ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Master thesis

A probabilistic approach to the classification of censored functional data

William Borgeaud dit Avocat

supervised by
Prof. Victor Panaretos

December 7, 2017

Contents

Chapter 1

Introduction to Functional Data Analysis

1.1 What are functional data?

1.1.1 Explain the structure of functional data, the spaces they live in, expectation, covariance, observable discretized versions.

1.2 Statistical analysis of functional data

1.2.1 Explain the basic principles of estimation of mean and covariance with convergence rates, etc.. Also mention smoothing methods.

Chapter 2

Censored functional data

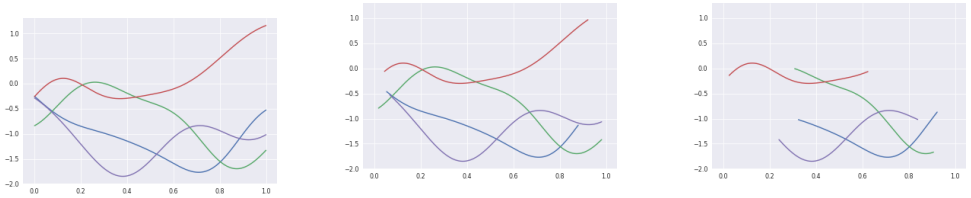
2.1 Functional fragments framework

2.1.1 Explain the functional fragments framework and and the methods used by Kraus and Delaigle and Hall

Censored functional data or *functional fragments* are functional data that are not observed in the full domain on which they are defined. If the data live in $\mathbb{L}^2(\mathcal{I})$ for some interval $\mathcal{I} \subset \mathbb{R}$, an example of functional fragment is a function $f \in \mathbb{L}^2(\mathcal{J})$ for some interval $\mathcal{J} \subset \mathcal{I}$.

By the *functional fragments framework*, we mean the statistical framework in which some or all of the observed data are in the form of fragments of some underlying, unobservable, functional data. In this case, the data at hand are pairs $\{(X_i, \mathcal{O}_i)\}_{i=1}^n$, where the X_i's are random functions in $\mathbb{L}(\mathcal{O}_i)$ for some subintervals $\mathcal{O}_i$. We will often assume that the subintervals $\{\mathcal{O}_i\}_{i=1}^n$ are themselves random, in order to make the asymptotic theory more tractable. This framework often arises in practice when an observation is unavailable before or after a certain time.
The main issue in the functional fragments framework is to know to what extent one can recover precise information on the underlying population from the observed fragments. For example, how precisely can we estimate the mean and covariance when no curve is fully observed.
Following [Descary and Panaretos, 2017], we distinguish between two ways in which the intervals $\{\mathcal{O}_i\}_{i=1}^n$ are distributed, see Figure 2.1:

1. A "blanket " regime, where the curves are typically observed on most or all of the domain. Then the number of observations at a given point of the domain is close to the total number of observations.
2. A "banded" regime, where the lengths of the $\mathcal{O}_i$ are bounded by some value $\delta > 0$. Then, we have no explicit information on the covariance of points that are at distance larger than δ.



(a) Full curves

(b) Fragments in the blanket regime

(c) Fragments in the banded regime

Figure 2.1: Example of functional fragments

In the rest of this chapter, we will present various methods used in the literature dealing with those issues.

2.2 Naive estimations

We present here the methods presented in [Kraus, 2015]. The data are i.i.d curves X_i in $\mathbb{L}^2[0,1]$ observed only on a random interval $\mathcal{O}_i \subset [0,1]$, $i = 1, ..., n$. To estimate the population mean $\mu = \mathbb{E}[X_1]$ and covariance operator $\mathcal{K} = \mathbb{E}[(X_1 - \mu) \otimes (X_1 - \mu)]$, the unobserved parts of the curves are ignored and the sample estimators are created naively as follows. The sample mean $\hat{\mu}$ is found by taking the mean of the pointwise observed values:

$$\hat{\mu}(t) = \frac{\mathbf{1}[\exists \mathcal{O}_i \ni t]}{\sum_{i=1}^n \mathbf{1}[t \in \mathcal{O}_i]} \sum_{i=1}^n \mathbf{1}[t \in \mathcal{O}_i] \cdot X_i(t).$$

The covariance operator is estimated in the same fashion via its associated covariance kernel $K(\cdot, \cdot)$. The sample kernel is given by:

$$\hat{K}(s,t) = \frac{\mathbf{1}[\exists \mathcal{O}_i \ni s,t]}{\sum_{i=1}^n \mathbf{1}[s,t \in \mathcal{O}_i]} \sum_{i=1}^n \mathbf{1}[s,t \in \mathcal{O}_i] \cdot \{X_i(s) - \hat{\mu}_{st}(s)\} \{X_i(t) - \hat{\mu}_{st}(t)\},$$

where $\hat{\mu}_{st}$ is an estimation of the mean using only the curves observed at s and t:

$$\hat{\mu}_{st}(s) = \frac{\mathbf{1}[\exists \mathcal{O}_i \ni s,t]}{\sum_{i=1}^n \mathbf{1}[s,t \in \mathcal{O}_i]} \sum_{i=1}^n \mathbf{1}[s,t \in \mathcal{O}_i] \cdot X_i(s).$$

