Representation learning with structured data

Florian Yger

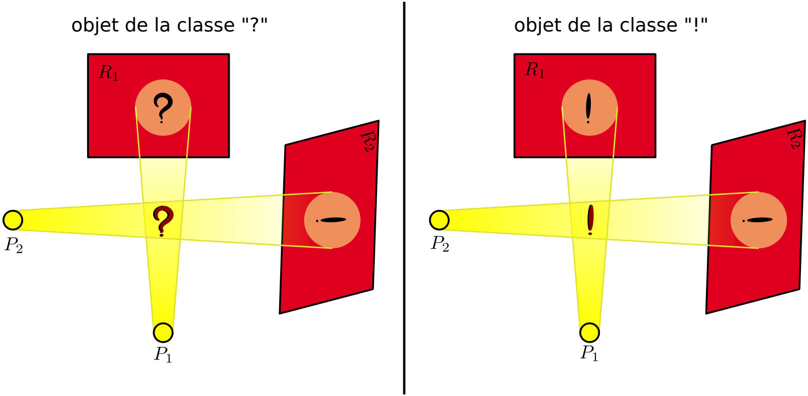
Mar. 20, 2023 - RIKEN AIP/PRAIRIE workshop

Université Paris-Dauphine, PSL Research University, LAMSADE, CNRS
PRAIRIE

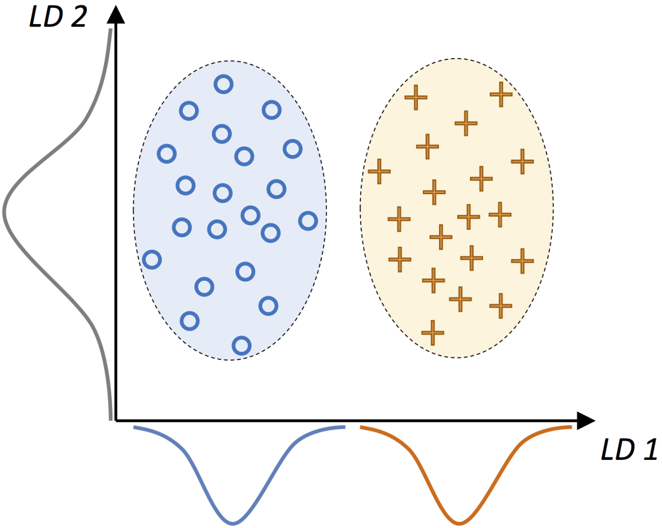
Table of Contents

Representation Learning

The issue with data representation



A well-known case



Representation is critical

A difficult task

- Representation is the first step of any data processing pipeline
- It has to be adapted to the downstream task
- Representation can be done explicitly or implicitly

but it can get harder

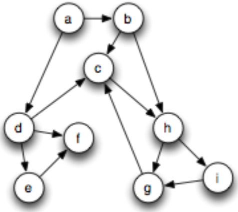
- when data are not tabular/numerical (e.g. structured data)
- when the data live on a particular space under some constraint or under a peculiar geometry (e.g. data on manifold)
- when some invariances are involved

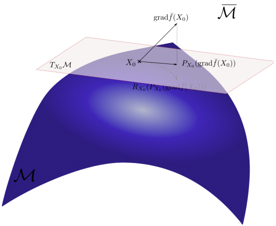
Focus of this talk

- Incorporate prior knowledge in a representation learning step
- Deep models will not be covered (or as promising extensions)

Motivation

- feasible solutions (e.g. averaging structured data)
- leveraging invariances in data (as permutations in graph data)
- incorporating prior knowledge
- accelerating optimization problem (by reducing the search space)





Learning with structures in data

Motivation

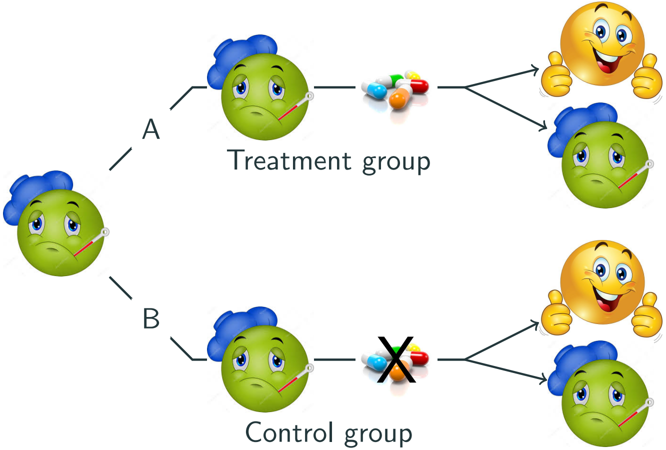
- feasible solutions (e.g. averaging structured data)
- leveraging invariances in data (as permutations in graph data)
- incorporating prior knowledge
- accelerating optimization problem (by reducing the search space)

Applications

- handling malicious applications as valued graphs (call graphs)
- electrodes covariance matrices to represent EEG signals (using Riemannian geometry)
- halving strategy in causal structure

Constraints in counterfactual application

Framework : controlled randomized experiment



Framework : controlled randomized experiment

Goal

- Check the efficiency of a treatment
- Find an optimal treatment strategy (?)

Limits

- no parallel universe to access to the counterfactual outcome

$$A \cap B = \varnothing$$

- A/B testing can give an answer for the whole population (but not at the level of the individual)

Framework : controlled randomized experiment

Goal

- Check the efficiency of a treatment
- **Find an optimal treatment strategy (?)**

Limits

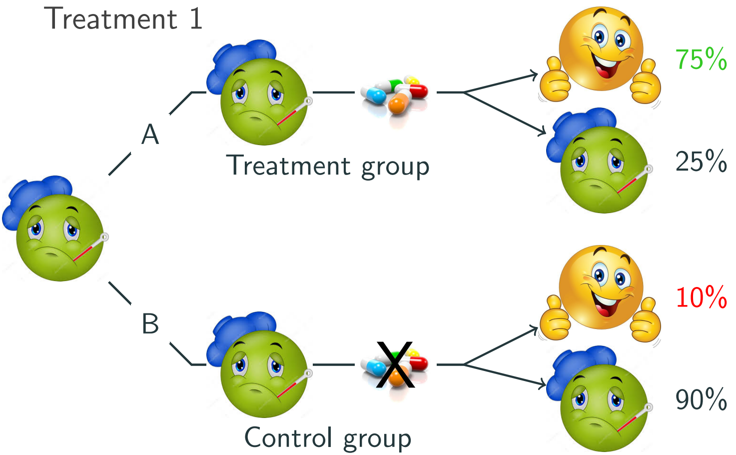
- no parallel universe to access to the counterfactual outcome

$$A \cap B = \varnothing$$

- A/B testing can give an answer for the whole population (but not at the level of the individual)

Uplift modelling aims at finding a strategy (given the features of the users/patients) for the treatment such it has the best overall effect.

