On the use of optimal transportation theory to merge databases

Valérie Garès*, Chloé Dimeglio*, Grégory Guernec, Romain Fantin, Benoit Lepage, Michael R. Kosorok, Nicolas Savy

12 january 2018

Conference Cimi on Statistics and Health, Toulouse

Problem

Optimal transportation

Simulation

Example

Current work

Problem

Problem

Optimal transportation

Simulation

Example

Current work

Context and problem

- Sharing and producing information from heterogeneous sources become a major issue in the Big Data context.

- The main issue is to **merge databases** from different sources in order to provide a strong knowledge database.

- The problem of **recoding** variables is very usual.
  - can occur when a categorical variable is not coded in the same scale in the two databases.

Context and problem

- Let two databases A and B
  - *In one study*
    - *a change in the survey questionnaire for the same information*
      two waves of recrutement (for different subjects)
      two waves at different ages (for same subjects)
  - *Two studies*
    - *a different questionnaire for the same information*
- n_j : the number of subjects in database $j = A, B$

Context and problem

- Let two databases A and B
  - *In one study*
    - *a change in the survey questionnaire for the same information*
      two waves of recrutement (for different subjects)
      two waves at different ages (for same subjects)
  - *Two studies*
    - *a different questionnaire for the same information*
- n_j : the number of subjects in database $j = A, B$

- Let Y the outcome **coded differently**
  - Y^A outcome coded *in the database A*
  - Y^B outcome coded *in the database B*
- *Example*
  - ***Outcome** : the mother's health status*
    - *"How would you rate your overall health ?"*
  - *During the first baseline data collection wave (January to April 2011)*
    - *five point ordinal scale : "excellent", "very well", "well", "fair", "bad"*
  - *During the second baseline data collection wave (May to December 2011)*
    - *five different point ordinal scale : "very well", "well", "medium" "bad" and "very bad"*

- $(C_1, C_2, \ldots, C_p)$: covariates

Context and problem

How to merge these two databases ?

Database A

	C_1	C_2	...	C_p	Y^A	Y^B
1					Observed	Unobserved
...						
...						
n_A						

Database B

	C_1	C_2	...	C_p	Y^A	Y^B
1					Unobserved	Observed
...						
...						
n_B						

Context and problem

How to merge these two databases ?

Database A

	C_1	C_2	...	C_p	Y^A	Y^B
1					Observed	Unobserved
...						
...						
n_A						

Database B

	C_1	C_2	...	C_p	Y^A	Y^B
1					Unobserved	Observed
...						
...						
n_B						

Aim : Complete Y^A on database B and/or complete Y^B on database A

- Ideas
  - Missing data problem (MAR)
  - Latent variables models (class latent analysis, trait latent analysis)
  - Estimation (polytomous regression) / Prediction
  - Optimal transportation

Optimal transportation

Optimal transportation

Transport map

We have two measures μ et ν such that distribution of Y^A is μ and distribution of Y^B is ν. We want to determine a measurable function T such that $\nu = T\mu$. T is a **transport application** from μ to ν.

Optimal transportation

μ

ν

Optimal transportation

y^A

$y^B = T\left(y^A\right)$

$\xrightarrow{T}$

μ

ν

Optimal transportation

y^A

$y^B = T\left(y^A\right)$

$\xrightarrow{T}$

μ

ν

Optimal transportation

- Let c a cost function measuring the displacement from y^A to y^B
- **Find a map T such that the average displacement is minimal**

Optimal transportation

- $\mathbb{Y}^A$ and $\mathbb{Y}^B$: two Radon spaces
- $c : \mathbb{Y}^A \times \mathbb{Y}^B \longrightarrow [0, \infty]$ a Borel-measurable function given probability measures μ on $\mathbb{Y}^A$ and ν on $\mathbb{Y}^B$ (**cost function**)

- **Monge's formulation** (1781) : Find a transport map $T : \mathbb{Y}^A \to \mathbb{Y}^B$ that realizes the infimum :

$$\left\{ \int_{\mathbb{Y}^A} c\left(y^A, T\left(y^A\right)\right) \mathrm{d}\mu\left(y^A\right) \;\middle|\; T(\mu) = \nu \right\},$$

  - **Optimal transportation map** : map T realizing this infimum
  - *Non-linear optimization problem, rigid assumptions on the regularity of T*

Optimal transportation

- $\mathbb{Y}^A$ and $\mathbb{Y}^B$: two Radon spaces
- $c : \mathbb{Y}^A \times \mathbb{Y}^B \longrightarrow [0, \infty]$ a Borel-measurable function given probability measures μ on $\mathbb{Y}^A$ and ν on $\mathbb{Y}^B$ (**cost function**)

- **Monge's formulation** (1781) : Find a transport map $T : \mathbb{Y}^A \to \mathbb{Y}^B$ that realizes the infimum :

$$\left\{ \int_{\mathbb{Y}^A} c\left(y^A, T\left(y^A\right)\right) \mathrm{d}\mu\left(y^A\right) \;\middle|\; T(\mu) = \nu \right\},$$

  - **Optimal transportation map** : map T realizing this infimum
  - *Non-linear optimization problem, rigid assumptions on the regularity of T*

- **Kantovorich's formulation** (1942) : Find a measure $\gamma \in \gamma(\mu, \nu)$ that realizes the infimum :

$$\left\{ \int_{\mathbb{Y}^A \times \mathbb{Y}^B} c\left(y^A, y^B\right) \mathrm{d}\gamma\left(y^A, y^B\right) \;\middle|\; \gamma \in \gamma(\mu, \nu) \right\},$$

  where $\gamma(\mu, \nu)$ denote the set of measures on $\mathbb{Y}^A \times \mathbb{Y}^B$ with marginals μ on $\mathbb{Y}^A$ and ν on $\mathbb{Y}^B$
  - *Linear problem, solution achievable with compacity (volume fitting) argument*

Optimal transportation

- $\mathbb{Y}^A$ and $\mathbb{Y}^B$: two Radon spaces
- $c : \mathbb{Y}^A \times \mathbb{Y}^B \longrightarrow [0, \infty]$ a Borel-measurable function given probability measures μ on $\mathbb{Y}^A$ and ν on $\mathbb{Y}^B$ (**cost function**)

