Functional Data Analysis

Valentin Patilea

MSc *Smart Data Science*, ENSAI

Fall 2024

- All the materials will be on Moodle
- Office 270
- Office hours: by appointment
- email: valentin.patilea@ensai.fr
- Evaluation:
  - home assignment (20%)
  - article presentation (30%)
  - written final exam (70%)

Avant-propos

- More and more applications produce data under the form of *curves* or/and *images*
  - sometimes they are called *signals*

- The stake is to extract the information carried by these curves/images/signals in a data-driven way, and proceed to inference, prediction, classification,...

- This is the purpose of Functional Data Analysis (FDA)

- A (closely) related field/community: Signal Processing
  - at the frontier of Computer Science and Applied Mathematics

Some resources

- Books: Ramsay and Silverman (2002, 2005), Ferraty & Vieu (2006), Ferraty & Romain (2010), Horváth & Kokoszka (2012), Kokoszka & Reimherr (2017)
- Reviews articles: Wang et al. (2016), Reiss et al. (2016)
- R packages: fda (Ramsay & Silverman (2005)), fda.usc (Febrero-Bande & Oviedo de la Fuente (2012)), ftsa (Shang et al. (2013)), refund (Crainiceanu et al. (2012)), GPFDA (Shi, Cheng & Konzen (2021))...
- https://cran.r-project.org/view=FunctionalData
- Useful books on smoothing: Wasserman (2006), Fan & Gijbels (1996)

Agenda

- Technology advances provide new type of data
  - sensors (cars, airplanes,...)
  - medical devices (cardiograms, fMRI, Blood Pressure, oxygen or glucose devices,...)
  - environemental devices (daily temperature, wind speed, solar radiation, pollution levels,...) and energy consumption
- The observation unit (entity), the datum, could be one or several curves, image(s), or several such objects
- In some situations one could be interested by the curves or/and the variations (derivatives) along the curves
- One can try to summarize such complex observation units by few indicators/statistics (mean/median, variance, sd,...), but one clearly faces a risk loose valuable information

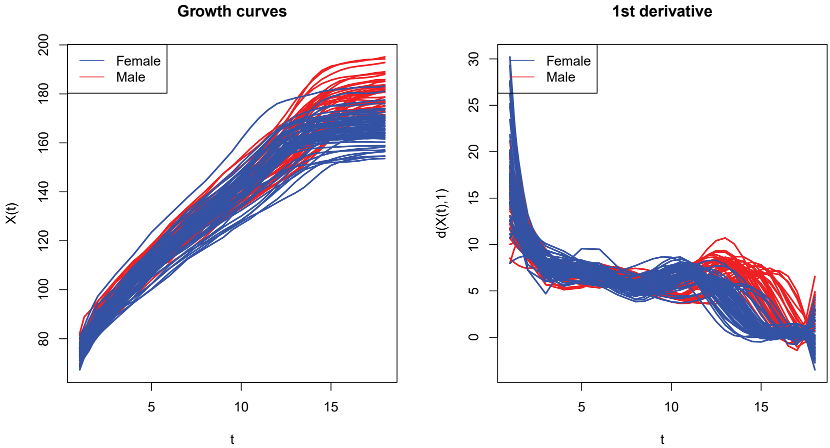
- A better idea is to keep the observation unit as collected, and to model your data as realizations in a suitable space of such complex objects (space of curves, images,...)
- Usual paradigm in statistics: data are (in)dependent realizations of some variable (function, map,...)

$$X : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{F})$$

- When $\mathcal{X}$ is a space of (vectors of) curves/images/signals, we have a Functional Data problem
- Functional Data Analysis (FDA) deals with the statistical description and modeling of samples of random variable taking values in spaces of functions

Examples

Growth Data, height vs. age : 54 females, 39 males (Berkeley Growth Study)



Tecator Data (levels)

215 spectrometric curves of meat samples also with Fat, Water and Protein contents obtained by analytic procedures.

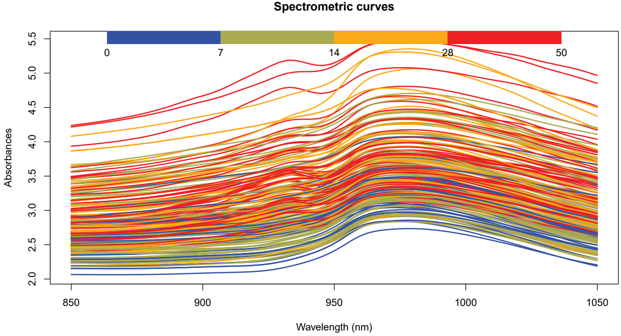


Figure: Coloured by fat content (blue=low, red=high)

Source: fda.usc

Tecator Data (levels and derivatives)

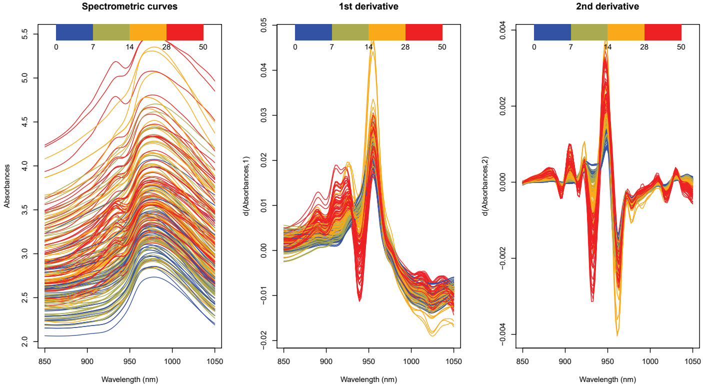
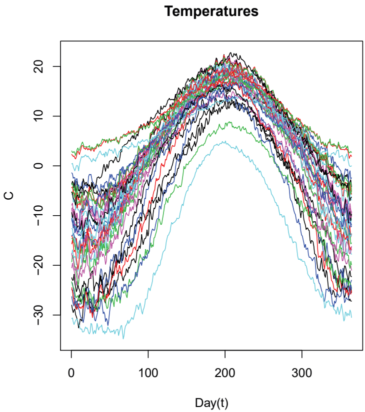
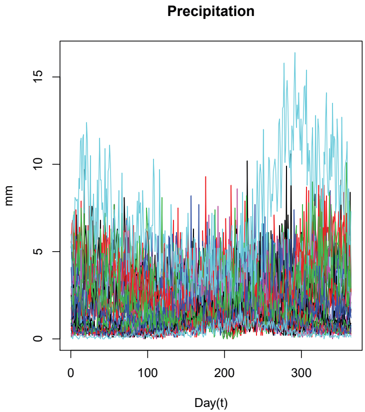