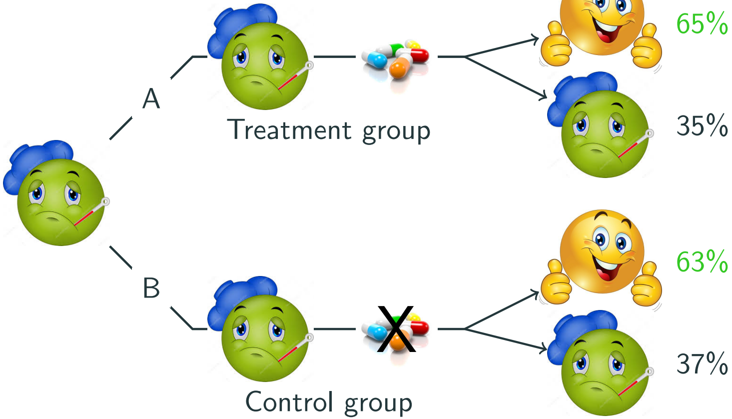
Impact of a treatment



Positive impact of the treatment

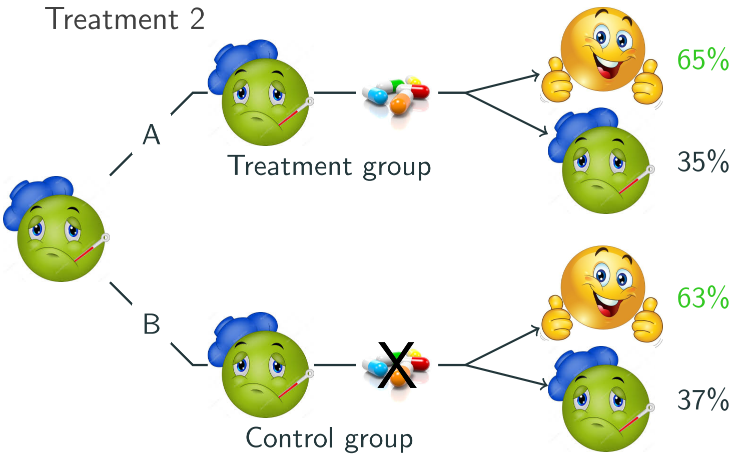
Impact of a treatment

Treatment 2



No significant impact

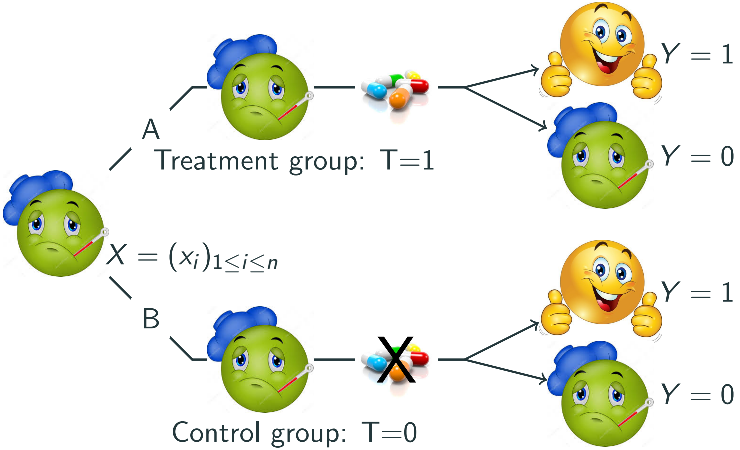
Impact of a treatment



No significant impact

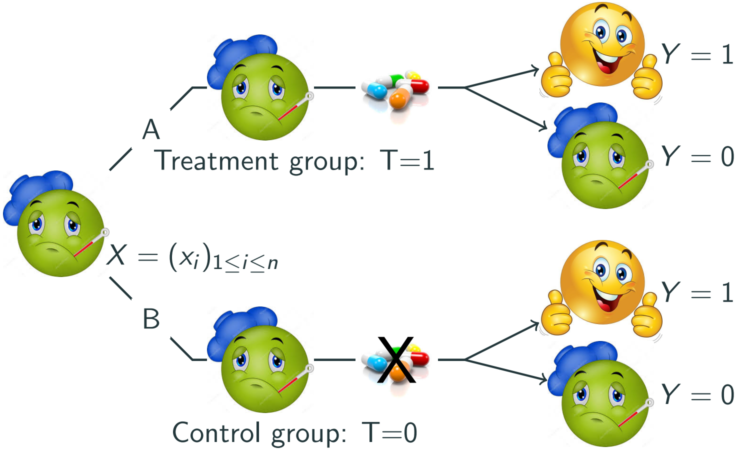
but it can be more complex than it looks as side-effect could compensate positive responses...

What is the uplift for a given individual ?



$$P(Y = 1|X = x, T = 1) - P(Y = 1|X = x, T = 0)$$

What is the uplift for a given individual ?



Classical uplift modeling:

$$\mathbb{E}[Y_i = 1 | X_i, T_i = 1] - \mathbb{E}[Y_i = 1 | X_i, T_i = 0]$$

Segmentation of the population

Given the outcome and the counter-factual outcome

- Responder positive outcome if treated (negative otherwise)
- Survivor positive outcome (whatever the treatment)
- Doomed negative outcome (whatever the treatment)
- Anti-responder negative outcome if treated (positive otherwise)

Consequences

- Unknown counter-factual outcome but partial information available
- Whole population modelled as a mixture of sub-populations

Segmentation of the population

Given the outcome and the counter-factual outcome

- Responder positive outcome if treated (negative otherwise)
- Survivor positive outcome (whatever the treatment)
- Doomed negative outcome (whatever the treatment)
- Anti-responder negative outcome if treated (positive otherwise)

Consequences

- Unknown counter-factual outcome but partial information available
- Whole population modelled as a mixture of sub-populations

From a counter-factual problem

Segmentation of the population

Given the outcome and the counter-factual outcome

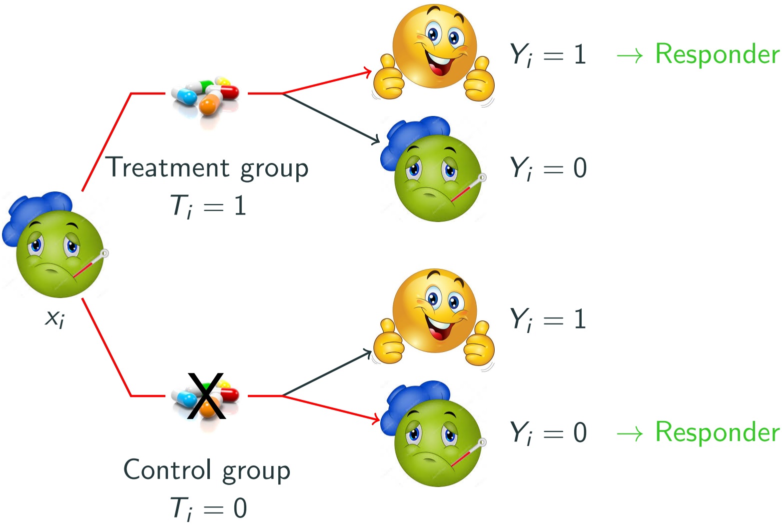
- Responder positive outcome if treated (negative otherwise)
- Survivor positive outcome (whatever the treatment)
- Doomed negative outcome (whatever the treatment)
- Anti-responder negative outcome if treated (positive otherwise)