- **Monge's formulation** (1781) : Find a transport map $T : \mathbb{Y}^A \to \mathbb{Y}^B$ that realizes the infimum :

$$\left\{ \int_{\mathbb{Y}^A} c\left(y^A, T\left(y^A\right)\right) \mathrm{d}\mu\left(y^A\right) \;\middle|\; T(\mu) = \nu \right\},$$

  - **Optimal transportation map** : map T realizing this infimum
  - *Non-linear optimization problem, rigid assumptions on the regularity of T*

- **Kantovorich's formulation** (1942) : Find a measure $\gamma \in \gamma(\mu, \nu)$ that realizes the infimum :

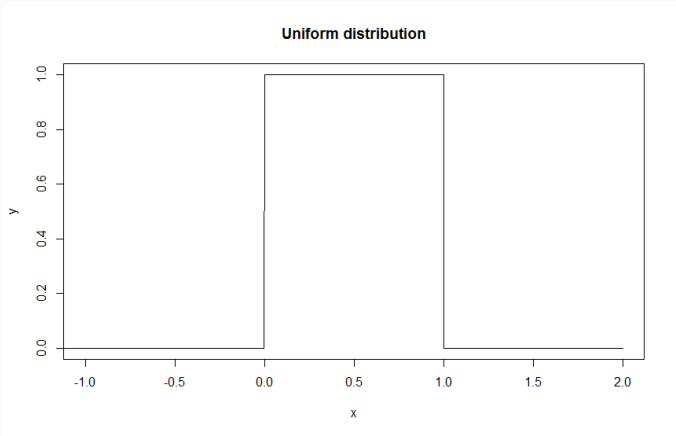
$$\left\{ \int_{\mathbb{Y}^A \times \mathbb{Y}^B} c\left(y^A, y^B\right) \mathrm{d}\gamma\left(y^A, y^B\right) \;\middle|\; \gamma \in \gamma(\mu, \nu) \right\},$$

  where $\gamma(\mu, \nu)$ denote the set of measures on $\mathbb{Y}^A \times \mathbb{Y}^B$ with marginals μ on $\mathbb{Y}^A$ and ν on $\mathbb{Y}^B$
  - *Linear problem, solution achievable with compacity (volume fitting) argument*

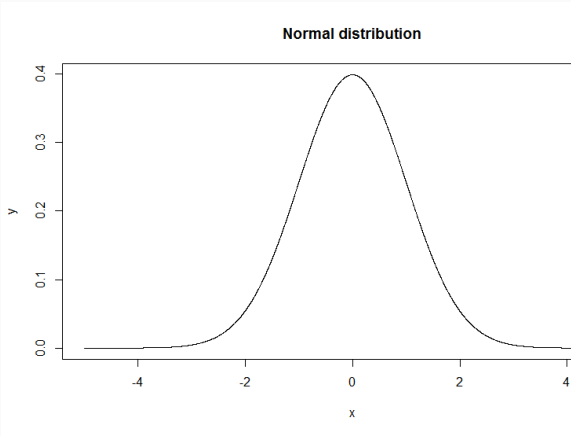
- **Villani (2009)** : a minimizer for this problem always exists when the cost function c is lower semi-continuous

Optimal transportation

$\mathcal{U}[0,1]$

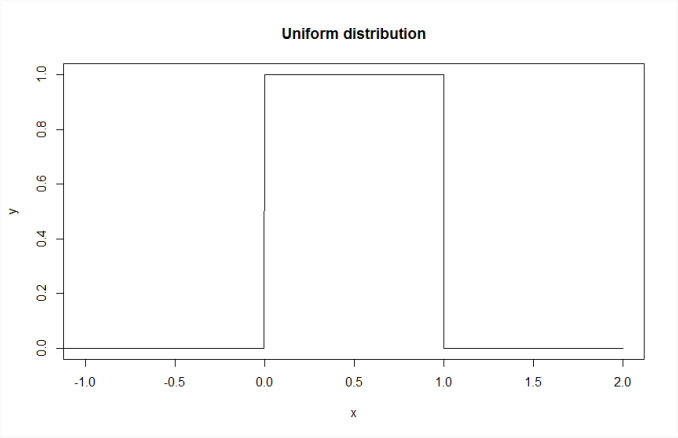


$\mathcal{N}(0,1)$



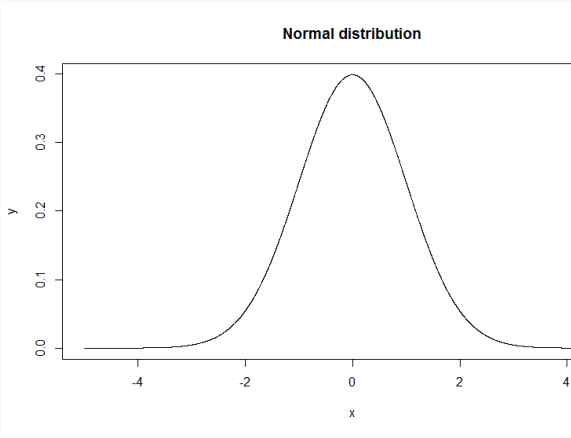
Optimal transportation

$\mathcal{U}[0,1]$



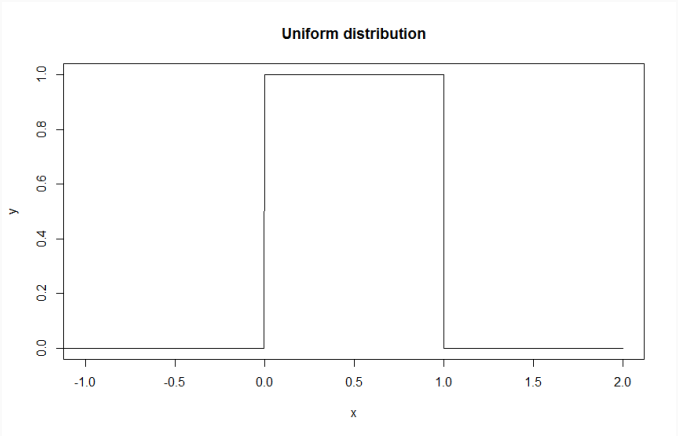
$$\xrightarrow[c(x,y)=(x-y)^2]{T^*(x)=\frac{1}{2\pi}\exp\left(-\frac{x^2}{2}\right)}$$

