# Seminaire CNAM

**8 Juin 2006**

# Uncertainty in Data Fusion

Tomàs Aluja-Banet

UNIVERSITAT POLITÈCNICA DE CATALUNYA

# Outline

- The 2 paradigms fo data fusion
- Uncertainty by data augmentation
- Uncertainty of T1DM

# General concepts about uncertainty in data fusion

- Data Fusion with single imputation makes up data.
- Fused data has a different nature, *it is uncertain*.

- Goals
  - Basic criteria for handling imputed and observed data
  - To report honest measures of variability

- Levels of uncertainty:
  - Macrodata or aggregated level (about the global statistics)
  - Microdata or individual level

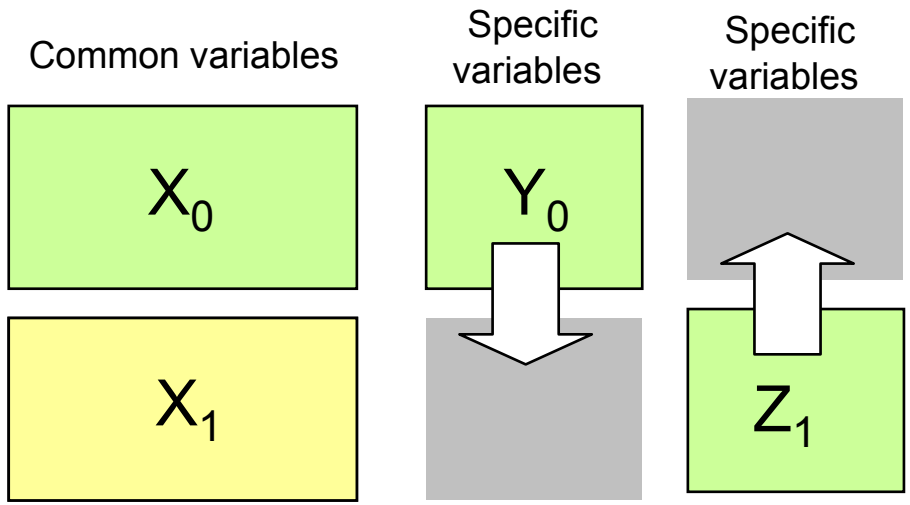
- Sources of uncertainty
  - If we know the data model, the unique source is the random fluctuation
  - If we don't know the data model, we have two sources: the random fluctuation and the uncertainty due to the lack of knowledge of the data model.

# Paradigms of Data Fusion

## *Bilateral fusion*

**Observations in both files are iid. random draws of the same distribution**

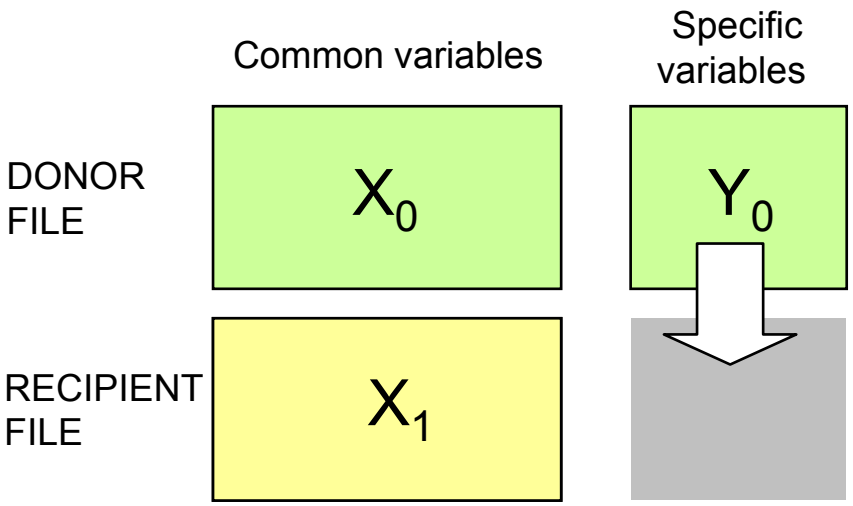
$$f(X_0, Y_0, Z_0) = g(X_1, Y_1, Z_1)$$




## *Unilateral fusion*

**Donor file is (must be) a representaqtive sample of the population. Recipient can be taken with a completely different design**

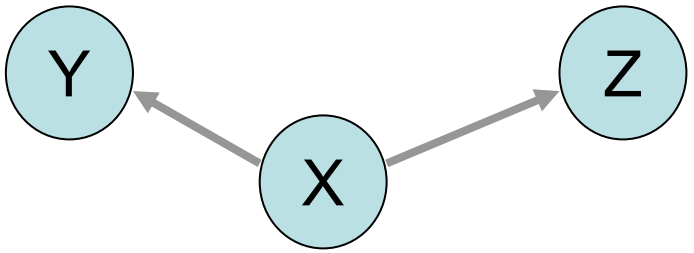
$$f(X_0, Y_0) \neq g(X_1, Y_1)$$



# Data Fusion basic assumptions

## *1.- The Conditional Independence Assumption (CIA):*

*The Y variables are independent of any other set of variables Z given the X.*




$$f(X,Y,Z/\theta) = f_{Y/X}(Y/X,\theta_{Y/X}) f_{Z/X}(Z/X,\theta_{Z/X}) f_X(X/\theta_X)$$

This hypothesis can't be tested.
but the CIA is the midpoint of all admissible *Cor(Y,Z/X)*

But we can approach the CIA with the
***Predictive Relevance Assumption (PRA):***

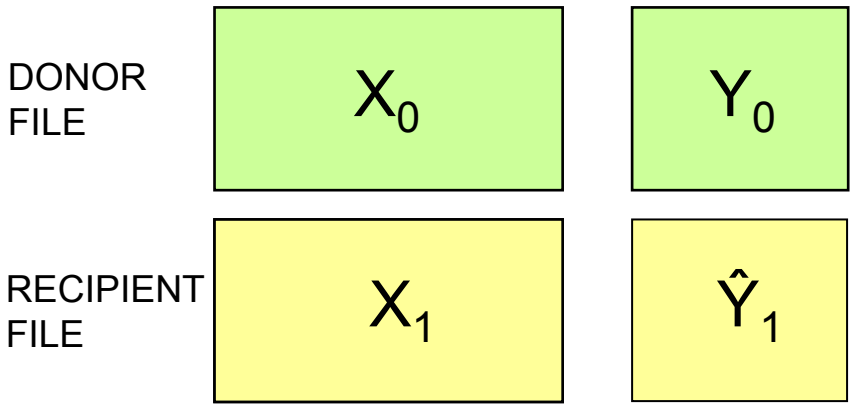
$$Y = i(X) + \varepsilon$$

*where **i(X)** represents the imputation model and ε represents white noise.*

# 2. Data Fusion basic assumption

## *Preservation of the conditional distribution f(Y/X):*

*We assume to have the same conditional distribution in both files $f(Y_1/X_1)=f(Y_0/X_0)$, even though the joint distributions may differ. The objective of data fusion is to transfer the specific variables of the donor file to the recipient file at individual level based on the f(Y/X).*




### *MAR or MCAR assumption:*

*since missing by design. $f(R/XY)=f(R)$ if $f=g$ or $f(R/XY)=f(R/XY_0)$.*
(*R* logical matrix indicating missingness)

# 3. Data Fusion basic assumption

## *Representativeness of the donor file:*

*We don't assume that both files are random samples of the same parent population $f(X_0,Y_0) \neq f(X_1,Y_1)$. We just assume that the donor file is a representative sample of the population $f(\theta/X)= f(\theta/X_0)$.*

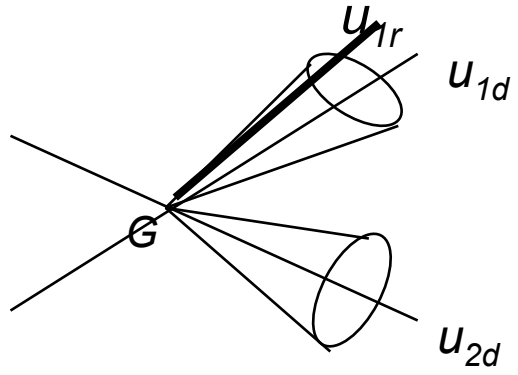
*This assumption need to be tested and assured by the usual way of weighing. Additionally it is interesting to test the equivalence of both data files respect the common information.*

***Bootstrap confidences cones***

Definition of a confidence cones for eigenvectors by bootstrap samples of the donor file with $n_r$ size..

$$V_d = U\Lambda U'$$

Extract *b* bootstrap samples of size $n_r$ of $X_d$
For each sample compute corr($u_{\alpha d}$,$u^b_{\alpha d}$)
Obtain the (1-$\alpha$) confidence cone
per each eigenvector
Plot the actual eigenvectors of $X_r$, $u_{\alpha r}$
Insensitive to inversions, but depending on the separation between eigenvalues.