Consequences

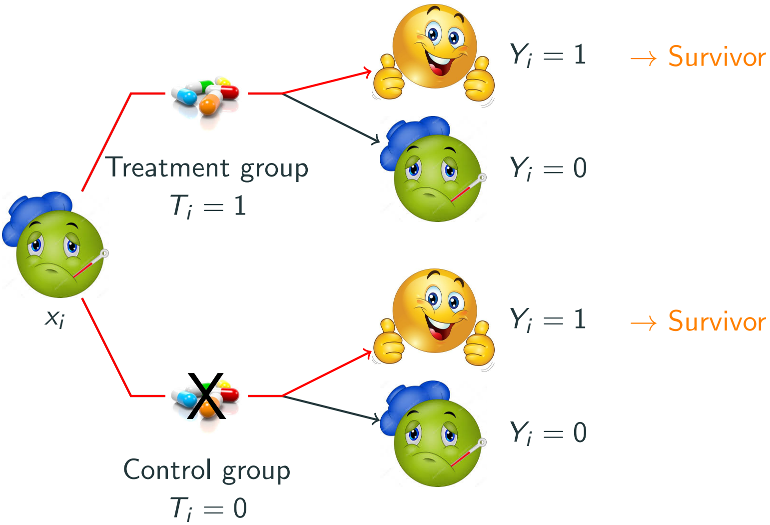
- Unknown counter-factual outcome but partial information available
- Whole population modelled as a mixture of sub-populations

From a counter-factual problem to density estimation with missing data

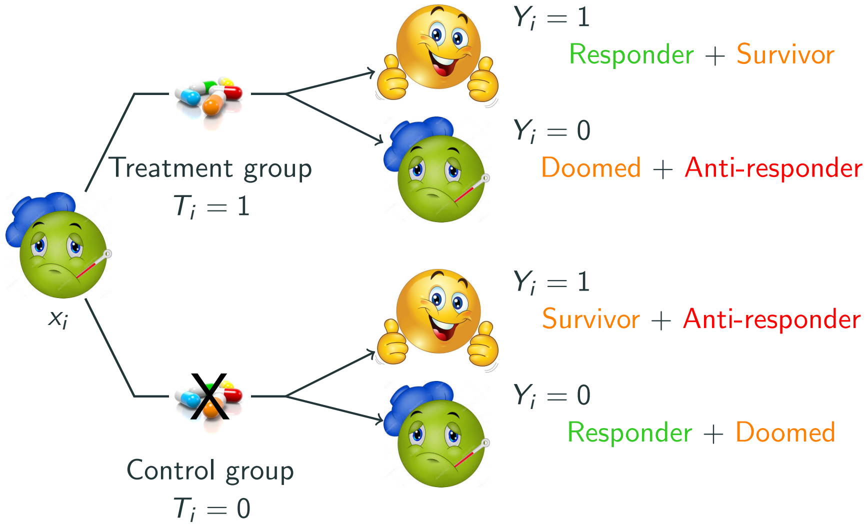
Observed outcome and constraints on the distributions



Observed outcome and constraints on the distributions



Observed outcome and constraints on the distributions



Density estimation for uplift modelling

Cost function : the log-likelihood

$$L(\{x_i\}, f_R, f_S, f_D, f_A) = \sum_{i=1}^{n} \sum_{g \in \{R,S,D,A\}} t_{ig} \log f_g(x_i)$$

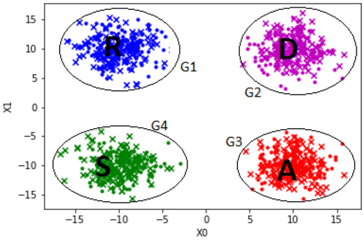
- t_{ig} membership level of x_i to the group g
- f_g density distribution of the group g (among Responder, Survivor, Doomed, Anti-responder)

On the way to a solution

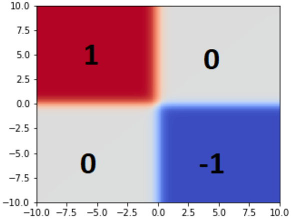
- in a parametric model t_{ig} and θ_g (parameter of f_g) are estimated
- EM algorithm is adapted to this problem of missing data
- compared to a mixture of distributions, we have some partial information

A parametric density estimation : MoG

Gaussian mixture
model estimation

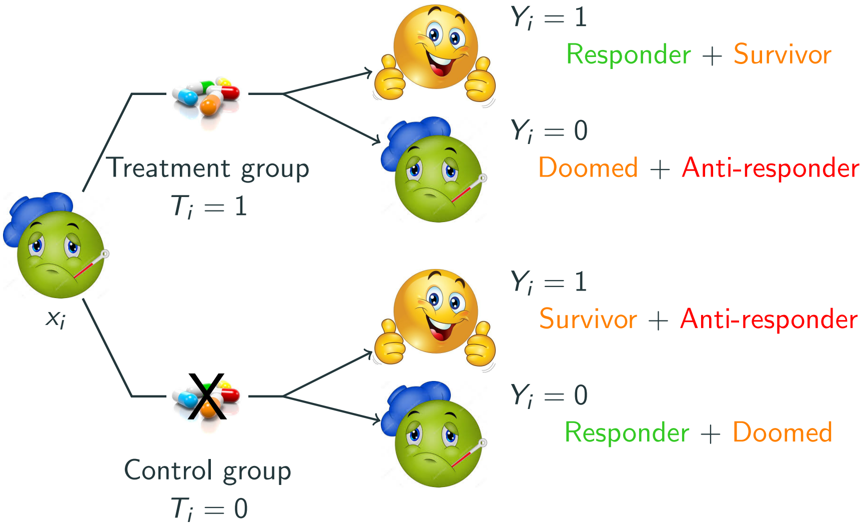


$$P(Y=1|X=x, T=1) - P(Y=0|X=x, T=$$
$$= P(R|X=x) - P(A|X=x)$$



$$argmax \sum_{i=1}^{n} \sum_{g \in \{R,S,D,A\}} t_{ig} \, log(\pi_g \mathcal{N}_g(x_i, \mu_g, \Sigma_g))$$

Observed outcome and constraints on the distributions



Constrained EM for MoG

Constraints on the distribution

T	Y	P(R)	P(D)	P(S)	P(A)
1	1	.	0	.	0
1	0	0	.	0	.
0	1	0	0	.	.
0	0	.	.	0	0

Constrained EM

- **E-step** (including a projection)

  -if $Y_i(1) = 1$ then $t_{iD} = t_{iA} = 0$

  -if $Y_i(1) = 0$ then $t_{iR} = t_{iS} = 0$

  -if $Y_i(0) = 1$ then $t_{iD} = t_{iR} = 0$

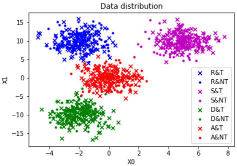
  -if $Y_i(0) = 0$ then $t_{iS} = t_{iA} = 0$