$\mathcal{N}(0,1)$

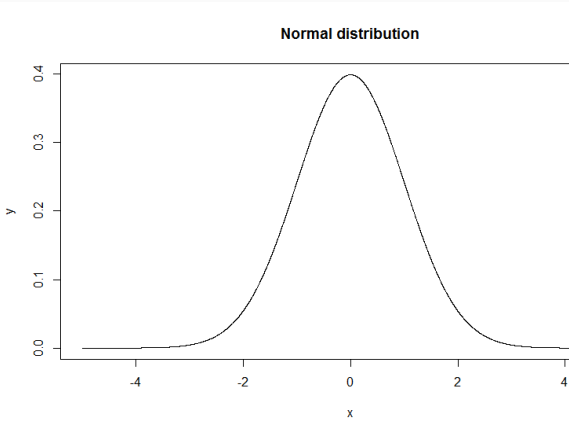


Optimal transportation

$\mathcal{U}[0,1]$ $\mathcal{N}(0,1)$



$$\xrightarrow[c(x,y)=(x-y)^2]{T^*(x)=\frac{1}{2\pi}\exp\left(-\frac{x^2}{2}\right)}$$



- The optimal transportation map **exists** and is **unique** if h is strictly convex with $c(x,y) = h(x-y)$

Optimal transportation

- Discrete case : Hitchcock's problem (1941)

- Y evaluated in both databases but not assessed on the same variable
  - Y^A the assessment of Y on database $D = A$
    - with distribution μ discrete with modalities $\{m_1^A, \ldots, m_R^A\}$
    - $a_r = \mathbb{P}(Y^A = m_r^A)$, $r = 1, \ldots, R$

  - Y^B the assessment of Y on database $D = B$
    - with distribution ν discrete with modalities $\{m_1^B, \ldots, m_S^B\}$
    - $b_s = \mathbb{P}(Y^B = m_s^B)$, $s = 1, \ldots, S$

$$\mu = \sum_{r=1}^{R} a_r \delta_{m_r^A} \qquad \text{and} \qquad \nu = \sum_{s=1}^{S} b_s \delta_{m_s^B}$$

- $C = (C_1, C_2, \ldots, C_p)$, covariates

- **Aim** :
  - **Complete Y^A on database B and/or complete Y^B on database A**

Optimal transportation

- ***Assumption 1 :*** $Y^j \sim Y^j|D = A \sim Y^j|D = B,\ j = A, B$
- *Example*
  - ***Outcome*** *: Self rated health*
  - ***Databases*** *: two waves of recruitment* ✓
  - ***Databases*** *: France (French National Health Study) and USA (North America NHANES study)* ×

Optimal transportation

- ***Assumption 1 :*** $Y^j \sim Y^j|D = A \sim Y^j|D = B,\ j = A, B$
- *Example*
  - ***Outcome*** *: Self rated health*
  - ***Databases*** *: two waves of recruitment* ✓
  - ***Databases*** *: France (French National Health Study) and USA (North America NHANES study)* ×
- We can estimate the distribution of Y^A and Y^B by the following estimators :

$$\hat{a}_r = \frac{\text{card}\,\{i|y_i^A = m_r^A\}}{n_A}, \qquad r = 1, \dots R,\ (\mathcal{L}(Y^A) \sim \mathcal{L}(Y^A|D = A))$$

$$\hat{b}_s = \frac{\text{card}\,\{j|y_j^B = m_s^B\}}{n_B}, \qquad s = 1, \dots S,\ (\mathcal{L}(Y^B) \sim \mathcal{L}(Y^B|D = B))$$

Optimal transportation

- ***Assumption 1 :*** $Y^j \sim Y^j|D = A \sim Y^j|D = B,\ j = A, B$
- *Example*
  - ***Outcome*** *: Self rated health*
  - ***Databases*** *: two waves of recruitment* ✓
  - ***Databases*** *: France (French National Health Study) and USA (North America NHANES study)* ×
- We can estimate the distribution of Y^A and Y^B by the following estimators :

$$\hat{a}_r = \frac{\text{card}\left\{i|y_i^A = m_r^A\right\}}{n_A}, \qquad r = 1, \dots R,\ (\mathcal{L}(Y^A) \sim \mathcal{L}(Y^A|D = A))$$

$$\hat{b}_s = \frac{\text{card}\left\{j|y_j^B = m_s^B\right\}}{n_B}, \qquad s = 1, \dots S,\ (\mathcal{L}(Y^B) \sim \mathcal{L}(Y^B|D = B))$$

- ***Assumption 2*** : $Y^j|C, D = A \sim Y^j|C, D = B,\ j = A, B$
- *Example*
  - ***Covariates*** C ***:*** *comorbidities, education level, age, sexe*
  - ***Outcome*** 1 *: functional limitation* ✓

Optimal transportation

- ***Assumption 1 :*** $Y^j \sim Y^j|D = A \sim Y^j|D = B,\ j = A, B$
- *Example*
  - ***Outcome*** *: Self rated health*
  - ***Databases*** *: two waves of recruitment* ✓
  - ***Databases*** *: France (French National Health Study) and USA (North America NHANES study)* ×
- We can estimate the distribution of Y^A and Y^B by the following estimators :

$$\hat{a}_r = \frac{\text{card}\left\{i|y_i^A = m_r^A\right\}}{n_A}, \qquad r = 1, \dots R,\ (\mathcal{L}(Y^A) \sim \mathcal{L}(Y^A|D = A))$$

$$\hat{b}_s = \frac{\text{card}\left\{j|y_j^B = m_s^B\right\}}{n_B}, \qquad s = 1, \dots S,\ (\mathcal{L}(Y^B) \sim \mathcal{L}(Y^B|D = B))$$

- ***Assumption 2*** : $Y^j|C, D = A \sim Y^j|C, D = B,\ j = A, B$
- *Example*
  - ***Covariates*** C ***:*** *comorbidities, education level, age, sexe*
  - ***Outcome*** 1 *: functional limitation* ✓
  - ***Covariates*** C ***:*** *functional limitation, education level, age, sexe*
  - ***Outcome*** 2 *: self rated health* ×

Optimal transportation

- ***Assumption 2*** : $\mathcal{L}(Y^j|C, D = A) \sim \mathcal{L}(Y^j|C, D = B)$, $j = A, B$

Cost function

$$c(m_r^A, m_s^B) = \frac{1}{\kappa_{r,s}} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(C_i, C_j)\, \mathbb{I}\left(y_i^A = m_r^A \ , \ y_j^B = m_s^B\right),$$

where

- $\kappa_{r,s} = \text{card}\left\{(i,j)|y_i^A = m_r^A \ , \ y_j^B = m_s^B\right\}$
- $d(C_i, C_j)$ is the distance between vectors of covariates

Optimal transportation

The choice of the distance depends on the type of the covariates.