Figure: Coloured by fat content (blue=low, red=high)

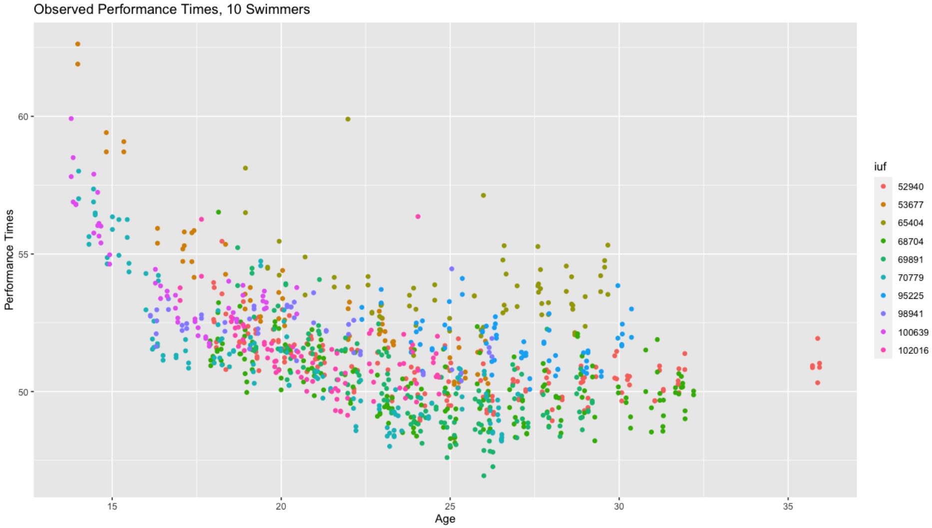
Source: fda.usc

Multivariate Functional Data: Canadian Weather data (35 stations, average over 30 years)

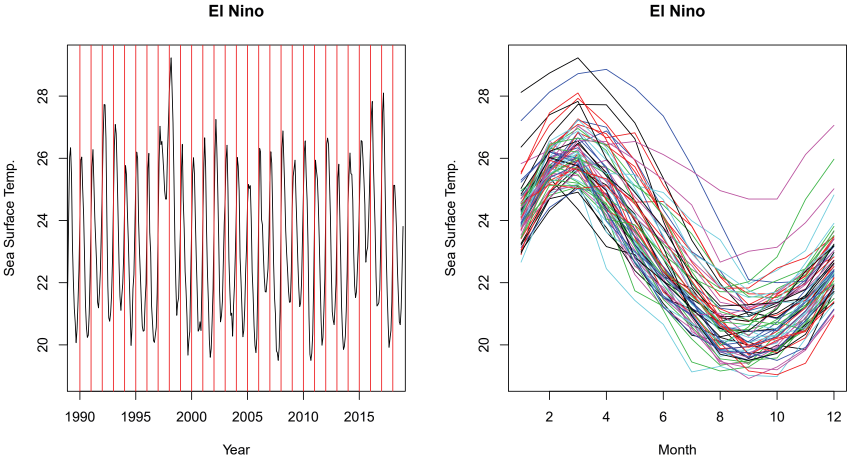




Swimming performance data (French athletes)

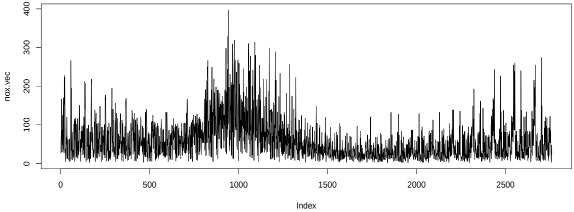


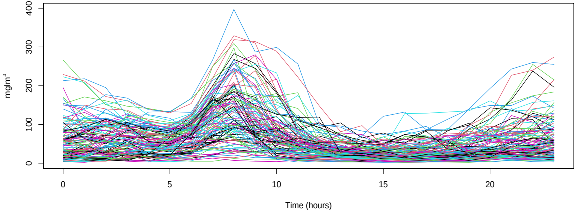
Functional Time Series Data



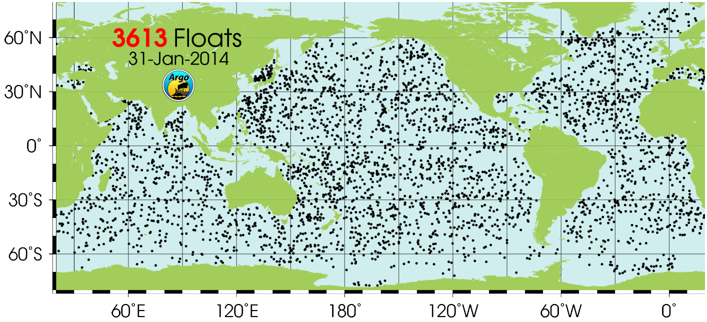
Hourly levels of nitrogen oxides in Poblenou (Barcelona)

127 daily records every hour (2005/01/06-2005/06/26)





Argo Ocean Temperature and Salinity Profiles (2D and 3D data)



Images as observation units: FDA vs. Signal Processing

Why FDA is a distinctive topic?