  - else $t_{ig} = \frac{p(x_i, \theta_g^c)}{\sum_{j \in \{R,D,S,A\}} p(X_i, \theta_j^c)}$

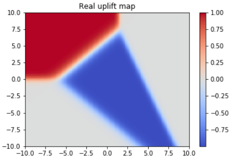
- **M-step**

$$\begin{cases} \pi_g = \frac{1}{n} \sum_{i=1}^{n} t_{ig} \\ \mu_g = \frac{\sum_{i=1}^{n} t_{ig} x_i}{\sum_{i=1}^{n} t_{ig}} \\ \Sigma_g = \frac{\sum_{i=1}^{n} t_{ig} (x_i - \mu_g)(x_i - \mu_g)^T}{\sum_{i=1}^{n} t_{ig}} \end{cases}$$

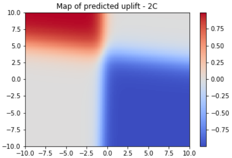
Some numerical results : toy data I



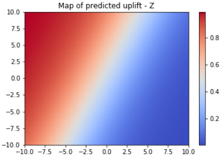
(a) Data distribution



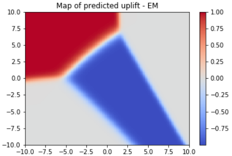
(b) Real uplift heatmap



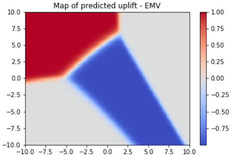
(c) Two classifiers



(d) Z transformation



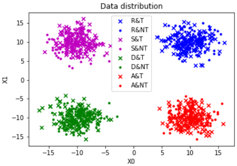
(e) EM uplift



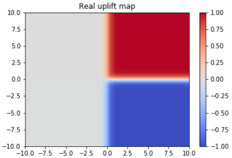
(f) V-EM uplift

Figure 1: Close but separable Gaussian distributions (Synthetic 1)

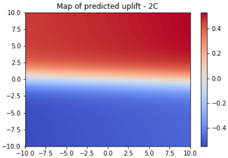
Some numerical results : toy data II



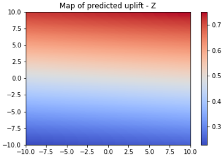
(a) Data distribution



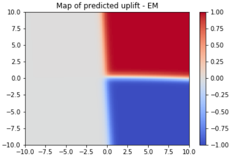
(b) Real uplift heatmap



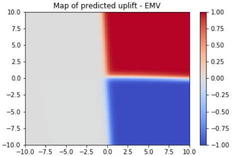
(c) Two classifiers



(d) Z transformation



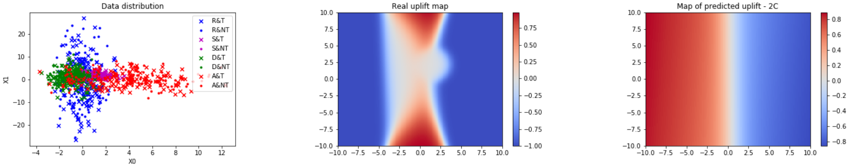
(e) EM uplift



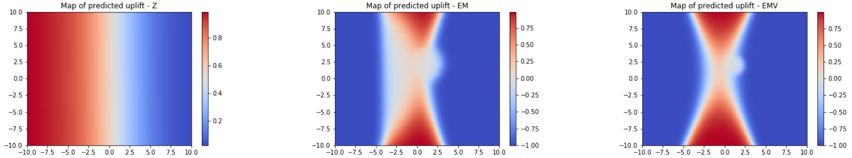
(f) V-EM uplift

Figure 2: Separable (but challenging) Gaussian distributions (Synthetic 2)

Some numerical results : toy data III



(a) Data distribution (b) Real uplift heatmap (c) Two classifiers



(d) Z transformation (e) EM uplift (f) V-EM uplift

Figure 3: Overlapping Gaussian distributions (Synthetic 3)

Constraints in preference aggregation

Application to computational social choice

Computationnal Social Choice

- at the interplay of social choice, computer science and multi-agents systems
- analyse the aggregation of preferences of a group of agents
- voting systems are the most common object of interest of the field (but not the only one : ranking, ressource allocation, crowdsourcing etc...)

The epistemic case

- votes considered as the realization of a random variable
- the probability distribution over the set of possible ballots is called a noise model
- aggregation is expressed as a Maximum Likelihood problem

Multi-winner approval voting

Example: Chord Transcription

Figure 4: Guitar Chords Transcription

A guitar chord contains at least 3 and at most 6 notes.

Problem Statement

Formally, we consider:

- **A set of** m **alternatives** $X = \{a_1, \ldots, a_m\}$**:** $\{A, A\#, B, C, C\#, D, Eb...\}$
- **A ground truth subset of alternatives** $S^* \subseteq X$**:** $C7 = \{C, E, G, B\}$
- **A set of** n **voters** N
- **A profile of** n **ballots** $A_i \subseteq X$**:** $\{C, E, G\}, \{C, Eb, E, G\}, \{A, C, E\}$

(+) Prior knowledge: $l \leq |S^*| \leq u$ for some l, u known to the central entity.

(+) Noise model.

Noise Model

The noise model will incorporate two types of errors:

$$P(a \in A_i | S^* = S) = \begin{cases} p_i & \text{if } a \in S \quad \text{TP} \\ q_i & \text{if } a \notin S \quad \text{FP} \end{cases}$$

We also suppose that:

(1) A voter's approvals of alternatives are mutually independent given the ground truth and parameters $(p_i, q_i)_{i \in N}$.

Noise Model

The noise model will incorporate two types of errors:

$$P(a \in A_i | S^* = S) = \begin{cases} p_i & \text{if } a \in S \quad \text{TP} \\ q_i & \text{if } a \notin S \quad \text{FP} \end{cases}$$

We also suppose that:

(1) A voter's approvals of alternatives are mutually independent given the ground truth and parameters $(p_i, q_i)_{i \in N}$.

(2) Voters' ballots are mutually independent given the ground truth.

The Likelihood (A Posterioi)

For now, our aim is to estimate the ground truth via Maximum a Posteriori:

$$\hat{S} = \arg\max_{S \subseteq X} P(S) \times P(A_1, \ldots, A_n | S) = \arg\max_{S \subseteq X} P(S) \prod_{i=1}^{n} P(A_i | S)$$

where:

$$P(A_i|S) = p_i^{|A_i \cap S|} q_i^{|A_i \cap \overline{S}|} (1 - p_i)^{|\overline{A_i} \cap S|} (1 - q_i)^{|\overline{A_i} \cap \overline{S}|}$$

General Model

We suppose that:

General Model

We suppose that:

- Voters answer multiple questions.
- The parameters (p_i, q_i) are unknown.

