and \( \breve{Y}_{ij} = 0 \) otherwise. We define \( Y_{\Omega}(Y) \) as the projection on the complement of \( \Omega \), that is \( P_{\Omega}^\Omega(Y) = Y - P_{\Omega_i}(Y) \).

We use \( \| \cdot \|_F \) to denote the Frobenius norm, that is the square root of the sum of squares of matrix elements, and \( \| \cdot \|_* \) to denote the nuclear norm, that is the sum of singular values.

As in Section 2 we imposed smoothness by using a continuous basis \( b(t) = [b_1(t), b_2(t), \ldots, b_K(t)]' \). When we evaluate the basis on a grid \( G \) we get a \( T \times K \) matrix \( B = [b(t_1), b(t_2), \ldots, b(t_T)]' \).

In our algorithms, we use diagonal thresholding operators defined as follows. Let \( D = \text{diag}(d_1, \ldots, d_p) \) be a diagonal matrix. We define soft-thresholding as

\[
D_\lambda = \text{diag}(\{d_1 - \lambda, (d_2 - \lambda)_{+}, \ldots, (d_p - \lambda)_{+}\}),
\]

where \((x)_{+} = \max(x, 0)\), and hard-thresholding as

\[
D_\lambda^H = \text{diag}(\{d_1, d_2, \ldots, d_q, 0, \ldots, 0\}),
\]

where \( q = \arg\min_k(d_k < \lambda) \).

3.2. Direct Approach

The optimization problem (1) of the direct approach described in Section 2.1 can be approximated in the matrix completion setting. Note that the bias introduced by the grid is negligible if the grid is sufficiently large, that is if each observed timepoint is relatively close to some grid point. We defined the basis as a set of continuous functions, so the finer the grid the closer we are to the actual value of the basis at a given timepoint. Eventually, the difference becomes negligible. Similarly, for the observed points, the coordinate corresponding to time is more accurate when the grid is finer, but again, due to continuity assumption this difference becomes negligible. The analysis of the properties of the grid is beyond the scope of the article. As a rule of thumb, we recommend choosing the grid based on a practical indication. For example, if we expect patients to visit a clinic 1–2 times per year, then a grid corresponding to 12 points per year might be sufficient.

We approximate each observation \( y_{ij} \) with a point on the selected grid \( \breve{y}_{ij} \). Next, we rewrite the optimization problem (1) as a matrix completion problem

\[
\arg\min_{[w_i]} \sum_{i=1}^N \sum_{j=1}^{n_i} \left| y_{ij} - w_{ij} b(g_j) \right|^2 = \arg\min_{[w_i]} \sum_{(i,k) \in \Omega} \left| y_{ik} - w_{ik} b(k) \right|^2 = \arg\min_{W} \| P_{\Omega}(Y - WB') \|_F^2,
\]

where by optimization over \([w_i]\) we mean optimization over all \([w_i] : i \in \{1, 2, \ldots, N\}\) and \( W \) is an \( N \times K \) matrix composed of vectors \([w_i]'\) stacked vertically.

The matrix formulation in (5) and the classic approach in Section 2.1 share multiple characteristics. In both cases, if data is dense, we may find an accurate representation of the underlying process simply by fitting least-squares estimates of \( W \) or \([w_i]\). Conversely, if the data is too sparse, the problem becomes ill-posed, and the least-squares estimates can overfit.

However, the representational difference between functional representation (1) and matrix representation (5) leads to different computational approaches. The matrix representation enables us to use the matrix completion framework and, in particular, in Section 3.4 we introduce convex optimization algorithms for solving (5).

Some low-rank constraints on the random effect from the mixed-effect model introduced in Section 2.2 can be expressed in terms of constraints on \( W \). In particular, in Section 3.3 we show that the linear mixed-effect model can be expressed by constraining the rank(\( W \)).

3.3. Low-Rank Matrix Approximation

In the low-rank approach described in Section 2.3 we assume that individual trajectories can be represented in a low-dimensional space, by constraining the rank of \( W \).

We use a similar approach in solving (5). One difficulty comes from the fact that optimization with a rank constraint on \( W \) turns the original least-squares problem into a non-convex problem. Motivated by the matrix completion literature, we relax
the rank constraint in (5) to a nuclear norm constraint and we attempt to solve
\[ \arg \min_W \| P_{12}(Y - WB')\|_F^2 + \lambda \| W \|_*, \] (6)
for some parameter \( \lambda > 0 \).

From the practical standpoint, low-rank representation corresponds to describing curves on a basis comprised of only a few functions which are the best functions for obtaining low reconstruction errors. In Figure 3 we illustrate two functions for describing processes introduced in Section 1 and Figure 1. Such low-dimensional representation of processes can also be used for clustering based on the type of their progression as we discuss in the data study (Section 6).

Cai, Candès, and Shen (2010) propose a first-order singular value thresholding algorithm (SVT), for solving a general class of problems involving a nuclear norm penalty and a linear form of \( Y \), which includes (6). Their algorithm can handle large datasets and has strong guarantees of convergence, but it requires tuning the stop size parameter, which can decrease computational efficiency as different steps have to be tried in practice. This limitation was addressed by Ma, Goldfarb, and Chen (2011), Mazumder, Hastie, and Tibshirani (2010), and Hastie et al. (2015) who introduced iterative procedures which do not depend on such tuning.

In the last decade, machine learning and statistical learning communities have introduced multiple algorithms for matrix completion. Many of them are suitable for solving (6). However, in this article, we focus on analyzing the benefits of framing the trajectory prediction problem (1) in the matrix completion framework, rather than on benchmarking possible solutions.

### 3.4. Low-Rank Approximation with Singular Value Thresholding

The low-rank probabilistic approach, introduced in Section 2.3, is based on the observation that the underlying processes for each subject can be approximated in a low-dimensional space. Here, we exploit the same characteristic using matrix-factorization techniques for solving the optimization problem (6).

For the sake of clarity and simplicity, we choose to solve the problem (6) with an extended version of the Soft-IMPUTE algorithm designed by Hastie et al. (2015) and Mazumder, Hastie, and Tibshirani (2010). We find it easy to implement because it only uses matrix multiplication and Singular Value Decomposition (SVD). Using standard matrix operations also reduces the computational cost of our method, since implementations of these operations leverage well-optimized algorithms. However, as discussed in Section 3.3, many other convex optimization algorithms can be applied.

The key component to the solution is the following property linking problem (6) with the SVD. Consider the fully observed case of (6). Using Theorem 2.1 in Cai, Candès, and Shen (2010), one can show that the optimization problem
\[ \arg \min_W \frac{1}{2} \| Y - WB'\|_F^2 + \lambda \| W \|_*, \] (7)
with \( B' = I \), has a unique solution \( W = S_k(YB) \), where \( S_k(X) = UD_kV' \) and \( X = UDV' \) is the SVD of \( X \), that is, \( U \) is an \( n \times n \) unitary matrix, \( D \) is an \( n \times K \) rectangular diagonal matrix, with nonnegative real numbers on the diagonal in decreasing order, while \( V \) is a \( K \times K \) matrix. We refer to \( S_k(X) \) as the singular value thresholding (SVT) of \( X \). Note that, the decomposition of \( YB \) is identifiable since the solution of SVD with constrained \( D \) is identifiable.

In order to solve (7) with a sparsely observed \( Y \), we modify the Soft-IMPUTE algorithm to account for the basis \( B \). Our Algorithm 1 iteratively constructs better approximations of the global solution for each \( \lambda \) in some predefined set \( \{\lambda_1, \lambda_2, \ldots, \lambda_k\} \) for \( k \in \mathbb{N} \). For a given approximation of the solution \( W^{\text{old}} \), we use \( W^{\text{old}}B' \) to impute unknown elements of \( Y \) obtaining \( \hat{Y} \). Then, we construct the next approximation \( W^{\text{new}} \) by computing SVT of \( \hat{Y} \).

As a stopping criterion, we compute the change between subsequent solutions, relative to the magnitude of the solution, following the methodology in Cai, Candès, and Shen (2010). We set a fixed threshold of \( \epsilon > 0 \) for this criterion, depending on the desired accuracy.

