Big Data - 3
Functional Data Analysis

—

Retail sales data: a functional perspective

—

Rémi Perrichon

Introduction

You have been hired by a promising Turkish retail company as a junior data scientist. Before launching new products for the coming years, the company expects to have a good summary of its pricing policy for the past years and a better understanding of the dynamics of the sales. Your manager asks you to analyze some retail data for the period 2017-2019. You have access to raw daily sales data and you are free to use any statistical tools that are relevant for your study.

Yet, upper management has been very clear: shareholders will not be pleased with a complex deep learning approach they won't understand. In any case, the company can't give you access to high computing power. Your approach should be motivated, clearly explained and elegantly implemented. The upper management trusts your intellectual honesty. You may fail to master some advanced methods. You can make some simplifying assumptions. Yet, there is no way to sweep things under the carpet. Rather, the limitation of the data should be fully taken into account. Critical thinking is key. Any external references should be included in a bibliography and accessible upon request.

This project is the perfect occasion to apply tools from functional data analysis and to make the most of your knowledge in multivariate statistics and machine learning.

Brief data description

You have asked the data engineer to make a first cleaning of the data. Because products may enter or exit the catalogue any time of the year, you focus on all products that have been on sale for the entire period.

In the `sales.csv` file, daily sales data covering 2017-2019 are recorded.

Name of the variable	Meaning
product_id	The unique identifier of a product
store_id	The unique identifier of a store
date	Sales date (YYYY-MM-DD)
sales	Sales quantity
price	Product sales price
promo_bin_1	Type of applied promotion (categorical from missing value (no promotion) to “veryhigh”)

In the `product_hierarchy.csv` file, there are data containing the sizes of the products.

Name of the variable	Meaning
product_id	The unique identifier of a product
product_length	Length of product (cm)
product_depth	Depth of product (cm)
product_width	Width of product (cm)

Mission statement and deadlines

You are expected to provide 3 elements: a R script with your code (1/3) with a companion Microsoft Word or Latex small report (2/3) and some slides for your oral presentation (3/3).

The R script and the short companion report (deadline: 31/03/2023 23:59:00 UTC+1)

Your R script (no R markdown nor R notebook) should be concise yet very well organised. Don't hesitate to make sections and subsections. Comments are expected.

Sections in the code are fully explained in the companion report (Microsoft Word or Latex). Interpretations are expected in the report. The companion report can only be few pages but special care must be given to spelling and syntax. Make short and insightful sentences.

It is advised to follow the outline that is presented in this document: doing so, you won't forget an important part of the statistical analysis. You are free to organise the sections as you like. The questions are only here to guide you.

Slides for the oral presentation (deadline: 02/04/2023 23:59:00 UTC+1)

You have to send your slides by email (PDF format).

Oral defence (date: 03/04/2023 from 14:00:00 UTC+1 to 17:00:00 UTC+1)

Your presentation should be 15 minutes and highlights some part(s) of your analysis. You should focus on the original aspects of your work.

Outline of the analysis

Exploring data

Usual exploration of the raw data.

- Time, place, scope, economic context.
- Assumptions on the data collection, possible biases and limitations.
- Missing values.
- Descriptive statistics.
- Interest of the functional approach.

Data smoothing

You may focus on a single product to first highlight the method you have chosen.

- Discussion of smoothing strategies for sales and for the prices. Interest of the smoothing approach.
- Illustration.
- Generalisation to all products.

Registration

- Would registration be helpful here?
- Is the type of applied promotion a good landmark? Why? Why not?
- Implementation.
- Do descriptive statistics change with registration? Why? Why not?

FPCA

- Would Functional Principal Component Analysis be helpful here?
- Implementation.

Clustering

- Cluster the products according to the sales. Interpretation?
- Is a the clustering of the products according to the prices relevant? Interpretation?
- Are the sizes of the products or the applied promotions any helpful for the interpretation of the clusters?