General Model

We suppose that:

- Voters answer multiple questions.
- The parameters (p_i, q_i) are unknown.
- The prior $P(S)$ is not uniform, but parameterized via $t = (t_1, \ldots, t_m)$ such that:

General Model

We suppose that:

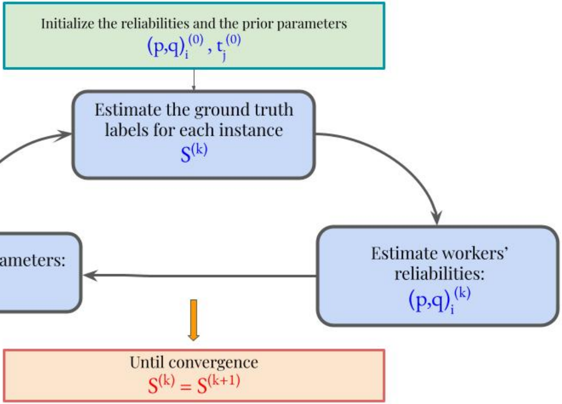
- Voters answer multiple questions.
- The parameters (p_i, q_i) are unknown.
- The prior $P(S)$ is not uniform, but parameterized via $t = (t_1, \ldots, t_m)$ such that:

$$P(S) = \begin{cases} \frac{1}{\beta(l,u,t)} \prod_{a_j \in S} t_j \prod_{a_j \notin S} (1 - t_j) & \text{if } S \in \mathcal{S}_{l,u} \\ 0 & \text{if } S \notin \mathcal{S}_{l,u} \end{cases}$$

where:

$$\beta(l, u, t) = \sum_{S \in \mathcal{S}_{l,u}} \prod_{a_j \in S} t_j \prod_{a_j \notin S} (1 - t_j)$$

Alternating Maximum Likelihood Estimations - Lloyd Heuristic



AMLE: Alternating Maximum Likelihood Estimations

To maximize the dataset's likelihood we proceed as follows (AMLE):

- Initialize $(\hat{p}_i^{(0)}, \hat{q}_i^{(0)}), (\hat{t}_j^{(0)})$.
- Alternate between:
  - Estimating the ground truth given the parameters.
  - Estimating the parameters given the ground truth.

AMLE: Alternating Maximum Likelihood Estimations

To maximize the dataset's likelihood we proceed as follows (AMLE):

- Initialize $(\hat{p}_i^{(0)}, \hat{q}_i^{(0)}), (\hat{t}_j^{(0)})$.
- Alternate between:
    - Estimating the ground truth given the parameters.
    - Estimating the parameters given the ground truth.

Theorem

For any initial values $(\hat{p}_i^{(0)}, \hat{q}_i^{(0)}), (\hat{t}_j^{(0)})$, AMLE increases the likelihood at each step, and it converges to a fixed point after a finite number of iterations.

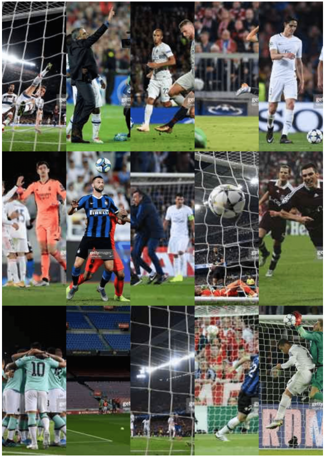


Figure 5: 15 football images

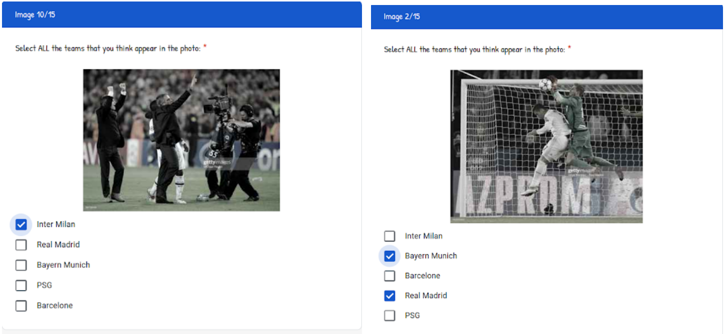
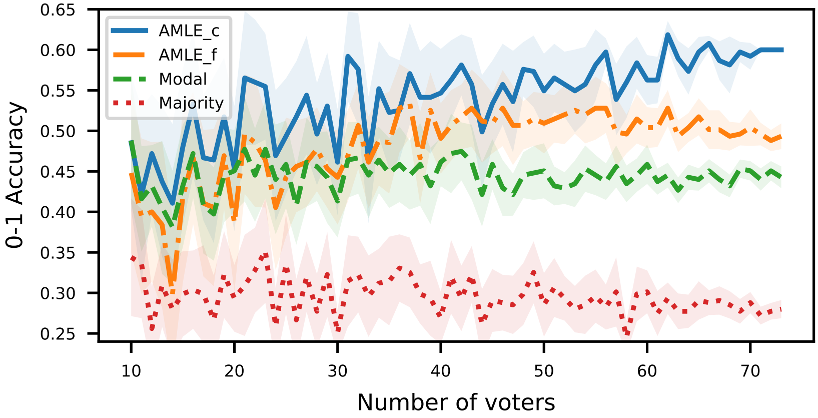


Figure 6: Image annotation datasets

We gathered the answers of 76 participants

0-1 Subset Accuracy with Different Groups of Voters



(a) 0/1 accuracy

Geometry for structured data

A problem of interest

Fréchet[1] averaging
Let (S, d) be a complete metric space. Let $x_1, \cdots, x_n \in S$, then, we define the problem of as :

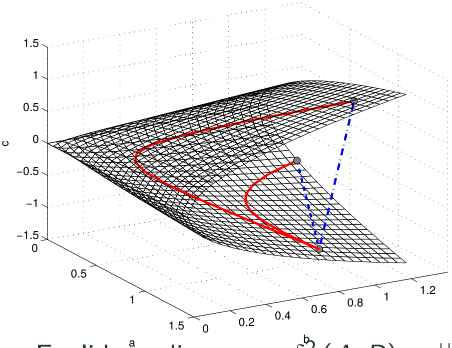
$$\min_{m \in S} \sum_{i=1}^{n} d^2(x_i, m)$$

Properties

- weighted variants exist and it can be extended for clustering
- invariances can be incorporated through d
- $m^\star$ is a representative point of the dataset and it belongs to S

[1] It is also sometimes referred as Karcher mean for Riemannian manifolds.

(Strictly) definite-positive matrices



$$C = \begin{vmatrix} a & b \\ b & c \end{vmatrix}$$

$$ac - b^2 > 0$$

- Euclidean distance : $\delta_E^2(A, B) = ||A - B||_{\mathcal{F}}^2$
  interpolation is possible but to the cost of the *swelling effect*.
- Riemannian distance (AIRM) :
  $\delta_R^2(A, B) = || \log(A^{-\frac{1}{2}} B A^{-\frac{1}{2}} ||_{\mathcal{F}}^2$.
  interpolation and extrapolation without any *swelling effect*.
- LogEuclidean distance : $\delta_L^2(A, B)_R = || \log_R(A) - \log_R(B)||_{\mathcal{F}}^2$

Where do we find those matrices ?