- **Continuous covariates**
  - Euclidian or Manhattan distance
- **Categorical covariates**
  - Hamming distance from the associated complete disjunctive tables
- **Mixed covariates**
  - distance for mixed data
    - Heterogeneous Euclidean-Overlap Metric
    - Value Difference Metric
    - Mahalanobis distance
  - distance for continuous covariates on the coordinates extracted from a factor analysis of mixed data (FMDA)

Optimal transportation

For $r = 1, \ldots, R$, $s = 1, \ldots, S$,

- $\gamma_{rs} = \mathbb{P}(Y^A = m_r^A, Y^B = m_s^B)$, the joint distribution of (Y^A, Y^B)
- $a_r = \mathbb{P}(Y^A = m_r^A)$, the distribution of Y^A
- $b_s = \mathbb{P}(Y^B = m_s^B)$, the distribution of Y^B

		Database A			Dist.
		m_1^A	$\ldots$	m_R^A	of Y^B
Database B	m_1^B	$\gamma_{1,1}$	$\ldots$	$\gamma_{1,R}$	b_1
	$\ldots$	$\ldots$	$\gamma_{r,s}$	$\ldots$	$\ldots$
	m_S^B	$\gamma_{S,1}$	$\ldots$	$\gamma_{S,R}$	b_S
Dist. of Y^A		a_1	$\ldots$	a_R	1

Table 1 – Joint distribution of (Y^A, Y^B)

OT algorithm

- Kantovorich's formulation : Find γ which realizes the infimum of

$$\int_{\mathbb{Y}^A \times \mathbb{Y}^B} c\left(y^A, y^B\right) \, \mathrm{d}\gamma\left(y^A, y^B\right)$$

OT algorithm

- Linear Programming problem : Find γ which realizes the infimum of

$$\sum_{r=1}^{R}\sum_{s=1}^{S} \gamma_{r,s}\ c\left(m_r^A, m_s^B\right)$$

OT algorithm

- Linear Programming problem : Find γ which realizes the infimum of

$$\sum_{r=1}^{R}\sum_{s=1}^{S} \gamma_{r,s}\ c\left(m_r^A, m_s^B\right)$$

  under the following constraints which must hold for any r and any s,

$$\gamma_{r,s} \geq 0, \quad \sum_{r=1}^{R}\gamma_{r,s} = b_s \quad \text{and} \quad \sum_{s=1}^{S} \gamma_{r,s} = a_r$$

- Exhibits γ **the joint distribution** of Y^A and Y^B

OT algorithm : affectation

- For each subjects i of database A, a predicted value for $\hat{y}_i^B$ can be specified by means of an **adapted nearest neighbor algorithm** accounting for the covariates with distance d.
  - $N_{r,s} = Ent(n_A \times \gamma_{r,s})$: number of subjects having modality m_r^A and m_s^B in database A
  - $N_{r',s'} = \max_{r,s} N_{r,s}$
  - For each subject i who was in modality $m_{r'}^A$ in database A, compute the average distance between this subject and all subjects having modality $m_{s'}^B$ in database B

$$c_i(m_{r'}^A, m_{s'}^B) = \frac{1}{\text{card}\,\{j|\ y_j^B = m_{s'}^B\}} \sum_{j=1}^{n_B} d(C_i, C_j)\ \mathbb{I}\left(y_j^B = m_{s'}^B\right),$$

  - Keep the first $N_{r',s'}$ subjects who have the minimum average distances. Assign them modality $m_{s'}^B$ for Y^B
  - Repeat the previous step with $max_{r,s}(N_{r,s} \backslash N_{r',s'})$

Simulation

Simulation

- (D_1, D_2, D_3), 3 covariates, $\sim \mathcal{N}(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & \rho & \delta \\ \rho & 1 & \mu \\ \delta & \mu & 1 \end{pmatrix}$$

  - $var(D_1) = var(D_2) = var(D_3) = 1$
  - $cor(D_1, D_2) = \rho$
  - $cor(D_1, D_3) = \delta$
  - $cor(D_2, D_3) = \mu$

- Y : an outcome defined by

$$Y = D_1 + D_2 + D_3 + \sigma\epsilon$$

  with $\epsilon \sim \mathcal{N}(0, 1)$

Simulation

We discretise and then observe only

- $C_1 = \mathbb{I}_{\{D_1 > t\}}$ in 2 categories
  - given a Bernoulli distribution $B(\pi_1)$ with $\pi_1 = \mathbb{P}(D_1 > t)$
- $C_2 = \mathbb{I}_{\{t_1 < D_2 < t_2\}} + \mathbb{I}_{\{D_2 > t_2\}}$ in 3 categories
  - given a multinomial distribution $\mathcal{M}(\pi_{21}, \pi_{22})$ with $\pi_{21} = \mathbb{P}(t_1 < D_2 < t_2)$ and $\pi_{22} = \mathbb{P}(D_2 > t_2)$
- $C_3 = D_3$ stays in continuous

- Y^A in 4 categories (given quartile of Y)
- Y^B in 3 categories (given tertile of Y)

- We divide the initial database (n subjects) in two databases
  - n_A number of subjects in database A
  - n_B number of subjects in database B
- And observe only
  - Y^A in database A
  - Y^B in database B

Simulation

	C_1	C_2	C_3	Y^A	Y^B
1				Observed	Unobserved
...					
...					
n_A					
1				Unobserved	Observed
...					
...					
n_B					

Simulation

- We defined R^2 theoretical as

$$\begin{aligned} R^2 &= \frac{Var(D_1 + D_2 + D_3)}{Var(Y)} \\ &= \frac{3 + 2\rho + 2\delta + 2\mu}{3 + 2\rho + 2\delta + 2\mu + \sigma^2} \end{aligned}$$

  - R^2 varies between 0 and 1
  - *"**association measure** between the covariates and the outcome"*