A common bottleneck of the algorithms introduced by Cai, Candès, and Shen (2010), Mazumder, Hastie, and Tibshirani (2010), and Ma, Goldfarb, and Chen (2011) as well as other SVT-based approaches is the computation of the SVD of large matrices. This problem is particularly severe in standard matrix completion settings, such as the Netflix problem, where the matrix size is over 400,000 \( \times \) 20,000. However, in our problem,
\[ \text{rank}(WB') \leq \text{rank}(B) = K \ll N, \] (8)
and \( K \sim 10 \) in our motivating data example. While algorithms for large matrices are applicable here, the property (8) makes the...
computation of SVD feasible in practice with generic packages such as PROPACK (Larsen 2004).

For estimating curves on new data, we need to find $w_t = V a$ minimizing the prediction error. To this end, we regress observed values $Y_t$ on $B_i V$ with a penalty $\frac{1}{2} \| a \|^2$ on the coefficients, that is we optimize

$$\min_a \| Y_t - B_i V a \|^2 + \frac{\lambda}{2} \| a \|^2,$$

where $B_i$ are basis functions evaluated at the corresponding timepoints.

While our algorithms for finding $W$ do not depend on the density of the grid, it is important to highlight that with the growing $T$ the norm of the solution $W_T$ grows to infinity and $W_T / \sqrt{T} \to W_0$, for some fixed $W_0$ (See Appendix A). In order to obtain comparable results for different $T$, parameters $\lambda$ need to be adjusted to this scale (e.g., by modifying the penalty to $\lambda \sqrt{T} \| W \|_a$).

The algorithm converges at a rate $\frac{1}{T}$, where $t$ is the number of iterations (this can be established by an elementary extension of the proofs in Mazumder, Hastie, and Tibshirani 2010) as we demonstrate in Appendix A. Each step of our algorithm requires matrix multiplications and an SVD. Thus, the complexity of each step is $O(\min(K n^2, n K^2))$. For EM-based algorithms, the minimization step requires multiple matrix multiplications and therefore we expect a larger constant in each step. For practical use, we report computing times on synthetic data in Section 5.

3.5. $l_0$ regularization

While the nuclear norm relaxation (6) is motivated by making the problem convex, Mazumder, Hastie, and Tibshirani (2010) argue that in many cases it can also give better results than the rank constraint. They draw an analogy to the relation between best-subset regression ($l_0$ regularization) and LASSO ($l_1$ regularization as in Tibshirani 1996; Friedman, Hastie, and Tibshirani 2001). In LASSO, by shrinking model parameters, we allow more parameters to be included, which can potentially improve predictions if the true subset is larger than the one derived through $l_0$ regularization. In the case of (6), shrinking the nuclear norm allows us to include more dimensions of $W$ again potentially improving the prediction if the true dimension is high.

In theory, estimating too many components can lead to problems if the true dimension is small. In such cases, shrinking may cause the inclusion of unnecessary dimensions emerging from the noise. To remedy this issue, following the analogy with penalized linear regression, we may consider another class of penalties. In particular, we may consider coming back to the rank constraint by modifying the nuclear norm penalty (6) to $\ell_0$. We define $\| W \|_0 = \text{rank}(W)$. The problem

$$\min_{W} \frac{1}{2} \| P_{\Omega} (Y - W^2 B^2) \|^2 + \lambda \| W \|_0,$$

has a solution $W = S^H \lambda (Y B)$, where $S^H \lambda (X) = U D \lambda V^T$ and $X = UD \lambda V^T$ is the SVD of $X$. We refer to $S^H \lambda (X)$ as the hard singular value thresholding (hard SVD) of $X$. We use Algorithm 2 to find the hard SVD of $Y B$.

Algorithm 2: Hard-Longitudinal-Impute

1. Initialize $W_0^{\text{old}}$ with solutions $W_{\lambda_i}$ from Soft-Longitudinal-Impute
2. For each $\lambda_1 > \lambda_2 > \cdots > \lambda_k$:
   (a) Repeat:
      i. Compute $W^{\text{new}} \leftarrow S_{\lambda_i} ((P_{\Omega} (Y) + P_{\Omega} (W^{\text{old}} B^2)) B)$
      ii. If $\| W^{\text{new}} - W^{\text{old}} \|^2 < \epsilon$ exit
      iii. Assign $W^{\text{old}} \leftarrow W^{\text{new}}$
   (b) Assign $W_{\lambda_i} \leftarrow W^{\text{new}}$
3. Output $W_{\lambda_1}, W_{\lambda_2}, \ldots, W_{\lambda_k}$

The problem (9) is non-convex, however, by starting from the solutions of Algorithm 1 we explore the space around the global minimum of the $l_1$ version of the problem. Ge, Lee, and Ma (2016) show that this strategy is successful empirically.

In practice, the continuous parameterization via $\lambda$ leads to similar results as rank-truncation methods for different ranks. There are two advantages of the former. First, it allows for a continuum of warm starts and, as such, it is a natural post-processor for Soft-Impute solutions. Second, it can be a basis for further generalizations with penalties between $\ell_0$ and $\ell_1$ (Mazumder, Hastie, and Tibshirani 2010). However, if the true rank can be estimated with high confidence and the signal is strong, then a rank-truncated method can lead to better predictive performance.

3.6. Dense Grid and Proximal Gradient Descent

Growing grid size parameter $T$ improves the accuracy of approximation of the data on the grid, but benefits in approximation are negligible when $T$ becomes sufficiently large. However, we show that $T \to \infty$ also leads to a continuous version of the algorithm, equivalent to the proximal gradient descent method applied to the initial formulation of the problem (1) with nuclear norm penalty.

We start by observing that in Algorithm 1 the Step 2(a)(i) depends on the grid and requires multiplication with all evaluations of the basis on that grid. We can rewrite the expression
in $S_q$, such that only a sparse number of evaluations of the basis is required. Indeed, we write

$$(P_\Omega(Y) + P_\Omega(W^{\text{old}}B'))B = (P_\Omega(Y) - P_\Omega(W^{\text{old}}B') + W^{\text{old}}B')B = (P_\Omega(Y - W^{\text{old}}B'))B + W^{\text{old}}$$

(10)

and to compute $W^{\text{old}}B'$ on $\Omega$ in (10) we only need $W^{\text{old}}$ and an evaluation of the basis on gridpoints corresponding to $\Omega$. Therefore, for computing $(P_\Omega(Y - W^{\text{old}}B'))B$ we also only need $B$ evaluated on gridpoints with a nonzero number of observations. Now, if we increase the number of grid points to infinity we derive a solution where the basis needs to be evaluated only at observed timepoints.

Since $B$ is orthogonal, when $T \to \infty$ the product $(P_\Omega(Y - W^{\text{old}}B'))B$ converges to 0 and the update in Step 2(a)(i) in Algorithm 1 also converges to 0 at the rate $1/T$. Therefore, in order to obtain comparable results for different $T$ we need to scale the update direction $(P_\Omega(Y - W^{\text{old}}B'))B$ with $\alpha/\sqrt{T}$, for some $\alpha > 0$. We refer to $\alpha$ as the step size.

We show that at the limit $T \to \infty$ our approach is equivalent to proximal gradient descent. Let us define the loss function as in (1), that is

$$f(W) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{n_i} |y_{ij} - w_i b_i(t_{ij})|^2,$$

where $t_{ij}$ are timepoints, $y_{ij}$ are observations. We optimize the loss function $f$ with a nuclear norm penalty for some $\lambda > 0$

$$\arg \min_W f(W) + \lambda \|W\|_n,$$  

(11)

where $W = [w_1, w_2, \ldots, w_N]'$. To solve problem (11) we define

$$G_i = \frac{\partial}{\partial w_i} f(W) = -2 \sum_{j=1}^{n_i} (y_{ij} - w_i b_i(t_{ij})) b_i'(t_{ij})$$

and the gradient of $f$ takes form

$$\nabla_W f(W) = [G_1', G_2', \ldots, G_N']'$$

Note that $\nabla_W f(W)$ is corresponding to $(P_\Omega(Y - W^{\text{old}}B'))B$ in (10) with infinitely dense grid. For the update step in the proximal gradient descent method, we use step size $\alpha > 0$ and singular value thresholding with a threshold $\alpha \lambda$ as recommended in Hastie, Tibshirani, and Wainwright (2015). We describe the full procedure in Algorithm 3.

Nesterov (2013) showed sufficient conditions for convergence of Algorithm 3. If $\alpha \in (0, 1/L]$ where $L$ is the Lipschitz constant of $\nabla_W f$, then the algorithm converges at the rate $1/t$, where $t$ is the number of steps.

