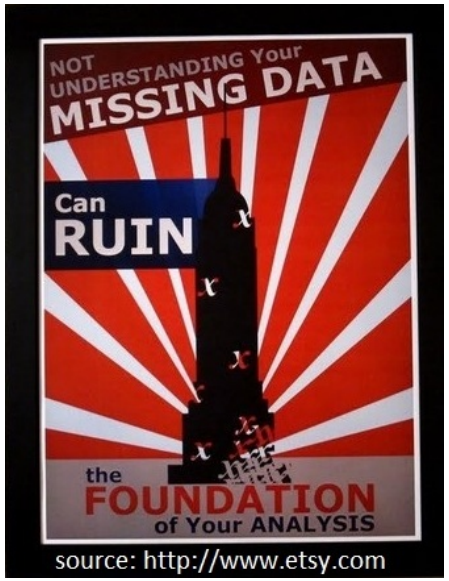
Distributed multilevel matrix completion for medical databases

Julie Josse
Ecole Polytechnique, INRIA
Séminaire Parisien de Statistique, IHP



January 15, 2018

Overview

Outline

Collaborators

Imputation of mixed data with multilevel SVD
Distributed computation

Genevieve Robin, François Husson, Balasubramanian Narasimhan

Polytechnique, Agrocampus, Stanford (stat/biomedical)



Public Assistance - Paris Hospitals

Traumabase: 15000 patients/ 250 variables/

	Center	Accident	Age	Sex	Weight	Height	BMI	BP	SBP
1	Beaujon	Fall	54	m	85	NR	NR	180	110
2	Lille	Other	33	m	80	1.8	24.69	130	62
3	Pitie Salpetriere	Gun	26	m	NR	NR	NR	131	62
4	Beaujon	AVP moto	63	m	80	1.8	24.69	145	89
6	Pitie Salpetriere	AVP bicycle	33	m	75	NR	NR	104	86
7	Pitie Salpetriere	AVP pedestrian	30	w	NR	NR	NR	107	66
9	HEGP	White weapon	16	m	98	1.92	26.58	118	54
10	Toulon	White weapon	20	m	NR	NR	NR	124	73
11	Bicetre	Fall	61	m	84	1.7	29.07	144	105

.....

	SpO2	Temperature	Lactates	Hb	Glasgow	Transfusion	.....
1	97	35.6	<NA>	12.7	12	yes	
2	100	36.5	4.8	11.1	15	no	
3	100	36	3.9	11.4	3	no	
4	100	36.7	1.66	13	15	yes	
6	100	36	NM	14.4	15	no	
7	100	36.6	NM	14.3	15	yes	
9	100	37.5	13	15.9	15	yes	
10	100	36.9	NM	13.7	15	no	
11	100	36.6	1.2	14.2	14	no	

.....

$\Rightarrow$ Predict the Glasgow score, whether to start a blood transfusion, to administer fresh frozen plasma, etc...

$\Rightarrow$ (Logistic) regressions with categorical/continuous values

Public Assistance - Paris Hospitals

Traumabase: 15000 patients/ 250 variables/ 6 hospitals

	Center	Accident	Age	Sex	Weight	Height	BMI	BP	SBP
1	Beaujon	Fall	54	m	85	NR	NR	180	110
2	Lille	Other	33	m	80	1.8	24.69	130	62
3	Pitie Salpetriere	Gun	26	m	NR	NR	NR	131	62
4	Beaujon	AVP moto	63	m	80	1.8	24.69	145	89
6	Pitie Salpetriere	AVP bicycle	33	m	75	NR	NR	104	86
7	Pitie Salpetriere	AVP pedestrian	30	w	NR	NR	NR	107	66
9	HEGP	White weapon	16	m	98	1.92	26.58	118	54
10	Toulon	White weapon	20	m	NR	NR	NR	124	73
11	Bicetre	Fall	61	m	84	1.7	29.07	144	105

.....

	SpO2	Temperature	Lactates	Hb	Glasgow	Transfusion	.....
1	97	35.6	<NA>	12.7	12	yes	
2	100	36.5	4.8	11.1	15	no	
3	100	36	3.9	11.4	3	no	
4	100	36.7	1.66	13	15	yes	
6	100	36	NM	14.4	15	no	
7	100	36.6	NM	14.3	15	yes	
9	100	37.5	13	15.9	15	yes	
10	100	36.9	NM	13.7	15	no	
11	100	36.6	1.2	14.2	14	no	

.....

$\Rightarrow$ Predict the Glasgow score, whether to start a blood transfusion, to administer fresh frozen plasma, etc...

$\Rightarrow$ (Logistic) regressions with missing categorical/continuous values

Public Assistance - Paris Hospitals

⇒ 6 hospitals: multilevel structure (patients within hospital)

Hopital effect: lack of standardization

⇒ Missing values

```
              Center       Accident Age Sex Weight Height   BMI BP SBP
1            Beaujon        Fall    54    m    85     NR    NR 180 110
2              Lille        Other   33    m    80    1.8 24.69 130  62
3  Pitie Salpetriere        Gun     26    m    NR     NR    NR 131  62
4            Beaujon      AVP moto  63    m    80    1.8 24.69 145  89
6  Pitie Salpetriere   AVP bicycle  33    m    75     NR    NR 104  86
7  Pitie Salpetriere AVP pedestrian 30    w    NR     NR    NR 107  66
9               HEGP   White weapon 16    m    98   1.92 26.58 118  54
...................

   SpO2 Temperature Lactates   Hb  Glasgow Transfusion ...........
1    97        35.6     <NA> 12.7       12          yes
2   100        36.5      4.8 11.1       15          no
3   100          36      3.9 11.4        3          no
4   100        36.7     1.66   13       15          yes
6   100          36       NM 14.4       15          no
7   100        36.6       NM 14.3       15          yes
9   100        37.5       13 15.9       15          yes
```

1) Handling missing values in multilevel data

Public Assistance - Paris Hospitals

⇒ 6 hospitals: multilevel structure (patients within hospital)

Hopital effect: lack of standardization

2) Data not aggregated but which stay on each site

⇒ Missing values

```
              Center       Accident Age Sex Weight Height   BMI BP SBP
1            Beaujon        Fall    54    m    85     NR    NR 180 110
2              Lille        Other   33    m    80    1.8 24.69 130  62
3  Pitie Salpetriere        Gun     26    m    NR     NR    NR 131  62
4            Beaujon      AVP moto  63    m    80    1.8 24.69 145  89
6  Pitie Salpetriere   AVP bicycle  33    m    75     NR    NR 104  86
7  Pitie Salpetriere AVP pedestrian 30    w    NR     NR    NR 107  66
9               HEGP   White weapon 16    m    98   1.92 26.58 118  54
...................

   SpO2 Temperature Lactates   Hb  Glasgow Transfusion ...........
1    97        35.6     <NA> 12.7       12          yes
2   100        36.5      4.8 11.1       15          no
3   100          36      3.9 11.4        3          no
4   100        36.7     1.66   13       15          yes
6   100          36       NM 14.4       15          no
7   100        36.6       NM 14.3       15          yes
9   100        37.5       13 15.9       15          yes
```

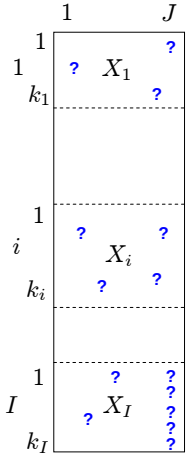
1) Handling missing values in multilevel data

Missing values

are everywhere: unanswered questions in a survey, lost data, damaged plants, machines that fail...

The best thing to do with missing values is not to have any" Gertrude Mary Cox.

$\Rightarrow$ Still an issue with "big data"
Data integration: data from different sources



Multilevel structure: sporadically - systematic missing values (one variable missing in one hospital)

Solutions to handle missing values

Litterature: Schaefer (2002); Little & Rubin (2002); Gelman & Meng (2004); Kim & Shao (2013); Carpenter & Kenward (2013); van Buuren (2015)

$\Rightarrow$ Modify the estimation process to deal with missing values.
Maximum likelihood: EM algorithm to obtain point estimates + Supplemented EM (Meng & Rubin, 1991) ; Louis for their variability
One specific algorithm for each statistical method...

$\Rightarrow$ Imputation (multiple) to get a completed data set on which you can perform any statistical method (Rubin, 1976)
Famous imputation based on SVD (PCA) - quantitative

Solutions to handle missing values

Litterature: Schaefer (2002); Little & Rubin (2002); Gelman & Meng (2004); Kim & Shao (2013); Carpenter & Kenward (2013); van Buuren (2015)

$\Rightarrow$ Modify the estimation process to deal with missing values.
Maximum likelihood: EM algorithm to obtain point estimates + Supplemented EM (Meng & Rubin, 1991) ; Louis for their variability
One specific algorithm for each statistical method...

