biosensors.usc: Distributional Data Analysis Techniques for Biosensor Data

Marcos Matabuena

CiTIUS (Centro Singular de Investigación en Tecnoloxías Intelixentes)
Universidad de Santiago de Compostela

Oct 20th, 2022

Summary

- Personalized and digital medicine
- Limitations of existing wearable-device metrics: Compositional metrics and other summary measures
- biosensors.usc
- Recent work

Personalized and digital medicine

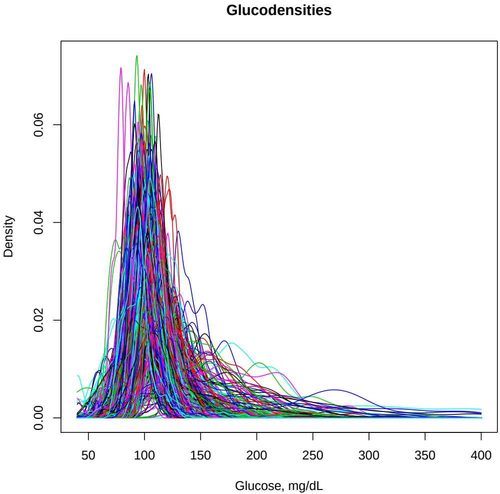
Motivation

- Despite enormous scientific and technological progress in the medical sciences in recent years, the prescription of treatments is far from being personalized in health systems worldwide. For example:
- Despite the known heterogeneity of type 2 diabetes and variable response to glucose-lowering medications, current evidence on optimal treatment is predominantly based on average effects in clinical trials rather than individual-level characteristics.
- Physical activity is the most effective, non-pharmacological, and low-cost intervention to combat a broad spectrum of diseases and decrease physiological decline with age. However, little research exists that introduces precision medicine approaches in this field.

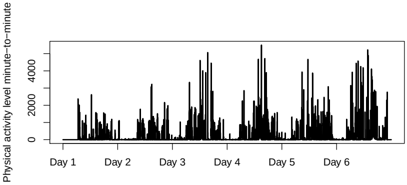
Motivation (new data opportunities)

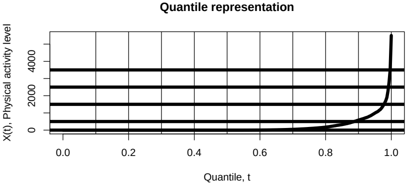
- For example, in the US, 99% of hospitals now use electronic health record systems (EHRs), compared to about 31% in 2003. In addition, in the US, 11% of office-based doctors used electronic records in 2006, whereas by 2017, 90% of office-based doctors used electronic records.
- Every day, wearable devices and smartphones generate billions of user-specific data points, but more than 95% of this newly created digital data remains unanalyzed.
- From a methodological and practical point of view, precision medicine is formalized as a dynamic optimization problem. However, this success depends on different scientific areas: biostatistics, applied mathematics, machine learning, operation research, and high-performance computing.

Statistical analysis in non-euclidean spaces appears in wearable-data analysis



Statistical analysis in non-euclidean spaces appears in wearable-data analysis





Limitations of existing wearable-device metrics: Compositional metrics and other summary measures

State-of-art wereable metrics

- Compositional metrics are probably the gold standard for summarizing information from biosensor data in multiple domains, such as in diabetes or in the case of physical activity with accelerometer devices.
- Traditional functional analysis methods are not applicable when patients are monitored under free-living conditions.
- Estimating specific moments of biological time series only provides a limited picture of the patient's behavior. Suppose that we summarize the information in the sample means!.

Limitations of compositional metrics

- Categorization of information into intervals defined by expert knowledge.
- The cut-off points may be highly dependent on the study population.
- Loss of information due to the restriction of summarizing the data in intervals.

biosensors.usc

C R package: Biosensors.usc

The R package biosensor.usc aims to provide a unified and user-friendly framework for using new distributional representations of biosensors data in different statistical modeling tasks: regression models, hypothesis testing, cluster analysis, visualization, and descriptive analysis. Distributional representations are a functional extension of compositional time-in-range metrics and we have used them successfully so far in modeling glucose profiles and accelerometer data. However, these functional representations can be used to represent any biosensor data such as ECG or medical imaging such as fMRI.