Simulation

- We want to test the effect of
  - the sample size n
  - the correlation between the 3 covariates Σ
  - the association measure between the covariates and the outcome : R^2

  on the effectiveness of OT and MICE (multiple imputation by chained equation) methods

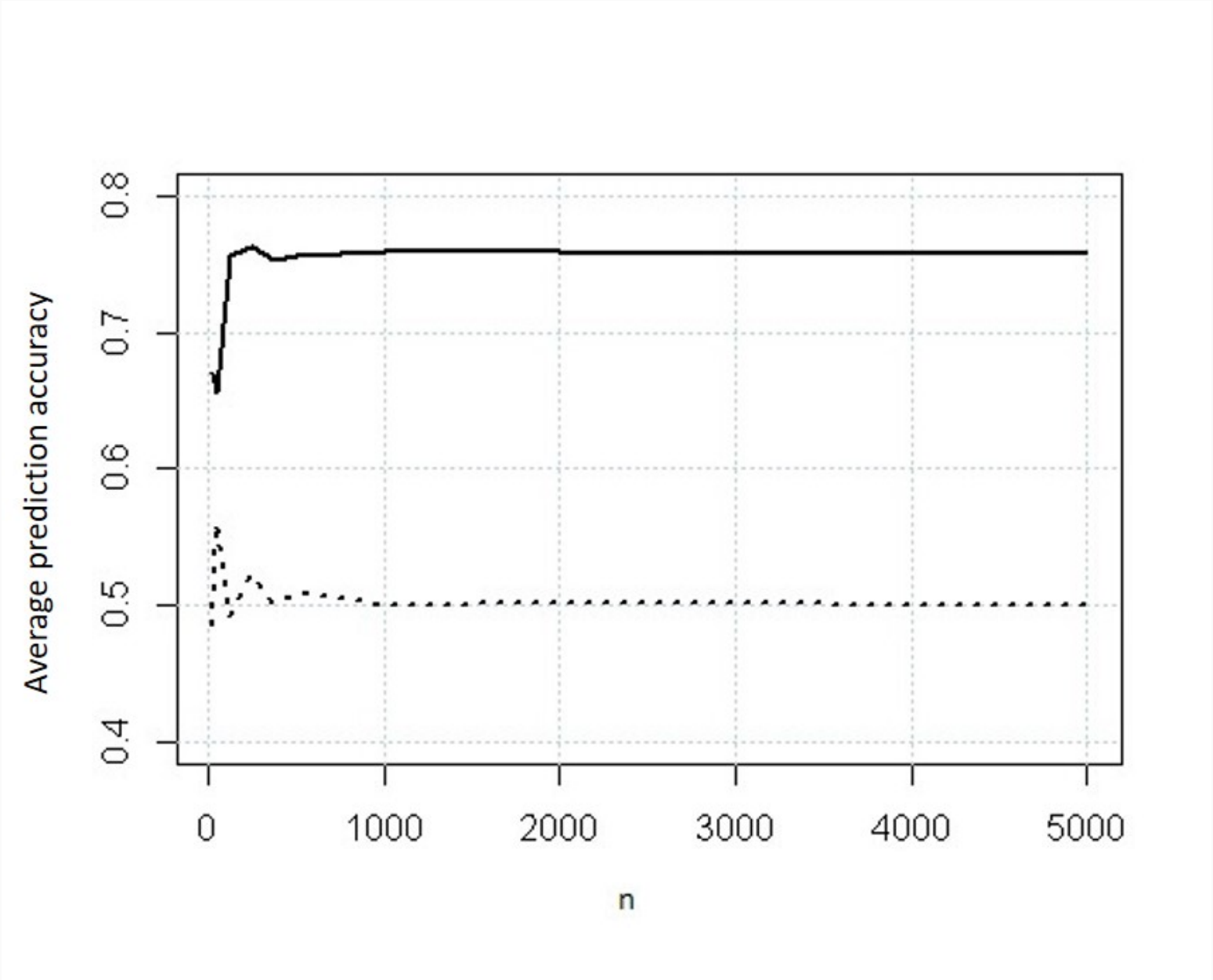
- We generate data depending on n, Σ and R^2

- On one database
  - We applied these different methods
  - We computed the % **of well classified subjects** *(Average prediction accuracy)*

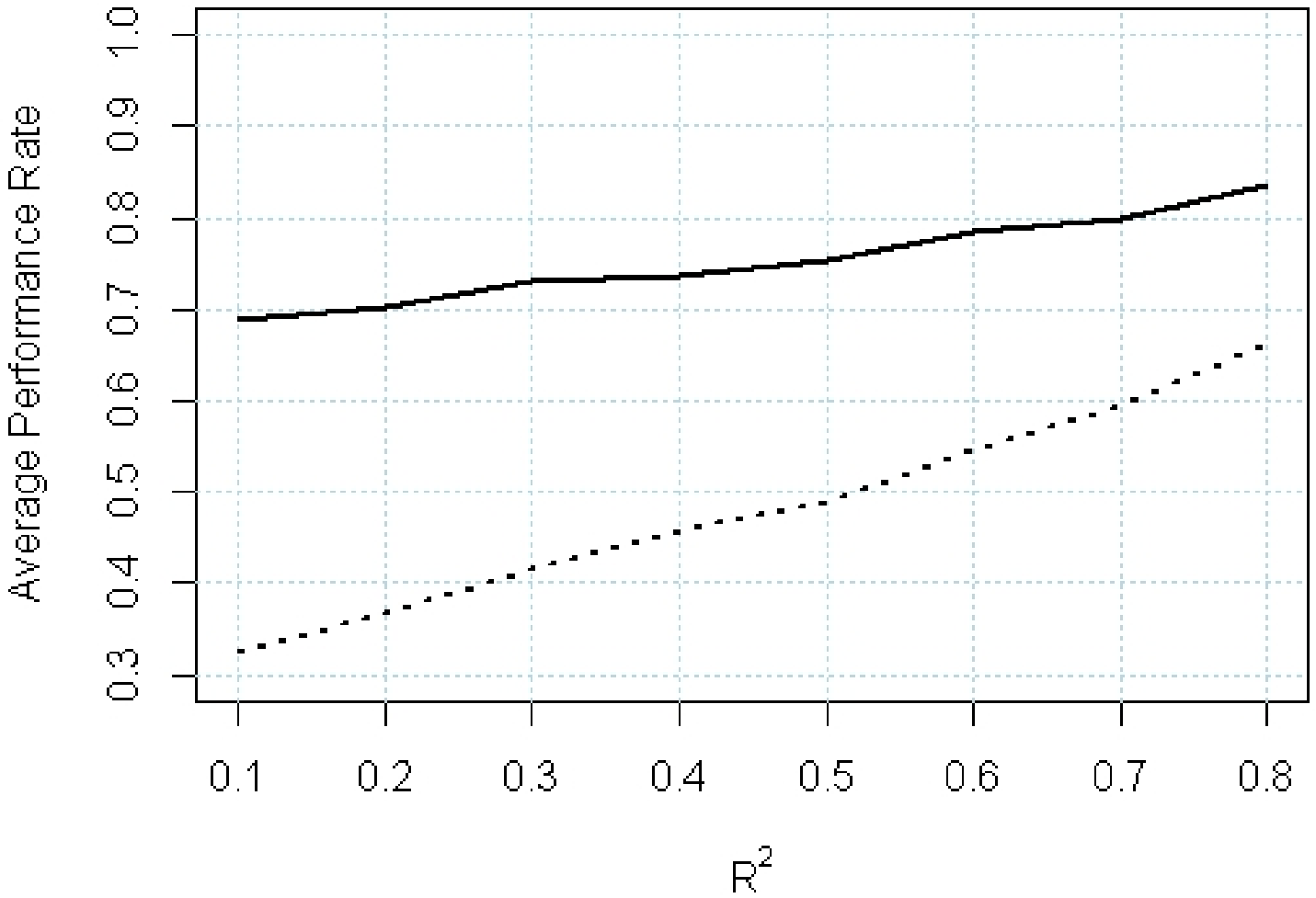
$$WC = \frac{1}{n_A}\sum_{i=1}^{n_A}\mathbb{I}(\hat{y}_i^B = y_i^B) + \frac{1}{n_B}\sum_{i=1}^{n_B}\mathbb{I}(\hat{y}_i^A = y_i^A)$$