$\Rightarrow$ Imputation (multiple) to get a completed data set on which you can perform any statistical method (Rubin, 1976)
Famous imputation based on SVD (PCA) - quantitative

Extension to imputation with multilevel SVD

Outline

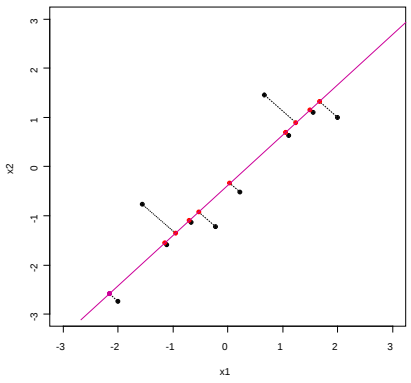
PCA reconstruction

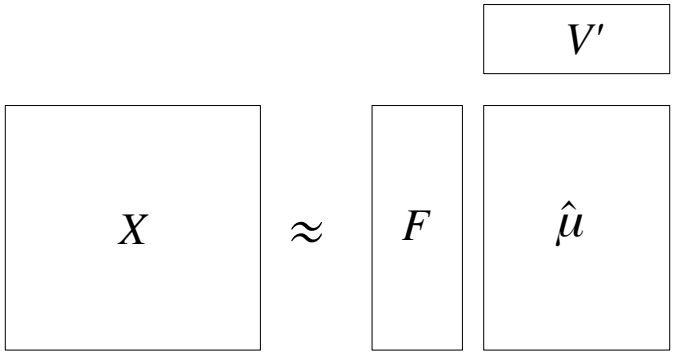
X

-2.00	-2.74
-1.56	-0.77
-1.11	-1.59
-0.67	-1.13
-0.22	-1.22
0.22	-0.52
0.67	1.46
1.11	0.63
1.56	1.10
2.00	1.00

$\hat{\mu}$

-2.16	-2.58
-0.96	-1.35
-1.15	-1.55
-0.70	-1.09
-0.53	-0.92
0.04	-0.34
1.24	0.89
1.05	0.69
1.50	1.15
1.67	1.33





$\Rightarrow$ Minimizes distance between observations and their projection

$\Rightarrow$ Approx $X_{n\times p}$ with a low rank matrix $k < p$ $\|A\|_2^2 = \mathrm{tr}(AA^\top)$:

$$\mathrm{argmin}_\mu \left\{ \|X - \mu\|_2^2 : \mathrm{rank}\,(\mu) \leq k \right\}$$

SVD X: $\hat{\mu}^{\mathrm{PCA}} = U_{n\times k} D_{k\times k} V'_{p\times k}$ $\quad F = UD$ PC - scores

$= F_{n\times k} V'_{p\times k}$ $\quad V$ principal axes - loadings

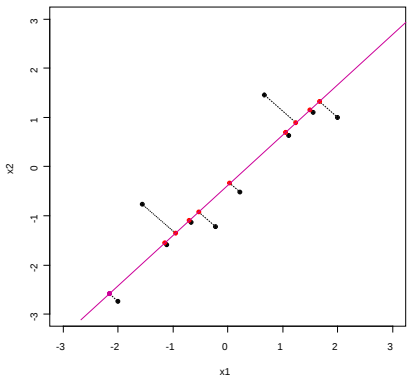
PCA reconstruction

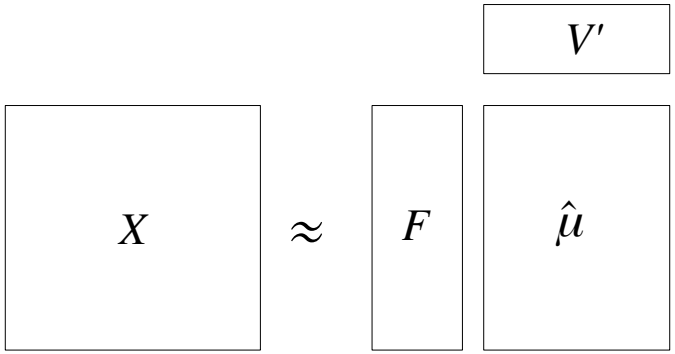
X

-2.00	-2.74
NA	-0.77
-1.11	-1.59
-0.67	-1.13
-0.22	NA
0.22	-0.52
0.67	1.46
NA	0.63
1.56	1.10
2.00	1.00

$\hat{\mu}$

-2.16	-2.58
-0.96	-1.35
-1.15	-1.55
-0.70	-1.09
-0.53	-0.92
0.04	-0.34
1.24	0.89
1.05	0.69
1.50	1.15
1.67	1.33





$\Rightarrow$ Minimizes distance between observations and their projection

$\Rightarrow$ Approx $X_{n\times p}$ with a low rank matrix $k < p$ $\|A\|_2^2 = \mathrm{tr}(AA^\top)$:

$$\mathrm{argmin}_\mu \left\{ \|X - \mu\|_2^2 : \mathrm{rank}\,(\mu) \leq k \right\}$$

SVD X: $\hat{\mu}^{\mathrm{PCA}} = U_{n\times k} D_{k\times k} V'_{p\times k}$ $\quad F = UD$ PC - scores

$= F_{n\times k} V'_{p\times k}$ $\quad V$ principal axes - loadings

Missing values in PCA

$\Rightarrow$ PCA: least squares

$$\text{argmin}_{\mu} \left\{ \|X_{n\times p} - \mu_{n\times p}\|_2^2 : \text{rank}\,(\mu) \leq k \right\}$$

$\Rightarrow$ PCA with missing values: weighted least squares

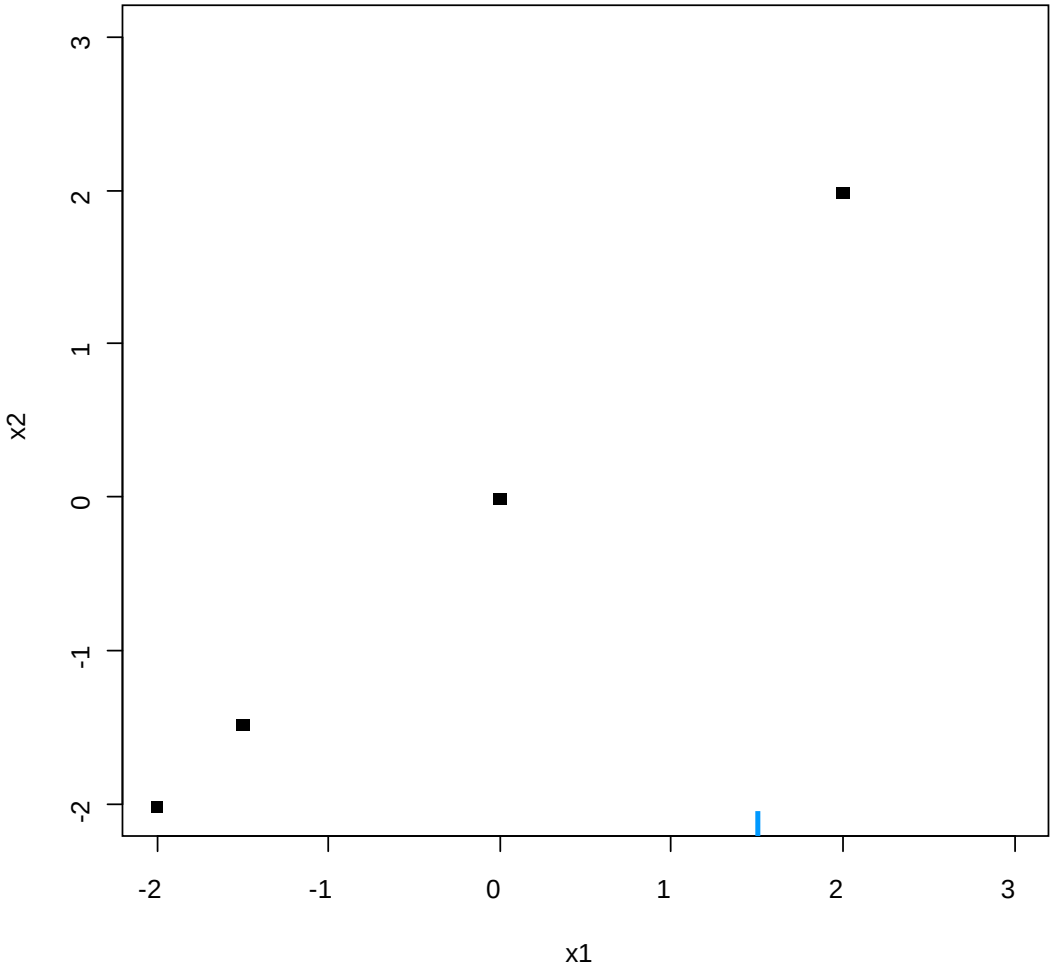
$$\text{argmin}_{\mu} \left\{ \|W_{n\times p} \odot (X - \mu)\|_2^2 : \text{rank}\,(\mu) \leq k \right\}$$

with $W_{ij} = 0$ if X_{ij} is missing, $W_{ij} = 1$ otherwise; $\odot$ elementwise multiplication

Many algorithms:
Gabriel & Zamir, 1979: weighted alternating least squares (without explicit imputation)
Kiers, 1997: iterative PCA (with imputation)

Iterative PCA

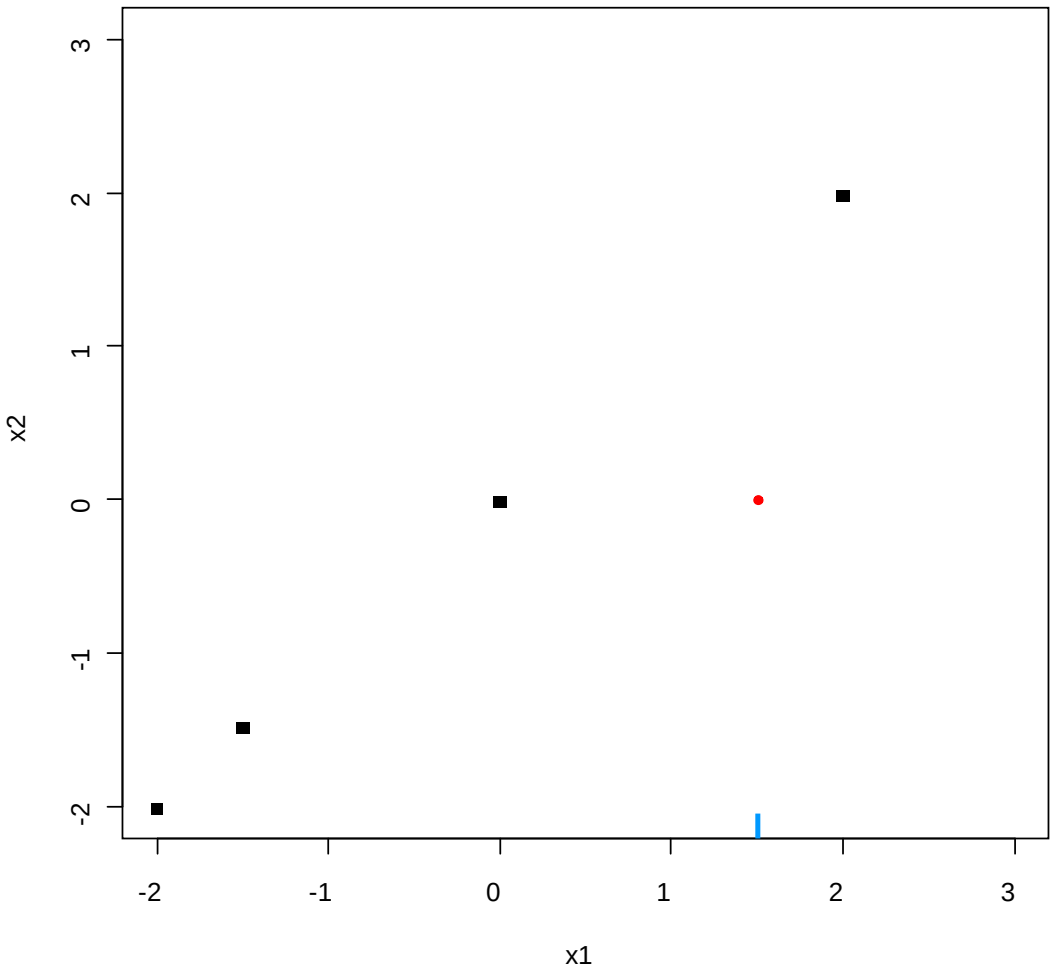
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98