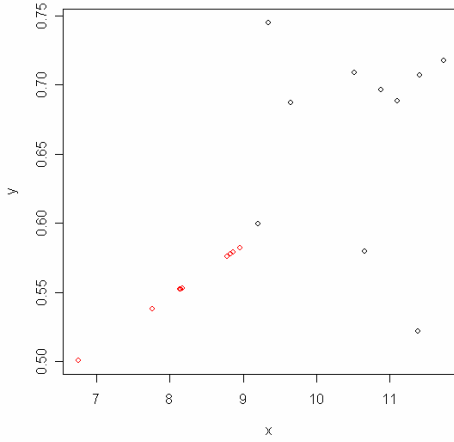


# Technical aspects of imputation

**Imputation must be based on $f(Y/X,\theta_{Y/X})$ (predictive distribution) to avoid bias**

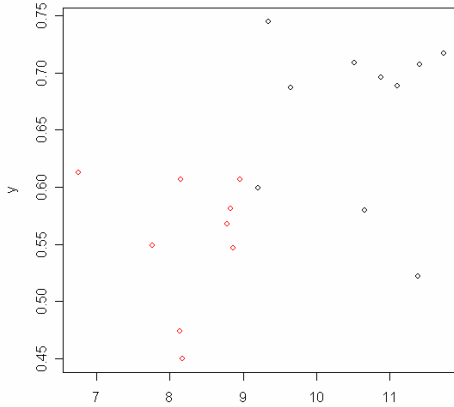
$f(Y/X, \theta_{Y/X})$ **can be stated parametrically**: multivariate normal, multinomial….
**or nonparametrically**: by local estimation (K-NN).

In either case, **it can be deterministic**: $\hat{Y}_1 = E\left[Y / X_1\right]$




preserves the marginal mean
reduces the variability of imputed data
increases the internal homogeneity of imputed data
increases the external homogeneity of imputed data
improves the accuracy of imputed values

**or stochastic**: $\hat{Y}_1 = E\left[Y / X_1\right] + \varepsilon \quad \varepsilon \sim rand\left(f(Y / X)\right)$



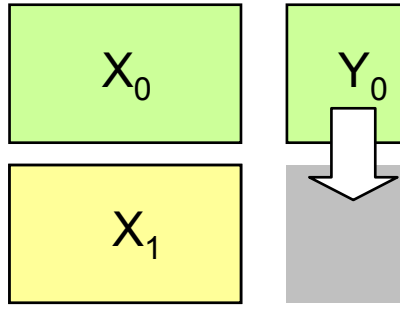

preserves the marginal mean
preserves the variability of imputed data
preserves the internal homogeneity of imputed data
preserves the external homogeneity of imputed data if no bias
deteriorates the accuracy of imputed values

# ML imputation with multivariate normal distribution

## Direct ML imputation

$$f(X,Y/\theta) = f_{Y/X}(Y/X,\theta_{Y/X}) f_X(X/\theta_X)$$

$$L(\theta/X,Y_0) = \prod^{n_0} f_{Y/X}(Y_0/X_0,\theta_{Y/X}) \prod^{n_0} f_X(X_0,\theta_X) \prod^{n_1} f_X(X_1,\theta_X)$$




***Hypothesis of multivariate normality***

$$f(X,Y/\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^{p+q}\,|\Sigma|}} e^{-\frac{1}{2}((X,Y)-\mu)'\Sigma^{-1}((X,Y)-\mu)}$$

$$\theta = (\mu,\Sigma), \quad \mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix}$$

$$f_X(X/\theta_X) \sim N_p(\mu_X,\Sigma_X)$$

$$f_{Y/X}(Y/X,\theta_{Y/X}) \sim N_q(\mu_{Y/X},\Sigma_{Y/X})$$

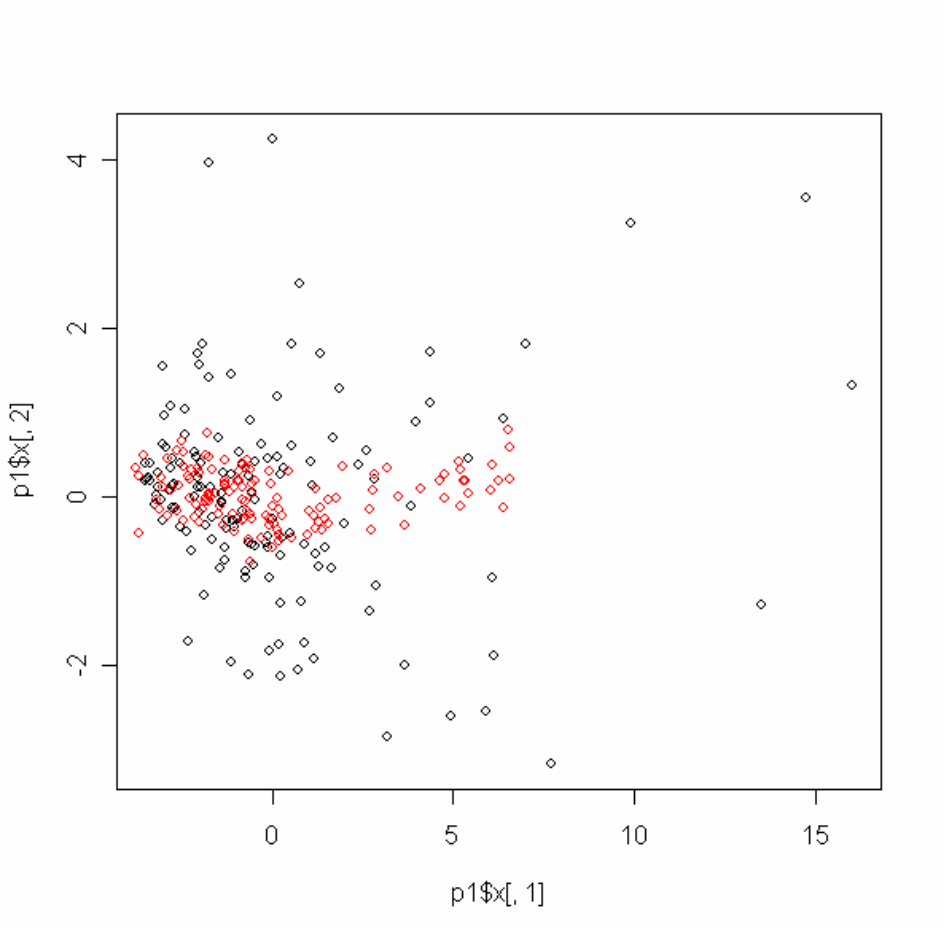
$$\mu_{Y/X} = \mu_Y - \Sigma_{YX}\Sigma_X^{-1}\mu_X$$

