# Merging databases

## - Big Data Project-

Chloe Dimeglio

UMR 1027 team 5

April $10^{th}$ 2015

Hôpitaux de Toulouse

UNIVERSITÉ TOULOUSE III PAUL SABATIER

# Outline

1. Introduction
2. Mathematical question
   - Modeling in the continuous case
   - Modeling in the dicrete case
3. Applied part
4. Recommendations

# Outline

# Definition

The problem when **merging databases** is in **associating**, **mixing** and **including** data from heterogeneous sources. The aim of this work is to provide a **strong knowledge base** to make decisions that may ultimately allow us to extract **more information from merged data** than we would get from using the databases seperately.

# Example

| Base A | | |
|---|---|---|
| Sexe | Age | Activité |
| M | 30 | 1 |
| M | 65 | 0 |
| M | 63 | 1 |
| F | 15 | 0 |
| M | 3 | 0 |
| F | 43 | 1 |

| Base B | | |
|---|---|---|
| Sexe | Age | Activité |
| M | 32 | 5 |
| F | 28 | 4 |
| F | 46 | 8 |
| M | 68 | 7 |
| M | 8 | 8 |
| M | 11 | 8 |

• We have two databases $A$ and $B$, and a common variable "Activité" coded in two different ways in each dataset.
• In each dataset, we have the same covariables linked to the target variable.

# Background

Example : longitudinal data

- If we change the mode of data collection during the same study
- If we create a common variable for the same people but at different times –> **merging** of longitudinal data.

**Problem** : **How to complete the cohort ?**

Example : cross-sectional data

- If we collect the same information in different ways during different studies.
- If we collect the same variable for the same people at the same time $t$ –> **merging** of cross-sectional data.

**Problem** : **How to consider all the information ?**

# Data fusion process

## Classical methods

- Bayesian networks
- Hidden Markov Models
- Probabilistic graphical models
- Least squares technique

Xu,L., Krzyzak, A. and Suen, C. (1992) : Méthods of combinig multiple classifiers and their application to handwriting recognition

Moravec, H. (1987) : Sensor fusion in certainty grids for mobile robots

Rabiner, L. (1989) :A tutorial on hidden Markov models and selected applications in speech recognition

Pearl, J. (1988) : Probabilistic reasoning in intelligent systems

Abidi, M and Gonzalez, R (1992) : Data fusion in robotics and machine intelligence

# New approach

**Merging databases** from a common variable using **optimal transport**

Ambrosio, L., Brenier, Y., Buttazzo, G., Caffarelli, L., Evans, L.C., Pratelli, A. and Villani, C. (2001) : Optimal transportation and applications

Villani, C. (2012) : Topics in optimal transportation

# Outline

# Prerequisites

## Framework

$A$ and $B$ are two databases.
We define $X$ and $Y$ the common variable which was coded in two different ways.

| | | | |
|---|---|---|---|
| X | x1 | x2 | ... |
| P(X=xi) | a1 | a2 | ... |
| | | | |
| Y | y1 | y2 | ... |
| P(Y=yj) | b1 | b2 | ... |

$cov(X)$ et $cov(Y)$ are the covariables associated with the common variable. Same covariables on the same scale in the two databases.

# Optimal transport

## Idea

We have two measures $\nu$ et $\mu$ such that law$(X) = \mu$ and law$(Y) = \nu$. We want to determine a measurable function $T$ such that $\nu = T\mu$. $T$ is a change of variables from $\mu$ to $\nu$.

## Continuous case

We have unicity of the function $T$ and we garantee the optimal transportation.

## Discrete case

The functions $T$ such that $T : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$ are all possible solutions. They are called transference plans from $A$ to $B$.

## The optimal transport

We introduce a **cost function** that can be interpreted as the cost of moving one unit of mass from a location in $A$ to a location in $B$.

# Outline

# Optimal transport : an example for continuous datasets

For instance if $X \sim \mathcal{N}(\mu_1, \sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2)$.
We can estimate $\mu_2$ and $\sigma_2$ in database $B$ and $\mu_1$ and $\sigma_1$ in database $A$.
Let $\hat{\mu}_1$ be the estimation of $\mu_1$ in database $A$ etc...
We have the following transportation :

$$X = (Y - \hat{\mu}_2)\frac{\hat{\sigma}_1}{\hat{\sigma}_2} + \hat{\mu}_1$$

We have **existence and uniqueness** of an optimal transport map for **continuous datasets**.

# Outline

# Measures and transference plans

• Let $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ the measure on base $A$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$ the measure on base $B$.