The sample covariance operator $\hat{\mathcal{K}}$ is then defined by

$$\hat{\mathcal{K}} f(t) = \int_0^1 \hat{K}(s,t) f(s) \, \mathrm{d}s.$$

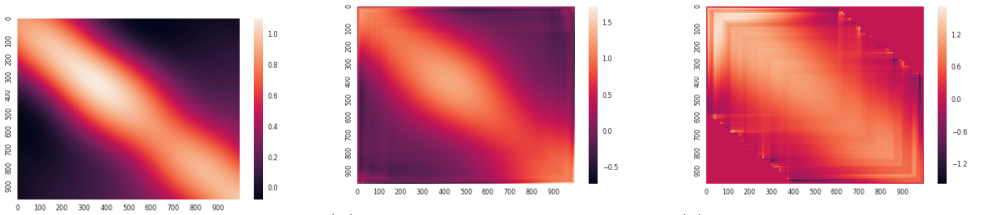
We note that this operator need not be positive-definite. This can be dealt with by clipping the negative eigenvalues to zero.
The following proposition, proved in [Kraus, 2015, Prop. 1], shows that under some assumptions on the random intervals $\{\mathcal{O}_i\}_{i=1}^n$, the above estimates enjoy the same asymptotic convergence rate as their counterparts when the curves are fully observed.

Proposition 1.

1. Suppose that $\mathbb{E}\|X_1\|^2 < \infty$ and the $\mathcal{O}_i$'s are i.i.d with $\inf_{t \in [0,1]} \mathbb{P}[t \in \mathcal{O}_1] > 0$. Then
$$\mathbb{E}\|\hat{\mu} - \mu\|^2 = O(n^{-1}) \text{ as } n \to \infty.$$
2. Suppose further that $\mathbb{E}\|X_1\|^4 < \infty$ and that $\inf_{s,t \in [0,1]} \mathbb{P}[s,t \in \mathcal{O}_1] > 0$. Then
$$\mathbb{E}\|\hat{\mathcal{K}} - \mathcal{K}\|_{HS}^2 = O(n^{-1}) \text{ as } n \to \infty.$$

□

For the covariance operator, even though the theoretical convergence rate is good, in practice the estimate is not adequate in the "banded" regime (see Section 2.1). The problem is that in this regime, the estimated kernel is necessarily zero in the region $\{(s,t) \in [0,1]^2 \mid |s-t| > \delta\}$. This problem that is not present in the "blanket" regime, as soon as one full curve is observed, see Figure .



(a) Full curves (b) Fragments in the blanket regime (c) Fragments in the banded regime

Figure 2.2: Sample estimates of the covariance kernel in the different regimes

2.3 Curve extension

In [Delaigle and Hall, 2013], the authors take the approach of manually extend the fragments by gluing some of their parts. This procedure is carried on in the context of classification, but it can readily be expended to the plain estimation of the population mean and covariance.
The extension of a fragment to the right is done by iteratively gluing a small section of another randomly chosen nearby fragment to the right endpoint of the original fragment. Likewise for the extension to the left. A few steps of this procedure are shown in Figure 2.3.