$$\Sigma_{Y/X} = \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$$

**Deterministic ML imputation** $E[y/x] = \mu_Y - \Sigma_{YX}\Sigma_X^{-1}(x-\mu_X)$

**Stochastic ML imputation** $y = E[y/x] + \varepsilon \quad \varepsilon \sim N(0,\Sigma_{Y/X})$

# PCA of $Y_1$ with deterministic ML imputation as illustrative (in red)




- true values
- imputed values

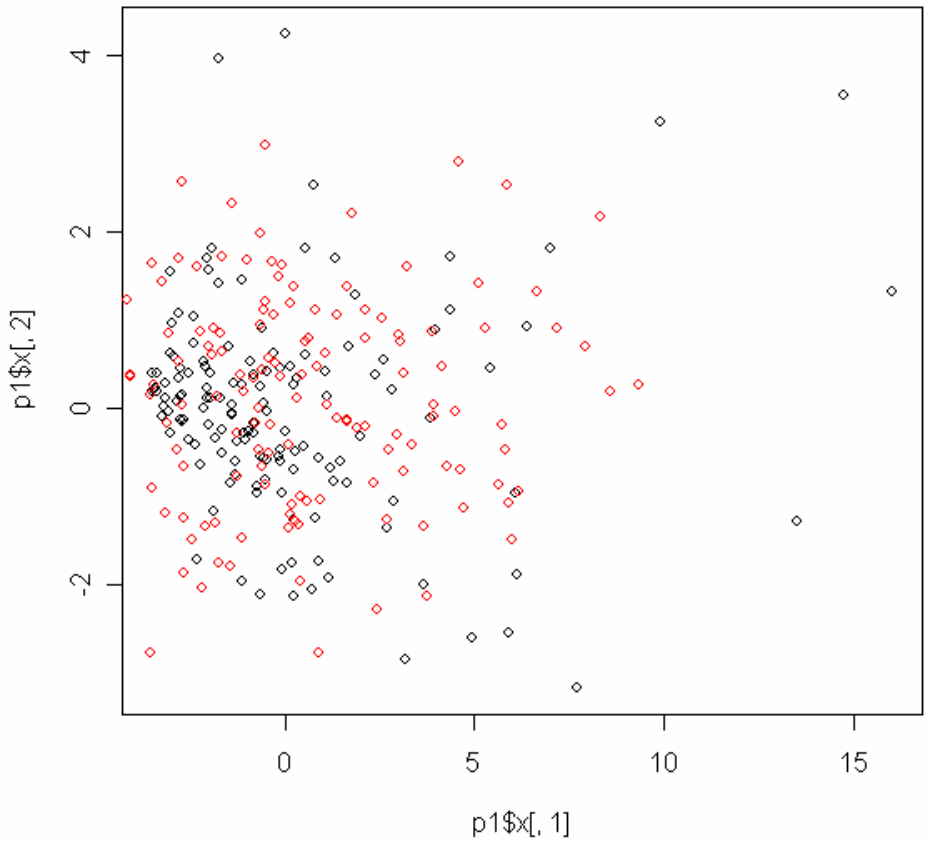
Validation statistics $Y_0 - \hat{Y}_1$

| | |
|---|---|
| aslm | 0.3624202 |
| asls | 2.38777e-07 |
| acdi | 0.435889 |
| acde | 0.1303550 |
| tau | NA |

Validation statistics $Y_1 - \hat{Y}_1$

| | |
|---|---|
| aslm | 0.2809857 |
| asls | 5.259313e-06 |
| acdi | 0.3987483 |
| acde | 0.1334511 |
| tau | 0.7011505 |

# PCA of $Y_1$ with stochastic ML imputation as illustrative (in red)




Validation statistics $Y_0 - \hat{Y}_1$

| | |
|---|---|
| aslm | 0.3167336 |
| asls | 0.3281321 |
| acdi | 0.04878506 |
| acde | 0.0667179 |
| tau | NA |

Validation statistics $Y_1 - \hat{Y}_1$

| | |
|---|---|
| aslm | 0.2321964 |
| asls | 0.2530406 |
| acdi | 0.07773023 |
| acde | 0.07590914 |
| tau | 1.577595 |

# ML imputation with multinomial distribution

## Direct ML imputation

$$f(X,Y/\theta) = f_{Y/X}(Y/X,\theta_{Y/X})\, f_X(X/\theta_X)$$

***Hypothesis of multinomial distribution***

(*X*, *Y* one categorical variable each one)

$$f(X,Y/\theta) = \theta_{ij} \quad \theta_{ij} \geq 0, \sum_{i,j} \theta_{ij} = 1 \qquad \theta_{ij} = \theta_i \times \theta_{j/i}$$

$$f_X(X/\theta_X) \sim M_p\left(\theta_{i\bullet} = \frac{n_{i\bullet}^0 + n_{i\bullet}^1}{n_0 + n_1}\right)$$

$$f_{Y/X}(Y/X,\theta_{Y/X}) \sim M_q\left(\theta_{j/i} = \frac{n_{ij}^0}{n_{i\bullet}^0}\right)$$

**Deterministic ML imputation** $\quad E[y/x] = \arg\max\{j, \theta_{j/i}\}$

**Stochastic ML imputation** $\quad y = rand\{\theta_{j/i}\}$

# EM imputation (Dempster, Laird and Rubin, 1997)

$$f(XY/\theta) = f(X/\theta_X) \times f(Y/X, \theta_{Y/X})$$

***X***: observed data ($X_0, X_1, Y$)
***Y***: missing data ($Y_1$)

$$l(\theta/XY) = l(\theta_X/X) + \log f(Y/X, \theta_{Y/X})$$

**Complete likelihood** **observed likelihood** **Predictive distribution**

$$Q(\theta/\theta^t) = \int l(\theta/XY) f(Y/X, \theta^t) dY$$

Intuitive basis. Given an initial estimate of $\theta^0$. Iterate until convergence

**Step E :**
Calculate the expected value $Q(\theta, \theta^t)$

$$E\left[l(\theta/XY)\right] = Q(\theta/\theta^t)$$

**Step M :**
Maximize $Q(\theta, \theta^t)$

$$\max_\theta \left[Q(\theta/\theta^t)\right]$$

# EM imputation for multivariate distribution

*Assuming* $f(X,Y/\theta) \sim N(\mu,\Sigma)$ $\quad \mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix}$

Initialize $\theta^0$, with the default values computed on observed data

**Step E :**

The complete likelihood is a linear function of the sufficient statistics *S(θ)*.

$E\left[ S(\theta)/\theta^t \right]$ $\quad S(\theta) = \left[ \sum_{i=1}^{n} (x_i, y_i), \sum_{i=1}^{n} (x_i, y_i)(x_i, y_i)' \right]$

$E\left[ y/x \right] = \mu_Y - \Sigma_{YX}\Sigma_X^{-1}(x - \mu_X)$ $\quad \Sigma_{Y/X} = \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$