### 3.7. A Link between the Reduced-Rank Model and Soft-Longitudinal-Impute

Intuitively, we might expect similarity between the principal directions derived using the probabilistic approach (3) and their counterparts derived from the SVT-based approach. We investigate this relation by analyzing the behavior of SVT for matrices sampled from the probabilistic model given by (3).

#### Algorithm 3 Grid-free soft-Longitudinal-Impute

1. Initialize $W^{\text{old}}$ with zeros
2. For each $\lambda_1 > \lambda_2 > \cdots > \lambda_k$:
   
   (a) Repeat:
      
      i. Compute $W^{\text{new}} := S_{\alpha \lambda}(W^{\text{old}} - \alpha \nabla_W f(W^{\text{old}}))$
      
      ii. If $\|W^{\text{new}} - W^{\text{old}}\|_F^2 < \varepsilon$ exit
      
      iii. Assign $W^{\text{old}} \leftarrow W^{\text{new}}$
   
   (b) Assign $\hat{W}_{\lambda_i} \leftarrow W^{\text{new}}$
3. Output $\hat{W}_{\lambda_1}, \hat{W}_{\lambda_2}, \ldots, \hat{W}_{\lambda_k}$

For simplicity, let us assume that $\mu = 0$ and the data is fully observed on a grid $G$ of $T$ time-points. Assume that observations $i \in \{1, 2, \ldots, N\}$ come from the mixed-effect model

$$y_i | w_i \sim N(Bw_i, \sigma^2 I_T),$$

$$w_i \sim N(0, \Sigma),$$

(12)

where $\Sigma$ is an unknown covariance matrix of rank $q < K$ and variables $[w_i, y_i]$ are independent. By the spectral decomposition theorem, we decompose $\Sigma = V \Lambda V'$, where $V$ is a $K \times K$ orthogonal and $\Lambda$ is a diagonal $K \times K$ matrix with $q$ positive coefficients in decreasing order. Since $y_i$ and $w_i$ are independent, the distribution (12) can be rewritten as

$$y_i \sim N(0, B\Lambda V' + \sigma^2 I_T).$$

The model (13) is a factor model with uncorrelated errors and $q$ factors—the first $q$ columns of $BV$.

The following theorem constitutes a link between the mixed-effect model and SVT (Theorem 9.4.1 in Mardia, Kent, and Bibby 1980).

**Theorem 1.** Let $Y = [y_1, \ldots, y_N]'$ be the observed matrix and let $S_{2q}(YB) = UD_{q^2}Q$. Then, $(D_{q^2}, Q)$ is the maximum likelihood estimator of $(\Lambda, V)$.

**Theorem 1** implies that in the fully observed case the estimator of $(\Lambda, V)$ converges to the true parameter at the rate $1/\sqrt{n}$, as the MLE.

Factor analysis methods give not only estimates of $\Lambda$ and $V$ but also predictors of the individual latent variables $W = [w_1, \ldots, w_N]'$. In the multivariate analysis literature, there are multiple ways to estimate factor scores, that is, a matrix $A$ such that $X \sim AD_{q^2}V'$. Most notably researchers introduce Spearman’s scores, Bartlett’s scores, and Thompson’s scores (Kim and Mueller 1978). Simply taking $W = U$ as the estimate of the scores corresponds to the solution of (7) as long as $\lambda = \sigma^2$.

In **Theorem 1** we assume that $\sigma^2$ is known, which is rarely the case in practice. However, the likelihood of $(V, \sigma)$ can be parametrized by $\sigma$, and we can find the optimal solution analytically. This corresponds to minimizing (7) for different $\lambda$.

A similar analogy is drawn between the classical principal component analysis and probabilistic principal component analysis by Tipping and Bishop (1999) and James, Hastie, and Sugar (2000).
4. Multivariate Longitudinal Data

In practice, we are often interested in the prediction of a univariate process in the context of other longitudinal measurements and covariates constant over time. Examples include the prediction of disease progression given patient’s demographics, data from clinical visits at which multiple blood tests or physical exams are taken, or measurements that are intrinsically multivariate such as gene expression or X-rays collected over time.

The growing interest in this setting stimulated research in latent models (Sammel and Ryan 1996) and multivariate longitudinal regression (Gray and Brookmeyer 1998, 2000). Diggle et al. (2002) present an example case study in which longitudinal multivariate modeling enables estimation of joint confidence region of multiple parameters changing over time, shrinking the individual confidence intervals.

In this section, we present an extension of our univariate framework to multivariate measurements sparse in time. We explore two cases: (a) dimensionality reduction, where we project sparsely sampled multivariate processes to a small latent space, and (b) linear regression, where we use a multivariate process and covariates constant over time for the prediction of a univariate process of interest. To motivate the methodology, we start with a regression involving two variables observed over time.

4.1. Illustrative Example: Univariate Regression

Suppose that the true processes are in a low-dimensional space of some continuous functions (e.g., splines) and that we observe them with noise. More precisely, let

\[ x_i = Bw_i + e_{x,i} \quad \text{and} \quad y_i = Bu_i + e_{y,i}, \]  

(14)

for \( 1 \leq i \leq N \), where \( x_i, y_i, e_{x,i}, e_{y,i} \) are \( T \times 1 \) vectors, \( w_i, u_i \) are \( K \times 1 \) vectors and \( B \) is a \( T \times K \) matrix of \( K \) splines evaluated on a grid of \( T \) points. We assume zero-mean independent errors \( e_{x,i}, e_{y,i} \) with fixed covariance matrices \( \Sigma_X, \Sigma_Y \), respectively, and that the true processes underlying the observed \( x_i \) and \( y_i \) follow a linear relation in the spline space, that is

\[ u_i = A'w_i, \]  

(15)

where \( A \) is an unknown \( K \times K \) matrix.

Let \( X, Y, U, W \) be matrices formed from \( x_i, y_i, w_i, u_i \) stacked vertically. From (15) we have \( U = WA \), while (14) implies

\[ X = WB' + E_X \quad \text{and} \quad Y = WAB' + E_Y, \]  

(16)

where \( E_X, E_Y \) are matrices of observation noise. Without loss of generality, we assume that \( B \) is orthonormal. We have full freedom to choose the basis \( B \), since any basis can be orthogonalized using, for example, the singular value decomposition.

Due to the orthogonality of \( B \) and after multiplying both expressions in (16) by \( B \) we can map the problem to the classical multivariate errors-in-variables models. Let

\[ \tilde{X} = XB = W + \tilde{E}_X \quad \text{and} \quad \tilde{Y} = YB = WA + \tilde{E}_Y, \]  

(17)

where \( \tilde{E}_X = EXB \) and \( \tilde{E}_Y = E_YB \). In errors-in-variables models it is assumed that the parameters \( W \) and \( A \) are unknown, and are to be estimated. Both regressors and responses are measured with noise (here \( \tilde{E}_X \) and \( \tilde{E}_Y \)). The parameter \( W \) can be interpreted as a latent representation of both \( \tilde{X} \) and \( \tilde{Y} \).

The problem of estimating parameters in (17) has been extensively studied in literature dating back to Adcock (1878). Two main methods for estimating parameters in (17) are maximum likelihood estimates (MLE) and generalized least squares estimators (GLSE). The estimators in MLE are derived under the assumption of certain distributions of errors. In GLSE, the only assumption about errors is that they are independent, zero-mean, and they have a common covariance matrix. Then, \( \tilde{X} - W \) and \( \tilde{Y} - WA \) are zero-mean and estimates for \( W \) and \( B \) can be found by optimizing some norm of these expressions. Glaser and Watson (1973) show that in the no-intercept model for \( \tilde{X} \) and \( \tilde{Y} \) of the same size (as in our case) and under the assumption of Gaussian errors, MLE and GLSE give the same estimates of \( A \), if GLSE is derived for the Frobenius norm.