Iterative PCA

```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5    NA
 2.0  1.98

  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.00
 2.0  1.98
```



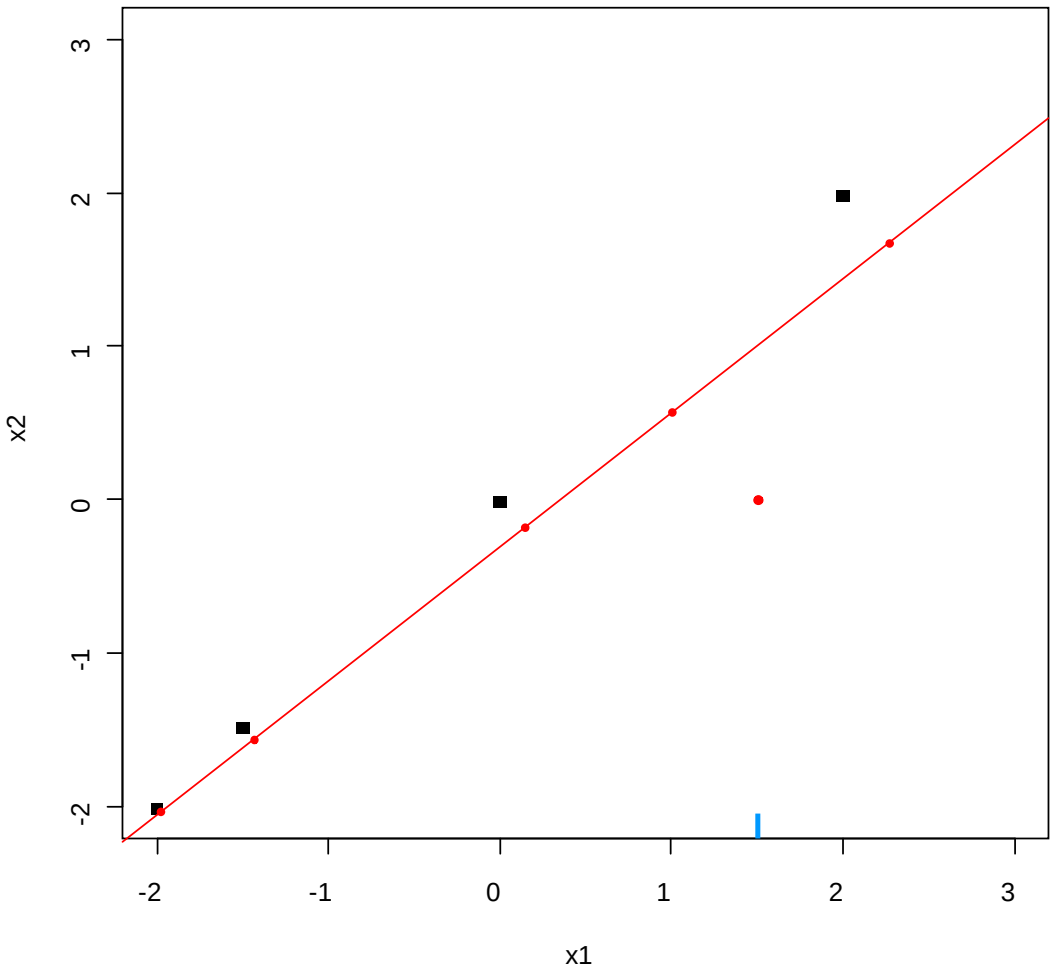
Initialization $\ell = 0$: X^0 (mean imputation)

Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	0.00
2.0	1.98

$\widehat{x1}$	$\widehat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	0.57
2.27	1.67



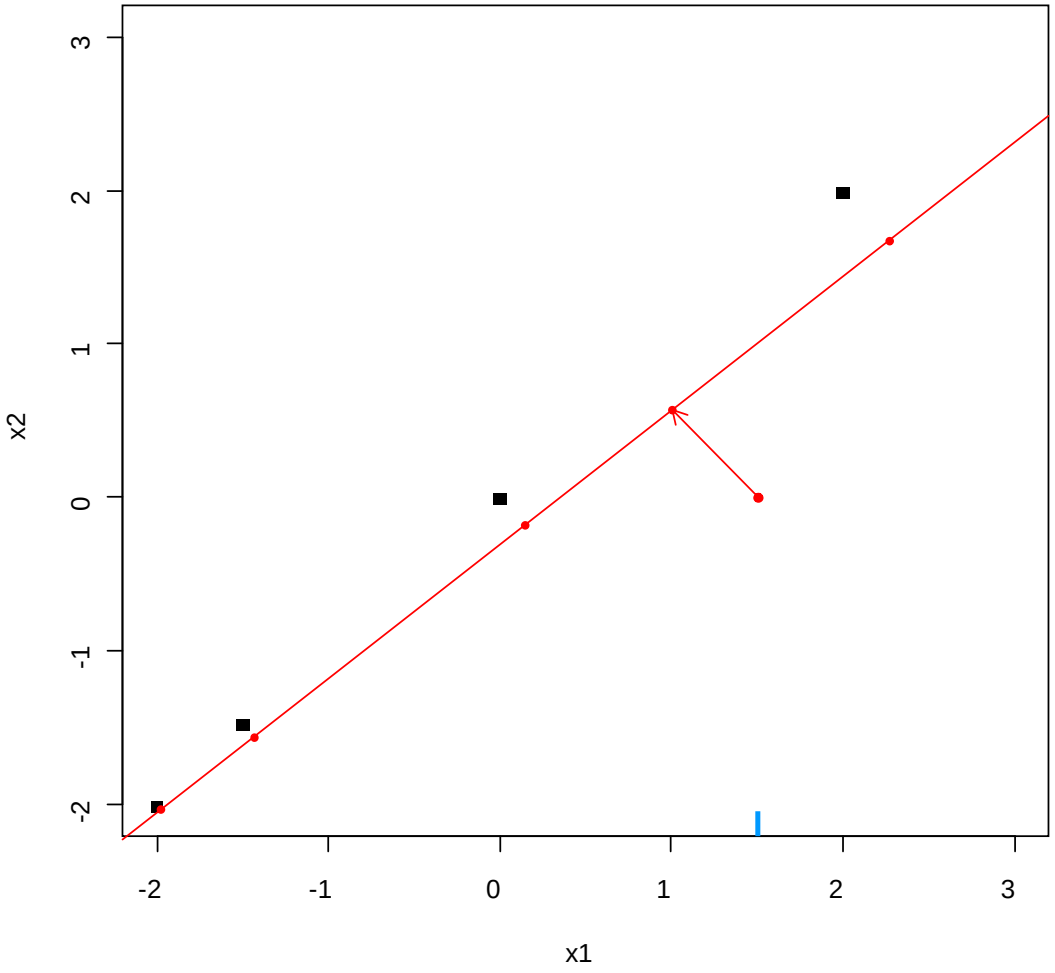
PCA on the completed data set $\rightarrow (U^{\ell}, \Lambda^{\ell}, D^{\ell})$;

Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	0.00
2.0	1.98

$\hat{x1}$	$\hat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	0.57
2.27	1.67



Missing values imputed with the fitted matrix $\hat{\mu}^{\ell} = U^{\ell} D^{\ell} V^{\ell\prime}$

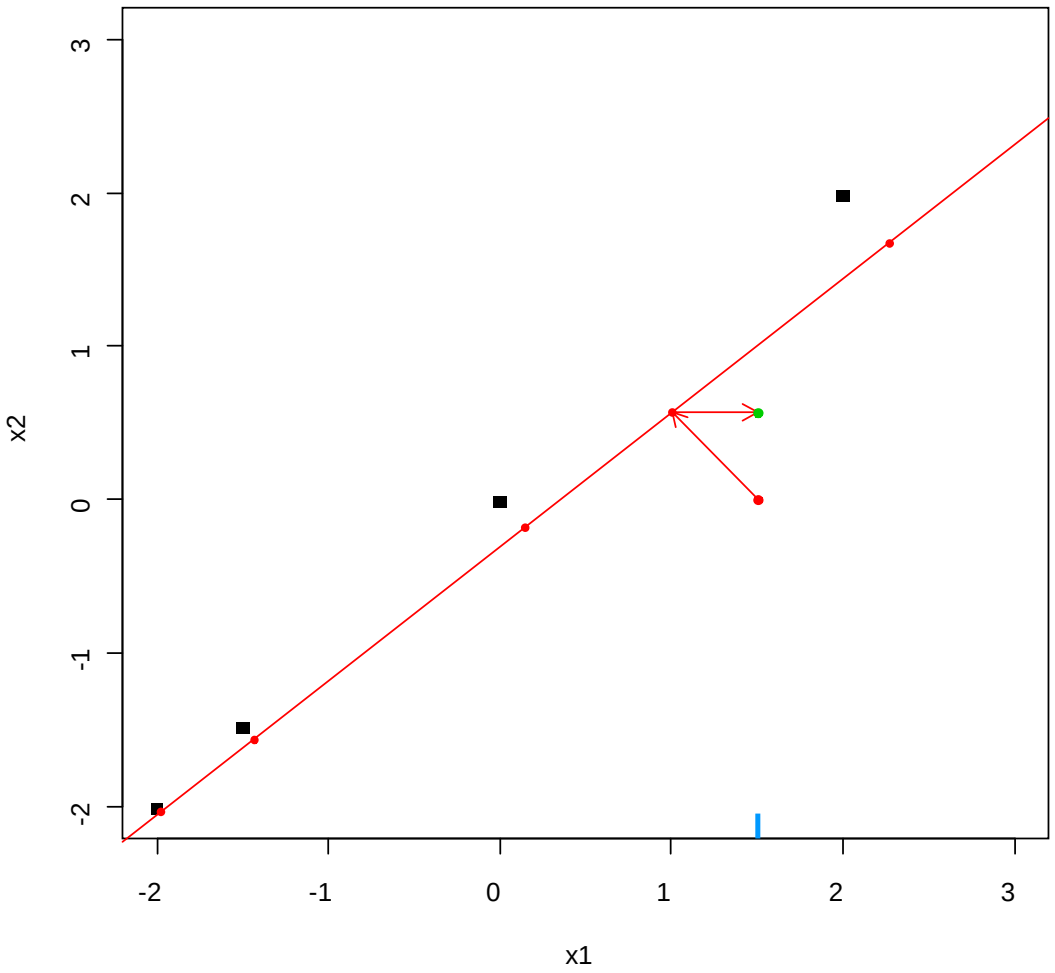
Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.00**
2.0	1.98