• The transference plans are the matrix $\gamma$ such as :

$$\gamma = \sum_{i,j} \gamma_{i,j} \delta_{(x_i, y_j)}$$

Where :

$$\sum_{j} \gamma_{i,j} = a_i$$

and

$$\sum_{i} \gamma_{i,j} = b_j$$

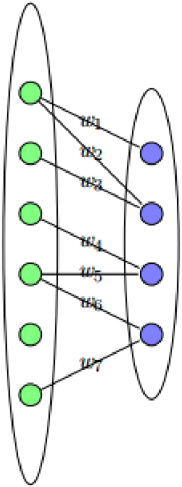
We have **not the uniqueness of the transport** –> Hitchcock's problem

# The cost function

- The cost function is defined as $c(\gamma) =$ **coupling risk** .
- Let $c(cov(x_i), cov(y_j))$ the distance between the **covariable distributions.**

$$c(\gamma) = \sum_{i,j} \gamma_{i,j} c(cov(x_i), cov(y_j))$$

# Risk of coupling




## How to define a risk ?

- We consider the distributions of covariables in the two bases. The more different the distributions in base $A$ and $B$, the greater the risk.
- The risk is defined from **the difference between the entropies of the covariable distributions**.

Our aim is to minimize this risk.

# Cost function

Let $K$ be the number of covariables.
Let $S$ be the number of modes.
Let the cost function be defined by :

$$c(\gamma) = \sum_{k=0}^{K} \sum_{i} \sum_{j} \sum_{s=0}^{S} \gamma_{i,j} \left| p_{i,s}^{k} \ln p_{i,s}^{k} - q_{j,s}^{k} \ln q_{j,s}^{k} \right|$$

Where $p_{i,s}^{k} = \mathbb{P}(\mathrm{cov}_k X = a_s | x_i)$ and $q_{j,s}^{k} = \mathbb{P}(\mathrm{cov}_k Y = b_s | y_j)$ with $p \ln(p) = 0$ when $p = 0$.

# Outline

1. Introduction
2. Mathematical question
   - Modeling in the continuous case
   - Modeling in the dicrete case
3. Applied part
4. Recommendations

# Practical case

## Background

- We are interested in the wage category of a sample of people.
- In dataset $A$, it's rated on a scale from 1 to 2.
- In dataset $B$, it's rated on a scale from 1 to 3.

## Data distribution

- In dataset $A$, 3 people were assessed as belonging to 1, and 5 people to 2.
- In dataset $B$, 4 people were assessed as belonging to 1, 2 to 2 and 2 to 3.

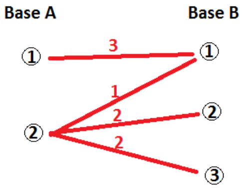
# Practical case

## Feasible solution

An application to transport the distribution of the variable from dataset $A$ to dataset $B$ satisfies the following transfer matrix :

$$\gamma = \begin{pmatrix} 3 & 0 & 0 \\ 1 & 2 & 2 \end{pmatrix}$$

## Corresponding graph



# Practical case

## Feasible solution

The following matrix is another solution : $\gamma = \begin{pmatrix} 1 & 1 & 1 \\ 3 & 1 & 1 \end{pmatrix}$

## Question

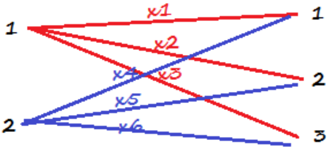
How to determine an optimal transfer ?

# Solving

## Flow of minimum cost

- We want to determine $\text{argmin}_{i,j} c(\gamma)$ under the constraints $Ax = b$
- Where $A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$, $b = \begin{pmatrix} 3 \\ 5 \\ 4 \\ 2 \\ 2 \end{pmatrix}$ and $x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix}$

## Outline



# Results

- When the variable is **completely determined by the covariables**, we have a **perfect coincidence** between the prediction and the "truth".
- We still **have to test situations closer to clinical reality**.

# Outline

# Recommendations

• Our work is based on a **strong assumption** : when you transport a distribution from one database to another, you have to ensure the **populations are comparable**.
If you force the behavior of a variable, you distort the information associated.
• When you transport a distribution from one database to another, you define a **reference population**. It's important to consider the **clinical reality** to ensure this definition is not too far from the objectives and the associated issues.

# To be continued

- To define an allocation rule for each person
- To test the validity when introducing a randomness in determining the variable using covariables
- To test the validity of the fusion when introducing missing data

Thank you