In this work, we focus on the GLSE approach as it can be directly solved in our matrix factorization framework and we find it easier to deploy and extend in practice. To account for different magnitudes of the noise in \( X \) and \( Y \), we consider the optimization problem with weights

\[
\begin{align*}
\minimize_{A,W} \quad & \frac{1}{\sigma_X^2} \| XB - W \|^2_F + \frac{1}{\sigma_Y^2} \| YB - WA \|^2_F, \\
\text{subject to} \quad & \text{rank}(W) \leq r, \\
\text{where} \quad & \sigma_X, \sigma_Y > 0. \quad \text{(18)}
\end{align*}
\]

where \((\cdot : \cdot)\) operator stacks horizontally matrices with the same number of rows. To solve (19), we show that the SVD of \((XB : yYB)\) truncated to the first \( K \) dimensions, can be decomposed to \( W(I : yA) \). Let \( USV^T \) be the SVD of \((XB : yYB)\), with

\[ U = [U_1 \ U_2], S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}, \]

where each \( S_{ij} \) and \( V_{ij} \) is a \( K \times K \) matrix for \( 1 \leq i, j \leq 2 \) and each \( U_i \) is a \( N \times K \) matrix for \( 1 \leq i \leq 2 \). By Lemma 2.1 and Lemma 2.2 in (Gleser 1981) matrix \( V_{11} \) is almost surely nonsingular. Therefore, \( V_{11}^{-1} \) almost surely exists and we can transform the decomposition such that \((I : yA) = (V_{11}^{-1})^{-1} [V_{11}^{-1} V_{21}] \) and \( W = U_1 S_{11}^{-1} V_{21}^{-1} \), solving (19).

For partially observed data, if they are very sparse, it might be essential to constrain the rank of the solution. The partially observed and rank-constrained version of the problem (19) takes the form

\[
\begin{align*}
\minimize_{A,W} \quad & \| P_{\tilde{\Omega}}((X : yY) - W(B' : yAB')) \|^2_F, \\
\text{subject to} \quad & \text{rank}(W(B' : yAB')) = k, \\
\text{where} \quad & k \text{ is the desired rank of the solution and } P_{\tilde{\Omega}} \text{ is a projection on } \\
\tilde{\Omega} = \{(q, r) : (q, r) \in \Omega \text{ or } (q, r - T) \in \Omega\}. \\
\text{As previously, for an unknown } k \text{ we can introduce a rank penalty using the nuclear norm}
\end{align*}
\]

\[
\begin{align*}
\minimize_{A,W} \quad & \| P_{\tilde{\Omega}}((X : yY) - W(B' : yAB')) \|^2_F + \lambda \| W(B' : yAB') \|_*. \\
\text{where} \quad & \lambda \text{ is a scalar that controls the penalty on the rank of } W, \lambda > 0, \quad \text{(20)}
\end{align*}
\]
The algorithm in the general case of multiple processes is derived in Section 4.2.

Although we motivate the problem as a regression of Y on X, X, and Y are symmetric in (19). The low-rank matrix W is, therefore, a joint low-rank representation of matrices X and Y and thus our method can be seen as a dimensionality reduction technique or as a latent space model. In Section 4.2 we extended this idea to a larger number of variables. In Section 4.3 we discuss how this approach can be used for regression.

The linear structure of (16) allows us to draw analogy not only to the errors-in-variables models but also to the vast literature on canonical correlation analysis (CCA), partial least squares (PLS), factor analysis (FA), and linear functional equation (LFE) models. Borga, Landelius, and Knutsson (1997) show that solutions of CCA and PLS can also be derived from the SVD of stacked matrices, as we did with (XB : γYB) in (19). Gleser (1981) thoroughly discusses the relation between errors-in-variables, FA, and LFE.

Finally, the method of using SVD for stacked matrices has also been directly applied in the recommender systems context. Condi et al. (1999) showed that for improving the prediction of unknown entries in some partially observed matrix Q one might consider a low-rank decomposition of (Q : R), where R are additional covariates for each row, for example, demographic features of individuals.

### 4.2. Dimensionality Reduction

Suppose that for every individual we observe multiple variables varying in time (e.g., results of multiple medical tests at different times in a clinic) and we want to find a projection on \( \mathbb{R}^d \) maximizing the variance explained for some \( d \in \mathbb{N} \). This projection would correspond to characterizing patients by their progression trends of multiple metrics simultaneously.

We extend the equation (19) to account for a larger number of covariates, and we do not impose decomposition of the solution yet. We formulate the following optimization problem

\[
\arg \min_W \left( \left\| (X_1 B : X_2 B : \cdots : X_p B) - W \right\|_F^2 + \lambda \left\| W \right\|_* \right), \tag{21}
\]

where \( X_i \) are some \( N \times T \) matrices corresponding to the processes measured on a grid of \( T \) points, \( B \) is a basis evaluated on the same grid and orthogonalized (a \( T \times K \) matrix), and \( W \) is a \( N \times pK \) matrix.

Let \( B = I_p \otimes B \) be the Kronecker product of \( p \times p \) identity matrix and \( B \), that is a \( pT \times pK \) matrix with \( B \) stacked \( p \) times on the diagonal, and let \( X = (X_1 : X_2 : \cdots : X_p) \). Matrix \( B \) is orthogonal and therefore results developed in Section 2.3 apply here. In particular, singular value thresholding solves

\[
\arg \min_W \left\| X - WB^* \right\|_F^2 + \lambda \left\| W \right\|_* \quad \text{argmin} \|P\Omega(Y - WAB)^\dagger\|_F + \lambda \|W\|_* \tag{22}
\]

and we can use Algorithm 1 for solving

\[
\arg \min_W \left\| P\Omega \left( X - WB^* \right) \right\|_F^2 + \lambda \left\| W \right\|_* \tag{22}
\]

where \( P\Omega \) is the projection on observed indices \( \Omega \).

The optimization problem (21) can be further extended. For example, we can choose a different basis for each process or scale individual contributions of processes using scaling factors. By modifying (22) correspondingly, we can extend our solution to these cases.

### 4.3. Regression

In practice, we are often interested in the relationship between the progression of an individual parameter (e.g., cancer growth) and some individual features constant over time (e.g., demographics) or progressions of other covariates (e.g., blood tests, vitals, biopsy results). We show how our framework can be used to exploit relations between curves in order to improve the estimation of the progression of the main parameter of interest.

We start with a regression problem with fixed covariates and sparsely observed response trajectories. Assume that for each subject \( 1 \leq i \leq N \) we observe covariates \( x_i \in \mathbb{R}^d \) and samples from trajectories \( y_i \in \mathbb{R}^{ni} \) following the distribution

\[
y_i | x_i \sim N(BiA, \sigma^2 I_{ni}),
\]

where \( Bi \) is the matrix of evaluations of the basis on the same evaluation points as \( y_i \) (see Section 2.2). We map the problem to the matrix completion framework as in Section 3. Let \( X \) be a \( N \times d \) matrix of observed covariates, \( Y \) be a sparsely observed \( N \times T \) matrix of trajectories, and \( B \) be a \( T \times K \) matrix representing a basis of \( K \) splines evaluated on a grid of \( T \) points. We consider the optimization problem

\[
\arg \min_A \|P\Omega(Y - XAB)^\dagger\|_F^2, \tag{23}
\]

where \( A \) is a \( d \times K \) matrix and \( P\Omega \) is a projection on the observed indices \( \Omega \). To solve (23) we propose an iterative Algorithm 4.

#### Algorithm 4 Sparse-Regression

1. Initialize \( A \) zeros
2. Repeat till convergence:
   (a) Compute \( \hat{Y} = P\Omega(Y + P\Omega(XAB)^\dagger) \)
   (b) Compute \( A_{\text{new}} \leftarrow (X'X)^{-1}X'\hat{Y}B \)
   (c) If \( \|A_{\text{new}} - A\|_F^2 < \varepsilon \) exit
   (d) Assign \( A \leftarrow A_{\text{new}} \)
3. Return \( A \)

Suppose we want to incorporate other variables varying in time for the prediction of the response process. We can directly apply the method proposed in Section 4.2 and model the response and regressors together. However, it might be suboptimal for prediction, as it optimizes the least-squares distance in all variables rather than only the response. This difference is analogous to the difference between the regression line of some univariate \( y \) on independent variables \( x_1, \ldots, x_p \) and the first principal component analysis of \((y, x_1, \ldots, x_p)\), used for the prediction of \( y \). While the first one minimizes the distance to \( y \), the latter minimizes the distance to \((y, x_1, \ldots, x_p)\), which is usually less efficient for predicting \( y \).

