A Smooth Reconstruction of Covariance Kernels from Fragmented Functional Data

Teresa Bortolotti[1], Roberta Troilo[1], Alessandra Menafoglio[1], and Simone Vantini[1]

MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milan, Italy.
teresa.bortolotti@polimi.it

Abstract. Estimating the covariance operator from incomplete observations is a known problem in functional data analysis. While existing methods provide consistent estimates of the operator when functional samples are partially observed on varying subintervals of their domain, they fail in settings where missing values follow a persistent pattern across all samples, and entire portions of the domain remain systematically unobserved. We propose a nonparametric covariance reconstruction method designed specifically for this setting. Our approach estimates the covariance kernel through the sample covariance matrix and extends matrix completion techniques by introducing a Laplacian regularization which enforces smoothness in the solution. Our method effectively mitigates instability and roughness issues associated with purely low-rank approaches, while preserving accuracy in the reconstruction. The proposed methodology is showcased on multi-temporal interferometric images of the Phlegraean Fields, Italy, where covariance reconstruction is key for reliable displacement estimation and natural hazard monitoring.

Keywords: Functional data analysis, incomplete data, covariance reconstruction, Laplacian regularization

1 Introduction

This work considers the problem of estimating the covariance operator of partially observed independent and identically distributed functional data. Recovering the covariance structure from incomplete observations is a well-known problem in the literature on partially observed functional data [8, 6, 4]. While the covariance operator plays a fundamental role in functional data analysis – enabling dimensionality reduction through principal component decomposition and facilitating prediction in forecasting problems – it becomes even more critical in the presence of incomplete functional data, as reconstruction methods for partially observed curves typically rely on its accurate estimation [8, 7].

The fundamental difficulty of estimating the covariance operator in this setting arises from the lack of observed functional pairs for certain domain locations, making standard empirical covariance estimation inconsistent or even infeasible.

Currently, solutions exist in certain regimes of data availability. Two commonly studied cases are the blanket regime and the banded regime. In the blanket regime [8], a sufficient number of functional samples are available over the entire domain, allowing for straightforward estimation of the covariance operator using classical estimators from functional data analysis. In the banded regime [4, 6], each functional sample is observed on a subinterval of fixed length δ, smaller than the length of the full domain. In this setting, the classical empirical estimator is a consistent estimate of the covariance kernel over a band of width δ that covers the main diagonal of the covariance matrix. Then, the covariance kernel can be effectively reconstructed in the missing cells using matrix completion techniques which leverage a low-rank structure of the space where the extension of the kernel is seeked [5, 6].

The solutions proposed for the banded regime do not directly apply when the functional observations are available only on the same, potentially disjoint subdomain O of the full domain. In this fragmented regime, entire portions of the domain remain persistently missing across all functional replicates. We propose a novel nonparametric approach to covariance reconstruction tailored to this regime. Building on ideas from [6], we propose to reconstruct the covariance operator using a regularized, fixed-rank matrix completion framework. By incorporating structural smoothness constraints via Laplacian regularization, our approach mitigates the instability and roughness associated with pure low-rank methods in the presence of persistently missing values. This allows for a more reliable estimation of the underlying covariance structure even in cases where conventional methods fail.

The proposed method is showcased on real data involving Synthetic Aperture Radar (SAR) images containing ground displacement measurements on the area of the Phlegraean Fields, in Italy. As SAR signals may be scattered, absorbed, or reflected away from the sensor by certain ground features –such as water, vegetation, or rocky surfaces – the analysed images present patterns of missing values that remain presistently unobserved over replicates. In these cases, reconstructing the covariance structure in every point of the image is crucial for recovering missing displacement values and improving monitoring of natural hazards.

2 Notation and Problem Definition

Consider a sequence of independent and identically distributed random functions, $X_1, \ldots, X_n$, taking values in $L^2([0,1])$, the space of square-integrable functions over the unit interval. We denote the mean function and covariance operator of the process by $\mu : [0,1] \to \mathbb{R}$ and $\mathcal{R} : L^2([0,1]) \to L^2([0,1])$, respectively. The covariance operator is an integral operator which acts on any function $f \in L^2([0,1])$ through its kernel $r(s,t) = \mathrm{Cov}(X_1(s), X_1(t))$; that is, $\mathcal{R}f = \int_0^1 r(\cdot,t)f(t)dt$. We suppose that each curve X_i is observed only on a subset $O \subset [0,1]$, which may in principle consist of multiple disjoint subintervals. The unobserved portion of the domain is denoted as $M = [0,1] \setminus O$.

When functional data are only partially observed, standard empirical estimators for the mean function and covariance operator cannot be directly applied. Formally, the estimator for the mean function μ at $t \in [0,1]$ reads

$$\hat{\mu}(t) = \frac{1_O(t)}{n} \sum_{i=1}^{n} X_i(t),$$

where $1_O(t)$ is the indicator function corresponding to the observed domain. The mean function is consistently estimated by $\hat{\mu}$ at locations where observations are available. However, in regions where no data are observed, the estimator takes value zero. It will be clear from what follows that having an estimate of μ on M is not necessary to get an estimate of the covariance operator everywhere on $[0,1]^2$. We argue, though, that extending $\hat{\mu}$ on M through a linear interpolation of its values on O could be a simple but effective solution for many practical problems. We hereafter focus on the estimation of the covariance operator.