$\hat{x1}$	$\hat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	**0.57**
2.27	1.67

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



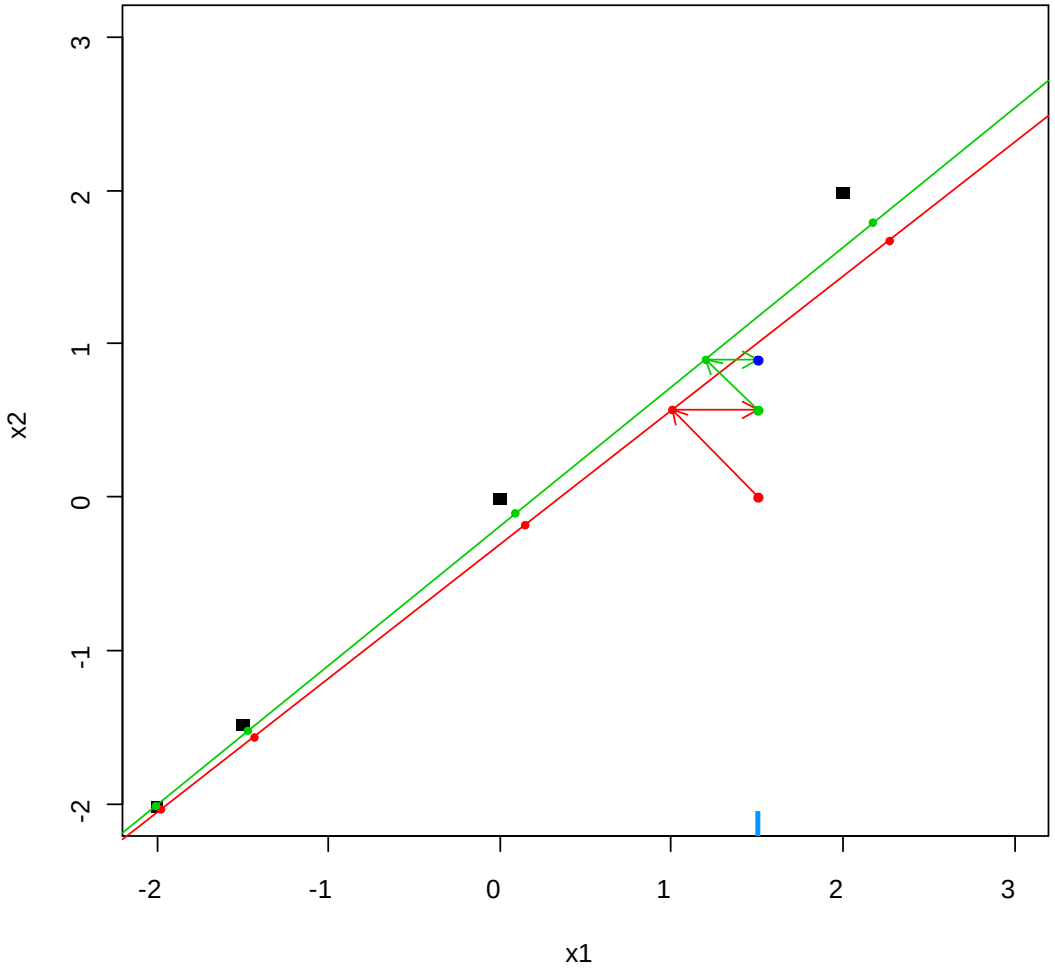
The new imputed dataset is $\hat{X}^\ell = W \odot X + (\mathbf{1} - W) \odot \hat{\mu}^\ell$

Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	0.57
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	0.57
2.0	1.98



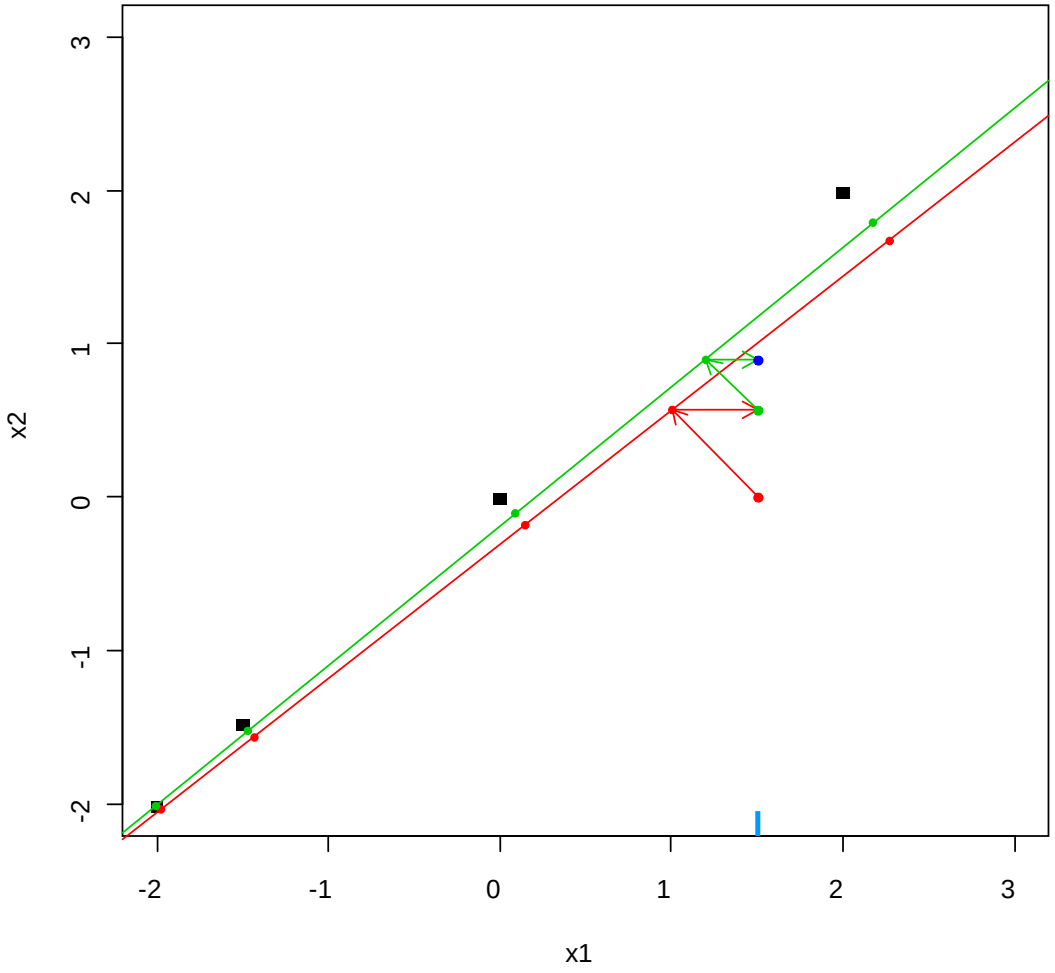
Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**NA**
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98

$\hat{x}1$	$\hat{x}2$
-2.00	-2.01
-1.47	-1.52
0.09	-0.11
1.20	**0.90**
2.18	1.78

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.90**
2.0	1.98



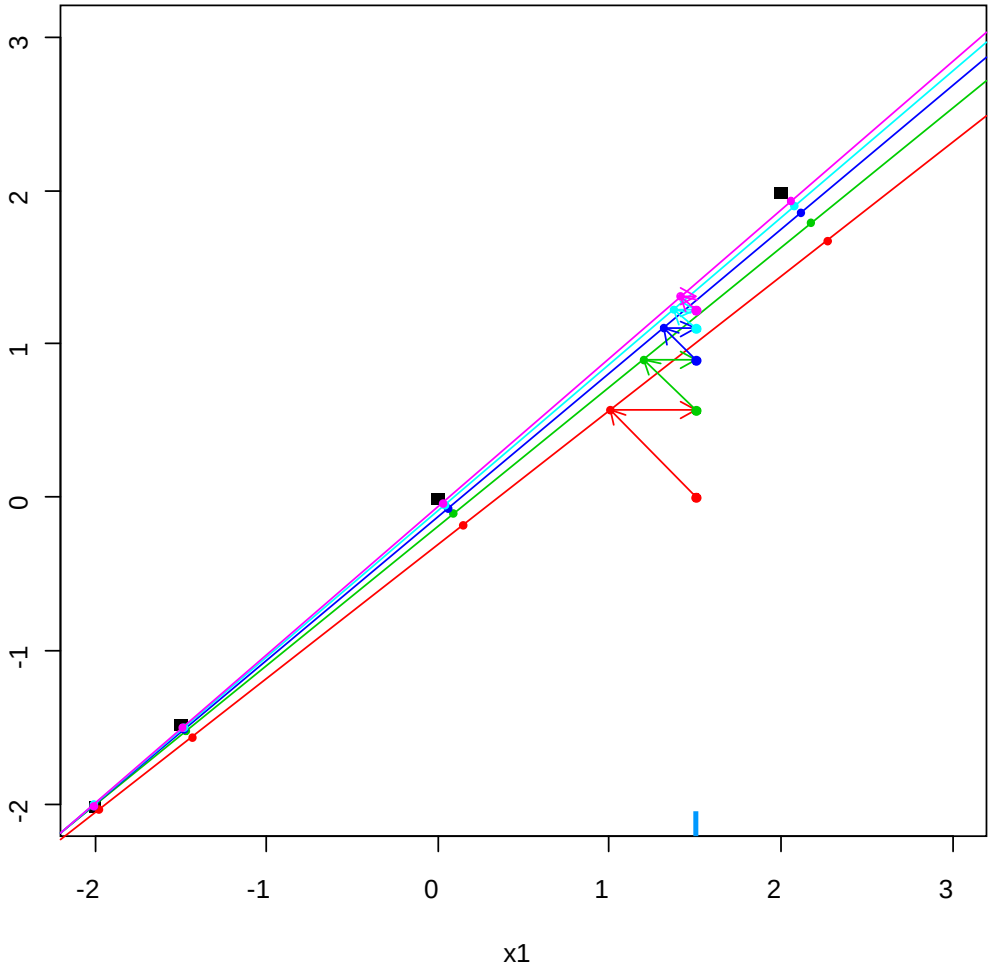
Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	0.00
2.0	1.98