Alternatively, we can use methodology from Section 4.2 only for covariates. The solution of (22) can be decomposed into \( W = USV' \), and we regress \( Y \) on \( U \) as in (23). Let \( U \) be an \( N \times d_2 \) orthogonal matrix derived from \((X_1, \ldots, X_p)\), where \( d_2 \) is the numbers of latent components. We search for a \( d_2 \times K \) matrix \( A \) solving

\[
\min_A \|P\Omega(Y - UAB)^\dagger\|_F^2 + \lambda \|A\|_* \tag{24}
\]

where \( P\Omega \) is a projection on a set of observed coefficients. We propose a two-step iterative procedure (Algorithm 5).
Algorithm 5 Sparse-Longitudinal-Regression

Step 1: Latent representation

1. For sparsely observed $X = (X_1, X_2, \ldots, X_p)$ find latent scores $U$ (Algorithm 1)

Step 2: Regression

1. For each $\lambda_1, \lambda_2, \ldots, \lambda_k$
   - Get $A_{\lambda_i}$ by solving the regression problem (24) with $Y, U, \lambda_i$ (Algorithm 4)
2. Return $A_{\lambda_1}, A_{\lambda_2}, \ldots, A_{\lambda_k}$

5. Simulations

We illustrate the properties of the multivariate longitudinal fitting in a simulation study. First, we generate curves with quickly decaying eigenvalues of covariance matrices. Then, we compare the performance of the methods in terms of the variance explained by the predictions.

5.1. Data Generation

We seek to simulate data similar to observations in Section 6. We want to parameterize the processes to draw insights depending on sample size and relation between variables. Let $G$ be a grid of $T$ equidistributed points and let $B$ be a basis of $K$ spline functions evaluated on the grid $G$. We simulate three $N \times K$ matrices using the same procedure $G(r, K, N)$, where $r, \in R^+_K$:

1. Define the procedure $M(r)$ generating symmetric matrices $K \times K$ for a given vector $r, \in R^+_K$:
   (i) Simulate $K \times K$ matrix $S$
   (ii) Use SVD to decompose $S$ to $UDV'$
   (iii) Return $Q = V \text{diag}[r]V'$, where $\text{diag}[r]$ is a diagonal matrix with $r$ on the diagonal
2. Let $\Sigma = M(r)$ and $\mu = N(0, I_K)$.
3. Draw $N$ vectors $v_i$ from the distribution $v_i \sim N(0, \Sigma)$
4. Return a matrix with rows $[v_i]_{1 \leq i \leq N}$.

We chose values of $r$ close to values on the diagonal of the spectral decomposition of covariance structure estimated in Section 6. Define

$$r = [1, 0.39, 0.19, \exp(-5), \ldots, \exp(-K - 1)]$$

Let $X_1, X_2, Z$ be generated following the procedure $G(r, K, N)$ and let $Y = Z + X_1 + X_2$. We consider $X_1, X_2,$ and $Y$ as coefficients in a spline space $B$. We derive corresponding functions by multiplying these matrices by $B^t$, that is, $X_{f,1} = X_1B^t$, $X_{f,2} = X_2B^t$ and $Y_f = YB^t$. We set $N = 100, K = 7$ and $T = 31$.

In order to reflect a realistic sampling of curves, we use timepoints at which subjects in the data study were observed (see Section 6). As we are particularly interested in predictive performance, we select subjects with at least 4 observed timepoints to enable leave-one-out estimation. For each simulated curve we randomly select a subject and use grid-aligned timepoints at which this subject visited the clinic. We acknowledge that specifics of sampling might affect relative performance of methods, however, a thorough analysis of sampling is beyond the scope of our work. Here we only focus on the case directly derived from our data study.

Each observed element of each matrix $X_{f,1}, X_{f,2},$ and $Y$ is drawn with Gaussian noise with mean 0 and standard deviation 0.25. Our goal is to recover $Y$ from such sparsely sampled noisy data.

5.2. Methods

We compare Soft-Longitudinal-Impute (SLI) defined in Algorithm 1, Proximal-gradient (PG) method defined in Algorithm 3, the fPCA procedure (James, Hastie, and Sugar 2000), implemented by Peng and Paul (2009), fdaplace (Zhou et al. 2022), and face (Xiao et al. 2018). The first three algorithms require a set of basis functions as one of the inputs. In all cases, we use a polynomial $B$. In SLI and PG we also need to specify the tuning parameter $\lambda$, while in fPCA we need to choose the rank $R$. We use cross-validation to choose $\lambda$ and $R$ by optimizing for the prediction error on held-out (validation) observations.

We divide the observed coefficients into training (81%), validation (9%), and test (10%) sets. We choose the best parameters of the three models on the validation set and then retrain on the entire training and validation sets combined. We compute the error by taking the mean squared Frobenius distance between $Y$ and estimated $\hat{Y}$, that is

$$\text{MSE}(\hat{Y}) = \frac{1}{T|S|} \sum_{i \in S} \|Y_i - \hat{Y}_i\|_F^2$$

on the test set $S$.

We train the algorithms with all combinations of parameters: regularization parameter for SLI and PG procedures $\lambda \in \{10, 15, 20, \ldots, 50\}$ and the rank for fPCA procedure $d \in \{2, 3, 4\}$. We compare the five methods fPCA, fdaplace, face, PG, and SLI, to the baseline null model which we define as the population means across all visits.

For testing our multivariate approach, we model $(Y, X_1, X_2)$ together and we compare our Sparse-Longitudinal-Regression (SLR) with mface package (Li, Xiao, and Luo 2020). We compare compute time and predictive performance in terms of recovering of $Y$.

5.3. Results and Discussion

The SLI and PG methods achieve superior performance at a good running time as presented in Figure 4. Fdaplace (Zhou et al. 2022) outperforms all methods in terms of speed, but it performs worse in terms of the mean squared error.

Results indicate that both our methods, SLI and PG, should be considered on par with fdaplace and face for practical applications since each method has its own advantages. A solution derived from fdaplace might also be used as a warm start for other algorithms to speed them up. We also observed that the performance gap decreases with the growing number of curves.
One additional advantage of SLI and PG is that they return a smooth space of solutions, parameterized with $\lambda$. In Figure 5, we present the estimated rank and cross-validation error of one of the simulation runs. Compared to fixed-rank alternatives, our methods allow for tuning the predictive accuracy and the rank simultaneously in a smooth manner.

In the multivariate case, both mface and our SLR successfully recovered additional information from variables $X_1$ and $X_2$ as their NMSE in modeling $Y$ was better compared to univariate counterpart methods (SLI and face respectively). SLR outperformed mface in terms of NMSE, but required more time to converge. We conclude that SLR should also be considered a viable candidate method for multivariate sparsely sampled functions on function regression.

6. Data Study

We present an application of our methods for understanding how longitudinal changes in gait patterns relate to subtypes of neurological disorders. First, we discuss how practitioners collect the data and use them to guide the decision process. Next, we describe our dataset and present how our methodology can improve current workflows.

In clinical gait analysis, at each visit movement of a child is recorded using optical motion capture. Optical motion capture allows estimating 3D positions of body parts using a set of cameras tracking markers positions on the subject's body. A set of at least three markers is placed at each analyzed body segment so that its 3D position and orientation can be identified.
Figure 6. Four joints were measured in clinical gait analysis: pelvis, hip, knee, and ankle. Each joint can be measured in three planes: sagittal plane (top row), frontal plate (middle row), and transverse plane (bottom row).

uniquely. These data are then used to determine the relative positions of body segments by computing the angle between the corresponding planes. Typically it is done using a biomechanical model for enforcing biomechanical constraints and improving accuracy.

In gait analysis practitioners usually analyze the movement pattern of seven joints in the lower limbs: ankle, knee, hip in each leg, and pelvis (Figure 6). Each joint angle is measured in time. For making the curves comparable between the patients, usually, the time dimension is normalized to the percentage of the gait cycle, defined as the time between two foot strikes.

While trajectories of joint angles are a piece of data commonly used by practitioners for taking decisions regarding treatment, their high-dimensional nature hinders their use as a quantitative metric of gait pathology or treatment outcome. This motivates the development of univariate summary metrics of gait impairment, such as questionnaire-based metrics Gillette Functional Assessment Walking Scale (FAQ) (Gorton III et al. 2011), Gross Motor Function Classification System (GMFCS) (Palisano et al. 2008) and Functional Mobility Scale (FMS) (Graham et al. 2004), or observational video analysis scores such as Edinburgh Gait Score (Read et al. 2003).