As it is typically done in functional data analysis, the empirical counterpart of the covariance operator is obtained through an estimator of its covariance kernel. Following the notation of [6], for any $(s,t) \in [0,1]^2$, we define the empirical covariance kernel $\hat{r}(s,t)$ as

$$\hat{r}(s,t) = \frac{1_O(s)1_O(t)}{n} \sum_{i=1}^{n} \left(X_i(s) - \hat{\mu}(s)\right)\left(X_i(t) - \hat{\mu}(t)\right).$$

Note that for pairs (s,t) where both points lie in M, the estimator is identically zero, indicating the need for reconstruction in these regions.

We specifically focus on the case where functional data are recorded on an equally spaced grid. Each function is therefore represented through its evaluations at K predefined points $(t_1,\ldots,t_K) \in T \subset O$, forming a discrete representation. The empirical covariance kernel can then be compactly represented by the $K \times K$ sample covariance matrix $R^K = \{\hat{r}(t_j,t_l)\}_{j,l=1}^K$. The goal is to recover the missing entries in this covariance matrix to retrieve a complete discrete estimate of $r(s,t)$ for any $(s,t) \in [0,1]^2$.

3 Methodology

Inspired by the covariance reconstruction method proposed in [6] for the banded regime, we formulate the covariance reconstruction as a fixed-rank matrix completion problem. First, we complete the missing diagonal entries of R^K by performing a simple interpolation of the observed diagonal entries. Treating the diagonal entries of R^K as observed is crucial in order to ensure the stability of the solutions to the optimization problem which we present below. In the space of positive semi-definite matrices with fixed rank r^*, we look for the matrix R which minimizes an objective function composed of two terms. The first term quantifies the error made in the approximation of the observed cells in R^K. The second term is a Laplacian regularization, which enforces smoothness by discouraging sharp variations in the estimated covariance kernel. This is particularly advantageous when the underlying functional process exhibits gradual changes over the domain. The balance between the adherence to the observed parts of R^K and the smoothness in the reconstructed covariance matrix is controlled by the regularization parameter α, which adjusts the trade-off between these two objectives.

To enhance computational efficiency, the problem is solved by leveraging the positive semidefiniteness of R^K. Specifically, the candidate covariance matrix R is parametrized as $R = \sigma\sigma^T$, where $\sigma \in \mathbb{R}^{r^* \times K}$. Then, the optimization is done over σ using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.

The Laplacian regularization is based on an adjacency-based Laplacian matrix that captures the connectivity between domain points and incorporates the domain structure into the optimization problem. Since the regularization term is computed over the entire reconstructed kernel, we introduce a weighting matrix to control the relative influence of observed and missing entries in the penalization term. Note that this strategy is coherent with the weighted approach to functional data analysis presented in [2]. The weighting matrix, parameterized by $m \in [0, 1]$, determines how much the observed and missing entries in the reconstructed covariance matrix contribute to the penalty. Higher values of m increase the influence of missing entries in the regularization term, while lower values prioritize smoothness over the observed data.

The rank r^* can be set as high as computational resources allow without affecting the smoothness of the reconstructed covariance matrix in the missing entries, as the Laplacian term enforces regularity and prevents instability or overfitting. The hyperparameters α and m are tuned according to the criteria established by conducting extensive simulation studies. The simulations show that the parameter α should be selected by identifying the elbow point in the reconstruction error made on the observed part of the covariance matrix. This elbow arises due to the trade-off in the optimization problem, where smaller α prioritizes observed data while larger α reduces its influence, leading to an increase in the error made when reconstructing the observed cells of the covariance matrix. For parameter m, on the other hand, the simulations prescribe that its selection should be based on cross-validation. The cross-validation scheme sequentially treats rows and columns near existing missing cells as fictitiously unobserved, selecting m that minimizes the reconstruction error in these left-out cells.

4 Case Study

Synthetic Aperture Radar (SAR) data are collected by Sentinel-1 satellites with a revisit time as short as six days. The Small Baseline Subsets (SBAS) method [1] leverages sequential SAR acquisitions to estimate ground displacement over time. Specifically, SBAS constructs a sequence of high-resolution images, where each pixel is associated to a time series of surface deformation for the corresponding ground location [3]. As SAR signals may be scattered, absorbed, or reflected away from the sensor by certain ground features – such as water, vegetation, or rocky surfaces – SBAS-processed images are typically characterized by large numbers of missing values, which remain absent over all instants of the time series.