$\hat{x1}$	$\hat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	0.57
2.27	1.67

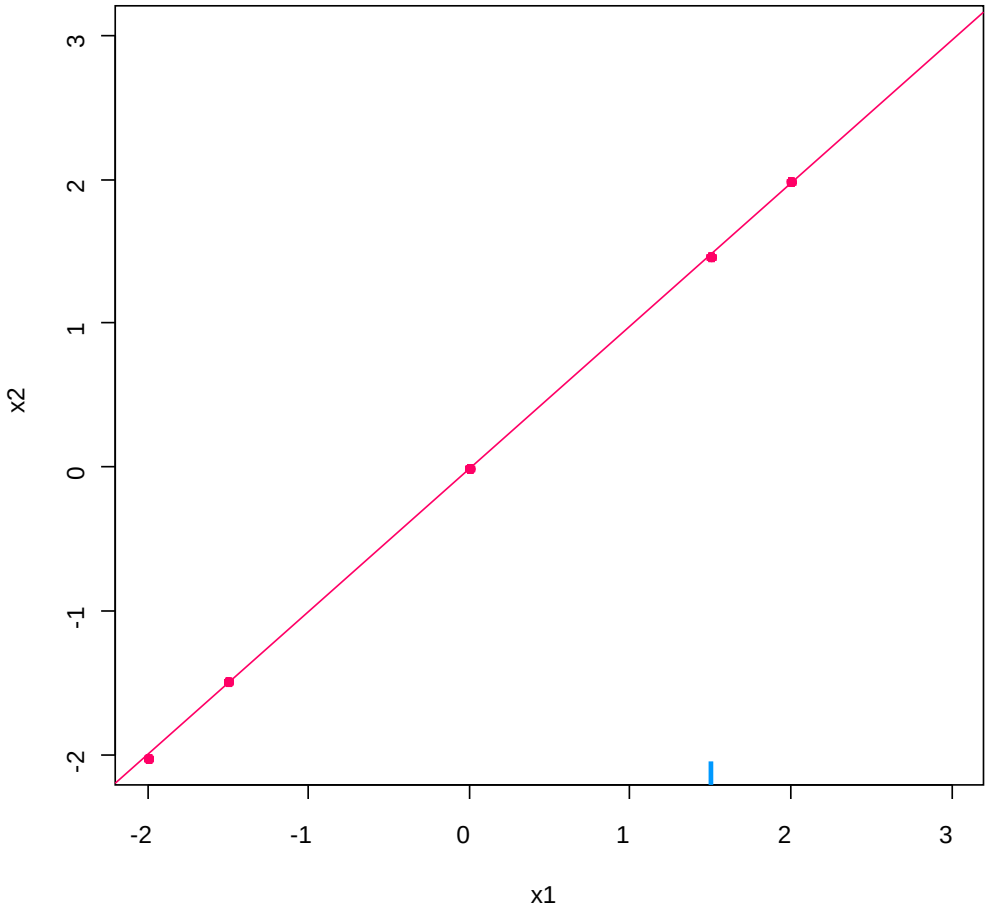
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	0.57
2.0	1.98



Steps are repeated until convergence

Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**NA**
2.0	1.98



x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**1.46**
2.0	1.98

PCA on the completed data set $\rightarrow (U^\ell, D^\ell, V^\ell)$

Missing values imputed with the fitted matrix $\hat{\mu}^\ell = U^\ell D^\ell V^{\ell\prime}$

Iterative PCA

1. initialization $\ell = 0$: X^0 (mean imputation)
2. step ℓ:
   (a) PCA on the completed data $\rightarrow (U^\ell, D^\ell, V^\ell)$; k dim kept
   (b) $(\hat{\mu}^\ell)^k = U^\ell D^\ell V^{\ell\prime}$ $\qquad X^\ell = W \odot X + (1 - W) \odot \hat{\mu}^\ell$
3. steps of estimation and imputation are repeated

Iterative PCA

1. initialization $\ell = 0$: X^0 (mean imputation)
2. step ℓ:
   (a) PCA on the completed data $\rightarrow (U^\ell, D^\ell, V^\ell)$; k dim kept
   (b) $(\hat{\mu}^\ell)^k = U^\ell D^\ell V^{\ell\prime}$ $\quad X^\ell = W \odot X + (1 - W) \odot \hat{\mu}^\ell$
3. steps of estimation and imputation are repeated

$\Rightarrow$ $\hat{\mu}$ from incomplete data: EM algo: $X = FV' + \varepsilon$
$X = \mu + \varepsilon,\ \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with μ of low rank

$\Rightarrow$ Completed data: good imputation (matrix completion, Netflix)
(Udell & Townsend Nice Latent Variable Models Have Log-Rank, 2017)

Selecting k? Generalized cross-validation (Josse & Husson, 2012)

Iterative PCA

1. initialization $\ell = 0$: X^0 (mean imputation)
2. step ℓ:
   (a) PCA on the completed data $\rightarrow (U^\ell, D^\ell, V^\ell)$; k dim kept
   (b) $(\hat{\mu}^\ell)^k = U^\ell D^\ell V^{\ell\prime}$ $\quad X^\ell = W \odot X + (1 - W) \odot \hat{\mu}^\ell$
3. steps of estimation and imputation are repeated

$\Rightarrow$ $\hat{\mu}$ from incomplete data: EM algo: $X = FV' + \varepsilon$

$X = \mu + \varepsilon,\ \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with μ of low rank

$\Rightarrow$ Completed data: good imputation (matrix completion, Netflix)

(Udell & Townsend Nice Latent Variable Models Have Log-Rank, 2017)

Selecting k? Generalized cross-validation (Josse & Husson, 2012)

Regularized iterative SVD

$\Rightarrow$ Overfitting issues of iterative PCA: many parameters ($U_{n\times k}$, $V_{k\times p}$)/observed values (k large - many NA); noisy data

$\Rightarrow$ Regularized versions. Init - estimation - imputation steps:

Imputation by $\hat{\mu}^{\text{PCA}} = \sum_{q=1}^{k} d_q u_q v_q^{'}$ is replaced by

$(\hat{\mu})_\lambda = \sum_{q=1}^{n} (d_q - \lambda)_+ u_q v_q^{'}$ $\text{argmin}_\mu \left\{ \|W \odot (X - \mu)\|_2^2 + \lambda \|\mu\|_* \right\}$

$\hat{\mu} = \sum_{q=1}^{k} \left(d_l - \frac{\hat{\sigma}^2}{d_q} \right) u_q v_q'$

Hastie et.al. (2015), Verbank, J. & Husson (2013); Gavish & Donoho (2014), J. & Wager (2015), J. & Sardy (2014)

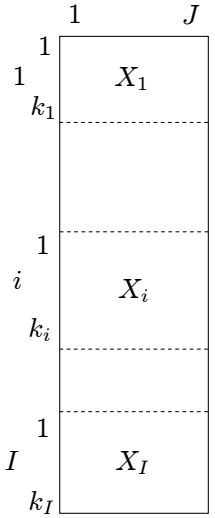
$\Rightarrow$ Iterative SVD algorithms good to impute data
$\Rightarrow$ Model makes sense: data = structure of rank k + noise

Outline

Multilevel component analysis

Ex: inhabitants nested within countries $X \in \mathbb{R}^{K \times J}$

- similarities between countries? level 1
- similarities between inhabitants within each country? level 2
- relationship between variables at each level



$$x_{ijk_i} = x_{.j.} + (x_{ij.} - x_{.j.}) \quad + \quad (x_{ijk_i} - x_{ij.})$$
$$\text{Between} \quad + \quad \text{Within}$$

Analysis of variance: split the sum of squares for each variable j

$$\sum_{i=1}^{I}\sum_{k=1}^{k_i}(x_{ijk_i})^2 = \sum_{i=1}^{I} k_i(x_{.j.})^2 + \sum_{i=1}^{I} k_i(x_{ij.} - x_{.j.})^2 + \sum_{i=1}^{I}\sum_{k=1}^{k_i}(x_{ijk_i} - x_{ij.})^2$$

Multilevel PCA MLPCA

$\Rightarrow$ Model for the between and within part $i = 1, ..., I$ groups, J var

$$X_{i_{(k_i \times J)}} = 1_{k_i} m' + 1_{k_i} F_i^{b'} V^{b'} + F_i^w V^{w'} + E_i$$

- F_i^b $(Q_b \times 1)$ between component scores of group i
- V^b $(J \times Q_b)$ between loading matrix
- F_i^w $(k_i \times Q_w)$ within component scores of group i
- V_w $(J \times Q_w)$ within loading matrix. Constant across groups

Fitted by solving the least squares (Timmerman, 2006)

$$\text{argmin}\, F(m, F_i^b, V^b, F_i^w, V^w) = \sum_{i=1}^{I} \left\| X_i - 1_{k_i} m' - 1_{k_i} F_i^{b'} V^{b'} - F_i^w V^{w'} \right\|^2$$

$\sum_{i=1}^{I} k_i F_i^b = 0_{Q_b}$ and $1'_{k_i} F_i^w = 0_{Q_w}$, $\forall i$ for identifiability.