C.1 Installation

You can install this package from source code using the devtools library:

```
devtools::install_github("glucodensities/biosensors.usc@main",

type = "source")
```

C.1.1 Quick start

The purpose of this section is to give users a general sense of the package, including the components, what they do and some basic usage. We will briefly go over the main functions, see the basic operations and have a look at the outputs. Users may have a better idea after this section what functions are available. More details are available in the package documentation.

First, we load the biosensors.usc package:

```
library(biosensors.usc)
```

C.1.2 Package example

This example is extracted from the paper [37].

We include part of this data set in the inst/exdata folder. This data set has two different types of files. The first one contains the functional data, which csv files must have long format with, at least, the following three columns: id, time, and value. The id identifies the individual, the time indicates the moment in which the data was captured, and the value is a monitor measure:

```
file1 = system.file("extdata", "data_1.csv", package = "biosensors.usc")
```

The second type contains the clinical variables. This csv file must contain a row per individual and must have a column id identifying this individual:

```
file2 = system.file("extdata", "variables_1.csv", package = "biosensors.usc")
```

From these files, biosensor data can be loaded as follow:

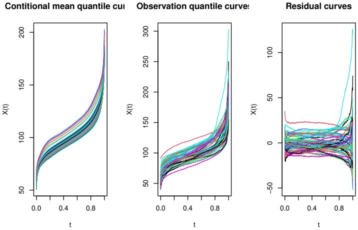
```
data1 = load_data(file1, file2)
class(data1)
#> [1] "biosensor"
names(data1)
#> [1] "data"      "densities" "quantiles" "variables"
```

The load_data function returns a biosensor object. This object contains a data frame with biosensor raw data, a functional data object (fdata) with a non-parametric density estimation, a functional data object (fdata) with the empirical quantile estimation, and a data frame with the covariates.

C.1.2.1 Wasserstein regression and prediction

You can call the Wasserstein regression, using as predictor the distributional representation and as response a scalar outcome. In this example, we use the previously loaded biosensor data and the BMI covariate:

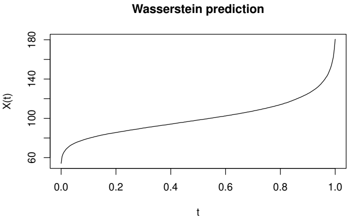
```
regm = regmod_regression(data1, "BMI")
```



As result, this function returns the fitted regression and plots the residuals of the curves against the fitted values. In addition, the function plots the confidance band of the mean values.

We can obtain the regression prediction from a kxp matrix of input values for regressors for prediction, where k is the number of points we do the prediction and p is the dimension of the input variables:

```
xpred = as.matrix(25)
pred = regmod_prediction(regm, xpred)
```



C.1.2.2 Ridge regression

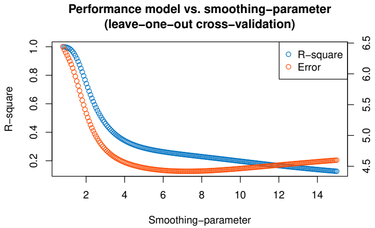
Call the ridge regression as follows, using as predictor the distributional representation and as response a scalar outcome:

```
ridg = ridge_regression(data1, "BMI")
```

C.1.2.3 Nadaraya-Watson regression and prediction

Use the following function to obtain the functional non-parametric Nadaraya-Watson regression with 2-Wasserstein distance, using as predictor the distributional representation and as response a scalar outcome:

```
nada = nadayara_regression(data1, "BMI")
```



Use the previously computed Nadaraya-Watson regression to obtain the regression prediction given the quantile curves:

```
npre = nadayara_prediction(nada, t(colMeans(data1quantilesdata)))
```

C.1.2.4 Hypothesis testing

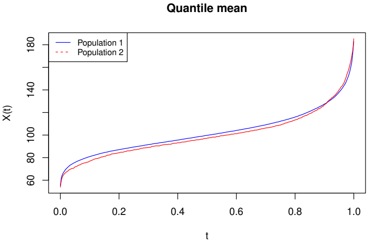
We can perform hypothesis testing between two random samples of distributional representations to detect differences in scale and localization (ANOVA test) or distributional differences (energy distance).