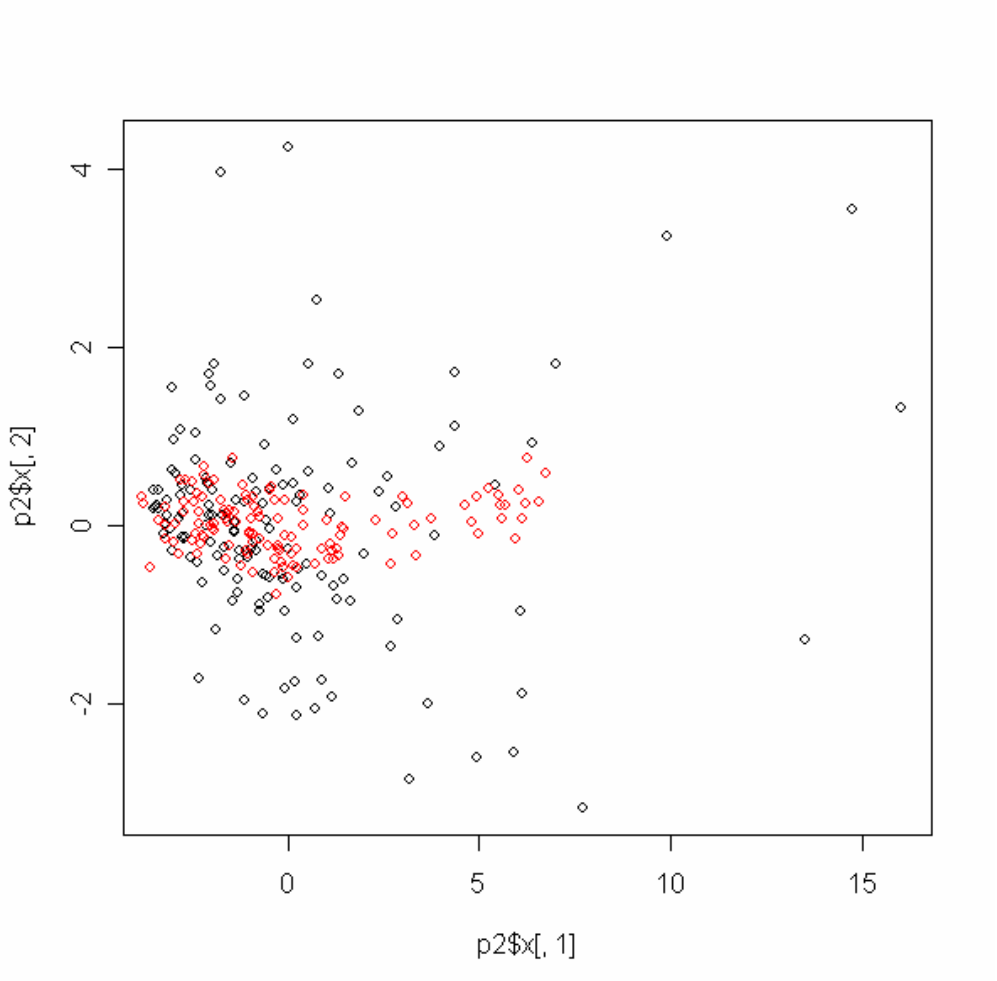
**Step M :**

Computation of the new parameters *θ* given the sufficient statistics

$$\mu^{t+1} = \frac{E\left[ \sum_i (x_i, y_i) \middle| X, \theta^t \right]}{n}$$

$$\Sigma^{t+1} = \frac{E\left[ \sum_i (x_i, y_i)(x_i, y_i)' \middle| X, \theta^t \right]}{n} - \mu^{t+1}\mu^{t+1\prime}$$

# PCA of $Y_1$ with EM deterministic imputation as illustrative (in red)



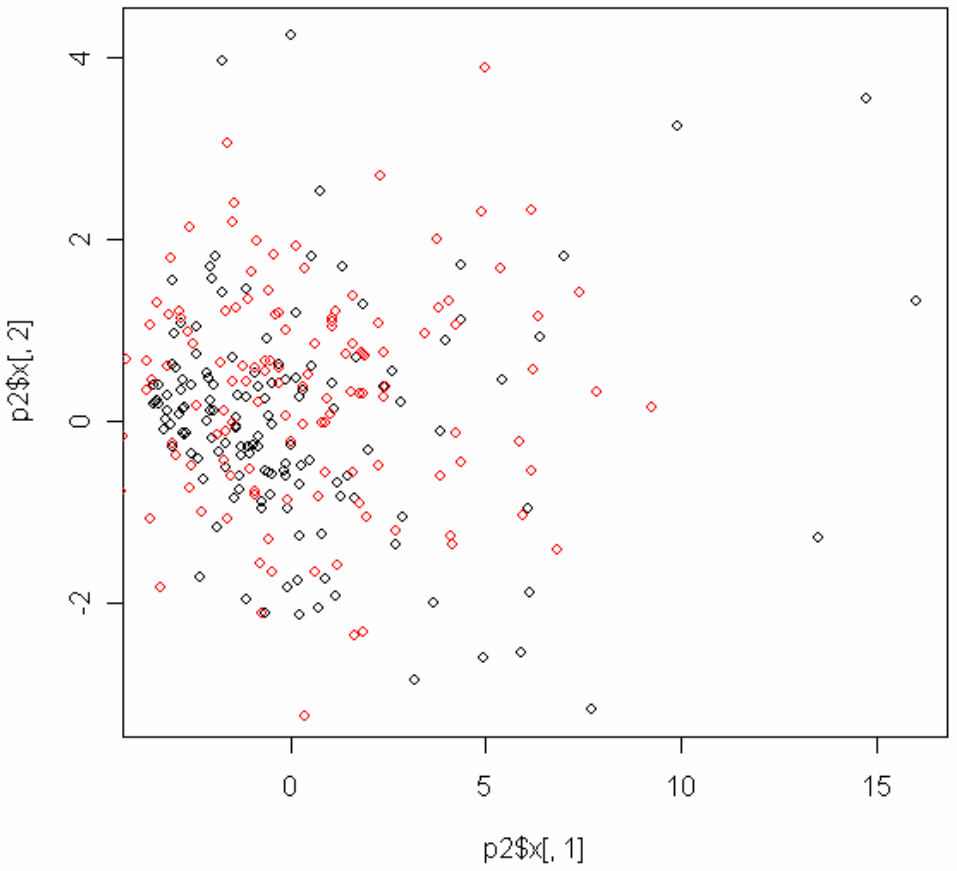

Validation statistics $Y_0 - \hat{Y}_1$

```
aslm  0.3101589
asls  3.772772e-07
acdi  0.4330134
acde  0.1464521
tau   NA
```

Validation statistics $Y_1 - \hat{Y}_1$

```
aslm  0.2772925
asls  8.9327e-06
acdi  0.3958727
acde  0.1380985
tau   0.7027498
```

# PCA of $Y_1$ with EM stochastic imputation as illustrative (in red)




Validation statistics $Y_0 - \hat{Y}_1$

| | |
|---|---|
| aslm | 0.2015052 |
| asls | 0.3555066 |
| acdi | 0.05353303 |
| acde | 0.07082791 |
| tau | NA |

Validation statistics $Y_1 - \hat{Y}_1$

| | |
|---|---|
| aslm | 0.2401951 |
| asls | 0.2318602 |
| acdi | 0.09328637 |
| acde | 0.07132088 |
| tau | 1.568504 |

# Data Augmentation imputation

(Tanner and Wong, 1987)

Extract random draws of the predictive distribution $f(Y/X) = \int f(Y/X,\theta) f(\theta) d\theta$

A class of MCMC algorithm

**I step:** $Y_1^{t+1} \sim f(Y/X,\theta^t)$ **Imputation step**

**P step:** $\theta^{t+1} \sim f(\theta/X,Y_1^{t+1})$ **Posterior step**

$$\left\{Y_1^1, Y_1^2, \ldots, Y_1^t\right\} \xrightarrow[t\to\infty]{} f(Y/X)$$

$$\left\{\theta^1, \theta^2, \ldots, \theta^t\right\} \xrightarrow[t\to\infty]{} f(\theta/X)$$

} stationary distributions

technical issue: assessing the convergence of the series

**DA ≈ EM double stochastic**

$$\mathrm{I}: \quad Y_1^{t+1} \leftarrow E\left[Y/X\right] + rand\left(f(Y/X,\theta^t)\right)$$

$$\mathrm{P}: \quad \theta^{t+1} \leftarrow \max_{\theta}\left(f(\theta/X,Y_1^{t+1})\right) + rand\left(f(\theta/X,Y_1^{t+1})\right)$$

# Data Augmentation for multivariate normal distributions

*Assuming* $f(X,Y/\theta) \sim N(\mu,\Sigma)$ $\theta = (\mu,\Sigma)$

I step: $y^{t+1} = \mu_Y - \Sigma_{YX}\Sigma_X^{-1}(x-\mu_X) + \varepsilon \quad \varepsilon \sim N_q(0,\Sigma_{Y/X})$

P step: $V_Y \sim Wishart(\nu,\Sigma_Y) \quad \mu^{t+1} \sim N\left(\overline{y},\frac{V_Y}{n}\right)$

**Algorithm**

(1) $\tilde{Y}_1 \leftarrow E\left[Y/X\right] + nrand\left(n_1,0,V_{Y/X}\right)$

(2) $\tilde{V}_{Y/X} \leftarrow \tilde{V}_Y - \tilde{V}_{YX}V_X^{-1}\tilde{V}_{XY}$