- Is FDA a special case of multivariate analysis?
- Yes, to a little extent! In many cases, at the end of the day, the curves (in a broad sense) are summarized by vectors of *scores*.
  - To these vectors one applies standard inference, prediction,..., multivariate statistics methods and ML algorithms
- *However,* the way the scores are built matters! Building the scores in an ‘optimal’ way remains an open problem
- Moreover, in FDA the derivatives of the curves are the quantities of interest. This is a specific feature for FDA

- ‘Vectorizing’ functional data (if possible!) is likely not a good idea!
  - part of the information is useless, the curves could be represented in much more parsimonious ways when there is an underlying structure
  - some methods and algorithms performs poorly when the components of the vectors are highly correlated
  - one may need very large amounts of curves to learn with modern algorithms
    - taking into account the underlying functional data structure, one may be more efficient with small and moderate sample sizes

FDA challenges and opportunities

- In real applications, functional data are :
  - discretely observed (possibly at random points, which may be sparsely distributed)
  - noisy measurements
- The challenge : find suitable representations (reconstruction) of the data
  - the quality of the representation (reconstruction) will influence the quality of the subsequent inference/prediction methods
  - parsimonious representations of the data are preferable
  - the 'optimal' representations/reconstruction could depend on the final purpose
- The opportunity, and also the main difference with respect to standard nonparametric statistics, come from the replication nature of the data
  - several (sometimes many) curves/images/signals are observed

Take away

- More and more applications produce functional data, where the unit of observation (the datum) is a curve, a surface or vectors of such objects
    - by abuse, an observation (a datum) will be usually called *curve*
- The nature of the functional data is distinct from that of time series and multivariate analysis
- Functional data carry information along the curves and among the curves
- Sometimes it can be difficult to recover the curves from the available data

Agenda

- Lecture 1:
  - Introduction
  - Representation of functional data
- Lecture 2:
  - Representation of functional data (cont'd)
- Lecture 3:
  - Mean and Covariance functions (ideal case)
- Lecture 4:
  - Excursion into the Smoothing world
- Lecture 5:
  - Nonparametric FDA
- Lecture 6:
  - Predictive models for functional data
- Lecture 7:
  - Predictive models for functional data (cont'd)

Agenda

- Lecture 8: Smarties' presentations
  - By group of 2 or 3
  - Articles available on Moodle
  - The groups and the article choice should be made the beginning of the second week of the module

Agenda

Functional data and stochastic processes

- To each observation unit corresponds a curve
- Let
$$X(t) \in \mathbb{R}, \qquad t \in \mathcal{T} \text{ with } \mathcal{T} \text{ compact domain,}$$
be such a generic curve. Typically, $\mathcal{T} = [0, 1]$
- The ideal sample is a set of independent random copies of $X(t)$: if N denotes the sample size, the ideal data are
$$X_1(t), \ldots, X_N(t), \quad t \in \mathcal{T}.$$
This are iid functional data, fully observed (for all $t \in \mathcal{T}$)
- First idea: each curve is a path (trajectory) of a stochastic process
$$X : \Omega \times \mathcal{T} \to \mathbb{R}, \qquad (\omega, t) \to X(\omega, t),$$
where $(\Omega, \mathcal{A})$ is a measurable space and $\mathcal{T}$ and $\mathbb{R}$ are considered with the Borel $\sigma-$field

Functional data in reality

- Although the textbooks usually consider ideal data, in applications the curves:
  - are **not** observed at any $t \in \mathcal{T}$;
  - are **often** observed with error.

- Usually, there is a gap between the methods described in the textbooks and the available data. This gap is quite often ignored, leading to incorrect or sub-optimal inference

- This aspect will be reconsidered later in the lecture.

- In addition, the realizations $X_1(\cdot), \ldots, X_n(\cdot)$ may not be independent! This lead researcher to study functional time series.
  - If not stated differently, herein we assume that data correspond to a iid sample of sample paths!

The standard representation of functional data

- The previous stochastic process representation is too general, it does not allow to take into the structure one usually expect in FDA (continuity of the curves, regularity,...)
- In FDA, it is usually supposed that the paths of the stochastic process X belong to some space of function defined on $\mathcal{T}$
- We will need these spaces of functions to be endowed with some suitable algebraic and topological structure

- Given a set $\mathcal{S}$, a map $d : \mathcal{S} \times \mathcal{S} \to [0, \infty)$ is called a *distance* or a *metric* (on $\mathcal{S}$) if $\forall x, y, z \in \mathcal{S}$,
  1. $d(x, y) = 0 \Leftrightarrow x = y$
  2. $d(x, y) = d(y, x)$
  3. $d(x, z) \leq d(x, y) + d(y, z)$

  - If $d(\cdot, \cdot)$ satisfies only conditions 2 and 3, we call it *semi-distance* or *semi-metric*

- Let $\mathcal{S}$ be a vector space over the real numbers. A function $\| \ \| : \mathcal{S} \to [0, \infty)$ is called a *norm* if $\forall x, y \in \mathcal{S}$ and $\alpha \in \mathbb{R}$,
  1. $\|x\| = 0 \Leftrightarrow x = 0$
  2. $\|\alpha x\| = |\alpha| \|x\|$
  3. $\|x + y\| \leq \|x\| + \|y\|$

  - If $\| \cdot \|$ satisfies only conditions 2 and 3, we call it *semi-norm*

- A (semi-)norm induces a (semi-)distance

Remember

- A space endowed with a metric is called a *metric space*
  - A metric defines open sets and thus a metric topology
- A vector space endowed with a norm is called a *normed space*
  - A norm induces a distance, which defines the open sets; one obtains a so-called topological vector space
- A complete normed space is called a *Banach space*
  - a normed space where the Cauchy sequences converge

Agenda

Spaces of continuous functions

- In many applications in FDA, it is supposed that the curves belong to the space of real-valued continuous functions defined on $\mathcal{T}$:

$$\mathcal{C}(\mathcal{T}) = \{f : \mathcal{T} \to \mathbb{R}, \quad f \text{ continuous function}\}$$

- Henceforth, if not stated differently, we suppose the curves belong to $\mathcal{C}(\mathcal{T})$
- The space $\mathcal{C}(\mathcal{T})$ could be endowed with several metrics or norms
  - uniform norm – it becomes a Banach space
  - the integral of the absolute value (L^1) norm – it becomes Banach
  - Hausdorff metric – defined as the greatest of all the Euclidean distances from a point in one curve to the closest point in the other curve
- The choice of the metric matters, it should be adapted to the application and the purposes of the analysis
  - different metric choices could lead to different results/conclusions !!