Let's load first another sample:

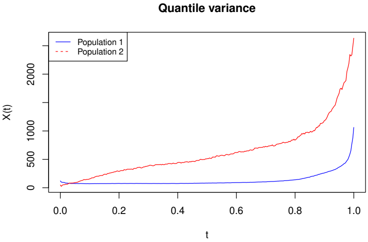
```
file3 = system.file("extdata", "data_2.csv", package = "biosensors.usc")
file4 = system.file("extdata", "variables_2.csv", package = "biosensors.usc")
data2 = load_data(file3, file4)
```

Then call the following function:

```
htest = hypothesis_testing(data1, data2)
```



```
#> Warning: executing %dopar% sequentially: no parallel backend registered
```



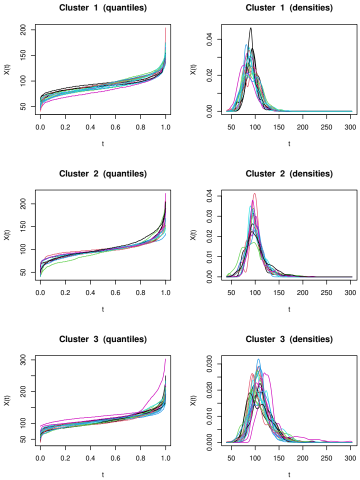
The function will plot the quantile mean and the quantile variance of the two populations. The corresponding p-values of the ANOVA test and distributional differences are stored in the following names:

```
print(htestenergy_pvalue)
#> [1] 0.00990099
print(htestanova_pvalue)
#> [1] 0.0003094763
```

C.1.2.5 Clustering

Call the energy clustering with Wasserstein distance using quantile distributional representations as covariates:

```
clus = clustering(data1, clusters=3)
```

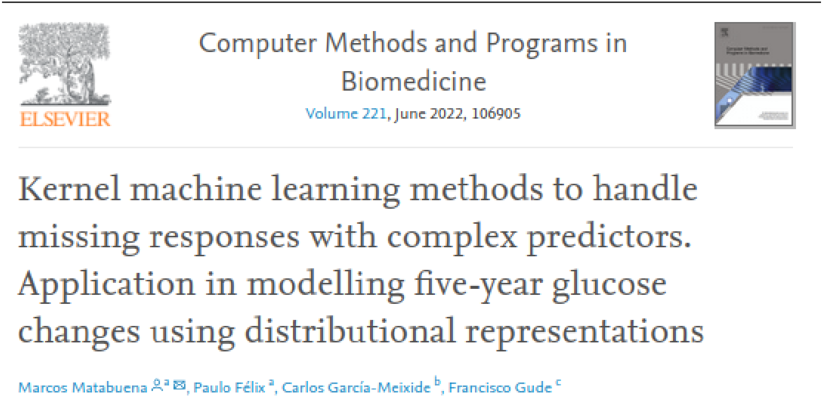


The function also plots the clusters of quantiles and densities.

We can also use the previously computed clustering to obtain the clusters of another set of objects calling the following function:

```
assignments = clustering_prediction(clus, data1quantilesdata)
print(assignments)
#>  [1] 1 1 1 1 1 2 2 1 1 1 1 2 2 3 1 2 1 3 2
#>  [20] 3 1 2 1 3 2 1 3 1 3 2 2 2 1 1 1 1 2 1
#> [39] 3 2 3 3 3 3 2 3 2 1 3 2 3 1
```