- We compute the means $\pm$ standard errors of WC over $M = 1000$ databases

Results


Average prediction accuracy
0.4
0.5
0.6
0.7
0.8
0
1000
2000
3000
4000
5000
n

Results


Average Performance Rate
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
R^2

Results

- The average prediction accuracy of OT and MICE
  - increases as the sample size n increases
  - decreases as the ratio n_A/n_B increases
  - increases as the R^2 increases
  - remains stable as ρ increases

- The standard error of prediction accuracy of OT and MICE methods are small and remain stable.

- In each case, the OT algorithm demonstrates a better performance than the MICE algorithm.

Example

- ELFE (Etude Longitudinale Francaise depuis l'Enfance)
  - french cohort of 18 000 children, followed from birth
  - study how various contextual factors affect children's developmental health and wellbeing over time, and into adulthood

- **Outcome** : the mother's `health status`
  - *"How would you rate your overall health ?"*

- During the first baseline data collection wave (January to April 2011)
  - five point ordinal scale : *"excellent", "very well", "well", "fair", "bad"*
- During the second baseline data collection wave (May to December 2011)
  - five different point ordinal scale : *"very well", "well", "medium" "bad" and "very bad"*

Example

- Three covariates coded in the same way in both databases are selected for their ability to predict the outcomes :
  - The mother's age at baby birth in years (continuous)
  - The `physical limitations` of the mother reported for a duration of at least six months (categorical with six modalities)
  - The `chronic health problem` of the mother at two months of baby age (categorical with three modalities)

- Only subjects with no missing information related to these covariate, were retained

- 13554 participants (2233 in database A, 11321 in database B)

Example

Database *A*		Database *B*	
Modalities	n (%)	Modalities	n (%)
"excellent"	950 (42.54)	"very well"	1834 (16.20)
"very well"	1047 (46.89)	"well"	4374 (38.64)
"well"	212 (9.49)	"medium"	4586 (40.51)
"passable"	22 (0.99)	"bad"	478 (4.22)
"bad"	2 (0.00)	"very bad"	49 (0.43)

Table 2 – ELFE study. Description of the modalities of the outcome mother's `health status` for both databases.

Example

		Database A	Database B	
Covariates	Modalities	n (%)	n (%)	p-value
`Physical limitations`	Severely limited	18 (0.81)	64 (0.57)	0.20
	Limited	140 (6.27)	657 (5.80)	
	No	2075 (92.92)	10600(93.63)	
`Chronic health problem`	Yes	285 (12.76)	1433 (12.66)	0.99
	No	1948 (87.24)	9888 (87.34)	
age (in years)		30.77 $\pm$4.68	31.10 $\pm$ 4.80	**0.002**

Table 3 – ELFE study. Description of covariates (mother) in both databases.

- Association between the outcome and each covariates tested independently in each dataset
  - **Each p-value is less than** 10^{-14}.

Example

		$\hat{y}^B$				
		"very well"	"well"	"average"	"bad"	"very bad"
$\hat{y}^A$	"excellent"	2196	588	0	0	0
	"very well"	2982	1666	773	0	0
	"well"	0	3917	801	80	0
	"passable"	0	0	405	75	20
	"bad"	0	0	0	51	0

Table 4 – Joint frequencies of predicted values $(\hat{y}^A, \hat{y}^B)$

Example

- NCDS (The National Child Development Study)
  - a continuing survey which follows the lives of over 17,000 people born in England, Scotland and Wales in a same week of the year 1958
  - collects specific information on many distinct fields
    - *physical and educational development, economic circumstances, employment, family life, health behaviour, well-being, social participation and attitudes*
  - 9 waves (0, 7, 11, 16, 22, 33, 42, 50 and 55 years old)

- **Outcome** : two measurements scales of the **social class** of the participants built from profession and collected at wave 5 :
  - *Goldthorp social class'90 scale* (`GSS90`) : a scale in 11 categories
  - *RGs social Class'91 scale* (`RGS91`) : a scale in 6 categories.

- The initial database was randomly divided in two databases of the same size and we kept
  - the `GSS90` scale in the first database
  - the `RGS91` scale in the second database

Example

Example

- Four covariates coded in the same way in both databases are selected for their ability to predict the outcomes :
  - the `sex` of participants
  - the `health status`
  - the `employment status` at wave 5
  - the `study level` at wave 4

- Association between the outcome and each covariates tested independently in each dataset
  - **Each p-value is less than** 10^{-14}.