One of the most widely adopted quantitative measurements of gait impairments in pediatrics is the gait deviation index (GDI) (Schwartz and Rozumalski 2008). GDI is derived from joint angle trajectories and measures the deviation of the first ten singular values from the population average of the typically developing population. GDI is normalized in such a way that 100 corresponds to the mean value of typically developing children, with the standard deviation equal 10. It is observed to be highly correlated with questionnaire-based methods. Thanks to its deterministic derivation from the motion capture measurements this method is considered more objective than questionnaires.

In medical practice, GDI has been adapted as a metric for diagnosing the severity of impairment, and it constitutes an integral part of the clinical decision-making process and evaluation of treatment outcomes. However, in order to correctly identify the surgery outcome, it is crucial to understand the natural progression of GDI. In particular, a positive outcome of surgery might be negligible when compared to natural improvement during puberty. Similarly, a surgery maintaining the same level of GDI might be incorrectly classified as a failure, if the decline in the patient's function over time is not accounted for.

Methods introduced in this article can be used to approximate individual progressions of GDI. First, we present how a prediction can be made solely based on the patient's GDI history and the histories of other patients. Next, using our regression procedure, we predict GDI trajectories using other sparsely observed covariates, namely O2 expenditure and walking speed.

6.1. Materials and Methods

We analyze a dataset of Gillette Children's Hospital patients visiting the clinic between 1994 and 2014, ages ranging from 4
to 19 years, mostly diagnosed with Cerebral Palsy. The dataset contains 84 visits of 36 patients without gait disorders and 6066 visits of 2898 patients with gait pathologies.

Motion capture data was collected at 120Hz and joint angles in time were extracted. These joint angles were then normalized in time to the gait cycle. Points from these curves were then subsampled (51 equidistributed points) for the downstream analysis established in the clinic.

In the dataset which we received from the hospital, for each patient we know their birthday and disease subtype. From each visit, we observe the following variables: patient ID, time of the visit, GDI of the left leg, GDI of the right leg, walking speed, and O2 expenditure. Other clinical variables that we received were not included in this study. Walking speed is related to information we lose during the normalization of the gait cycle in time. O2 expenditure is a measure of a subject’s energy expenditure during walking. Pathological gait is often energy inefficient and reduction of O2 expenditure is one of the objectives of treatments. Finally, GDI is computed for two legs while in many cases the neurological disorder affects only one limb. To simplify the analysis, we focus on the more impaired limb by analyzing the minimum of the left and the right GDI.

Our objective is to model individual progression curves. We test two methods: functional principal components (fPCA) and Soft-Longitudinal-Impute (SLI). We compare the results to the null model—the population mean across all visits (mean). In SLR, we approximate GDI using latent variables of sparsely observed covariates O2 expenditure and walking speed, following the methodology from Section 4.3.

Let us denote the test set as $\Omega \subset \{1, 2, \ldots, N\} \times \{1, 2, \ldots, T\}$. We validate each model $M$ on held-out indices by computing the mean squared error as defined in (25). We select the parameters of each of the three methods using cross-validation, using the same validation set.

In our evaluation procedure, for the test set, we randomly select 5% of observations of patients who visited the clinic at least 4 times. We chose the cutoff at 4 in order to have at least 3 observations to fit the curve, and keep 1 observation for testing. We verified with the collaborators at the hospital that three visits are reasonable for the practical use of our methods. We split the remaining 95% of observations into training and validation sets in 90:10 proportion. We also ran the procedures using a smaller test set and did not observe any relevant difference in results, likely because the training set is already relatively large for the task. We train the algorithms with the following combinations of parameters: the regularization parameter for SLI and SLR procedures $\lambda \in \{0, 0.1, 0.2, \ldots, 2.0\}$ and the rank for fPCA procedure $d \in \{2, 3, 4, \ldots, K\}$. We define the grid of $T = 51$ points. We repeat the entire evaluation procedure 20 times.

### 6.2. Results

Compared to the null model, fPCA and SLI explain around $\sim 30\%$ of the variance. We present detailed results in Table 1. SLR using additional predictors, O2 expenditure and walking speed, yielded a mean MSE of 0.68 with a standard deviation of 0.8. We conclude that O2 expenditure and walking speed provide additional information for the prediction of GDI progression.

Both fPCA and Sparse-Longitudinal-Impute provide latent representations of patients’ progression curves. We analyze the singular value vectors from our SVD solution which we refer to as principal components. In the left plot in Figure 3 we show the first two estimated principal components. We found that the first component estimates the change between GDI before and after age of 20. The second component model changes around the age of 10 and around the age of 18. In the right plot in Figure 3, by adding a principle component to the population mean curve, we illustrate how differences in the first component are reflected in the patient’s trajectory. By visual investigation of curves returned by our Sparse-Longitudinal-Impute and by fPCA we found similar trends in the first two components.

Since our SVD decomposition defines a low-rank representation of the progression trends, we can also use it to gain insights into progression in different groups of patients. In Cerebral Palsy, we divide paralysis into subtypes depending on which limbs are affected: monoplegia (one leg), diplegia (two legs), hemiplegia (one side of the body), triplegia (three limbs), quadriplegia (four limbs). Hemiplegia is the most prevalent in our population and it might be divided depending on severity, from type 1 (weak muscles, drop foot) to type IV (severe spasticity). We find differences between trends of progression for different subtypes of paralysis of patients ($F_{6,541} = 17.17, p < 10^{-15}$). We illustrate these distributions in Figure 7.

### 7. Discussion

Results presented in Section 6.2 imply that our Sparse-Longitudinal-Impute and Sparse-Longitudinal-Regression methods can be successfully applied to understand trends of the variability of disease progression. We show how to incorporate progressions of O2 expenditure and walking speed in the prediction of the progression of GDI. We present how low-rank representation can be leveraged to gain insights into subtypes of impairment.

While a large portion of variance remains unexplained, it is important to note that in practice the individual progression is not accounted for explicitly in the current decision-making process. Instead, practitioners only use the population-level characteristics of the dependence between age and impairment severity. Our model can greatly improve this practice.

Despite successful application, we identify limitations that could be potentially addressed in the extensions of our model. First, the method is meant to capture the natural continuous progression of GDI, while in practice there are many discrete events, such as surgeries that break continuity assumptions and render the mean trajectories less interpretable. Second, our methodology does not address the “cold start problem,” that is we do not provide tools for predictions with only one or zero observations.
Third, we do not provide explicit equations for the confidence bounds of predicted parameters. Fourth, in our methods, we focus on finding a solution of (7) regardless of the distribution of the data (e.g., the nature of missingness of the data). While we see it as an advantage if distributions are unknown, in certain situations we could leverage the distribution and improve results.

While these and other limitations can constrain the applicability of the method in the current form, they can be addressed using existing techniques of matrix completion. The focus of this article is to introduce a computational framework rather than build a full solution for all cases. Elementary formulation of the optimization problem as well as the fully-functional implementation can foster the development of new tools using existing techniques of matrix completion. The focus of the optimization problem as well as the fully-functional implementation can foster the development of new tools using existing techniques of matrix completion.

In our R package fcomplete available at https://github.com/kidzik/fcomplete, we provide implementations of all algorithms described in this article as well as helper functions for transforming the data, sampling training and test datasets, and plotting functions. For convenience, we also provided an interface for using the fpca package implementing Sparse Functional Principal Components algorithms (James, Hastie, and Sugar 2000; Peng and Paul 2009). The analysis was performed on a desktop PC with 64 GB RAM memory and an Intel® Core™ Intel(R) Core(TM) i9-10900K CPU @ 3.70GHz, operating on an Ubuntu 18.04 system with R version 4.1.0.

Appendix A. Proofs

We prove convergence by mapping our problem into the framework of Mazumder, Hastie, and Tibshirani (2010). Following their notation we define

\[ f_\lambda(W) = \frac{1}{2} \|P_\Omega(Y) - P_\Omega(WB')\|_F^2 + \lambda \|W\|_* \]  

(A1)

Figure 7. Progression trends in different subsets of diseases. Negative values of the score, such as most of the quadriplegic group, correspond to individual trends where the first component (the red curve Figure 3 left) is subtracted from the mean (the red curve in Figure 3 right). Positive values of the score, such as most of the hemiplegic group, correspond to individual trends where the first component is added (the green curve in Figure 3 right).