Missing data I



ELSEVIER

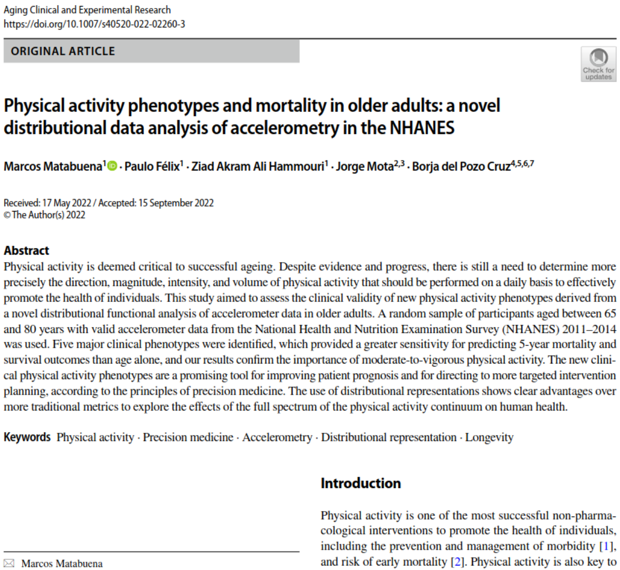
Computer Methods and Programs in Biomedicine

Volume 221, June 2022, 106905

Kernel machine learning methods to handle missing responses with complex predictors. Application in modelling five-year glucose changes using distributional representations

Marcos Matabuena [a], Paulo Félix [a], Carlos García-Meixide [b], Francisco Gude [c]

Novel applications



Aging Clinical and Experimental Research
https://doi.org/10.1007/s40520-022-02260-3

ORIGINAL ARTICLE

Check for updates

Physical activity phenotypes and mortality in older adults: a novel distributional data analysis of accelerometry in the NHANES

Marcos Matabuena[1] · Paulo Félix[1] · Ziad Akram Ali Hammouri[1] · Jorge Mota[2,3] · Borja del Pozo Cruz[4,5,6,7]

Received: 17 May 2022 / Accepted: 15 September 2022

Abstract

Physical activity is deemed critical to successful ageing. Despite evidence and progress, there is still a need to determine more precisely the direction, magnitude, intensity, and volume of physical activity that should be performed on a daily basis to effectively promote the health of individuals. This study aimed to assess the clinical validity of new physical activity phenotypes derived from a novel distributional functional analysis of accelerometer data in older adults. A random sample of participants aged between 65 and 80 years with valid accelerometer data from the National Health and Nutrition Examination Survey (NHANES) 2011–2014 was used. Five major clinical phenotypes were identified, which provided a greater sensitivity for predicting 5-year mortality and survival outcomes than age alone, and our results confirm the importance of moderate-to-vigorous physical activity. The new clinical physical activity phenotypes are a promising tool for improving patient prognosis and for directing to more targeted intervention planning, according to the principles of precision medicine. The use of distributional representations shows clear advantages over more traditional metrics to explore the effects of the full spectrum of the physical activity continuum on human health.

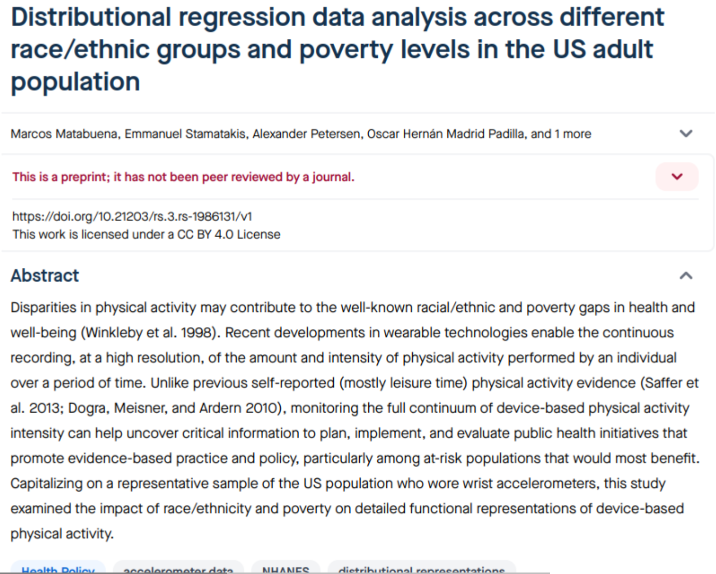
Keywords Physical activity · Precision medicine · Accelerometry · Distributional representation · Longevity

Introduction

Physical activity is one of the most successful non-pharmacological interventions to promote the health of individuals, including the prevention and management of morbidity [1], and risk of early mortality [2]. Physical activity is also key to

✉ Marcos Matabuena

Novel applications



Distributional regression data analysis across different race/ethnic groups and poverty levels in the US adult population

Marcos Matabuena, Emmanuel Stamatakis, Alexander Petersen, Oscar Hernán Madrid Padilla, and 1 more

This is a preprint; it has not been peer reviewed by a journal.