MLPCA - quantitative data

$i = 1, ..., I$ groups, J var, k_i nb obs in group i
$\Rightarrow$ Estimation: minimize the RSS

$$\text{argmin}\, F() = \sum_{i=1}^{I} \left\| X_i - 1_{k_i} m' - 1_{k_i} F_i^{b'} V^{b'} - F_i^w V^{w'} \right\|^2,$$

$\sum_{i=1}^{I} k_i F_i^b = 0_{Q_b}$ and $1'_{k_i} F_i^w = 0_{Q_w}$, $\forall i$ for identifiability.

$(\hat{F}^b, \hat{V}^b)$: Weigthed PCA on the between part: SVD on WX_m; X_m $(I \times J)$ the means of the variables per group, W $(I \times I)$ $w_{ii} = \sqrt{k_i}$

$(\hat{F}^w, \hat{V}^w)$ PCA on the within part: SVD on the centered data per group X^w $(K \times J)$, $K = \sum_i k_i$

$\Rightarrow$ With missing values: Weighted Least Squares
$\Rightarrow$ Iterative imputation algorithm (imputation - estimation)

Iterative MLPCA

2. iteration ℓ: *estimation of the between structure*
   - SVD $WX_m^\ell = PDQ'$; Q_b eigenvectors are kept:
     $\hat{F}_i^b = [W^{-1}P_{Q_b}]_i$, $\hat{F}^b$ concatenation by row of $[\mathbf{1}_{k_i}\hat{F}_i^b]$
     $\hat{V}^b = Q_{Q_b}D_{Q_b}$, $(J \times Q_b)$
   - the between hat matrix is computed: $(\hat{X}^b)^\ell = \hat{F}^b\hat{V}^{b'}$
3. iteration ℓ: *imputation of the missing values with the fitted values*
   - $\hat{X}^\ell = \mathbf{1}_K\hat{m}^{(\ell-1)'} + (\hat{X}^b)^\ell + (\hat{X}^w)^{(\ell-1)}$. The newly imputed dataset is
     $X^\ell = W \odot X + (\mathbf{1}_K \times \mathbf{1}'_J - W) \odot \hat{X}^\ell$
   - $\hat{m}^\ell$ is computed on X^ℓ
4. iteration ℓ: *estimation of the within structure*
   - SVD $(X^w)^\ell = PDQ'$; Q_w eigenvectors are kept:
     $F^w = P_{Qw}$ $(K \times Q_w)$
     $V^w = Q_{Qw}D_{Qw}$ $(J \times Q_w)$
   - the within hat matrix is computed $(\hat{X}^w)^\ell = \hat{F}^w\hat{V}^{w'}$
5. iteration ℓ: *imputation of the missing values with the fitted values*
   - $X^{\ell+1} = M \odot X + (\mathbf{1}_K \times \mathbf{1}'_J - M) \odot \left(\mathbf{1}_K\hat{m}^{(\ell)'} + (\hat{X}^b)^\ell + (\hat{X}^w)^\ell\right)$
   - $\hat{m}^{\ell+1}$ is computed on $X^{\ell+1}$

Regularized iterative MLSCA

Estimation SVD - Imputation $\hat{X} = \mathbf{1}_K \hat{m}' + \hat{F}^b \hat{V}^{b'} + \hat{F}^w \hat{V}^{w'}$

$\hat{V}^b = Q_{Q_b} D_{Q_b} = Q_{Q_b} \mathrm{diag}\,(d_1, ..., d_{Q_b})$, as:

$$Q_{Q_b} \mathrm{diag}\left(d_1 - \frac{\hat{\sigma}_b^2}{d_1}, ..., d_{Q_b} - \frac{\hat{\sigma}_b^2}{d_{Q_b}} \right), \ \hat{\sigma}_b^2 = \frac{1}{J - Q_b} \sum_{q=Q_b+1}^{J} d_q$$

$$\hat{V}^w = Q_{Q_w} \mathrm{diag}\left(d_1 - \frac{\hat{\sigma}_w^2}{d_1}, ..., d_{Q_w} - \frac{\hat{\sigma}_w^2}{d_{Q_w}} \right)$$

Multiple Correspondence Analysis (MCA)

$X_{n\times m}$ m categorical variables coded with indicator matrix A

$$X = \begin{array}{|c|c|c|}\hline y & \dots & attack \\ y & \dots & attack \\ y & \dots & attack \\ n & \dots & suicide \\ & & \\ & & \\ n & \dots & accident \\ n & \dots & suicide \\ \hline \end{array} \quad A = \begin{array}{|cc|c|ccc|}\hline 1 & 0 & \dots & 1 & 0 & 0 \\ 1 & 0 & \dots & 1 & 0 & 0 \\ 1 & 0 & \dots & 1 & 0 & 0 \\ 0 & 1 & \dots & 0 & 1 & 0 \\ & & & & & \\ & & & & & \\ 0 & 1 & \dots & 0 & 0 & 1 \\ 0 & 1 & \dots & 0 & 1 & 0 \\ \hline \end{array} \quad D_p = \begin{array}{|ccc|}\hline p_1 & & 0 \\ & \ddots & \\ 0 & & p_J \\ \hline \end{array}$$

For a category c, the frequency of the category: $p_c = n_c/n$.

A SVD on weighted matrix: $Z = \frac{1}{\sqrt{mn}}(A - 1p^T)D_p^{-1/2} = UDV'$

The PC ($F = UD$) satisfies: $\arg\max_{F_q\in\mathbb{R}^n} \quad \frac{1}{m}\sum_{j=1}^m \eta^2(F_q, X_j)$

$$\eta^2(F, X_j) = \frac{\sum_{c=1}^{C_j} n_c(F_{.c} - F_{..})^2}{\sum_{i=1}^n \sum_{c=1}^{C_j}(F_{ic})^2} = \frac{\text{RSS between}}{\text{RSS tot}}$$

Benzecri, 1973 :*"In data analysis the mathematical problems reduces to computing eigenvectors; all the science (the art) is in finding the right matrix to diagonalize"*

Multilevel MCA

$\Rightarrow$ Start with the matrix of dummy variables A and define a between and a within part
$\Rightarrow$ Then, MCA is applied on each part

Between: Apply MCA on the matrix with the mean of A per group i (proportion of obs taking each category in group i) (proportion of some disease in a particular hospital). $\hat{A}^b = F^b V^{b'} D_p^{1/2} + 1_n p'$

Within part Apply MCA on the data where the between part has been swept out (SVD is applied to $\frac{1}{np}\left(A - \hat{A}^b\right) D_p^{-1/2}$)
$\hat{A}^w = (np) F^w V^{w'} D_p^{1/2}$.

$$\hat{A} = \hat{A}^b + \hat{A}^w$$

Regularized iterative Multilevel MCA

	V1	V2	V3	...	V14
ind 1	a	**NA**	g	...	u
ind 2	**NA**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**NA**		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	**NA**	**NA**	1	0	...
ind 2	**NA**	**NA**	**NA**	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	**NA**	**NA**	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

Regularized iterative Multilevel MCA

- Initialization: imputation of the indicator matrix (proportions)

	V1	V2	V3	...	V14
ind 1	a	**NA**	g	...	u
ind 2	**NA**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**NA**		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	0.41	0.59	1	0	...
ind 2	0.20	0.30	0.50	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	0.27	0.78	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

Regularized iterative Multilevel MCA

- Initialization: imputation of the indicator matrix (proportions)
- Iterate until convergence
  1. estimation: Multilevel MCA on the completed data $\rightarrow$ $\hat{F}^b, \hat{V}^b, \hat{F}^w, \hat{V}^w$

	V1	V2	V3	...	V14
ind 1	a	**NA**	g	...	u
ind 2	**NA**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**NA**		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	0.41	0.59	1	0	...
ind 2	0.20	0.30	0.50	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	0.27	0.78	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

Regularized iterative Multilevel MCA

- Initialization: imputation of the indicator matrix (proportions)
- Iterate until convergence
  1. estimation: Multilevel MCA on the completed data $\rightarrow$ $\hat{F}^b, \hat{V}^b, \hat{F}^w, \hat{V}^w$
  2. imputation with the fitted matrix $\hat{A} = \hat{A}^b + \hat{A}^w$

	V1	V2	V3	...	V14
ind 1	a	**NA**	g	...	u
ind 2	**NA**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**NA**		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	0.65	0.35	1	0	...
ind 2	0.11	0.20	0.69	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	0.30	0.40	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

Regularized iterative Multilevel MCA

- Initialization: imputation of the indicator matrix (proportions)
- Iterate until convergence
  1. estimation: Multilevel MCA on the completed data $\rightarrow$ $\hat{F}^b, \hat{V}^b, \hat{F}^w, \hat{V}^w$
  2. imputation with the fitted matrix $\hat{A} = \hat{A}^b + \hat{A}^w$
  3. column margins are updated

	V1	V2	V3	...	V14
ind 1	a	**NA**	g	...	u
ind 2	**NA**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**NA**		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	0.65	0.35	1	0	...
ind 2	0.11	0.20	0.69	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	0.30	0.40	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

Regularized iterative Multilevel MCA

- Initialization: imputation of the indicator matrix (proportions)
- Iterate until convergence
  1. estimation: Multilevel MCA on the completed data $\rightarrow$ $\hat{F}^b, \hat{V}^b, \hat{F}^w, \hat{V}^w$
  2. imputation with the fitted matrix $\hat{A} = \hat{A}^b + \hat{A}^w$
  3. column margins are updated

	V1	V2	V3	...	V14
ind 1	a	**NA**	g	...	u
ind 2	**NA**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**NA**		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	**0.71**	**0.29**	1	0	...
ind 2	**0.12**	**0.29**	**0.59**	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	**0.37**	**0.63**	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

$\Rightarrow$ the imputed values can be seen as degree of membership

Regularized iterative Multilevel MCA

- Initialization: imputation of the indicator matrix (proportions)
- Iterate until convergence
  1. estimation: Multilevel MCA on the completed data $\rightarrow$ $\hat{F}^b, \hat{V}^b, \hat{F}^w, \hat{V}^w$
  2. imputation with the fitted matrix $\hat{A} = \hat{A}^b + \hat{A}^w$
  3. column margins are updated