- Only subjects with no missing information related to these covariates were retained

- 8030 participants (4015 by database)

Example

Social class GSS90	Database *A*	Database *B*		Social class RGS91	Database *A*	Database *B*	
Modalities	n (%)	n (%)	p-value	Modalities	n (%)	n (%)	p-value
Not applicable	116 (2.89)	85 (2.12)	0.72	Not applicable	129 (3.21)	102 (2.54)	0.37
I	646 (16.09)	697 (17.36)		I	201 (5.01)	207 (5.16)	
II	761 (18.95)	702 (17.48)		II	1241 (30.91)	1214 (30.24)	
IIIa	650 (16.19)	683 (17.01)		IIIN	930 (23.16)	982 (24.46)	
IIIb	349 (8.69)	311 (7.75)		IIIM	736 (18.33)	765 (19.05)	
IVa	13 (0.32)	12 (0.30)		IV	617 (15.37)	580 (14.45)	
IVb	146 (3.64)	146 (3.64)		V	161 (4.01)	165 (4.11)	
IVc	27 (0.67)	31 (0.77)					
V	161 (4.01)	182 (4.53)					
VI	426 (10.61)	435 (10.83)					
VIIa	699 (17.41)	705 (17.56)					
VIIb	21 (0.52)	26 (0.65)					

Table 5 – NCDS study. Description of the modalities of the outcomes for both databases.

Example

OT	MICE
63.5%	29.3%

Table 6 – NCDS study. % of well classified subjects.

Conclusion

- Summary
  - Our method is always more accurate than a particular multiple imputation process
  - This accuracy is highly linked to the R-square value

- Limitations
  - needs a lot of run times
  - Strong assumption : comparable populations.
    - $\mathcal{L}(Y^j) \sim \mathcal{L}(Y^j|D = A) \sim \mathcal{L}(Y^j|D = B)$, $j = A, B$
    - Non identifiable

Current work

Current work

- Missing data on covariates *(Guernec G., Gares V.)*
  - Multiple imputation on covariates and OT
  - OT with Gower distance

- Comparison with other models, based on statistical learning

  (Vuillemenot D., Lepage B., Saint-Pierre P.)
  - Logistic regression
  - Neuron network
  - Linear Discriminant Analysis
  - Naive Baise
  - Support vector machine
  - Decision tree
  - Random forest
  - Latent class analysis

Current work

- Missing data on covariates *(Guernec G., Gares V.)*
  - Multiple imputation on covariates and OT
  - OT with Gower distance

- Comparison with other models, based on statistical learning

  (Vuillemenot D., Lepage B., Saint-Pierre P.)
  - Logistic regression ✓
  - Neuron network ✓
  - Linear Discriminant Analysis ✓
  - Naive Baise ✓
  - Support vector machine ✓
  - Decision tree
  - Random forest
  - Latent class analysis

Current work

- Missing data on covariates *(Guernec G., Gares V.)*
  - Multiple imputation on covariates and OT
  - OT with Gower distance

- Comparison with other models, based on statistical learning

  (Vuillemenot D., Lepage B., Saint-Pierre P.)
  - Logistic regression ✓
  - Neuron network ✓
  - Linear Discriminant Analysis ✓
  - Naive Baise ✓
  - Support vector machine ✓
  - Decision tree
  - Random forest
  - Latent class analysis

- Extension of OT *(Omer J., Gares V.)*
  - Intersection : Outcome (Y^A, Y^B) observed in both databases
  - Different distributions $Y^j|D = A$ and $Y^j|D = B, j = A, B$

Optimal transportation

- ***Assumption 1*** : $Y^{)} \sim Y^{j}|D = A \sim \mathcal{L}(Y^{j}|D = B)$, $j = A, B$
- We **can** estimate the distribution of Y^{A} and Y^{B} by the following estimators :

$$\hat{a}_r = \frac{\text{card}\left\{i|y_i^A = m_r^A\right\}}{n_A}, \qquad r = 1, \dots R, \ (\mathcal{L}(Y^A) \sim \mathcal{L}(Y^A|D = A))$$

$$\hat{b}_s = \frac{\text{card}\left\{j|y_j^B = m_s^B\right\}}{n_B}, \qquad s = 1, \dots S, \ (\mathcal{L}(Y^B) \sim \mathcal{L}(Y^B|D = B))$$

- ***Assumption 2*** : $\mathcal{L}(Y^{j}|C, D = A) \sim \mathcal{L}(Y^{j}|C, D = B)$, $j = A, B$

Cost function

$$c(m_r^A, m_s^B) = \frac{1}{\kappa_{r,s}} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(C_i, C_j)\, \mathbb{I}\left(y_i^A = m_r^A \ , \ y_j^B = m_s^B\right),$$

where $\kappa_{r,s} = \text{card}\left\{(i,j)|y_i^A = m_r^A \ , \ y_j^B = m_s^B\right\}$ and $d(C_i, C_j)$ is the distance between vectors of covariates.

Optimal transportation

- ~~***Assumption 1*** : $Y^j \sim Y^j|D = A \sim Y^j|D = B,\ j = A, B$~~
- We **can't** estimate the distribution of Y^A and Y^B by the following estimators :

$$\hat{a}_r = \frac{\text{card}\left\{i|y_i^A = m_r^A\right\}}{n_A}, \qquad r = 1, \dots R,\ (\mathcal{L}(Y^A) \sim \mathcal{L}(Y^A|D = A))$$

$$\hat{b}_s = \frac{\text{card}\left\{j|y_j^B = m_s^B\right\}}{n_B}, \qquad s = 1, \dots S,\ (\mathcal{L}(Y^B) \sim \mathcal{L}(Y^B|D = B))$$

- ***Assumption 2*** : $\mathcal{L}(Y^j|C, D = A) \sim \mathcal{L}(Y^j|C, D = B),\ j = A, B$

Cost function

$$c(m_r^A, m_s^B) = \frac{1}{\kappa_{r,s}} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(C_i, C_j)\, \mathbb{I}\left(y_i^A = m_r^A\ ,\ y_j^B = m_s^B\right),$$

where $\kappa_{r,s} = \text{card}\left\{(i,j)|y_i^A = m_r^A\ ,\ y_j^B = m_s^B\right\}$ and $d(C_i, C_j)$ is the distance between vectors of covariates.