Classical ways of extracting features for EEG data

- signal energy-based features (for Motor Imagery, SSVEP,...)
- sample based features (for ERP)

Covariance-based features

$X \in \mathbb{R}^{n \times s}$ a an epoch of signal and $T \in \mathbb{R}^{n \times s}$ a template

- spatial covariance matrix: $C_s = \frac{1}{s} X X^\top$ - with the variance/power of electrodes on the diagonal,
- template-signal covariance: $C_T = \begin{pmatrix} TT^\top & TX^\top \\ XT^\top & XX^\top \end{pmatrix}$
- filtered signal covariance: $C_f = \begin{pmatrix} X_{f_1} X_{f_1}^\top & \cdots & X_{f_1} X_{f_F}^\top \\ \vdots & \ddots & \vdots \\ X_{f_F} X_{f_1}^\top & \cdots & X_{f_F} X_{f_F}^\top \end{pmatrix}$

  with the X_f filtered versions of the original signal.

Where it all started from

A new golden standard

- introduced in **Multi-class Brain Computer Interface Classification by Riemannian Geometry**, A. Barachant, S. Bonnet, M. Congedo, C. Jutten, *IEEE TBME* (2012)
- average improvement of 5% (from 65.1% to 70.2%) on the BCI competition IV (dataset IIa) over SOTA (CSP + LDA)
- introduction of MDRM (Minimum Distance to the Riemannian Mean) and Tangent Space Linear Discriminant Analysis (TSLDA)

Riemannian PCA as a variant of Fréchet averaging

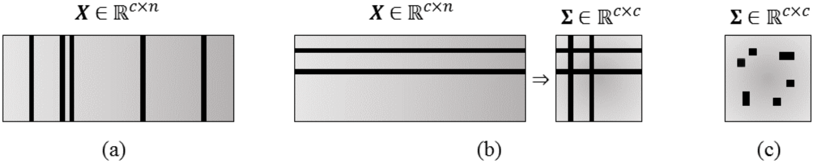
From $\mathcal{P}_n$ to $\mathcal{P}_m$ - geometry-aware dimensionality reduction

- $\forall W \in \mathbb{R}^{n\times m}$ (full column rank), $\forall i$, $W^\top C_i W \in \mathcal{P}_m$
- similar (in spirit) to previous work with a nice flavour of dimensionality reduction
- based on the maximization of a generalization of the notion of variance (without any invariance)

$$\max_W \sum_i \delta_R^2 \left(W^\top C_i W, W^\top \bar{C} W \right)$$

Taxonomy for missing data problems

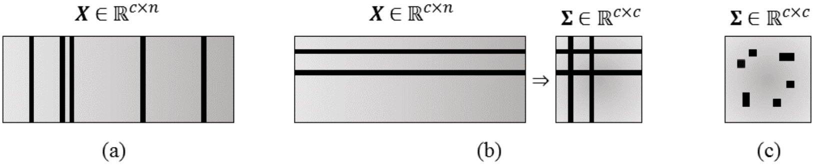
Types of missing data



- a missing samples / observations in matrix X
- b missing variables / channels in matrices X and $\Sigma = \frac{1}{n}XX^\top$ (under the hypothesis that X is centered)
- c missing elements (at random) in the matrix Σ

Taxonomy for missing data problems

Types of missing data



- a missing samples / observations in matrix X
- b **missing variables / channels in matrices X and Σ**
- c missing elements (at random) in the matrix Σ

Setup

- when a whole channel of EEG is noisy/missing, then the spatial covariance matrix is badly affected (on the corresponding row and columns)
- in a Riemannian framework, the whole covariance would be discarded (which is bad in a case of scarse data as in BCI)
- the trusted information in a matrix C with $\mathcal{S}$ the set of the indices of retained channels can be written as :

$$\hat{C} = M^\top C M$$

with M a matrix of mask (i.e., an identity matrix with only the columns indexed by $\mathcal{S}$) - a submatrix of C

How to use the Riemannian geometry in this context ?

Handling missing data as a variant of Fréchet averaging

A Féchet average with missing data
With a (possibly) different mask for each covariance :

$$\min_X \sum_i \delta_R^2(M_i^\top C_i M_i, M_i^\top X M_i)$$

- encouraging early results on synthetic experiments (channels are hidden randomly on a clean dataset)
- potential application for transfer learning between datasets recorded with different sets of electrodes
- possibly generalised to other loss functions
- generalization to any orthonormal matrix $M_i^\top M_i = \mathbb{I}_p$ for compressing sensing approach on covariance matrices

Working with graph data

Setup

- dataset made of graphs (for which the ordering of the labels is unknown)
- each graph is represented with its adjacency matrix A_i (or its laplacien L_i), $\mathcal{D} = A_1, \cdots, A_n$
- not completely unrelated to previous work : another instance of non-Euclidean data (e.g. covariance as weighted graphs)

$$d_m(A, B) = \min_{P \in \mathbb{P}_m} ||P^\top AP - B||_{\mathcal{F}}^2$$

- comparing 2 observations leads to an NP-hard problem (graph isomorphism)

Fréchet averaging of graph data

Formulation

- another instance of Fréchet averaging, with $\forall i,\ P_i \in \mathbb{P}_m$:

$$\min_{P_1,\cdots,P_n,B} \sum_i ||P_i^\top A_i P_i - B||_{\mathcal{F}}^2$$

- relaxation from the set of permutation matrices $\mathbb{P}_m$ to the set of bi-stochastic matrices $\mathbb{B}_m$ (permutation matrices are obtained through sampling) and the elements of B are then naturally in $[0,1]$
  problem convex in P_i and B (for some formulation)

Fréchet averaging of graph data

Formulation

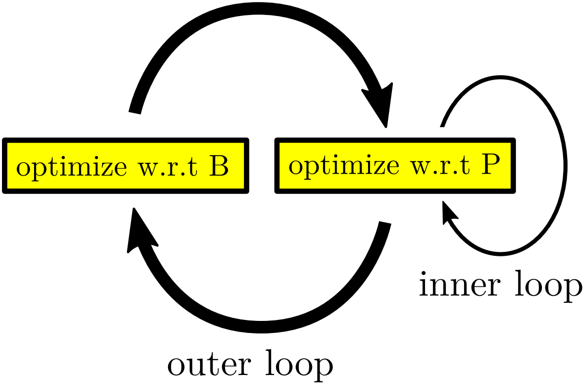
- another instance of Fréchet averaging, with $\forall i,\ P_i \in \mathbb{P}_m$:

$$\min_B \sum_i \underbrace{\min_{P_i} ||P_i^\top A_i P_i - B||_{\mathcal{F}}^2}_{d_m^2(A_i, B)}$$

- relaxation from the set of permutation matrices $\mathbb{P}_m$ to the set of bi-stochastic matrices $\mathbb{B}_m$ (permutation matrices are obtained through sampling) and the elements of B are then naturally in $[0, 1]$
problem convex in P_i and B (for some formulation)

Algorithm

Adapt the alternate optimization by tuning the number of optimization steps in the inner and outer loops