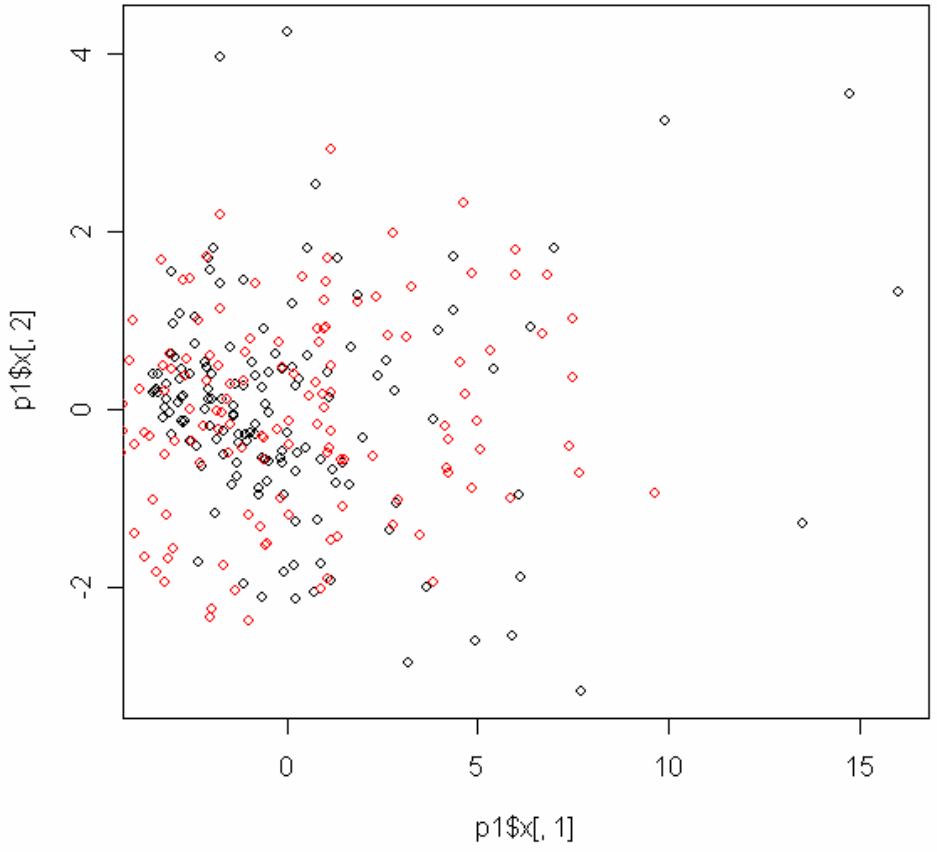
(3) $\tilde{\mu}_Y \leftarrow nrand(1,\mu_Y^t,\tilde{V}_Y/n-1)$

(4) $Y_1^t \leftarrow \tilde{\mu}_Y + \tilde{V}_{YX}\tilde{V}_X^{-1}(X_1-\mu_X) + nrand(n_1,0,\tilde{V}_{Y/X})$

(5) $(\mu_Y^t,V_{Y/X}^t) \leftarrow (\tilde{\mu}_Y,\tilde{V}_{Y/X})$

starting $\theta^0$ values: EM estimates.

# PCA of $Y_1$ with Data Augmentation imputation as illustrative (in red)




Validation statistics $Y_0 - \hat{Y}_1$

| | |
|---|---|
| aslm | 0.2210767 |
| asls | 0.3045797 |
| acdi | 0.05242198 |
| acde | 0.07104441 |
| tau | NA |

Validation statistics $Y_1 - \hat{Y}_1$

| | |
|---|---|
| aslm | 0.1599606 |
| asls | 0.245129 |
| acdi | 0.07719836 |
| acde | 0.07481653 |
| tau | 1.4022857 |

# Results

## 100 simulations of

- Default deter.
- Default stoch.
- T1DM uncons
- T1DM constr.
- T1SM
- TkDM
- TkSM
- EM deter
- EM stoch
- PLS deter
- PLS stoch
- DA

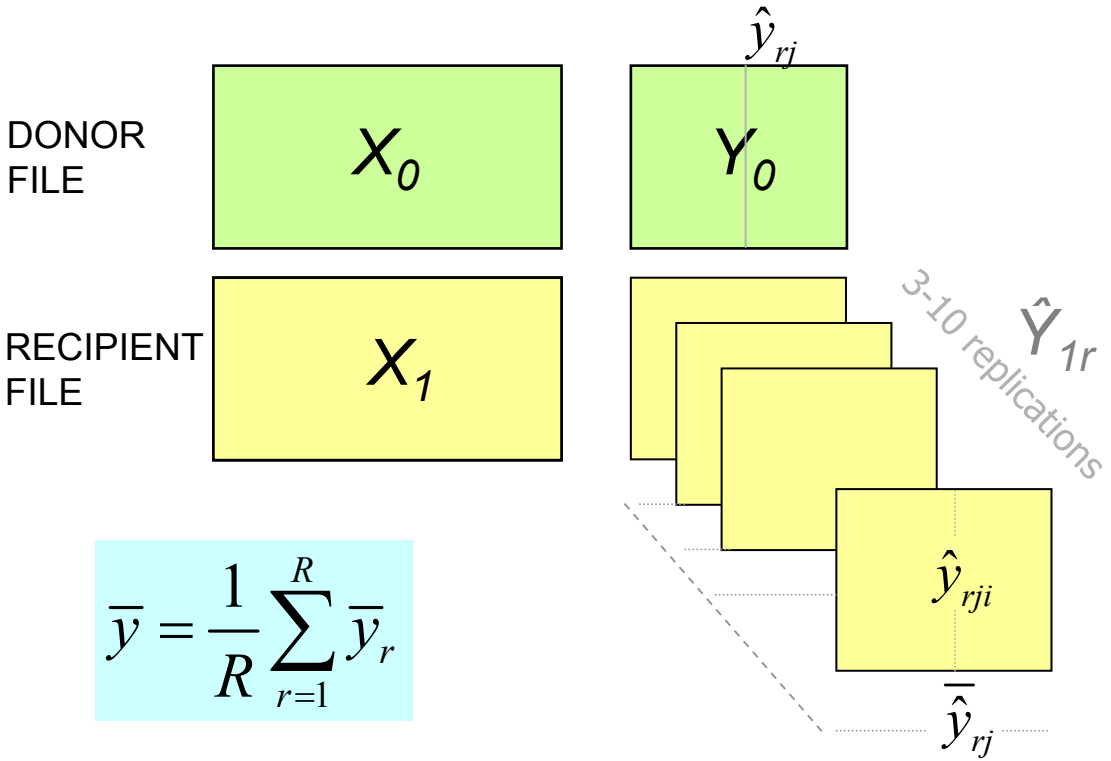
| | | | | | | |
|---|---|---|---|---|---|---|
| $\hat{Y}_1$ vs $Y_0$ | 0.500 | NA | 0.434 | 0.169 | 1.000 | |
| $\hat{Y}1$ vs $Y_1$ | 0.230 | NA | 0.435 | 0.189 | 1.013 | |
| *Default stoch.* | aslm | asls | acdi | acde | tau | bias |
| $\hat{Y}_1$ vs $Y_0$ | 0.455 | 0.280 | 0.064 | 0.101 | 2.009 | |
| $\hat{Y}1$ vs $Y_1$ | 0.352 | 0.131 | 0.121 | 0.111 | 2.174 | |
| *T1DM unsconst* | aslm | asls | acdi | acde | tau | bias |
| $\hat{Y}_1$ vs $Y_0$ | 0.347 | 0.139 | 0.090 | 0.069 | 1.152 | |
| $\hat{Y}1$ vs $Y_1$ | 0.357 | 0.128 | 0.123 | 0.069 | 1.235 | 0.05 |
| *T1DM constrair* | aslm | asls | acdi | acde | tau | bias |
| $\hat{Y}_1$ vs $Y_0$ | 0.392 | 0.176 | 0.061 | 0.059 | 1.156 | |
| $\hat{Y}1$ vs $Y_1$ | 0.377 | 0.139 | 0.112 | 0.065 | 1.272 | 0.06 |
| *T1SM* | aslm | asls | acdi | acde | tau | bias |
| $\hat{Y}_1$ vs $Y_0$ | 0.362 | 0.156 | 0.085 | 0.071 | 1.252 | |
| $\hat{Y}1$ vs $Y_1$ | 0.348 | 0.127 | 0.124 | 0.074 | 1.383 | 0.22 |
| *TkDM* | aslm | asls | acdi | acde | tau | bias |
| $\hat{Y}_1$ vs $Y_0$ | 0.284 | 0.008 | 0.148 | 0.085 | 0.822 | |
| $\hat{Y}1$ vs $Y_1$ | 0.291 | 0.027 | 0.168 | 0.079 | 0.795 | 0.26 |
| *TkSM* | aslm | asls | acdi | acde | tau | bias |
| $\hat{Y}_1$ vs $Y_0$ | 0.374 | 0.158 | 0.080 | 0.070 | 1.268 | |
| $\hat{Y}1$ vs $Y_1$ | 0.358 | 0.134 | 0.124 | 0.074 | 1.385 | 0.23 |
| *EM deter.* | aslm | asls | acdi | acde | tau | bias |
| $\hat{Y}_1$ vs $Y_0$ | 0.293 | 0.000 | 0.424 | 0.136 | NA | |
| $\hat{Y}1$ vs $Y_1$ | 0.251 | 0.007 | 0.415 | 0.144 | 0.736 | -0.00 |
| *EM stoch* | aslm | asls | acdi | acde | tau | bias |
| $\hat{Y}_1$ vs $Y_0$ | 0.235 | 0.285 | 0.053 | 0.069 | NA | |
| $\hat{Y}1$ vs $Y_1$ | 0.216 | 0.143 | 0.113 | 0.087 | 1.518 | -0.07 |
| *PLS deter.* | aslm | asls | acdi | acde | tau | bias |
| $\hat{Y}_1$ vs $Y_0$ | 0.320 | 0.000 | 0.421 | 0.133 | 0.665 | |
| $\hat{Y}1$ vs $Y_1$ | 0.253 | 0.012 | 0.427 | 0.144 | 0.742 | 0.05 |
| *PLS stoch.* | aslm | asls | acdi | acde | tau | bias |
| $\hat{Y}_1$ vs $Y_0$ | 0.245 | 0.291 | 0.052 | 0.068 | 1.353 | |
| $\hat{Y}1$ vs $Y_1$ | 0.216 | 0.140 | 0.112 | 0.084 | 1.601 | 0.06 |
| *DA* | aslm | asls | acdi | acde | tau | bias |
| $\hat{Y}_1$ vs $Y_0$ | 0.222 | 0.277 | 0.056 | 0.071 | | |