Hausdorff metric

- Let $f, g \in \mathcal{C}(\mathcal{T})$ and let $\|\cdot\|$ denote the Euclidean norm in $\mathbb{R}^2$
- The Hausdorff distance between f and g is defined as

$$d_H(f,g) = \max\left\{\sup_{t\in\mathcal{T}}\inf_{s\in\mathcal{T}} \|(t,f(t)) - (s,g(s))\|\ ,\right.$$
$$\left.\sup_{t\in\mathcal{T}}\inf_{s\in\mathcal{T}} \|(s,f(s)) - (t,g(t))\|\right\}$$

$L^p(\mathcal{T})$ spaces

- Consider μ a measure on the real line
  - typically μ is the Lebesgue measure
- Let $1 \leq p < \infty$
- For any measurable function $f : \mathcal{T} \to \mathbb{R}$, consider

$$\|f\|_p = \left(\int_{\mathcal{T}} |f|^p d\mu \right)^{1/p}$$

- Then we define the space

$$L^p(\mathcal{T}) = L^p(\mathcal{T}; \mu) = \{f : \mathcal{T} \to \mathbb{R}, \quad f \text{ measurable and } \|f\|_p < \infty\}$$

- Similarly, the space $L^\infty(\mathcal{T})$ is defined using

$$\|f\|_\infty = \inf\{M \geq 0 : |f(t)| \leq M \text{ for } \mu\text{−almost all } t \in \mathcal{T}\}$$

Properties

- Hereafter, if not stated differently, $\mathcal{T} \subset \mathbb{R}$ is a compact interval and μ is the Lebesgue measure
- For any $1 \leq p \leq \infty$, the space $L^p(\mathcal{T})$ is a Banach space, and $\mathcal{C}(\mathcal{T}) \subset L^p(\mathcal{T})$
- The case $p = 2$ is particularly important
  - In almost all applications, the paths (trajectories) of the process X, which was introduced to model the functional data, are considered as elements of $L^2(\mathcal{T})$

Agenda

Inner product, Hilbert space

- Let $\mathcal{H}$ a vector space

- A (real) *inner product* (also called *scalar product*) is a function $\langle , \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ such that, $\forall x, y, z \in \mathcal{H}$ and $\forall \alpha, \beta \in \mathbb{R}$,
  1. $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0$ iff $x = 0$
  2. $\langle x, y \rangle = \langle y, x \rangle$
  3. $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$

- Any x and y such that $\langle x, y \rangle = 0$ are called *orthogonal* : $x \perp y$

- An inner product induces a norm:

$$\|x\| = \sqrt{\langle x, x \rangle}$$

- A vector space $\mathcal{H}$ endowed with an inner product is a *Hilbert space* if $\mathcal{H}$, endowed with the norm induced by the inner space, is a Banach space

Examples of inner products

- Finite dimensional Euclidean spaces $\mathbb{R}^d$:

$$\langle x, y \rangle = x^\top y, \qquad x, y \in \mathbb{R}^d$$

is an inner product. Exercise!

- Sequence space ℓ^2, that is the space of infinite vectors $x = (x_1, x_2, \ldots)$ such that $\sum_{j\geq 1} x_j^2 < \infty$. Let $x, y \in \ell^2$, then

$$\langle x, y \rangle = \sum_{j\geq 1} x_j y_j$$

is an inner product. Exercise!

Parallelogram identity

- Exercise: Show that if $\mathcal{H}$ is a Hilbert space, then

$$\forall x, y \in \mathcal{H}, \qquad \|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2) \qquad (1)$$

- Exercise: Let $\mathcal{S}$ be a Banach space such that the parallelogram identity (1) holds true. Then

$$\langle x, y \rangle = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right)$$

is an inner product. Moreover, $\mathcal{S}$ endowed with this inner product is a Hilbert space

The Hilbert space $L^2(\mathcal{T})$

- For any f, g real-valued functions defined on $\mathcal{T}$ such that $\|f\|_2, \|g\|_2 < \infty$, let
$$\langle f, g\rangle = \int_{\mathcal{T}} f\, g\, d\mu \tag{2}$$
- Exercise: show that (2) defines an inner product
- Exercise: show that
$$|\langle x, y\rangle| \leq \|x\|\|y\| \qquad \text{(Cauchy-Schwarz inequality.}$$
When the equality holds true?
- The Hilbert space $L^2(\mathcal{T})$ is the main space used in FDA!

Agenda

Bases in Banach spaces

- Let $\mathcal{S}$ be a Banach space
- A *basis* in $\mathcal{S}$ is a sequence $\{e_j\}_{j\geq 1} \subset \mathcal{S}$ such that, $\forall x \in \mathcal{S}$, there exists a unique sequence $\{a_j\}_{j\geq 1} \subset \mathbb{R}$ such that

$$x = \sum_{j\geq 1} a_j e_j$$

- The convergence of the series means[1]

$$\lim_{J\to\infty} \left\| x - \sum_{j=1}^{J} a_j e_j \right\| = 0$$

- One can also ask each e_j to be of norm 1, in which case one gets a *normalized* basis

[1] In functional analysis, the sequence sequence $\{e_j\}_{j\geq 1}$ is usually called a *Schauder basis*

Bases in Hilbert spaces

- The concept of basis extends to Hilbert spaces
- In Hilbert spaces one could even ask the e_j be orthogonal, with norm equal to 1. Then one gets an *orthonormal basis*
- When the e_j be orthogonal, not necessarily with norm 1, one has a *orthogonal basis*
- Given a basis in a Hilbert space, it is possible to build an orthonormal basis from it by the so-called *Gram-Schmidt Orthogonal Procedure*