Our objective is to find \( W_\lambda = \arg\min_W f_\lambda(W) \). We define

\[ Q_\lambda(W|\tilde{W}) = \frac{1}{2} \|P_\Omega(Y) + P_\Omega(\tilde{W}B') - WB'\|_F^2 + \lambda \|W\|_* \]  

(A2)

Algorithm 1 in the step \( k \) computes \( W^{k+1}_\lambda = \arg\min_W Q_\lambda(W|W^k_\lambda) \). We show that \( W^k_\lambda \) converges to the solution of (A1).

Lemma 2. Let \( W \) be an \( N \times K \) matrix of rank \( r \leq K \) and \( B \) is an orthogonal \( T \times K \) matrix. The solution to the optimization problem

\[ \min_W \frac{1}{2} \|Y - WB'||_F^2 + \lambda \|W\|_* \]  

(A3)

is given by \( \tilde{W} = S_\lambda(YB) \) where

\[ S_\lambda(YB) \equiv WD_\lambda V' \] 

with \( D_\lambda = \text{diag}[(d_1 - \lambda)_{+}, \ldots, (d_r - \lambda)_{+}] \).

\( WDV' \) is the SVD of \( YB \), \( D = \text{diag}[d_1, \ldots, d_r] \), and \( t_+ = \max(t, 0) \).

Proof. By Lemma 1 from Mazumder, Hastie, and Tibshirani (2010) we know that \( S_\lambda(YB) \) solves

\[ \min_W \frac{1}{2} \|Y - W||_F^2 + \lambda \|W\|_* \]

Now, since we have \( \|Y - WB'||_F = \|YB - W||_F, S_\lambda(YB) \) also solves our Lemma 2.

Lemma 3. For every fixed \( \lambda \geq 0 \), define a sequence \( W^k_\lambda \) by

\[ W^{k+1}_\lambda = \arg\min_W Q_\lambda(W|W^k_\lambda) \]

with any starting point \( W^0_\lambda \). The sequence \( W^k_\lambda \) satisfies

\[ f_\lambda(W^{k+1}_\lambda) \leq Q_\lambda(W^{k+1}_\lambda|W^k_\lambda) \leq f_\lambda(W^k_\lambda) \]
Proof. By Lemma 2 and the definition $(A2)$, we have:

\[
\begin{align*}
& f_\lambda(W^k) = Q_\lambda(W^k \mid W^\infty) \\
& = \frac{1}{2} \left( \| P_\Omega(Y) + P_\Omega(W^k B') - W^k B' \|^2_F + \lambda \| W^k \|_* \right) \\
& \geq \min_w \frac{1}{2} \left( \| P_\Omega(Y) + P_\Omega(W^k B') - W^k B' \|^2_F + \lambda \| W^k \|_* \right) \\
& = Q_\lambda(W^{k+1} \mid W^k) \\
& = \frac{1}{2} \left( \| (P_\Omega(Y) - P_\Omega(W^{k+1} B')) + (P_\Omega(W^k B') - P_\Omega(W^{k+1} B')) \|^2_F \\
& + \lambda \| W^{k+1} \|_* \right) \\
& = \frac{1}{2} \left( \| P_\Omega(Y) - P_\Omega(W^{k+1} B') \|^2_F + \frac{1}{2} \| P_\Omega(W^k B') - P_\Omega(W^{k+1} B') \|^2_F \\
& + \lambda \| W^{k+1} \|_* \right) \\
& \geq \frac{1}{2} \left( \| P_\Omega(Y) - P_\Omega(W^{k+1} B') \|^2_F + \lambda \| W^{k+1} \|_* \right) \\
& = Q_\lambda(W^{k+1} \mid W^k) = f(W^{k+1}).
\end{align*}
\]

Note that proofs of Lemmas 2 and 3 are just extentions of their counterparts in Mazumder, Hastie, and Tibshirani (2010) with the basis $B$ included. Similarly, we get equivalent results for their Lemma 3–5 and the main theorem.

Theorem 4. The sequence of $W^k$ defined in Lemma 3 converges to a limit $W^\infty$ that solves

\[
\min_w \frac{1}{2} \left( \| P_\Omega(Y) - P_\Omega(WB) \|^2_F + \lambda \| W \|_* \right).
\]

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

Łukasz Kidziński was supported by the Mobilize Center grant U54 EB020405 from the National Institute of Health. Trevor J. Hastie was partially supported by grants DMS-2013736 And IIS, 1837931 from the National Science Foundation, and grant 5R01 EB, 001988-21 from the National Institutes of Health.

ORCID

Łukasz Kidziński https://orcid.org/0000-0002-0986-3078

References

Adcock, R. J. (1878), "A Problem in Least Squares," The Analyst, 5, 53–54. [558]
Bennett, J., and Lanning, S. (2007), "NetFlix: The Netflix Prize," in KDD Cup and Workshop in conjunction with KDD. [551]
Berky, C. S., and Kent, R. L. (1983), "Longitudinal Principal Components and Non-linear Regression Models of Early childhood Growth," Annals of Human Biology, 10, 523–536. [551,553]
Besse, P., and Ramsay, J. O. (1986), "Principal Components Analysis of Sampled Functions," Psychometrika, 51, 285–311. [551,553]
Besse, P. C., Cardot, H., and Ferraty, F. (1997), "Simultaneous Nonparametric Regressions of Unbalanced Longitudinal Data," Computational Statistics & Data Analysis, 24, 255–270. [553]
Bigelow, J. L., and Dunson, D. B. (2009), "Bayesian Semiparametric Joint Models for Functional Predictors," Journal of the American Statistical Association, 104, 26–36. [551,553]
Borga, M., Landelius, T., and Knutsson, H. (1997), "A Unified Approach to pca, pls, mlr and cca," Linköping University, Department of Electrical Engineering. [559]
Cai, I.-F., Candès, E. J., and Shen, Z. (2010), "A Singular Value Thresholding Algorithm for Matrix Completion," SIAM Journal on Optimization, 20, 1956–1982. [555]
Candès, E. J., and Recht, B. (2009), "Exact Matrix completion via Convex Optimization," Foundations of Computational Mathematics, 9, 717–772. [551]
Cnaan, A., Laird, N. M., and Slasor, P. (1997), "Tutorial in Biostatistics: Using the General Linear Mixed Model to Analyse Unbalanced Repeated Measures and Longitudinal Data," Statistics in Medicine, 16, 2349–2380. [553]
Condil, M. K., Lewis, D. D., Madigan, D., and Posse, C. (1999), "Bayesian Mixed-Effects Models for Recommender Systems," in ACM SIGIR (Vol. 99). [559]
Descary, M.-H., and Panaretos, V. M. (2016), "Functional Data Analysis by Matrix Completion," arXiv preprint arXiv:1609.00834. [552,553]
Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002), Analysis of Longitudinal Data, Oxford: Oxford University Press. [558]
Fithian, W., and Mazumder, R. (2018), "Flexible Low-Rank Statistical Modeling with Missing Data and Side Information," Statistical Science, 33, 238–260. [551]
Friedman, J., Hastie, T., and Tibshirani, R. (2001), The Elements of Statistical Learning (Vol. 1), Springer Series in Statistics, New York: Springer. [556]
Ge, R., Lee, J. D., and Ma, T. (2016), "Matrix Completion has no Spurious Local Minimum," in Advances in Neural Information Processing Systems, pp. 2973–2981. [556]
Gleser, L. J. (1981), "Estimation in a Multivariate "Errors in Variables" Regression Model: Large Sample Results," The Annals of Statistics, 9, 24–44. [558,559]
Gleser, L. J., and Watson, G. S. (1973), "Estimation of a Linear Transformation," Biometrika, 60, 525–534. [558]
Gorton III, G. E., Stout, J. L., Bagley, A. M., Bevans, K., Novacheck, T. F., and Tucker, C. A. (2011), "Gillette Functional Assessment Questionnaire 22-Item Skill Set: Factor and Rasch Analyses," Developmental Medicine & Child Neurology, 53, 250–255. [562]
Graham, H. K., Harvey, A., Rodda, J., Nattrass, G. R., and Pirpiris, M. (2004), "The Functional Mobility Scale (fms)," Journal of Pediatric Orthopaedics, 24, 514–520. [556]
Gray, S. M., and Brookmeyer, R. (1998), "Estimating a Treatment Effect from Multidimensional Longitudinal Data," Biometrics, 54, 976–988. [558]
Heidelberger, P. (2000), "Multidimensional Longitudinal Data: Estimating a Treatment Effect from Continuous, Discrete, or Time-to-Event Response Variables," Journal of the American Statistical Association, 95, 396–406. [558]
Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2011), "Longitudinal Functional Principal Component Analysis," in Recent Advances in Functional Data Analysis and Related Topics, ed. F. Ferraty, pp. 149–154, Heidelberg: Springer. [551,553]
Hall, P., Müller, H.-G., and Wang, J.-L. (2006), "Properties of Principal Component Methods for Functional and Longitudinal Data Analysis," The Annals of Statistics, 34, 1493–1517. [551,553]
Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015), "Matrix Completion and Low-Rank svd via Fast Alternating Least Squares," Journal of Machine Learning Research, 16, 3367–3402. [555]
Hastie, T., Tibshirani, R., and Wainwright, M. (2015), Statistical Learning with Sparsity: The Lasso and Generalizations, Boca Raton, FL: CRC Press. [557]
Henderson, C. R. (1950), "Estimation of Genetic Parameters," in Biométrie, 6, 186–187. [553]
Horváth, L., and Kokoszka, P. (2012), Inference for Functional Data with Applications (Vol. 200), New York: Springer. [553]
James, G. M., Hastie, T. J., and Sugar, C. A. (2000), "Principal Component Models for Sparse Functional Data," Biometrika, 87, 587–602. [551,553,557,560,561,564]

