*5th International Conference on Advanced Research Methods and Analytics (CARMA2023)*
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16467

# Density modelling with functional data analysis

**Stefano A. Gattone[1], Tonio Di Battista[1]**
[1]DISFIPEQ Department, University "G. d'Annunzio" of Chieti-Pescara, Italy

***Abstract***

*Recent technological advances have eased the collection of big amounts of data in many research fields. In this scenario density estimation may represent an important source of information. One dimensional density functions represent a special case of functional data subject to the constraints to be non-negative and with a constant integral equal to one. Because of these constraints, a naive application of functional data analysis (FDA) methods may lead to non-valid results. To solve this problem, by means of an appropriate transformation densities are embedded in the Hilbert space of square integrable functions where standard FDA methodologies can be applied.*

***Keywords:*** *Bayes space; Density; Functional data analysis; Transformation approach.*

## 1. Introduction

This work deals with density modeling using functional data analysis. (Ramsay and Silvermann, 2005). In scenario with big amounts of data collection, probability density functions can provide more information than single summary statistics.

One of the main goals in statistical data analysis is to associate the change of (a function of) some response variable $y$ with a set of covariates $x$. The most common tool for this is mean regression which focuses on the conditional expectation *of* $y$ *given* $x$. Quantile regression models investigate specific quantiles of the conditional distribution of the response. By modelling the entire probability distribution of the response, density regression methods consider the impact of the covariates on the entire distribution. Densities could represent the data atoms of interest such as yearly income distribution, population age and mortality distributions across different countries.

Probability density functions (pdfs) represent a special case of functional data since they must satisfy the constraints of being non-negative everywhere and present a constant integral equal to one. Standard functional data analysis (FDA) methods cannot be naively applied without considering such constraints. To address this issue several strategies can be found in the literature.

One strand of literature represents densities as elements of the so-called Bayes space starting from the Aitchison geometry valid for compositional data (Aitchison, 1982). In this setting, pdfs are represented by a centred log-ratio transformation which represents an isometric isomorphism between the Bayes space of pdfs and the Hilbert space (Hron et al, 2016).

Another approach is envisaged by Petersen and Muller (2016) where the pdfs are mapped into a linear functional space through a suitably chosen transformation. Established methods for Hilbert space valued data can be applied to the transformed functions and the results are moved back into the density space by means of the inverse transformation. Examples of transformations are the log-hazard transformation and the log-quantile density transformation. The view is completed by considering the objected-oriented analysis of densities where spaces are equipped with metrics such as the Wasserstein or the Fisher-Rao providing a manifold structure on probability distributions. Within this framework, tangent space structures need to be defined to facilitate computations (Petersen and Muller, 2019).

## 2. Density functions as constrained functional data

A functional variable is defined as a random variable $f$, taking values in an infinite functional space, the Hilbert space of square integrable functions equipped with the usual inner product and norm:

$$H(t) = \left\{ f: T \to \mathbb{R} \ such\ that \int_T f(t)^2\, dt < \infty \right\} \quad (1)$$

with $\langle f, g \rangle = \int f(t)g(t)\, dt$ and $\|f\| = \left\{\int_T f^2(t)\, dt\right\}^{\frac{1}{2}}$.

We are interested in the case where the observed functions are density functions. We denote with *D* the functional space of density functions. In this space functions are positive and integrate up to 1 as described in equation (2):

$$D(t) = \left\{ f: T \to \mathbb{R} \ such\ that\ f(t) > 0 \text{ and } \int_T f(t)\, dt = 1 \right\} \quad (2)$$

We assume the data consists of a sample of *n* random density functions. In many situations, the densities themselves will not be directly observed. Instead, a sample of data that are generated by the random density is available. Thus, there are two random mechanisms at work: the first generates the sample of densities and the second generates the samples of data. Typically the first step in working with functional data is the use of basis expansion and penalized smoothing. Estimation is developed, for example, in the natural cubic splines framework:

$$\sum_j \left[y_j - f(t)\right]^2 + \lambda \int [D^2 f(t)]^2 dt \quad (3)$$

where $y_j$ are the observed discrete data points which must be converted to a functional data object *f*. The constant lambda is the smoothing parameter with larger values resulting in smoother fits. Now, imagine imposing on the estimated function *f* some constraints. The constrained curves cannot be treated as vectors in the Hilbert space since a plain basis expansion of the curves does not guarantee the fulfilment of the constraints. In other words, the problem is to simultaneously smooth nonlinear structure in data and incorporate constraints.

## 3. The w-transform

Let Y have an un unknown positive density function. Following Ramsay and Silvermann (2005) we can write its log-density function in the form $w - C(w)$ where

$$C(h) = log \int exp[w(y)] dy \quad (4)$$

The corresponding log-likelihood function is given by

$$l(w, \boldsymbol{Y}) = w(\boldsymbol{Y}) - C(w) \tag{5}$$

Note that $w(\boldsymbol{y})$ is not constrained in any way. In this way a constrained problem is transformed into an unconstrained one that reduces to the modelling of $w(\boldsymbol{y})$. The modeling of $w(\boldsymbol{y})$ can be obtained by using a flexible nonparametric estimator based on spline basis functions. Once the estimator is obtained, we are able to map the densities into the Hilbert space since the functions $w(\boldsymbol{y})$ are free of constraints. Our proposal is to apply linear FDA methods in the transformed linear space and eventually results on the linear space are mapped back into the density space by means of an appropriate inverse map.

## 4. Applications

In many application fields, densities are the data atoms of interest such as yearly income distribution, population age and mortality distributions across different countries or distribution of cross-sectional financial returns of different firms or different markets.

Data analysis frequently concerns itself with associating the change in a function of some response variable *y* with a set of covariates *x*. The most common tool for this is mean regression which focuses on the conditional expectation *of y given x*. This prevents inference about other parts of the conditional density. Quantile regression models investigate specific quantiles of the conditional distribution of the response. In such circumstances, individual quantiles are being targeted as proxies of the distribution. By modelling the entire probability distribution of the response, density regression methods perform a substantially harder task than mean and quantile regression. In doing so, one can consider the impact of the covariates on the entire distribution.

A naïve application of the function-on-scalar regression or the function-on-function regression model (Ramsay and Silvermann, 2005) would not guarantee the estimated response to fulfill the definition of a density. Similarly, to compositional regression (Talskà et al., 2018), an alternative could be applying the functional regression model on the unconstrained functions w(t) in eq. (5) and then the parameters estimates are mapped back to the density space applying the inverse transformation. In contrast to the estimates resulting from the naïve functional regression model, the estimates are bona fide density function.

## References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B*, 44, 139-177.

Hron, K., Menafoglio, M. Templ, M., Hruzova, K. & Filzmoser, P. (2016). Simplicial principal component analysis for density functions in Bayes space. *Computational Statistics & Data Analysis*, 94, 330-350.

Petersen, A., & Muller, H. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics*, 32 (1), 183-218.

Petersen, A., & Muller, H. (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics*, 47 (2), 691-719.

Ramsay, J.O. & Silverman, B.W. (2005). *Functional Data Analysis*, 2nd edn. New York: Springer.

Talskà, R, Menafoglio, A., Machalovà, J., Hron, K. & Fiserovà, E. (2018). Compositional regression with functional response. *Computational Statistics and Data Analysis*, 123, 66-85.