Basis decomposition coefficients

- Let $\{e_j\}_{j\geq 1} \subset \mathcal{H}$ be an orthonormal basis in the Hilbert space $\mathcal{H}$
- Let $x \in \mathcal{H}$, and let
$$x = \sum_{j\geq 1} a_j e_j$$
be the decomposition of x in the basis $\{e_j\}_{j\geq 1}$
- Exercise. Show that
$$a_j = \langle x, e_j \rangle, \qquad j \geq 1,$$
and thus
$$x = \sum_{j\geq 1} \langle x, e_j \rangle e_j$$

Remember Pythagoras

- If $x, y \in \mathcal{H}$, with $\mathcal{H}$ Hilbert space, it is easy to check that

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x, y\rangle$$

- Pythagoras' Theorem:

$$x \text{ and } y \text{ are orthogonal iff } \|x+y\|^2 = \|x\|^2 + \|y\|^2$$

- Let $\{x_j\}_{j\geq 1} \subset \mathcal{H}$, with $\mathcal{H}$ Hilbert space. Assume that the x_j are orthogonal and the series of $\{x_j\}_{j\geq 1}$ convergences to some limit $x_\infty \in \mathcal{H}$.
  - Exercise. Then

$$\text{the series of } \{\|x_j\|^2\}_{j\geq 1} \subset \mathbb{R} \text{ converges and } \left\|\sum_{j\geq 1} x_j\right\|^2 = \sum_{j\geq 1} \|x_j\|^2$$

Parseval's identity

- Let $\{e_j\}_{j\geq 1} \subset \mathcal{H}$ be an orthonormal basis in the Hilbert space $\mathcal{H}$
- Exercise. Show that

$$\forall x \in \mathcal{H}, \qquad \sum_{j\geq 1} \langle x, e_j\rangle^2 = \|x\|^2$$

Projections and best approximations in Hilbert spaces

- Hilbert Projection Theorem. Let $F \subset \mathcal{H}$ be a closed vector subspace of the Hilbert space $\mathcal{H}$. For any $x \in \mathcal{H}$, there exists a unique point $x_F \in F$ which is the closest to x in F, *i.e.*

$$\|x - x_F\| < \|x - y\|, \quad \forall y \in F$$

- The point x_F is characterized by

$$x_F \in F \text{ and } x - x_F \perp F$$

- The typical application of this general result in FDA is for F the finite-dimensional space spanned by a finite subset of a basis $\{e_j\}_{j\geq 1}$: for some $J \geq 1$,

$$F = \overline{\mathrm{sp}}\{e_1, \ldots, e_J\}$$

Basis coefficients and distances in Hilbert spaces

- In FDA we need to compute distances between curves
- Assume that the curves belong to a Hilbert space $\mathcal{H}$ and let $\{e_j\}_{j\geq 1} \subset \mathcal{H}$ be a basis
- Let $x, y \in \mathcal{H}$,

$$x = \sum_{j\geq 1} a_j e_j, \qquad y = \sum_{k\geq 1} b_k e_k$$

- Exercise. Show that
  - $$\langle x, y\rangle = \sum_{j,k\geq 1} a_j b_k \langle e_j, e_k\rangle$$
  - When $\{e_j\}_{j\geq 1}$ is orthogonal,
  $$\langle x, y\rangle = \sum_{j\geq 1} a_j b_j \|e_j\|^2$$
  - For an orthonormal basis
  $$\langle x, y\rangle = \sum_{j\geq 1} a_j b_j$$

- Assume that the curves belong to a Hilbert space $\mathcal{H}$ and let $\{e_j\}_{j\geq 1} \subset \mathcal{H}$ be a **orthonomal basis**
- Let $x, y \in \mathcal{H}$,
$$x = \sum_{j\geq 1} a_j e_j, \qquad y = \sum_{j\geq 1} b_j e_j$$
- Exercise. Show that
$$\|x - y\|^2 = \sum_{j\geq 1} (a_j - b_j)^2$$

Why considering a basis?

- In FDA, the paths (trajectories) of the process X are supposed to belong to a space of function, typically the Hilbert space $L^2(\mathcal{T})$
- Given a basis $\{e_j\}_{j\geq 1}$, orthonormal or not, each realization (curve, path, trajectory) of X could be decomposed in basis such as

$$X(t) = \mathbb{E}[X(t)] + \sum_{j\geq 1} a_j e_j(t), \qquad t \in \mathcal{T}.$$

- The realization being random, the coefficients $\{a_j\}_{j\geq 1}$ are also random.
  - Centering X is equivalent to centering the coefficients $\{a_j\}_{j\geq 1}$.
- Each *centered* realization of X is represented by its *zero-mean* coefficients $\{a_j\}_{j\geq 1}$

Truncated representations

- In practice, approximated representations of the realizations of X are obtained using only a finite number among the $\{a_j\}_{j\geq 1}$

$$X - \mathbb{E}[X] \approx \sum_{j=1}^{J} a_j e_j$$

- Then each realization is approximately represented by
  - the mean curve (the same for all realizations);
  - a finite set coefficients $a_1, \ldots, a_J$ (one set for each realization).

- *It seems that* the statistician is now back in the classical multivariate statistics

Remember multivariate data analysis

- Consider a functional linear model with a scalar response Y :

$$Y = c + \langle X, \beta \rangle + \varepsilon,$$

with $\beta \in L^2[0,1]$. The unknown parameters of the model are $c \in \mathbb{R}$, β (and the variance of ε)

- Without loss of generality, assume $\mathbb{E}[X(t)] = 0$, $\forall t$. If

$$X \approx \sum_{j=1}^{J} a_j e_j \quad \text{and} \quad \beta \approx \sum_{j=1}^{J} c_j e_j,$$

then

$$Y \approx c + \sum_{i=1}^{J} c_j a_j + \varepsilon$$

- A similar idea applies to nonlinear context (e.g., functional logistic model)

Not so fast/easy!