# Assessing variability by Data Augmentation: Multiple Imputation

- Once we have performed the data fusion operation, the completed data file $(X_0,X_1;Y_0,\hat{Y}_1)$ or $(X_1;\hat{Y}_1)$ would serve to infer macrodata $(\mu_Y, Cor(X,Y), \ldots)$ or microdata. Then, how can we obtain reliable values or representations incorporating the uncertainty of imputed data.
- By independent draws from the predictive distribution $f(Y/X)$.
- Several stochastic imputations gives us idea of the variability of the imputation method.
- Since we are only interested in macrodata, a few draws (3-10) are enough.

# Multiple imputation

(Rubin, 1987)




$$\hat{y}_{rj} = \begin{pmatrix} y_1 \\ \vdots \\ \hat{y}_{rn} \end{pmatrix}$$

$$\hat{y}_r = \begin{pmatrix} \hat{y}_{r1} & \cdots & \hat{y}_{rq} \end{pmatrix}$$

$$\bar{y}_r = \begin{pmatrix} \bar{\hat{y}}_{r1} & \cdots & \bar{\hat{y}}_{rq} \end{pmatrix}$$

$$\bar{y} = \frac{1}{R}\sum_{r=1}^{R} \bar{y}_r$$

$$V[Y] \approx E[\mathrm{var}(\hat{y}_r)] + V[\bar{y}]$$

$$E[\mathrm{var}(\hat{y}_r)] = \frac{1}{R}\sum_{r=1}^{R} \mathrm{var}(\hat{y}_r)$$

$$V[\bar{y}] = \frac{1}{R-1}\sum_{r=1}^{R} (\bar{y}_r - \bar{y})(\bar{y}_r - \bar{y})' = B$$

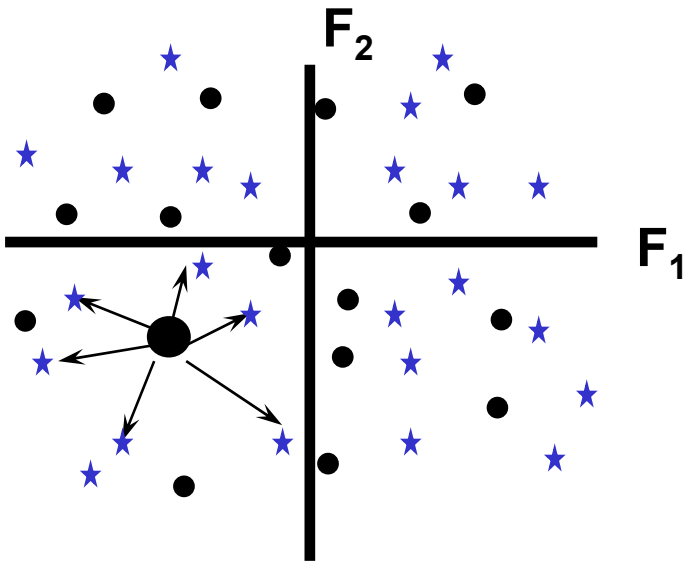
$$\hat{V}[Y] = \overline{\mathrm{var}(\hat{y}_r)} + (1 + \frac{1}{R})B$$

$$\frac{y - \bar{y}}{\sqrt{\widehat{\mathrm{var}}(y)}} \sim t_\nu$$

$$\nu = (R-1)\left(1 + \frac{\overline{\mathrm{var}(\hat{y}_r)}}{(1 + 1/R)B}\right)^2$$

# Uncertainty in T1DM

- We assume that for every recipient $x_{1i}$ the closest neighbor may fluctuate randomly according the kernel K[$x_{1i}$ ,$x_{0j}$].




We can simulate different T1DM imputations by means of T1SM fusion.

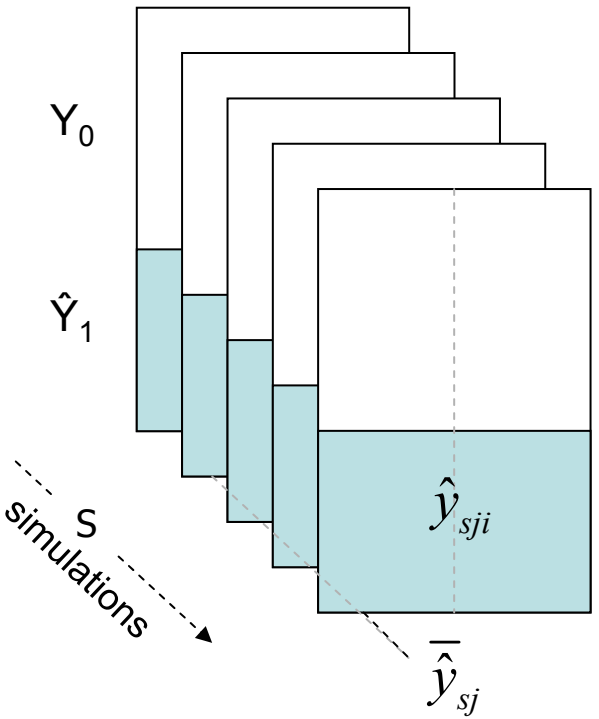
Hence we will perform *S* T1SM imputations to assess the uncertainty of T1DM.

But the attained variability by the *S* T1SM imputations is conveyed by the TkDM imputation.

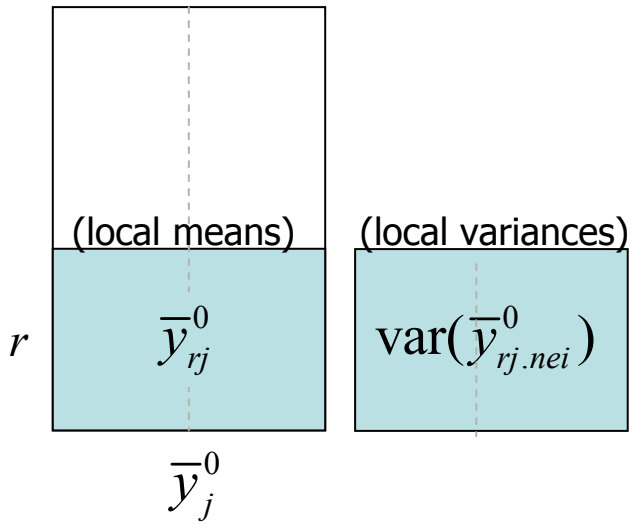
# Measuring the uncertainty of T1DM

- We perform *S* T1SM imputations, we compute the variance of every variable $y_j$ and we compare it with the corresponding TkDM variance.