Results

Properties

- the learned weighted graphs has a nice probabilistic interpretation
- underlying generative model (generalized ERG)

Potential extension

- each graph can have a different size (as each P_i can compress the graph to a given size)
- relaxing on orthogonal matrices (instead of bistochastic) could enable to learn an embedding for each graph (at the cost of the probabilistic interpretation of the learned average)

Conclusion

Many thanks to my collaborators

- Application to counterfactual learning : Céline Béji & Jamal Atif (ESANN 2020)
- Application to epistemic social choice : Tahar Allouche & Jérôme Lang (UAI 2022)
- Riemannian PCA : Inbal Horev & Masashi Sugiyama (ACML 2015)
- RG with missing data : Quentin Barthélemy, Sylvain Chevallier & Suvrit Sra (ACML 2020)
- Graph averaging : Nicolas Boria & Benjamin Negrevergne (ESANN 2020)

Wrap up

Take home message

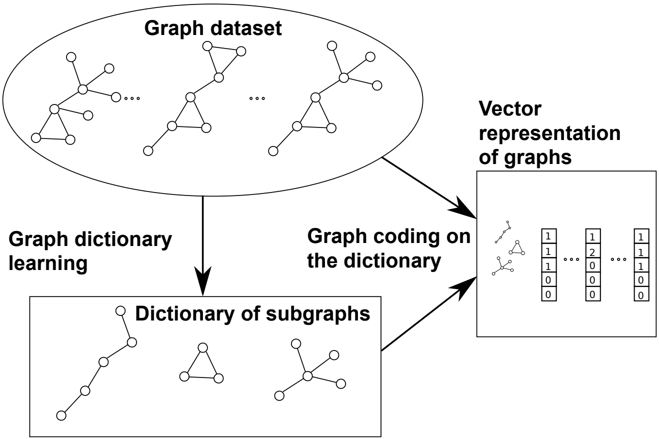
- There are many (sometimes surprinsingly simple) ways to incorporate prior knowledge or structural constraints on data
- Riemannian geometry is a practical tool for many problems with a rich theory for optimization and many libraries

What's next ?

- from metric learning on non-Euclidean data to dictionary and then deep models
- potential methodological pitfall (scarse data for deep models)
- averaging trajectories on manifold (for modelling dynamic in EEG or fatigue)

The end

Dictionary learning on graphs

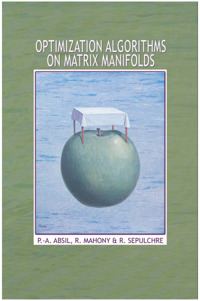


Some core references on manifolds

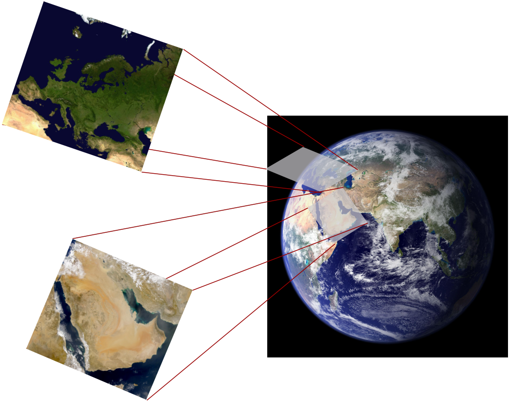
1. Edelman, A., Arias, T.A. and Smith, S.T., 1998. **The geometry of algorithms with orthogonality constraints**. SIAM journal on Matrix Analysis and Applications.
2. Absil, P.A., Mahony, R. and Sepulchre, R., 2009. **Optimization algorithms on matrix manifolds**. Princeton University Press.

Some core references on manifolds

1. Edelman, A., Arias, T.A. and Smith, S.T., 1998. **The geometry of algorithms with orthogonality constraints**. SIAM journal on Matrix Analysis and Applications.
2. Absil, P.A., Mahony, R. and Sepulchre, R., 2009. **Optimization algorithms on matrix manifolds**. Princeton University Press.



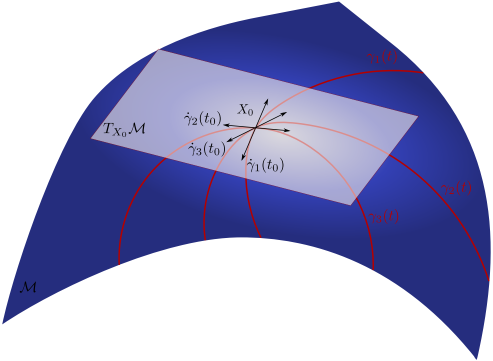
A well-known manifold



Intuitions :

- "curved" subspace of $\mathbb{R}^N$
- locally approximated by hyperplanes (i.e. maps)
- from one point to another through geodesics

Different manifolds

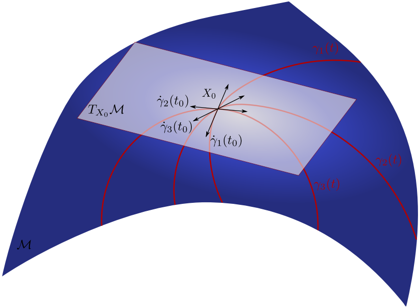


A russian doll structure

- topological manifold
- + differentiable structure (differentiable manifold)
- + riemannian metric (Riemannian manifold)

Summary : a curved space can locally be linearly approximated.

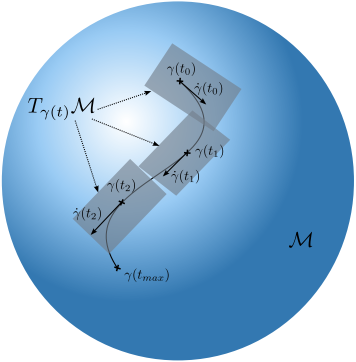
Local approximation



Tangent space $T_{X_0}\mathcal{M}$
For $\mathcal{M}$, at X_0 :

- the set of gradients at X_0 of every curve $\gamma_i(t)$ passing through this point (tangent plane)
- equipped with a scalar product (riemannian metric).

Riemannian distance

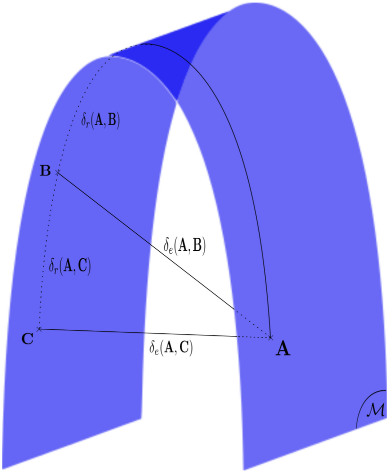


- the lenght of a curve is deduced by integrating the norm of its gradient in the tangent spaces

$$L_g(\gamma) = \int_a^b ||\gamma'(t)||_g dt$$

- depends on the chosen geometry (i.e. the riemannian metric)

Riemannian distance



- the lenght of a curve is deduced by integrating the norm of its gradient in the tangent spaces

$$L_g(\gamma) = \int_a^b ||\gamma'(t)||_g dt$$

- depends on the chosen geometry (i.e. the riemannian metric)