	V1	V2	V3	...	V14
ind 1	a	**e**	g	...	u
ind 2	**c**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**g**		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	**0.71**	**0.29**	1	0	...
ind 2	**0.12**	**0.29**	**0.59**	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	**0.37**	**0.63**	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

Two ways to impute categories: majority or draw

Regularized iterative Multilevel MCA

- Initialization: imputation of the indicator matrix (proportions)
- Iterate until convergence
  1. estimation: Multilevel MCA on the completed data $\rightarrow$ $\hat{F}^b, \hat{V}^b, \hat{F}^w, \hat{V}^w$
  2. imputation with the fitted matrix $\hat{A} = \hat{A}^b + \hat{A}^w$
  3. column margins are updated

	V1	V2	V3	...	V14
ind 1	a	**e**	g	...	u
ind 2	**c**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**g**		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	**0.71**	**0.29**	1	0	...
ind 2	**0.12**	**0.29**	**0.59**	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	**0.37**	**0.63**	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

Two ways to impute categories: majority or draw

Public Assistance - Paris Hospitals

Traumabase: 15000 patients/ 250 variables/ 6 hospitals

	Center	Accident	Age	Sex	Weight	Height	BMI	BP	SBP
1	Beaujon	Fall	54	m	85	NR	NR	180	110
2	Lille	Other	33	m	80	1.8	24.69	130	62
3	Pitie Salpetriere	Gun	26	m	NR	NR	NR	131	62
4	Beaujon	AVP moto	63	m	80	1.8	24.69	145	89
6	Pitie Salpetriere	AVP bicycle	33	m	75	NR	NR	104	86
7	Pitie Salpetriere	AVP pedestrian	30	w	NR	NR	NR	107	66
9	HEGP	White weapon	16	m	98	1.92	26.58	118	54
10	Toulon	White weapon	20	m	NR	NR	NR	124	73
11	Bicetre	Fall	61	m	84	1.7	29.07	144	105

.....

	SpO2	Temperature	Lactates	Hb	Glasgow	Transfusion	.....
1	97	35.6	<NA>	12.7	12	yes	
2	100	36.5	4.8	11.1	15	no	
3	100	36	3.9	11.4	3	no	
4	100	36.7	1.66	13	15	yes	
6	100	36	NM	14.4	15	no	
7	100	36.6	NM	14.3	15	yes	
9	100	37.5	13	15.9	15	yes	
10	100	36.9	NM	13.7	15	no	
11	100	36.6	1.2	14.2	14	no	

.....

Imputed Paris Hospitals data

Traumabase: 15000 patients/ 250 variables/ 6 hospitals

```
               Center       Accident   Age Sex Weight Height   BMI BP SBP
1            Beaujon        Fall         54    m 85.00   1.84 27.04  83  13
2              Lille        Other        33    m 80.00   1.80 24.69  33  98
3  Pitie Salpetriere        Gun          26    m 81.78   1.85 24.33  34  98
4            Beaujon        AVP moto     63    m 80.00   1.80 24.69  48 125
6  Pitie Salpetriere        AVP bicycle  33    m 75.00   1.83 24.53   6 122
7  Pitie Salpetriere        AVP pedestri 30    m 81.89   1.82 25.24   9 102
9               HEGP        White weapon 16    m 98.00   1.92 26.58  21  90
10            Toulon        White weapon 20    m 81.68   1.82 25.05  27 109
11           Bicetre        Fall         61    m 84.00   1.70 29.07  47   8
   SpO2 Temperature Lactates Hb Glasgow............
1    46          61   289.07 33      14
2     2          72   464.00 16      14
3     2          65   416.00 19       7
4     2          74   130.00 36       6
6     2          65   285.91 50       6
7     2          73   244.99 49       6
9     2          83   196.00 65       6
10    2          76   262.44 43       6
11    2          73    84.00 48       5
```

Design

The simulated data:

- $X_{i_{(k_i \times J)}} = 1_{k_i} m' + 1_{k_i} F_i^{b'} V^{b'} + F_i^w V^{w'} + E_i$, with $E_{ijk_i} \sim \mathcal{N}(0, \sigma)$
- 5 groups, 10 variables, $Q_b = 2$, $Q_w = 2$

Many scenarios are considered:

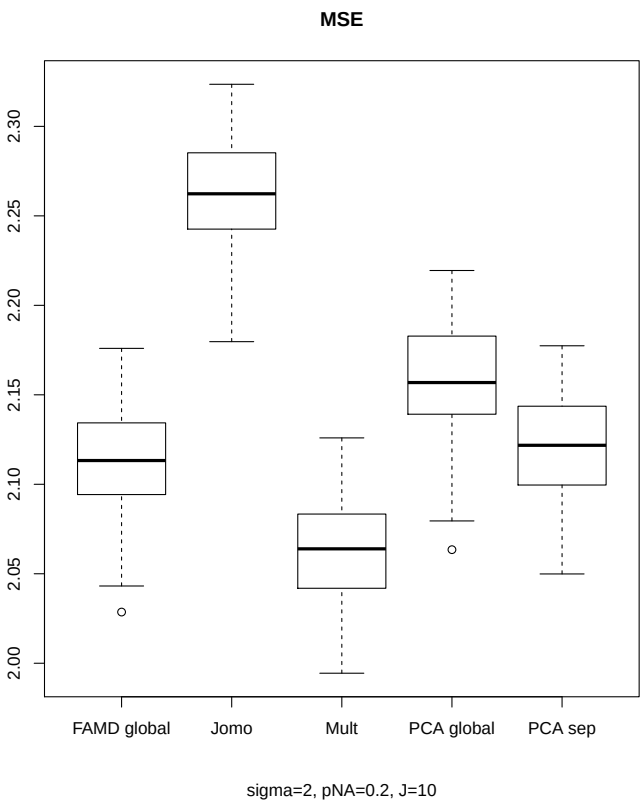
- number of measurement occasions: 10-20, 70-100
- level of noise: low, strong
- percentage of missing values: 10%, 25%, 40%
- missing values mechanism: MCAR, 2 MAR situations

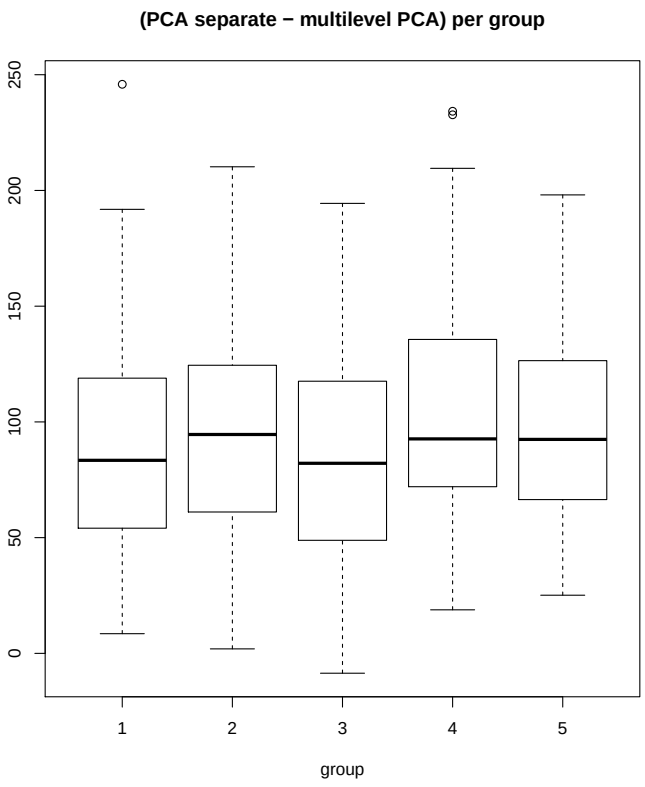
$\Rightarrow$ Prediction error: $\frac{1}{KJ} \sum (x_{ijk_i} - \hat{x_{ijk_i}})^2$

Results

Competitors:

- Conditional model with random effect regression (mice)
- Random forests (bühlmann, 2012) (not designed)
- Global PCA - Separate PCA
- Global mixed PCA (with hospital)





Results

	$J = 10$	$J = 30$	$5cat$	$5cat$
Global PCA	0.09	0.3		
mice	11	282		
Multilevel SVD	1.5	1.2	2	7
Global mixed PCA	0.4	0.7	1	4
Random forest	59	200	27	246

Table: Time in seconds for a dataset with 20% NA, $I = 5$ $k_i = 200$

- PCA mixed as Random Forest
- mice (random effect model): difficulties with large dimensions
- Separate PCA: pb with many missing values
- Multilevel SVD = global SVD when no group effect
- Imputation properties depends on the method (linear)
- Other methods do not handle categorical variables

Outline

6 hospital databases

Combining data from different institutional databases promises many advantages in personalizing medical care (large n, more chance for finding patients like me)

$\Rightarrow$ NIH requires sharing of data from funded projects

6 hospital databases

Combining data from different institutional databases promises many advantages in personalizing medical care (large n, more chance for finding patients like me)

$\Rightarrow$ NIH requires sharing of data from funded projects

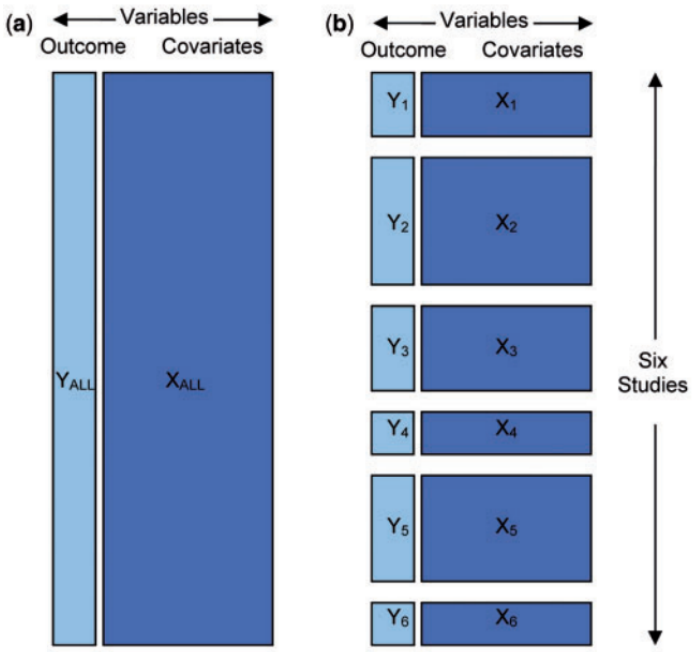
$\Rightarrow$ The problem: high barriers to aggregation of medical data