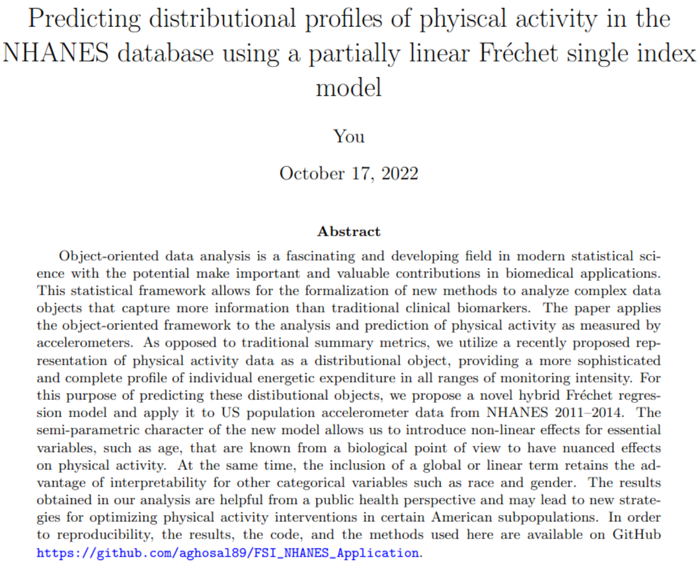
https://doi.org/10.21203/rs.3.rs-1986131/v1

Abstract

Disparities in physical activity may contribute to the well-known racial/ethnic and poverty gaps in health and well-being (Winkleby et al. 1998). Recent developments in wearable technologies enable the continuous recording, at a high resolution, of the amount and intensity of physical activity performed by an individual over a period of time. Unlike previous self-reported (mostly leisure time) physical activity evidence (Saffer et al. 2013; Dogra, Meisner, and Ardern 2010), monitoring the full continuum of device-based physical activity intensity can help uncover critical information to plan, implement, and evaluate public health initiatives that promote evidence-based practice and policy, particularly among at-risk populations that would most benefit. Capitalizing on a representative sample of the US population who wore wrist accelerometers, this study examined the impact of race/ethnicity and poverty on detailed functional representations of device-based physical activity.

Health Policy accelerometer data NHANES distributional representations

New regression models



Predicting distributional profiles of phyiscal activity in the NHANES database using a partially linear Fréchet single index model