Exp/Log map

Exponential map
The exponential map sends points from a given tangent plan to the manifold. $\gamma_v(t)$ is the geodesic from p with initial acceleration v

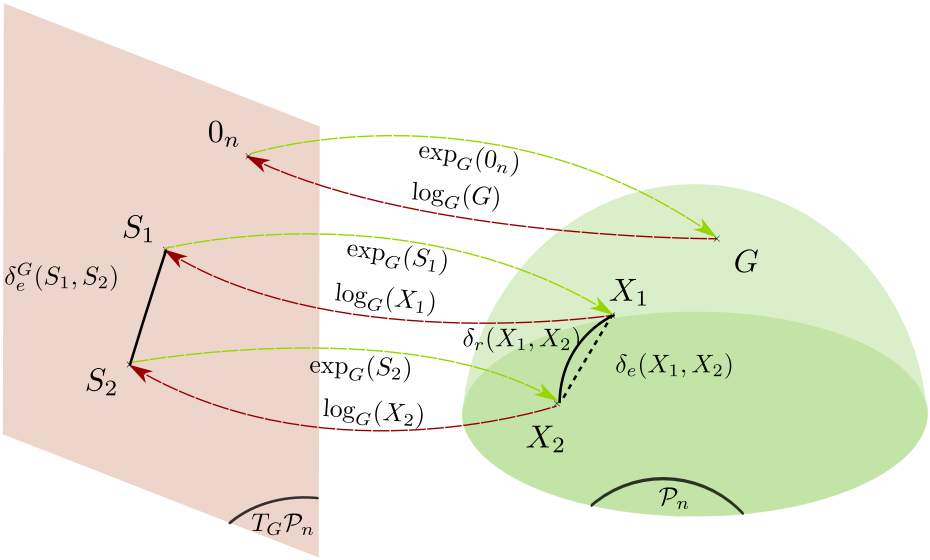
$$T_p\mathcal{M} \to \mathcal{M}$$
$$\exp_p(v) = \gamma_v(1)$$

Logarithmic map
The logarithmic map flattens the manifold around one point (i.e. the tangent plan tangent at this point).

Remarks

- back and forth between a manifold $\mathcal{M}$ and (euclidean) $T_p\mathcal{M}$
- usually computationally expensive operations (retractions).

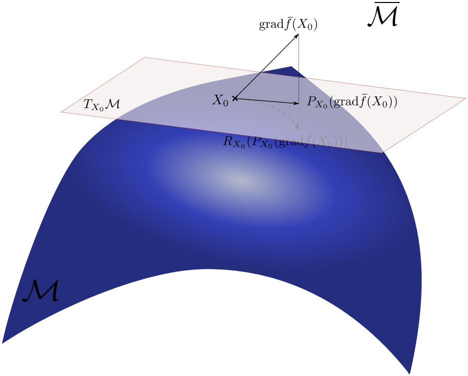
Goings and comings between a manifold and a tangent plane



Riemannian optimization in one slide

Key step

1. (euclidean) gradient
2. projection on the tangent plan (riemannian gradient)
3. exponential map (or retraction)



Philosophy of this approach

Move on a geodesic in the direction of the gradient (i.e. geodesic minimising the cost).

Bestiary of manifolds

Stiefel $St(p,n) = \{X \in \mathbb{R}^{n\times p} : X^\top X = \mathbb{I}_p\}$

ex : the sphere : $St(1,3)$

orthogonal group $\mathcal{O}(p) = \{X \in \mathbb{I}_p^{p\times p} : X^\top X = \mathbb{I}_p\}$

ex : $\mathcal{O}(p) = St(p,p)$

dp matrices $\mathcal{P}_n = \{X \in \mathbb{R}^{n\times n} : X \succ 0\}$

ex : covariance matrices

Bestiary of manifolds

Stiefel $St(p,n) = \{X \in \mathbb{R}^{n\times p} : X^\top X = \mathbb{I}_p\}$

ex : the sphere : $St(1,3)$

orthogonal group $\mathcal{O}(p) = \{X \in \mathbb{I}_p^{p\times p} : X^\top X = \mathbb{I}_p\}$

ex : $\mathcal{O}(p) = St(p,p)$

dp matrices $\mathcal{P}_n = \{X \in \mathbb{R}^{n\times n} : X \succ 0\}$

ex : covariance matrices

and many others...

Simple yet efficient approach I

MDRM

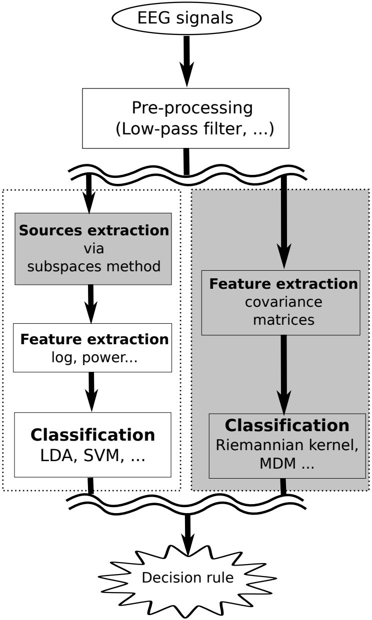
- algorithm
  For each class j $\quad \min_{X_j \in \mathcal{P}_n} \sum_{i \in \text{Class } j} \delta_R^2(C_i, X_j)$
- j independent problems of Fréchet/Karcher averaging
- prediction for C: $argmin_j \delta_R^2(C, X_j)$

Simple yet efficient approach II

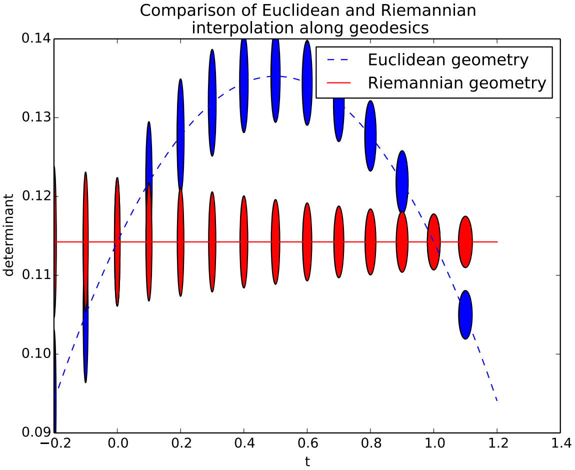
TSLDA

- algorithm
  compute M the Fréchet mean of the whole dataset of covariances
  map each covariance C_i to S_i in $T_G\mathcal{P}_n$ with
  $\log_M(C_i) = \log(M^{-\frac{1}{2}} C_i M^{-\frac{1}{2}})$
  apply a simple LDA on the extracted feature
- interpretation in term of withening for $M^{-\frac{1}{2}} C_i M^{-\frac{1}{2}}$
- linearization of the manifold around M

A simpler framework



Swelling effect



It can occur that $det(\frac{A+B}{2}) \geq \max(det(A), det(B))$, which is an artifact of the Euclidean framework.