- lack of standardization of ontologies
- privacy concerns
- proprietary attitude towards data, reluctance to cede control
- complexity/size of aggregated data, updates problems

Solution: distributed computation

⇒ Data aggregation is not always necessary

⇒ NIH splits the storage of aggregated data across several centers

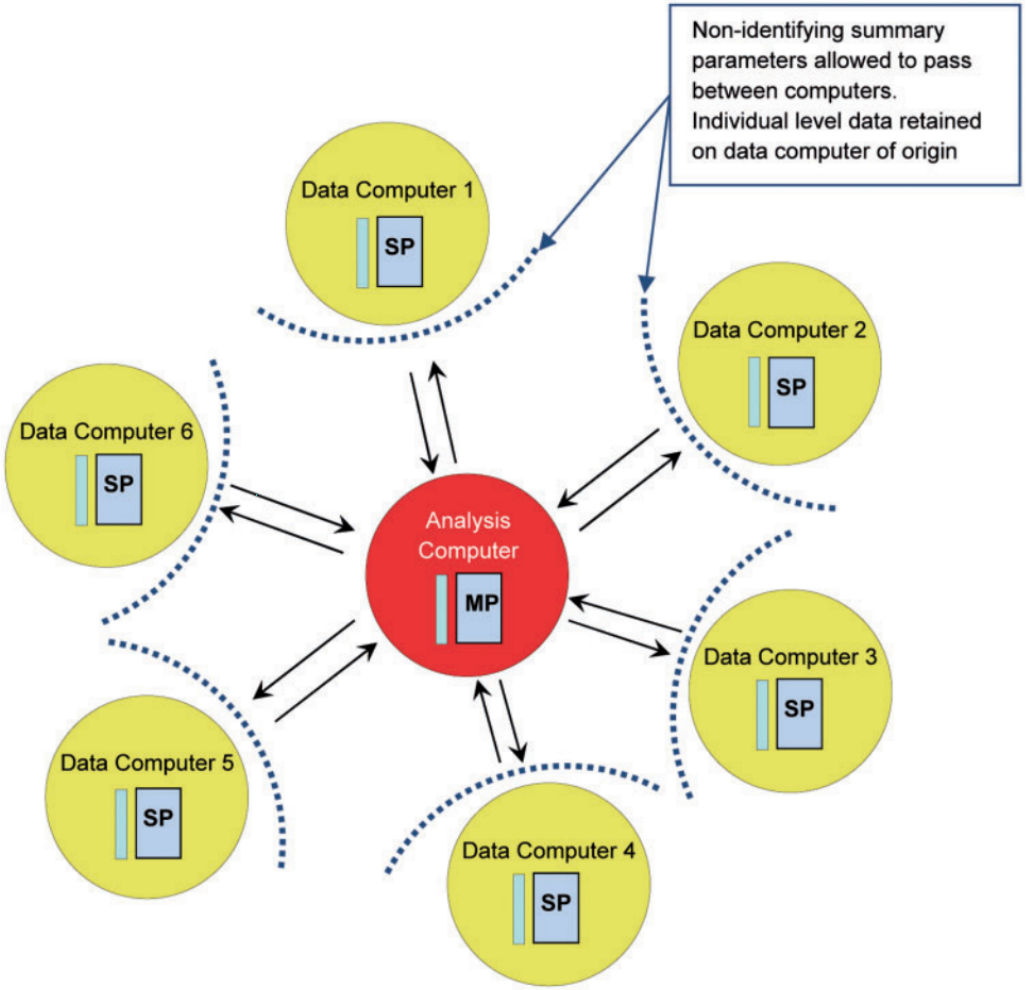


⇒ Data can stay at site

⇒ Computations can be distributed (share burden)

⇒ Hospitals only share intermediate results instead of the raw data

Topology: master-workers (Wolfson, et. al (2010))



$\Rightarrow$ Ex: Each site share the sum of age $\tilde{X}_i$ and the number of patients n_i. The master computes $\bar{X} = \sum n_i \tilde{X}_i / \sum n_i$

Solution: distributed computation

⇒ Many models fitting can be implemented:

- Maximizing a likelihood. Intermediate computations break up into sums of quantities computed on local data at sites. Log-likelihood, score function and Fisher information can partition into sums. (OK for logistic regression)
- Singular Value Decomposition. Iterative algorithms available for SVD using quantities computed on local data at sites.
- And more.

Implemented in the R package discomp (Narasimhan et. al., 2017)

Singular value decomposition

SVD: $X_{n\times p} : U_{n\times k} D_{k\times k} V'_{p\times k}$

Power method to get the first direction:

Data: $\boldsymbol{X} \in \mathcal{R}^{n\times p}$
Result: $u \in \mathcal{R}^n$, $v \in \mathcal{R}^p$, and $d > 0$
$u \leftarrow (\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$;
repeat
  $v \leftarrow \boldsymbol{X}^\top u$;
  $v \leftarrow v/\|v\|$;
  $u \leftarrow \boldsymbol{X} v$;
  $d \leftarrow \|u\|$;
  $u \leftarrow u/\|u\|$;
until *convergence*;

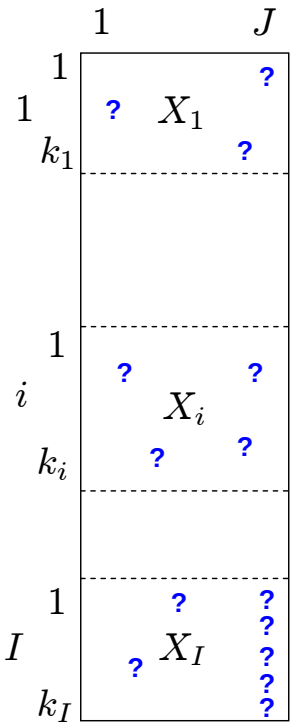
Other dims: "deflation", same procedure in the residuals $(X - udv')$

$\Rightarrow$ Involves inner products and sums: distributed

Privacy preserving rank k SVD

Data: each worker has private data $\boldsymbol{X}_i \in \mathcal{R}^{n_i \times p}$
Result: $V \in \mathcal{R}^{p \times k}$, and $d_1 \geq \ldots d_k \geq 0$
$V \leftarrow 0$, $d \leftarrow 0$ **foreach** *worker site* j **do**
  $U^{[j]} = 0$;
  transmit n_j to master;
end
for $i \leftarrow 1$ **to** k **do**
  foreach *worker site* j **do** $u^{[j]} \leftarrow (1, 1, \ldots, 1)$ of length n_j;
  $\|u\| \leftarrow \sqrt{\sum_j n_j}$;
  transmit $\|u\|, V$, and D to workers;
  repeat
    foreach *worker site* j **do**
      $u^{[j]} \leftarrow u^{[j]} / \|u\|$;
      calculate $v^{[j]} \leftarrow (\boldsymbol{X}^{[j]} - U^{[j]} D V^\top)^\top u^{[j]}$;
      transmit $v^{[j]}$ to master;
    end
    $v \leftarrow \sum_j v^{[j]}$;
    $v \leftarrow v / \|v\|$;
    transmit v to workers;
    foreach *worker site* j **do**
      calculate $u^{[j]} \leftarrow \boldsymbol{X}^{[j]} v$;
      transmit $\|u^{[j]}\|$ to master;
    end
    $\|u\| \leftarrow \sum_j \|u^{[j]}\|$;
    transmit $\|u\|$ to workers;
    $d_i \leftarrow \|u\|$;
  until *convergence*;
  $V \leftarrow \texttt{cbind}(V, v)$;
  foreach *worker site* j **do** $U^{[j]} \leftarrow \texttt{cbind}(U^{[j]}, u^{[j]})$;
end

Multilevel imputation



⇒ Impute multilevel data with Multilevel SVD
⇒ Distributed multilevel imputation
⇒ Impute the data of one hospital using the data of the others
⇒ Incentive to encourage the hospitals to participate in the project
⇒ Apply other statistical methods on the imputed data (logistic regression)

Take home message - On going work

Multilevel PCA powerful for single imputation of continuous & categorical multilevel data: reduce the dimensionality - capture the similarities between rows and relationship between variables at both levels.

Method without missing values

$\Rightarrow$ Computationaly fast - distributed - Implemented R package missMDA

- Numbers of components Q_b and Q_w ?

- Inference after imputation. Underestimation of the variance with single imputation

Take home message - On going work

Multilevel PCA powerful for single imputation of continuous & categorical multilevel data: reduce the dimensionality - capture the similarities between rows and relationship between variables at both levels.

Method without missing values

$\Rightarrow$ Computationaly fast - distributed - Implemented R package missMDA

- Numbers of components Q_b and Q_w ?

cross-validation?

- Inference after imputation. Underestimation of the variance with single imputation

Take home message - On going work

Multilevel PCA powerful for single imputation of continuous & categorical multilevel data: reduce the dimensionality - capture the similarities between rows and relationship between variables at both levels.

Method without missing values

$\Rightarrow$ Computationaly fast - distributed - Implemented R package missMDA

• Numbers of components Q_b and Q_w ?
cross-validation?

• Inference after imputation. Underestimation of the variance with single imputation
Multiple imputation: bootstrap + drawn from the predictive distribution
$\mathcal{N}\left(\mathbf{1}_K \hat{m}' + \hat{F}^b \hat{B}^{b'} + \hat{F}^w \hat{B}^{w'}, \hat{\sigma}^2\right)$

Research activities

- Dimensionality reduction methods to visualize complex data (PCA based): multi-sources, textual, arrays, questionnaire
- Missing values - matrix completion
- Low rank estimation, selection of regularization parameters
- Fields of application: bio-sciences (agronomy, sensory analysis), health data (hospital data)
- R community: book R for Statistics, R foundation, R taskforce, R packages and JSS papers:

  FactoMineR explore continuous, categorical, multiple contingency tables (correspondence analysis), combine clustering and PC, ..
  MissMDA for single and multiple imputation, PCA with missing
  denoiseR to denoise data