- Indeed
  - the a_j may be easily calculable from the data, for each observation;
  - c and c_j can be then estimated by least squares.
- However, the statistical properties of the resulting estimator are not obvious, because
  - the estimated model is just a proxy of the original one;
  - quite often J is random;
  - a_j are usually calculated with error (both statistical and numerical).
- There is still a lot to be done to make FDA practice completely rigorous!

Agenda

What type of basis?

- Since we will use approximated representations of the realizations of X by considering only a finite number among the $\{a_j\}_{j\geq 1}$, we need "economic" bases
- "Economic" = provides an accurate representation with only few coefficients $\{a_j\}_{j\geq 1}$
- Two types of basis are used
  - fixed
  - data-driven
- Fixed bases are preferred in engineering, computer science,...
- Data-driven approaches are preferred by statisticians

Examples of fixed bases

- Some 'basic' choices
  - Step-function
  - Polynomials/power basis: $\{1, t, t^2, t^3, \ldots\}$ or $\{1, t^{\lambda_1}, t^{\lambda_2}, t^{\lambda_3}, \ldots\}$ for some unbounded sequence of λ's
  - Exponentials: $\{1, \exp^{\lambda_1 t}, \exp^{\lambda_2 t}, \exp^{\lambda_3 t}, \ldots\}$ for some unbounded sequence of λ's
  - Hartley: $\{1, \sin t + \cos t, \sin 2t + \cos 2t, \sin 3t + \cos 3t \ldots\}$
- Some more elaborated choices
  - (B-)Splines
  - Fourier
  - Wavelets

Examples of fixed bases

- (B-)Splines. Set of polynomials (of some order, say, m) defined in subintervals and constructed in such a way that and the endpoints of the subintervals the polynomials, and their up to $m-2$ derivatives, coincide

```
spline.basis=create.bspline.basis(rangeval=c(0,10), nbasis=5)
plot(spline.basis, lty=1, lwd=2)
```

- See Section 5.5 in Wasserman (2006) for a short but comprehensive description of the B-Splines.

Splines: a quick introduction (1/6)

- Consider a univariate nonparametric regression model

$$Y = r(X) + \epsilon, \tag{3}$$

  where
  - X is a regressor (predictor/covariate/...) with values in a bounded interval
  - $r(\cdot)$ is some unknown function
  - ϵ is the error term with $\mathbb{E}(\epsilon \mid X) = 0$ and finite variance

- The independent draws $(Y_1, X_1), \ldots, (Y_n, X_n)$ are observed

- One may consider minimizing, w.r.t. to $r(\cdot)$ in a set of functions, the penalized sum of squares

$$M(r;\lambda) = \sum_{i=1}^{n} \{Y_i - r(X_i)\}^2 + \lambda J(r), \tag{4}$$

  where
  - $J(r)$ is some roughness penalty; typically $J(r) = \int (r''(x))^2 dx$
  - $\lambda \geq 0$ is a regularization parameter; controls the fit and the penalty

Splines: a quick introduction (2/6)

- A $M^{th}-$order spline is a piecewise $M-1$ degree polynomial, with $M-2$ continuous derivatives at the knots $\xi_1, \ldots, \xi_k \in (a, b)$
  - Exercise: How many parameters have a $M^{th}-$order spline defined by the knots $\xi_1 < \xi_2 < \cdots < \xi_k$. How many of them are free?
- A spline that is linear beyond the boundary knots (vanishing second derivative) is called a natural spline
- The most commonly used are the cubic splines ($M = 4$)

Theorem (Wasserman (2006)[2], Theorem 5.73)

The function $\widehat{r}_n(\cdot)$ solution of (4), for a given λ, is a natural cubic spline with knots at the data points. The estimator $\widehat{r}_n(\cdot)$ called a ***smoothing spline****.*

[2] Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media. See Moodle.

Splines: a quick introduction (3/6)[3]

- Let $a < \xi_1 < \dots \xi_k < b$ be the knots

- **Truncated power basis**
  - Let $h_1(x) = 1$, $h_2(x) = x$, $h_3(x) = x^2$, $h_4(x) = x^3$, and
$$h_j(x) = (x - \xi_{j-4})^3_+, \qquad 5 \leq j \leq k + 4.$$
  - Then any cubic spline $r(x)$ with these knots can be written as
$$r(x) = \sum_{j=1}^{k+4} \beta_j h_j(x).$$

[3] Discover a comprehensive study of splines, due to T. Luche & K. Morken, here
https://www.uio.no/studier/emner/matnat/math/MAT4170/v18/pensumliste/splinebook-2018.pdf

B-Splines: a quick introduction (4/6)

- Let $a < \xi_1 < \ldots \xi_k < b$ be the knots
- **B-spline basis**
  - Define $2M$ extra knots (not necessarily distinct)

    $$\tau_1 \leq \tau_2 \cdots \leq \tau_M \leq \xi_0 = a \qquad \text{and} \qquad b = \xi_{k+1} \leq \tau_{M+1} \leq \cdots \tau_{2M}$$
  - Let

    $$B_{i,0}(x) = \begin{cases} 1 & \text{if } \tau_i \leq x < \tau_{i+1} \\ 0 & \text{otherwise} \end{cases}, \qquad i = 1, \ldots, k + 2M - 1$$
  - Using the rule "$0/0 = 0$", for $1 \leq m \leq M$, define

    $$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+1+m} - x}{\tau_{i+1+m} - \tau_{i+1}} B_{i+1,m-1}(x)$$
- Then any cubic spline $r(x)$ with these knots can be written as

$$r(x) = \sum_{j=1}^{k+4} \beta_j B_{j,4}(x).$$

(B-)Splines: a quick introduction (5/6)

- Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an independent sample generated according to (3)
- Let $B_1, \ldots, B_{n+4}$ be a basis for the natural splines with the knots X_i
  - e.g., the B-splines
- We then know that the solution of (4) can be written

$$\widehat{r}_n(x) = \sum_{j=1}^{n+4} \widehat{\beta}_j B_j(x)$$

- It can be shown that $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{n+4})^\top \in \mathbb{R}^{n+4}$ is the solution pf

$$\min \left(\mathbb{Y} - \mathbb{B}\boldsymbol{\beta}\right)^\top \left(\mathbb{Y} - \mathbb{B}\boldsymbol{\beta}\right) + \lambda \boldsymbol{\beta}^\top \Omega \boldsymbol{\beta}, \tag{5}$$

  where
  - $\mathbb{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$
  - $\mathbb{B}_{ij} = B_j(X_i)$, $1 \leq i \leq n$, $1 \leq j \leq n+4$
  - $\Omega_{jk} = \int B_j''(x) B_k''(x) dx$, $1 \leq j, k \leq n+4$

(B-)Splines: a quick introduction (6/6)

Theorem (Wasserman (2006), Theorem 5.79)

The value of β that minimizes (5) *is*

$$\widehat{\boldsymbol{\beta}} = \left(\mathbb{B}^\top \mathbb{B} + \lambda \Omega\right)^{-1} \mathbb{B}^\top \mathbb{Y}$$

- The splines, in particular B-splines, are linear smoothers.
- In the case $\lambda = 0$, the spline smoother interpolates the data, while larger λ yields smoother fit. The smoothing parameter can be selected by CV.

Examples of fixed bases

- Fourier. Let $\mathcal{T} = [0, T]$. The orthonormal Fourier basis is given by

$$\frac{1}{\sqrt{T}} \quad \text{and} \quad \left\{ e_{2k-1}(t) = \frac{\sin(2k\pi t/T)}{\sqrt{T/2}}, e_{2k}(t) = \frac{\cos(2k\pi t/T)}{\sqrt{T/2}} \right\}_{k \geq 1}$$

```
fourier.basis=create.fourier.basis(rangeval=c(0,10), nbasis=5)
plot(fourier.basis, lty=1, lwd=2)
```

- Wavelets Orthonormal basis are constructed for $L^2(\mathbb{R})$, based on translation and/or dilation of a function called *mother wavelet* $\psi(\cdot)$:

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k), \quad j, k \in \mathbb{Z}.$$

The construction of orthonormal wavelet bases on $[0, 1]$ is more involved...

Other examples of data-driven bases

- Functional Principal Components basis (see below)
- Partial Least Squares basis (appropriate for regression context)

Basis decomposition: curves observed without error

- Consider the $i-$th observation, corresponding to the realization $X_i(\cdot)$ of the stochastic process X

- Assume that an estimator $\widehat{\mu}(\cdot)$ of the mean curve is available. Say that $\mathcal{T} = [0, 1]$

- Let $e_1, e_2, \ldots$ be the basis (given or estimated from the data)

- Since $a_j = \langle X - \mu, e_j \rangle$, the coefficients of the $i-$th observation can be obtained as

$$a_{ij} = \int_0^1 \{X_i(t) - \widehat{\mu}(t)\} e_j(t) dt \approx \frac{1}{M_i} \sum_{k=1}^{M_i} \{X_i(T_{ik}) - \widehat{\mu}(T_{ik})\} e_j(T_{ik}),$$

with $1 \leq j \leq J$, where T_{ik}, $1 \leq k \leq M_i$, are points (here, for simplicity, equally spaced) where $X_i(\cdot)$ is observed

Example: zero-mean Gaussian sample paths

- Consider some large K (e.g., $K = 50000$),

$$S_i = \frac{1}{\sqrt{K}} \sum_{k=1} N_k, \quad N_k \sim iid\ N(0,1), \quad 1 \leq k \leq K$$

and

$$t_i = i/K.$$

- Use the following code to generate a (discretized) sample path using S_i and decompose its truncated representation in the B-spline basis

```
library(fda); K=40000; nb_base = 25 #nb of elements in the base
Wiener=cumsum(rnorm(K))/K**.5 # random walk on [0,K]
plot.ts(Wiener, xlab="", ylab="")
B.tbasis=create.bspline.basis(rangeval=c(0,K), nbasis=nb_base)
Wiener.fd=smooth.basis(y=Wiener, fdParobj=B.tbasis)
lines(Wiener.fd, lwd=3)
```

Basis decomposition: curves observed with error

- Consider the data corresponding to the realization $X_i(\cdot)$ are the pairs
$$(Y_{ik}, T_{ik}), \qquad 1 \leq k \leq M_i,$$
where
$$Y_{ik} = X_i(T_{ik}) + \varepsilon_{ik}$$

- The coefficients of the $i-$th observation (curve) can be obtained by least-squares regression

$$\arg \min_{a_{i1},\ldots,a_{iJ}} \sum_{k=1}^{M_i} \{Y_{ik} - \widehat{\mu}(T_{ik}) - a_{i1} e_1(T_{ik}) - \cdots a_{iJ} e_J(T_{ik})\}^2$$

- Quite often a regularization is needed (due to colinearity).

Take away ideas

- Functional data are usually modeled as the realizations of random elements taking values in a space of functions
- It is of often convenient to represent the realizations in a suitable basis of the image space, and to identify the realization with the (infinite) vector of coefficients in the basis
- As functional data are discretely observed, sometimes with noise, the vector of coefficients is always corrupted by error; the packages do not take into account this error.
- Many alternative basis can be used, the choice should be driven by the nature of the data and the purpose of the analysis
- For statistical inference, the realizations are approximated by truncated representations, which next allow to apply multivariate analysis tools
- Since the representation in a basis is usually corrupted by error, the truncated representations are also corrupted, and performing correct multivariate analysis is challenging

Agenda

From standard normal to Gaussian vectors

- A random vector $\mathbf{X} \in \mathbb{R}^d$ is Gaussian if any linear combination of its components has a Gaussian (normal) distribution (possibly degenerated)
- A random vector $\mathbf{X} \in \mathbb{R}^d$ has a multivariate Gaussian distribution, and we write $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if it admits the following density (w.r.t. the Lebesgue measure) :