**S stochastic imputations T1SM**




**TkDM imputation**




$$\mathrm{var}(\bar{\hat{y}}_j) + \frac{1}{S}\sum_{s=1}^{S}\mathrm{var}(\hat{y}_{sj}) \simeq \mathrm{var}(\bar{y}^0_j) + \frac{1}{n}\sum_{l=1}^{nmiss}\mathrm{var}(y^0_{rj.nei}$$

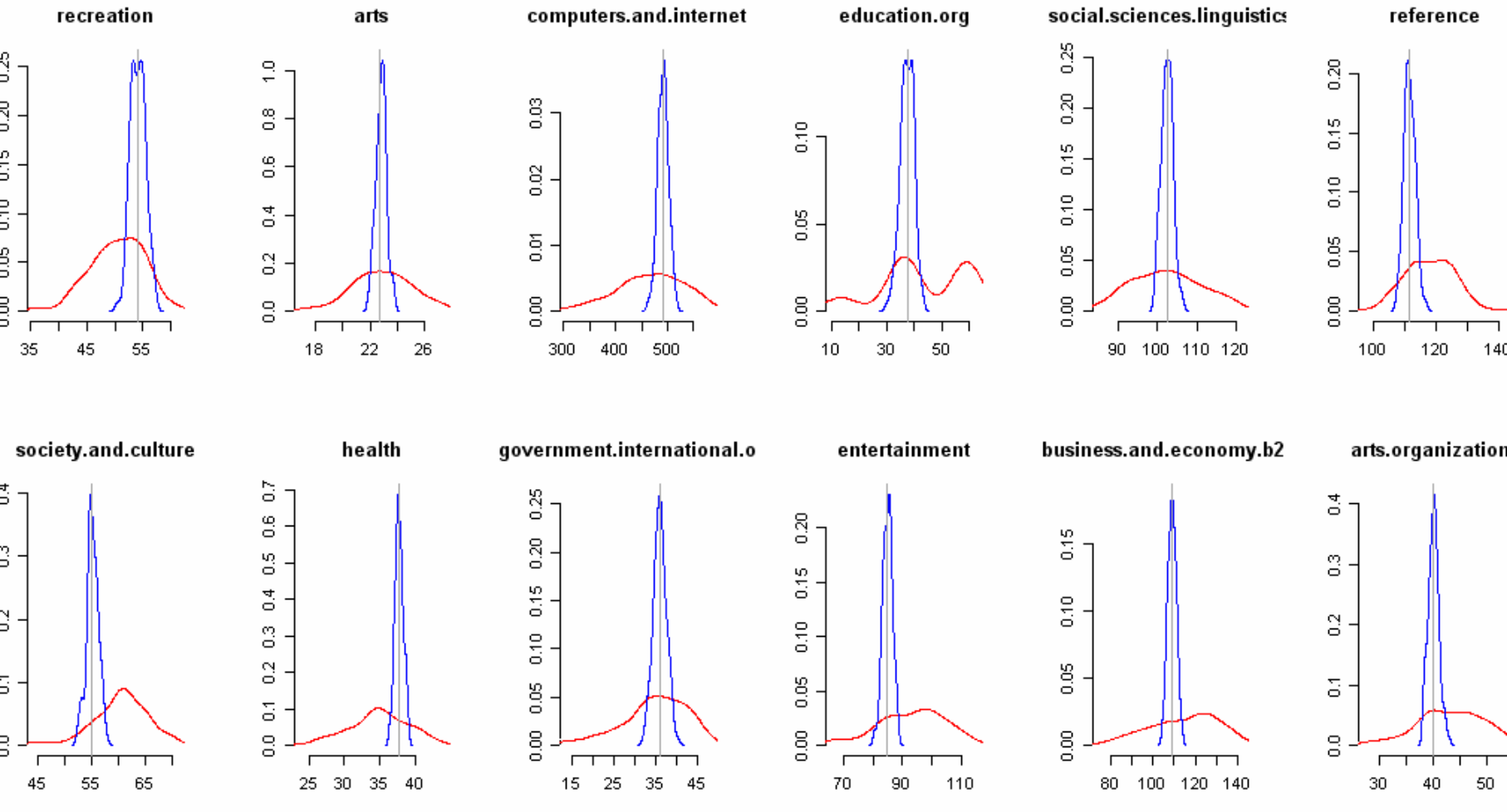
Thus, it is equivalent to perform *S* stochastic simulations or to perform TkDM symbolic imputation

# Comparison of T1SM and MI with TkDM

| *Results of 100 simulations* | T1SM | *TkDM* | MI | | T1SM | *TkDM* | MI | |
|---|---|---|---|---|---|---|---|---|
| | | *mean* | mean | sdev | var | *var* | var | sdev |
| *recreation* | 5.29 | *5.29* | 5.16 | 0.21 | 53.95 | *54.06* | 50.46 | 4.86 |
| *arts* | 2.33 | *2.33* | 2.37 | 0.15 | 22.74 | *22.75* | 22.89 | 2.21 |
| *computers_and_internet* | 25.27 | *25.29* | 25.07 | 0.60 | 489.46 | *490.04* | 466.77 | 63.84 |
| *education_organizations* | 1.93 | *1.94* | 2.05 | 0.24 | 37.65 | *37.65* | 42.90 | 15.38 |
| *social_science_linguistics_and_hun* | 10.37 | *10.38* | 9.91 | 0.30 | 102.29 | *102.45* | 102.47 | 8.83 |
| *reference* | 13.76 | *13.77* | 13.75 | 0.23 | 111.34 | *111.61* | 118.18 | 8.05 |
| *society_and_culture* | 6.65 | *6.65* | 6.98 | 0.22 | 55.12 | *55.05* | 60.37 | 5.13 |
| *health* | 4.26 | *4.26* | 4.22 | 0.21 | 37.69 | *37.76* | 34.92 | 4.22 |
| *government_international_organizati* | 2.38 | *2.39* | 2.40 | 0.19 | 36.33 | *36.37* | 34.94 | 7.76 |
| *entertainment* | 7.03 | *7.02* | 7.36 | 0.28 | 85.05 | *84.89* | 93.21 | 11.28 |
| *business_and_economy_business_* | 12.46 | *12.47* | 12.29 | 0.28 | 108.68 | *108.89* | 114.45 | 15.95 |
| *arts_organizations* | 3.75 | *3.75* | 3.91 | 0.21 | 40.04 | *40.03* | 43.11 | 6.10 |
| *government* | 14.20 | *14.21* | 13.80 | 0.28 | 133.67 | *133.76* | 118.64 | 10.29 |
| *entertainment_music* | 2.50 | *2.51* | 2.56 | 0.17 | 21.34 | *21.35* | 21.46 | 2.29 |
| *news_and_media* | 14.87 | *14.88* | 14.69 | 0.37 | 184.78 | *185.06* | 191.31 | 20.11 |
| *business_and_economy* | 9.85 | *9.85* | 9.85 | 0.30 | 104.74 | *104.86* | 105.24 | 14.09 |
| *education* | 4.48 | *4.48* | 4.52 | 0.19 | 45.95 | *46.00* | 49.67 | 5.48 |
| *recreation_outdoors* | 3.73 | *3.73* | 3.48 | 0.18 | 40.65 | *40.63* | 35.51 | 7.24 |
| *business_and_economy_organizatio* | 3.61 | *3.61* | 3.72 | 0.21 | 52.15 | *52.27* | 47.17 | 7.29 |
| *social_science* | 5.33 | *5.33* | 5.73 | 0.21 | 36.07 | *36.12* | 43.16 | 3.79 |
| *science* | 7.60 | *7.60* | 7.46 | 0.24 | 69.69 | *69.69* | 64.70 | 4.97 |
| *recreation_sports* | 7.58 | *7.58* | 7.39 | 0.26 | 61.77 | *61.87* | 59.78 | 8.11 |