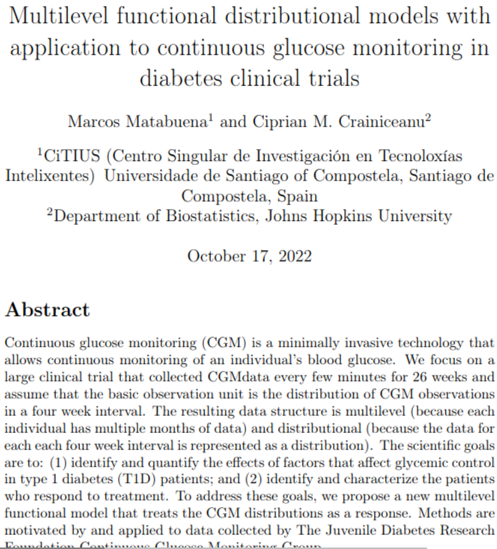
You

October 17, 2022

Abstract

Object-oriented data analysis is a fascinating and developing field in modern statistical science with the potential make important and valuable contributions in biomedical applications. This statistical framework allows for the formalization of new methods to analyze complex data objects that capture more information than traditional clinical biomarkers. The paper applies the object-oriented framework to the analysis and prediction of physical activity as measured by accelerometers. As opposed to traditional summary metrics, we utilize a recently proposed representation of physical activity data as a distributional object, providing a more sophisticated and complete profile of individual energetic expenditure in all ranges of monitoring intensity. For this purpose of predicting these distibutional objects, we propose a novel hybrid Fréchet regression model and apply it to US population accelerometer data from NHANES 2011–2014. The semi-parametric character of the new model allows us to introduce non-linear effects for essential variables, such as age, that are known from a biological point of view to have nuanced effects on physical activity. At the same time, the inclusion of a global or linear term retains the advantage of interpretability for other categorical variables such as race and gender. The results obtained in our analysis are helpful from a public health perspective and may lead to new strategies for optimizing physical activity interventions in certain American subpopulations. In order to reproducibility, the results, the code, and the methods used here are available on GitHub https://github.com/aghosal89/FSI_NHANES_Application.

Multilevel models



Multilevel functional distributional models with application to continuous glucose monitoring in diabetes clinical trials

Marcos Matabuena[1] and Ciprian M. Crainiceanu[2]

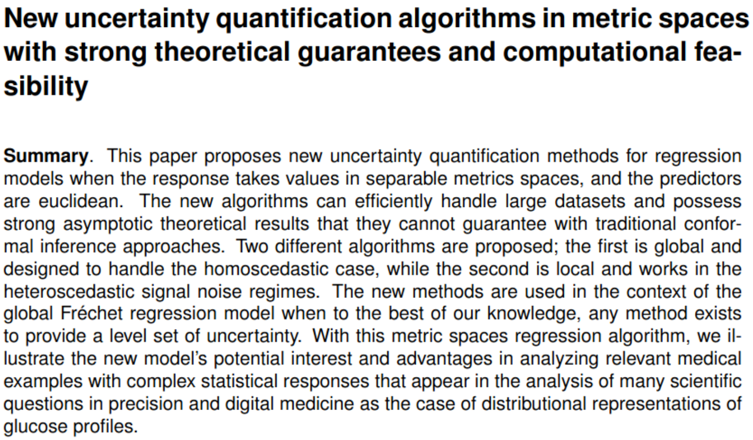
[1]CiTIUS (Centro Singular de Investigación en Tecnoloxías Intelixentes) Universidade de Santiago of Compostela, Santiago de Compostela, Spain
[2]Department of Biostatistics, Johns Hopkins University

October 17, 2022

Abstract

Continuous glucose monitoring (CGM) is a minimally invasive technology that allows continuous monitoring of an individual's blood glucose. We focus on a large clinical trial that collected CGMdata every few minutes for 26 weeks and assume that the basic observation unit is the distribution of CGM observations in a four week interval. The resulting data structure is multilevel (because each individual has multiple months of data) and distributional (because the data for each each four week interval is represented as a distribution). The scientific goals are to: (1) identify and quantify the effects of factors that affect glycemic control in type 1 diabetes (T1D) patients; and (2) identify and characterize the patients who respond to treatment. To address these goals, we propose a new multilevel functional model that treats the CGM distributions as a response. Methods are motivated by and applied to data collected by The Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Group

Uncertainty quantification



New uncertainty quantification algorithms in metric spaces with strong theoretical guarantees and computational feasibility

Summary. This paper proposes new uncertainty quantification methods for regression models when the response takes values in separable metrics spaces, and the predictors are euclidean. The new algorithms can efficiently handle large datasets and possess strong asymptotic theoretical results that they cannot guarantee with traditional conformal inference approaches. Two different algorithms are proposed; the first is global and designed to handle the homoscedastic case, while the second is local and works in the heteroscedastic signal noise regimes. The new methods are used in the context of the global Fréchet regression model when to the best of our knowledge, any method exists to provide a level set of uncertainty. With this metric spaces regression algorithm, we illustrate the new model's potential interest and advantages in analyzing relevant medical examples with complex statistical responses that appear in the analysis of many scientific questions in precision and digital medicine as the case of distributional representations of glucose profiles.

The End