Optimal transportation

Conditional to the database $i = A, B$,

- $\gamma^i_{rs} = \mathbb{P}(Y^A = m^A_r, Y^B = m^B_s | D = i)$, $r = 1, \dots, R$, $s = 1, \dots, S$ the joint distribution of (Y^A, Y^B)
- $a^i_r = \mathbb{P}(Y^A = m^A_r | D = i)$, $r = 1, \dots, R$ the distribution of Y^A
- $b^i_s = \mathbb{P}(Y^B = m^B_s | D = i)$, $s = 1, \dots, S$ the distribution of Y^B

		Database A			Dist.
		m^A_1	...	m^A_R	of Y^B
Database B	m^B_1	$\gamma^A_{1,1}$	...	$\gamma^A_{1,R}$	b^A_1
	...	...	$\gamma^A_{r,s}$	...	...
	m^B_S	$\gamma^A_{S,1}$	...	$\gamma^A_{S,R}$	b^A_S
Dist. of Y^A		a^A_1	...	a^A_R	1

Table 7 – Joint distribution of $(Y^A, Y^B)|D = A$

		Database A			Dist.
		m^A_1	...	m^A_R	of Y^B
Database B	m^B_1	$\gamma^B_{1,1}$	...	$\gamma^B_{1,R}$	b^B_1
	...	...	$\gamma^B_{r,s}$	...	...
	m^B_S	$\gamma^B_{S,1}$	...	$\gamma^B_{S,R}$	b^B_S
Dist. of Y^A		a^B_1	...	a^B_R	1

Table 8 – Joint distribution of $(Y^A, Y^B)|D = B$

- $\gamma_{r,s} = \mathbb{P}(D = A)\gamma^A_{r,s} + \mathbb{P}(D = B)\gamma^B_{r,s}$

OT algorithm

- Linear Programming problem : Find γ which realizes the infimum of

$$\sum_{r=1}^{R}\sum_{s=1}^{S} \gamma_{r,s}\ c(m_r^A, m_s^B)$$

under the following constraints which must hold for any r and any s,

$$\gamma_{r,s} \geq 0, \quad \sum_{r=1}^{R} \gamma_{r,s}^A = b_s^A, \qquad \sum_{s=1}^{S} \gamma_{r,s}^A = a_r^A \quad \text{and} \quad \sum_{s=1}^{S}\sum_{r=1}^{R} \gamma_{r,s}^A = 1$$

$$\gamma_{r,s} \geq 0, \quad \sum_{r=1}^{R} \gamma_{r,s}^B = b_s^B, \qquad \sum_{s=1}^{S} \gamma_{r,s}^B = a_r^B \quad \text{and} \quad \sum_{s=1}^{S}\sum_{r=1}^{R} \gamma_{r,s}^B = 1$$

OT algorithm

- Linear Programming problem : Find γ which realizes the infimum of

$$\sum_{r=1}^{R}\sum_{s=1}^{S} \gamma_{r,s}\; c\left(m_r^A, m_s^B\right)$$

under the following constraints which must hold for any r and any s,

$$\gamma_{r,s} \geq 0, \quad \sum_{r=1}^{R} \gamma_{r,s}^A = b_s^A, \qquad \sum_{s=1}^{S} \gamma_{r,s}^A = a_r^A \quad \text{and} \quad \sum_{s=1}^{S}\sum_{r=1}^{R} \gamma_{r,s}^A = 1$$

$$\gamma_{r,s} \geq 0, \quad \sum_{r=1}^{R} \gamma_{r,s}^B = b_s^B, \qquad \sum_{s=1}^{S} \gamma_{r,s}^B = a_r^B \quad \text{and} \quad \sum_{s=1}^{S}\sum_{r=1}^{R} \gamma_{r,s}^B = 1$$

b_s^A and a_r^B are unknow

OT algorithm

- Linear Programming problem : Find γ which realizes the infimum of

$$\sum_{r=1}^{R}\sum_{s=1}^{S} \gamma_{r,s}\; c\left(m_r^A, m_s^B\right)$$

under the following constraints which must hold for any r and any s,

$$\gamma_{r,s} \geq 0, \quad \sum_{r=1}^{R} \gamma_{r,s}^A = b_s^B + \epsilon_s^A, \quad \sum_{s=1}^{S} \gamma_{r,s}^A = a_r^A \quad \text{and} \quad \sum_{s=1}^{S}\sum_{r=1}^{R} \gamma_{r,s}^A = 1$$

$$\gamma_{r,s} \geq 0, \quad \sum_{r=1}^{R} \gamma_{r,s}^B = b_s^B, \quad \sum_{s=1}^{S} \gamma_{r,s}^B = a_r^A + \epsilon_r^B \quad \text{and} \quad \sum_{s=1}^{S}\sum_{r=1}^{R} \gamma_{r,s}^B = 1$$

$$\sum_{s=1}^{S} \epsilon_s^A = 0 \quad \text{and} \quad \sum_{r=1}^{R} \epsilon_r^B = 0$$

OT algorithm

- Linear Programming problem : Find γ which realizes the infimum of

$$\sum_{r=1}^{R}\sum_{s=1}^{S} \gamma_{r,s} \left(c\left(m_r^A, m_s^B\right) + \epsilon_c \right)$$

under the following constraints which must hold for any r and any s,

$$\gamma_{r,s} \geq 0, \quad \sum_{r=1}^{R} \gamma_{r,s}^A = b_s^B + \epsilon_s^A, \quad \sum_{s=1}^{S} \gamma_{r,s}^A = a_r^A \quad \text{and} \quad \sum_{s=1}^{S}\sum_{r=1}^{R} \gamma_{r,s}^A = 1$$

$$\gamma_{r,s} \geq 0, \quad \sum_{r=1}^{R} \gamma_{r,s}^B = b_s^B, \quad \sum_{s=1}^{S} \gamma_{r,s}^B = a_r^A + \epsilon_r^B \quad \text{and} \quad \sum_{s=1}^{S}\sum_{r=1}^{R} \gamma_{r,s}^B = 1$$

$$\sum_{s=1}^{S} \epsilon_s^A = 0 \quad \text{and} \quad \sum_{r=1}^{R} \epsilon_r^B = 0$$

Thank you for your attention