We apply our methodology to SBAS-processed SAR images of the Phlegraean Fields, in Italy, focusing on a study area of 101×101 pixels, where each pixel represents an 80m $\times$ 80m section of terrain. The dataset consists of $n = 391$ temporal observations. Due to significant autocorrelation between images, we preprocess the data to remove temporal dependencies.

To reconstruct the covariance structure of the image, we extend the approach presented in Section 3 to accommodate functions defined over a two-dimensional domain. Specifically, we modify the Laplacian regularization term to incorporate the four-dimensional proximity of elements within the covariance kernel. We determine α by analyzing the curve of the reconstruction error made on the observed entries of the covariance matrix for fixed values of m, and then select m using the cross-validation strategy outlined in Section 3. Although our method operates in a four-dimensional space, Figure 1 provides a simplified visualization by vectorizing three consecutive columns of the reconstructed empirical covariance kernel for illustration.

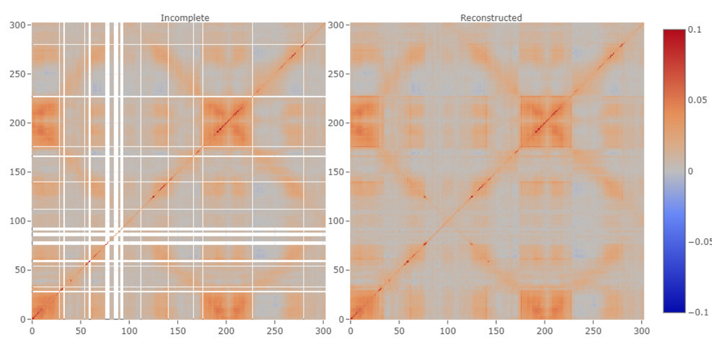

Fig. 1. Empirical covariance kernel (left) and reconstructed covariance kernel (right) for 79^{th}, 80^{th} and 81^{st} column of the considered area. The rank of the solution is $r^* = 130$, and the selected hyperparameters are $\alpha = 0.01$ and $m = 0.1$.

5 Conclusions

We introduce a nonparametric approach for reconstructing the covariance kernel of partially observed functional data, specifically addressing the newly introduced fragmented regime, where missing values follow persistent patterns. Building on the low-rank matrix completion framework of [6], which enforces a low-rank constraint while preserving observed entries, our method incorporates instead a Laplacian regularization within a fixed-rank formulation. This regularization encourages the smoothness of the reconstructed covariance kernel by promoting consistency across neighboring cells of the covariance matrix. We provide systematic criteria for selecting the regularization parameter α and weighting parameter m. Finally, applying our approach to SBAS-processed ground displacement data, we successfully reconstruct the covariance structure, demonstrating its practical applicability. The recovered covariance kernel can then be integrated into existing methods for functional data reconstruction [8, 7], facilitating the estimation of the ground deformation in the regions where the information is missing.

Acknowledgements This work is partially supported by ACCORDO Attuativo ASI-POLIMI "Attività di Ricerca e Innovazione" n. 2018-5-HH.0, collaboration agreement between the Italian Space Agency and Politecnico di Milano, and by the initiative "Dipartimento di Eccellenza 2023–2027", MUR, Italy, Dipartimento di Matematica, Politecnico di Milano. The authors gratefully acknowledge the financial support of IREA-CNR (Istituto per il Rilevamento Elettromagnetico dell'Ambiente del Consiglio Nazionale delle Ricerche) for funding a PhD grant in cooperation with Politecnico di Milano.

Bibliography

[1] Berardino, P., Fornaro, G., Lanari, R., Sansosti, E.: A new algorithm for surface deformation monitoring based on small baseline differential SAR interferograms. IEEE Transactions on Geoscience and Remote Sensing **40**, 2375–2383 (2002).
[2] Bortolotti, T., Peli, R., Lanzano, G., Sgobba, S., Menafoglio, A.: Weighted Functional Data Analysis for the Calibration of a Ground Motion Model in Italy. Journal of the American Statistical Association, **119**, 1697–1708 (2024).
[3] Casu, F., Manzo, M., Lanari, R.: A quantitative assessment of the SBAS algorithm performance for surface deformation retrieval from DInSAR data. Remote Sensing of Environment, **102**, 195–210 (2006).
[4] Delaigle, A., Hall, P., Huang, W., Kneip, A.:Estimating the covariance of fragmented and other related types of functional data. Journal of the American Statistical Association, **116**, 1383–1401 (2021).
[5] Descary, M.-E., Panaretos, V. M.: Functional data analysis by matrix completion. The Annals of Statistics **47**, 1–38 (2016).
[6] Descary, M.-E., Panaretos, V. M.: Recovering Covariance from Functional Fragments. Biometrika **106**, 145–160 (2019).
[7] Kneip, A., Liebl, D.: On the optimal reconstruction of partially observed functional data. The Annals of Statistics **48**, 1692–1717 (2020).
[8] Kraus, D.: Components and completion of partially observed functional data. Journal of the Royal Statistical Society **77**, 777–801 (2015).