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \quad \mathbf{x} \in \mathbb{R}^d,$$

  where
  - $\boldsymbol{\mu} \in \mathbb{R}^d$ – mean vector
  - $\boldsymbol{\Sigma}$ is a $d \times d-$positive definite matrix – (co)variance matrix
  - $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$
- If $Z_1, \ldots, Z_d$ are i.i.d. $N(0,1)$, then $\mathbf{Z} = (X_1, \ldots, X_d)^\top \in \mathbb{R}^d$ is a standard Gaussian (normal) random vector; the distribution is denoted $N_d(\mathbf{0}, \boldsymbol{I}_d)$

Some properties of Gaussian vectors

- The multivariate Gaussian distribution is determined by the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
- If $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{A}$ is a $d \times r-$matrix, then $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{A}\mathbf{Z}$ is a Gaussian vector
- The (i,j) entry of $\boldsymbol{\Sigma}$ is equal to the covariance of the $i-$th and $j-$th components of $\mathbf{X}$, *i.e.*, $\mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$.
- **Conditional distribution:** let $\mathbf{X}_1 \in \mathbb{R}^{d_1}$, $\mathbf{X}_2 \in \mathbb{R}^{d_2}$ be random vectors such that
$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim N_{d_1+d_2}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right).$$
The conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$ is then a multivariate Gaussian, with parameters

$$\boldsymbol{\mu}_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathbf{X}_1|\mathbf{X}_2} = \boldsymbol{\Sigma}_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Agenda

Gaussian process (1/3)

- A continuous time stochastic process $\{X(t) : t \in \mathcal{T}\}$, where $\mathcal{T}$ is (say) a compact interval, is Gaussian if and and only if, for any positive integer k and any

$$\min \mathcal{T} \leq t_1 < \cdots < t_k \max \leq \mathcal{T},$$

the random vector $(X(t_1), \ldots, X(t_k))$ is Gaussian.[4]

- A Gaussian process (GP) is completely defined by the mean function and the covariance function, that are

$$\mu(t) = \mathbb{E}[X_t] \quad \text{and} \quad c(s,t) = \mathbb{E}[(X_s - \mu(s))(X_t - \mu(t))]$$

- If a GP has zero mean, properties such as stationary, smoothness of the sample paths, *etc*, can be defined through the covariance function

[4] *In general, the distributional properties of a stochastic process are determined by the finite-dimensional distributions*

Gaussian process (2/3): examples of zero-mean GP

- (Fractional) Brownian motion: usually denoted B_H, is a zero-mean Guassian process with covariance function

$$c(s,t) = \frac{1}{2}\left(|s|^{2H} + |t|^{2H} - |s-t|^{2H}\right),$$

where $H \in (0,1)$ is the so-called Hurst index (parameter). Brownian motion corresponds to $H = 1/2$.

- Multifractional Brownian motion (mfBm): is a centered Gaussian process with covariance function

$$c(s,t) = D(H_s, H_t)\left[|s|^{H_s+H_t} + |t|^{H_s+H_t} - |t-s|^{H_s+H_t}\right],$$

where $t \mapsto H_t \in (0,1)$ is the so-called Hurst function and

$$D(x,y) = \frac{\sqrt{\Gamma(2x+1)\Gamma(2y+1)\sin(\pi x)\sin(\pi y)}}{2\Gamma(x+y+1)\sin(\pi(x+y)/2)}, \qquad D(x,x) = 1/2$$

- Exercise: build a code which generates functional data from a mfBm

Gaussian process (3/3)

- Exercise: Let $c(s,t) = \mathbb{E}[(X_s - \mu(s))(X_t - \mu(t))]$, $s, t \in \mathcal{T}$, be the covariance function of a (Gaussian) process. Then, $c(s,t) = c(t,s)$ and, for any positive integer k and any $t_1, \ldots, t_k \in \mathcal{T}$, we have

$$\sum_{l=1}^{k} \sum_{l'=1}^{k} t_l t_{l'} c(t_l, t_{l'}) \geq 0.$$

A function satisfying the condition in the display os also called a *positive definite kernel*

- On the other hand, it can be shown that for any positive definite kernel $c(\cdot,\cdot)$ and function $\mu(\cdot)$, there exists a corresponding Gaussian process with mean μ and covariance function c.

More on Gaussian processes

- Gaussian or de Moiviran?

  “Dear listeners, please smile indulgently at this brief introductory phase of somewhat metered, sing-song words; Not poetry, perhaps, yet not quite prose, I hope. To speak of "Gaussian" measures, now is first of all, to credit C. F. Gauss for objects he was not the first to find. Some three quarters of a century before Carl Friedrich took them up we find that Abraham de Moivre already had described them. And so to him should go, I think, the fame of finding out the laws we call by Gauss’s name, those bell-shaped curves of density whose formulas we now write with ease in terms of e to minus half x squared, but which de Moivre more laboriously did call the number which answers to the hyperbolic logarithm minus half x times x. The central limit theorem, too, is found in Abe de Moivre’s book, *Doctrine of Chances*, and if it’s only for binomial distributions, well now only after Fourier and then by Paul Levy is rendered easy such a proof. Without their tools perhaps, dear listener, you’d demonstrate it as a tour de force. I’ve tried without success. And yet to say "de Moivrian" rather twists the tongue and it’s too late to change the name, so we’ll have to find some other way of remembering the founder of this line of work. Let’s dedicate now to him, de Moivre a few moments of our kindest thoughts.”

  (Dudley, R. M. (1975). The Gaussian process and how to approach it. Proceedings of the International Congress of Mathematicians. Vol. 2. pp. 143–146.)

- GP are intensively studied in Bayesian statistics. GP are related to RKHS spaces, kernel ridge regression,...
  - See the nice review Kanagawa et al. (2018). Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences. arXiv 1807.02582