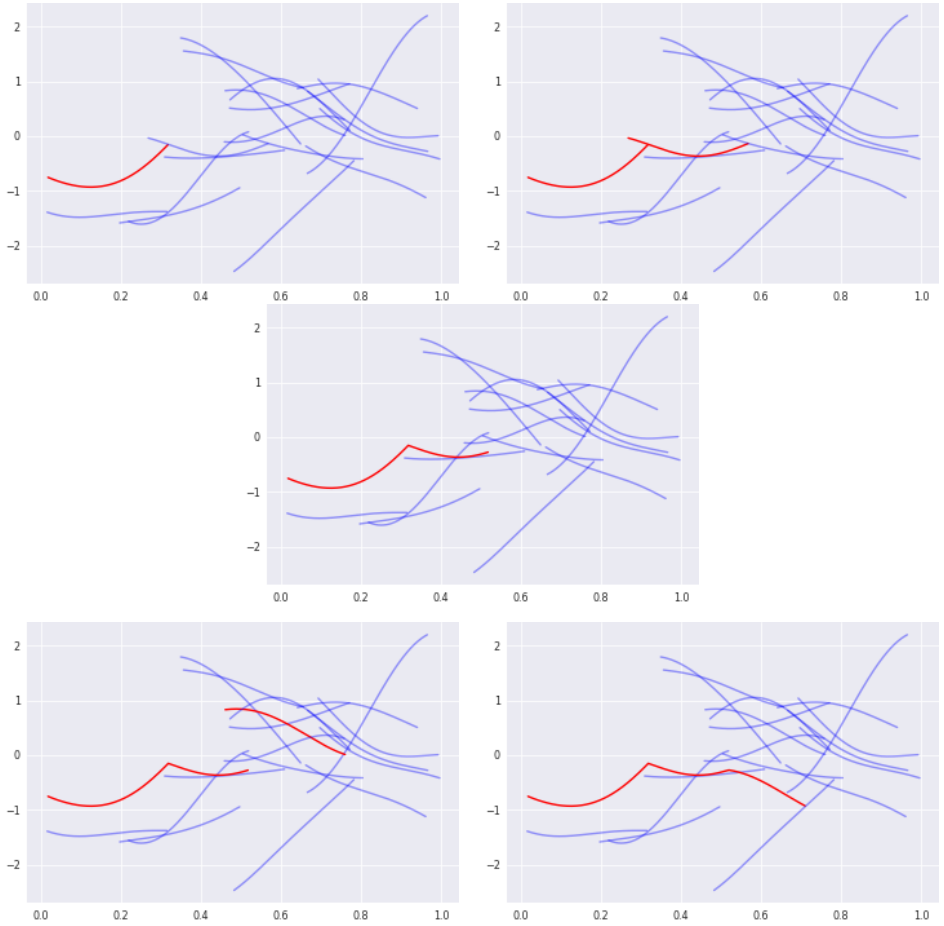
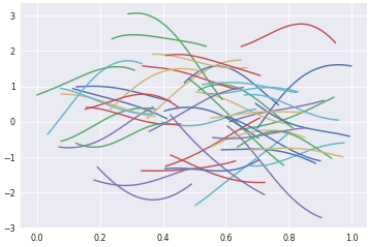
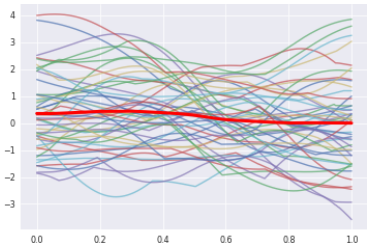


Figure 2.3: Illustration of a few steps of the gluing procedure

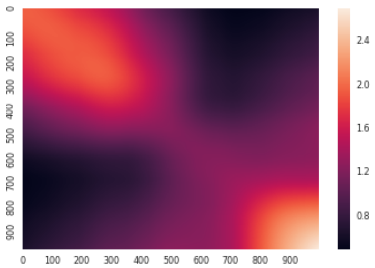
In this way, for each fragment $(X_i, \mathcal{O}_i)$, one gets a curve $\tilde{X}_i$ defined on the entire domain. From these full curves, the sample mean and covariance can be estimated. This estimation method works both in the "blanket" and the "banded" regime, since in both cases full curves are constructed. Examples of those estimates are shown in Figure 2.4.



(a) Fragments



(b) Reconstructed curves with their mean in red



(c) Estimated covariance kernel form the reconstructed curves

Figure 2.4: Example of mean and covariance estimation using the gluing procedure

2.4 Covariance recovery

2.4.1 Explain the method by Descary and Panaretos to recover the covariance from functional fragments

Chapter 3

Gaussian measures in Hilbert space

3.1 Gaussian measures in finite dimensions

3.1.1 Basic definitions

3.2 Gaussian measures in infinite dimensions

3.2.1 Definition and construction of gaussian measures in infinite dimensions. Basic properties.

3.3 The Feldman-Hajek theorem

3.3.1 Stating the F-H theorem with some intuition

Chapter 4

Classification of functional data

4.1 Linear and quadratic discriminant analysis for functional data

4.1.1 Explain LDA and QDA

4.2 Classification of functional fragments

4.2.1 Explain the work done by Kraus and Delaigle and Hall on the classification of functional fragments and my work

4.3 A probabilistic approach to classification

4.3.1 Explain the perfect classification phenomenon through the F-H theorem. If I manage to find something, also the classification of fragments.

Chapter 5

Numerical experiments

Bibliography

[Delaigle and Hall, 2013] Delaigle, A. and Hall, P. (2013). Classification using censored functional data. *Journal of the American Statistical Association*, 108(504):1269–1283.

[Descary and Panaretos, 2017] Descary, M.-H. and Panaretos, V. M. (2017). Recovering covariance from functional fragments.

[Kraus, 2015] Kraus, D. (2015). Components and completion of partially observed functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):777–801.