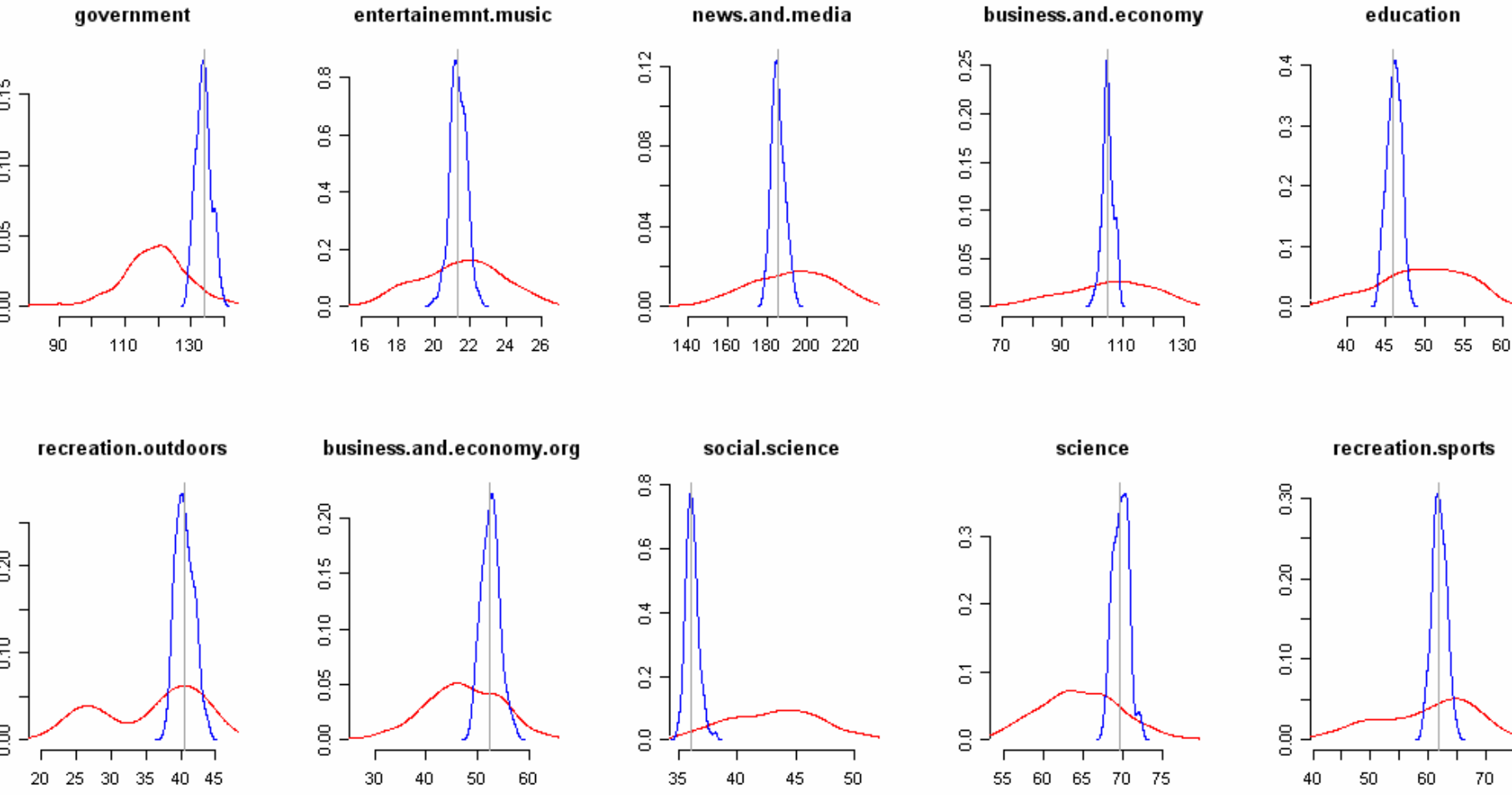
T1SM: results of 100 simulations of 10 runs of T1SM
MI: results of 100 simulations of 10 DA runs

# Estimation of the variance in 100 sim. with T1SM (k=10), TkDM and MI (10 replications)




Densities of the variances of T1SM, of MI and the TkDM variance

# Estimation of variance in 100 simulations T1SM, k=10 with the variance from TkDM (in gray)




Densities of the variances of T1SM, of MI and the TkDM variance

# TkDM fuzzy imputation

- TkDM provides a fuzzy imputation (with local mean and local variance) equivalent to several stochastic imputations
  - For continuous variables

$$\mu_{V_i} = \sum_{k \in V_i} \omega_{ki} y_{k0} \quad \sigma^2_{V_i} = \sum_{k \in V_i} \omega_{ki} (y_{k0} - \mu_{V_i})^2$$

  - For categorical variables

$$(\alpha_j, j = 1 \ldots q) \quad with \quad p_i = (p_{ij}, j = 1 \ldots q)$$

$$where \quad p_{ij} = \sum_{l \in V_i} p_{lj}$$

- Moreover, it handles individual microdata uncertainty as well as macrodata uncertainty.

# Open research lines

- Justify the equivalence of TkDM with MI (continuous and categorical variables).
- Report honest measures of variances, correlations, …
- To set a suitable fuzzy TkDM data set able to be used with complete data analysis techniques.
- To assess the equivalence (or not) between the PRA and the CIA assumption.
- To experiment with more sophisticated kernels
- …

# References


- Aluja T. Thió S. (2001). Evaluation de campagnes publicitaires par fusion de fichiers. *Traitement des fichiers d'enquêtes*, ed. Michel Lejeune. Presses Universitaires de Grenoble.
- Co V. (1997) Méthodes statistiques et informatiques pour le traitement des données manquantes. Thesis, *Conservatoire National des arts and Métiers*. Paris.
- Comyn M. (1999) Modélisation et validation des raprochements et fusions de fichiers d'enquêtes. Thesis, *Ecole Nationale Supérieure des Télécommunications*. Paris.
- Dempster A.P., Laird N.M., Rubin D.B. (1977) Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B* 39, 1-38.
- D'Orazio M., Di Zio M, Scanu M. (2006) Statistical Matching. Theory and Practice. Wiley.
- Fischer N. (2004) Fusion Statistique de Fichiers de Données. Ph.D. *CNAM*, Paris.
- Juárez-Alonso C.A. (2005). Fusión de Datos. Imputación y Validación. Ph.D. *UPC*, Barcelona.
- Lebart L., Lejeune M. (1995). Assessment of data fusions and injections. *Encuentro Internacional AIMC sobre Investigación de Medios*. Madrid, pp. 208-225.
- Little R.J.A., Rubin D.B. (1987) *Statistical Analysis with Missing Data*. John Wiley & Sons. New York.
- Martinez-Abarca M-J., Aluja T. (1999) Fusión de datos de audiencia: metodologia y su aplicación en el mercado publicitario. *II Seminario sobre nuevas tecnologias*. *AEDEMO*, pp. 129-146.
- Putten P.V.D. , Kok J.N., Gupta A. (2002) Data Fusion Through Statistical Matching. Paper 185. *ebusiness@MIT*.
- Rässler S. (2004) Data Fusion: Identification problems, Validity and Multiple Imputation. *Austrian Journal of Statistics*. Vol 33, n 1&2, pp 153-171.
- Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*, John Wiley. New York.
- Santini, G. (1988) Validation of data fusion techniques: what can statistical theory do for us?. *Readership Research: Theory and Practice. Proceedings of the Fourth International Symposium*. Pp. 384-393. H. Henry (ed.). Barcelona
- Schafer J.L. (1997) *Analysis of Incomplete Multivariate Data*. Chapman&Hall.
- Tanner M.A., Wong W.H. (1987) The calculation of posterior distributions by data augmentation. JASA 82, 528-550.
- Tutorial GRAFT SPAD, www.esisproject.com