Kim, J.-O., and Mueller, C. W. (1978), Factor Analysis: Statistical Methods and Practical Issues (Vol. 14), London: Sage. [557]

Kosambi, D. D. (2016), "Statistics in Function Space," in D.D. Kosambi, pp. 115–123, New Delhi: Springer. [553]

Laird, N. M. (1988), "Missing Data in Longitudinal Studies," Statistics in Medicine, 7, 305–315. [553]

Laird, N. M., and Ware, J. H. (1982), "Random-Effects Models for Longitudinal Data," Biometrics, 38, 963–974. [553]

Larsen, R. M. (2004), "Propack-Software for Large and Sparse svd Calculations," pp. 2008–2009. Available at http://sun.stanford.edu/rmunk/PROPACK. [556]

Lawrence, N. D. (2004), "Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data," in Advances in Neural Information Processing Systems, pp. 329–336. [553]

Li, C., Xiao, L., and Luo, S. (2020), "Fast Covariance Estimation for Multivariate Sparse Functional Data," Stat, 9, e245. [560,561]

Liu, L., and Huang, X. (2009), "Joint Analysis of Correlated Repeated Measures and Recurrent Events Processes in the Presence of Death, with Application to a Study on Acquired Immune Deficiency Syndrome," Journal of the Royal Statistical Society, Series C, 58, 65–81. [551]

Ma, S., Goldfarb, D., and Chen, L. (2011), "Fixed Point and Bregman Iterative Methods for Matrix Rank Minimization,” Mathematical Programming, 128, 321–353. [555]

MacLehose, R. F., and Dunson, D. B. (2009), "Nonparametric Bayes Kernel-based Priors for Functional Data Analysis," Statistica Sinica, 19, 611–629. [553]

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1980), Multivariate Analysis. Probability and Mathematical Statistics, New York: Academic Press. [557]

Mazumder, R., Hastie, T., and Tibshirani, R. (2010), "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," Journal of Machine Learning Research, 11, 2287–2322. [555,556,564,565]

McCulloch, C. E., and Neuhaus, J. M. (2001), Generalized Linear Mixed Models. Wiley Online Library, Hoboken, NJ: Wiley. [551]

Nesterov, Y. (2013), "Gradient Methods for Minimizing Composite Functions," Mathematical Programming, 140, 125–161. [557]

Palisano, R. J., Rosenbaum, P., Bartlett, D., and Livingston, M. H. (2008), "Content Validity of the Expanded and Revised Gross Motor Function Classification System," Developmental Medicine & Child Neurology, 50, 744–750. [562]

Peng, J., and Paul, D. (2009), "A Geometric Approach to Maximum Likelihood Estimation of the Functional Principal Components from Sparse Longitudinal Data," Journal of Computational and Graphical Statistics, 18, 995–1015. [560,564]

Ramsay, J. O. (2006), Functional Data Analysis. Wiley Online Library. [553]

Ramsay, J. O., and Dalzell, C. J. (1991), "Some Tools for Functional Data Analysis," Journal of the Royal Statistical Society, Series B, 53, 539–572. [553]

Read, H. S., Elizabeth Hazlewood, M., Hillman, S. J., Prescott, R. J., and Robb, J. E. (2003), "Edinburgh Visual Gait Score for Use in Cerebral Palsy," Journal of Pediatric Orthopaedics, 23, 296–301. [562]

Rennie, J. D. M., and Srebro, N. (2005), "Fast Maximum Margin Matrix Factorization for Collaborative Prediction," in Proceedings of the 22nd International Conference on Machine Learning, pp. 713–719, ACM. [551]

Rice, J. A., and Wu, C. O. (2001), "Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves," Biometrics, 57, 253–259. [551,553]

Rizopoulos, D., Hatfield, L. A., Carlin, B. P., and Takkenberg, J. J. M. (2014), "Combining Dynamic Predictions from Joint Models for Longitudinal and Time-to-Event Data Using Bayesian Model Averaging," Journal of the American Statistical Association, 109, 1385–1397. [551]

Robinson, G. K. (1991), "That Blup is a Good Thing: The Estimation of Random Effects," Statistical Science, 6, 15–32. [553]

Samuel, M. D., and Ryan, L. M. (1996), "Latent Variable Models with Fixed Effects," Biometrics, 52, 650–663. [558]

Schulam, P., and Arora, R. (2016), "Disease Trajectory Maps," in Advances in Neural Information Processing Systems, pp. 4709–4717. [553]

Schwartz, M. H., and Rozumalski, A. (2008), "The Gait Deviation Index: A New Comprehensive Index of Gait Pathology," Gait & Posture, 28, 351–357. [562]

Song, X., Davidian, M., and Tsiatis, A. A. (2002), "A Semiparametric Likelihood Approach to Joint Modeling of Longitudinal and Time-to-Event Data," Biometrics, 58, 742–753. [551]

Srebro, N., Alon, N., and Jaakkola, T. S. (2005), "Generalization Error Bounds for Collaborative Prediction with Low-Rank Matrices," in Advances In Neural Information Processing Systems, pp. 1321–1328. [552]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society, Series B, 58, 267–288. [556]

Tipping, M. E., and Bishop, C. M. (1999), "Probabilistic Principal Component Analysis," Journal of the Royal Statistical Society, Series B, 61, 611–622. [557]

Verbeke, G. (1997), "Linear Mixed Models for Longitudinal Data," in Linear Mixed Models in Practice, pp. 63–153, New York: Springer. [551,553]

Watanabe, S. (1965), "Karhunen-Loeve Expansion and Factor Analysis," in Transactions of the 4th Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes, Prague. [553]

Xiao, L., Li, C., Checkley, W., and Crainiceanu, C. (2018), "Fast Covariance Estimation for Sparse Functional Data," Statistics and Computing, 28, 511–522. [560,561]

Yan, F., Lin, X., Huang, X. (2017), "Dynamic Prediction of Disease Progression for Leukemia Patients by Functional Principal Component Analysis of Longitudinal Expression Levels of an Oncogene," The Annals of Applied Statistics, 11, 1649–1670. [551,553]

Yao, F., and Lee, T. (2006), "Penalized Spline Models for Functional Principal Component Analysis," Journal of the Royal Statistical Society, Series B, 68, 3–25. [551,553]

Yao, F., Müller, H.-G., Wang, J.-L. (2005), "Functional Linear Regression for Longitudinal Data," The Annals of Statistics, 33, 2873–2903. [553]

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach," Biometrics, 44, 1049–1060. [551,553]

Zhou, Y., Bhattacharjee, S., Carroll, C., Chen, Y., Dai, X., Fan, J., Gajardo, A., Hadjipantelis, P. S., Han, K., Ji, H., Zhu, C., Müller, H.-G., and Wang, J.-L. (2022), fdapace: Functional Data Analysis and Empirical Dynamics. R package version 0.5.9. Available at https://CRAN.R-project.org/package=fdapace [560,561]