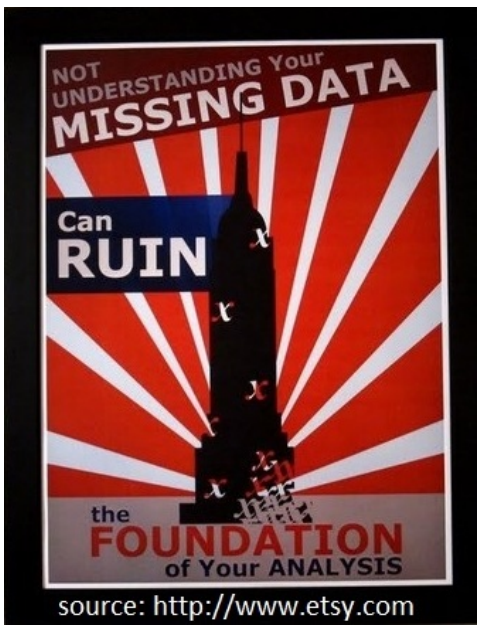
Inference with missing values using principal components methods

Julie Josse
Mathematical Methods of Modern Statistics, Luminy, CIRM



July 10, 2017

Outline

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 in the "big data" area

Data integration: data from different sources

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 missing categorical/continuous values

Recommended methods

$\Rightarrow$ Modify the estimation process to deal with missing values.
Maximum likelihood: EM algorithm to obtain point estimates + Supplemented EM (Meng & Rubin, 1991) 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)

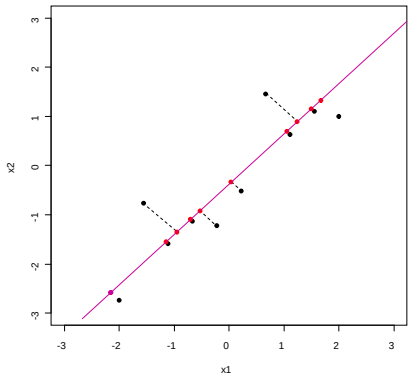
Litterature: Schaefer (2002); Little & Rubin (2002); Gelman & Meng (2004); Kim & Shao (2013); Carpenter & Kenward (2013); van Buuren (2015)

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



$\Rightarrow$ Minimizes distance between observations and their projection

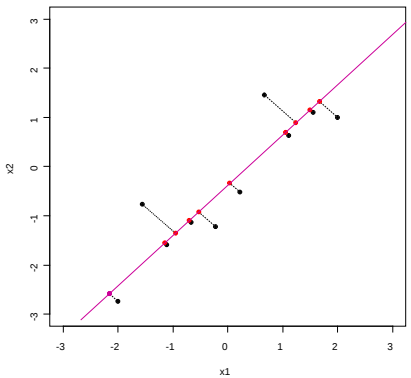
PCA reconstruction

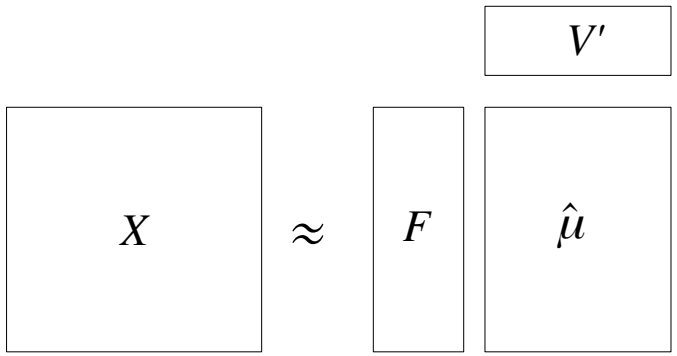
X

-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 $k < p$ $\|A\|_2^2 = \text{tr}(AA^\top)$:

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

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

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

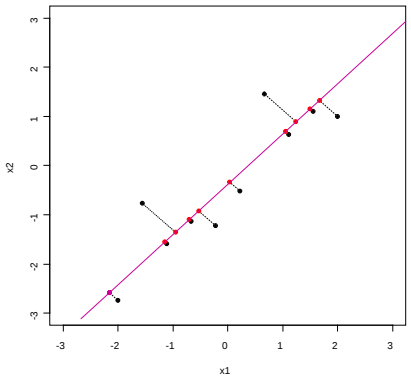
PCA reconstruction

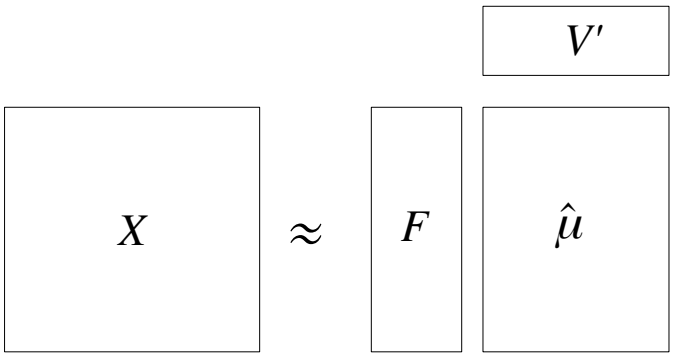
X

-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 $k < p$ $\|A\|_2^2 = \text{tr}(AA^\top)$:

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

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

$= F_{n\times k} V'_{p\times k}$ 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

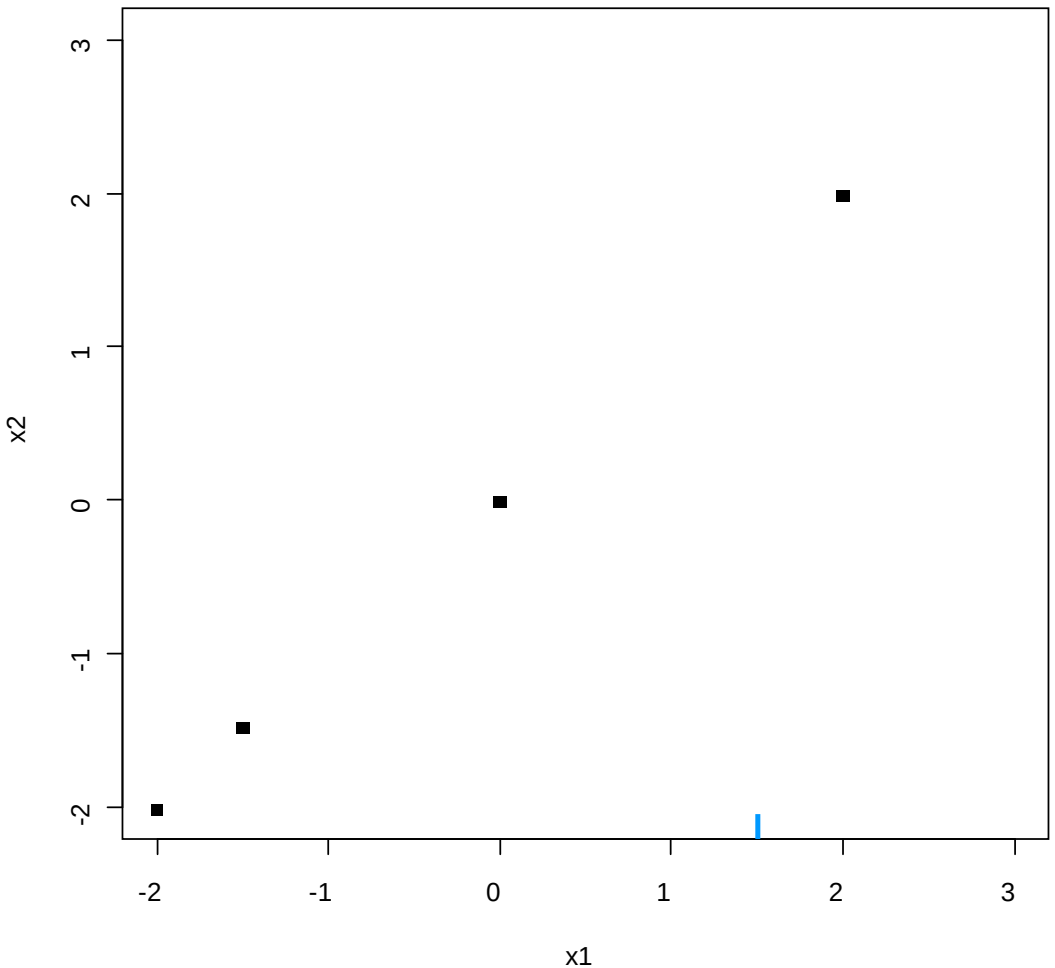
$$\text{argmin}_{\mu} \left\{ \|W_{n\times p} \odot (X - \mu)\|_2^2 : \text{rank}\,(\mu) \leq k \right\}$$

with $W_{ij} = 0$ if X_{ij} is missing, $W_{ij} = 1$ otherwise

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

Iterative PCA

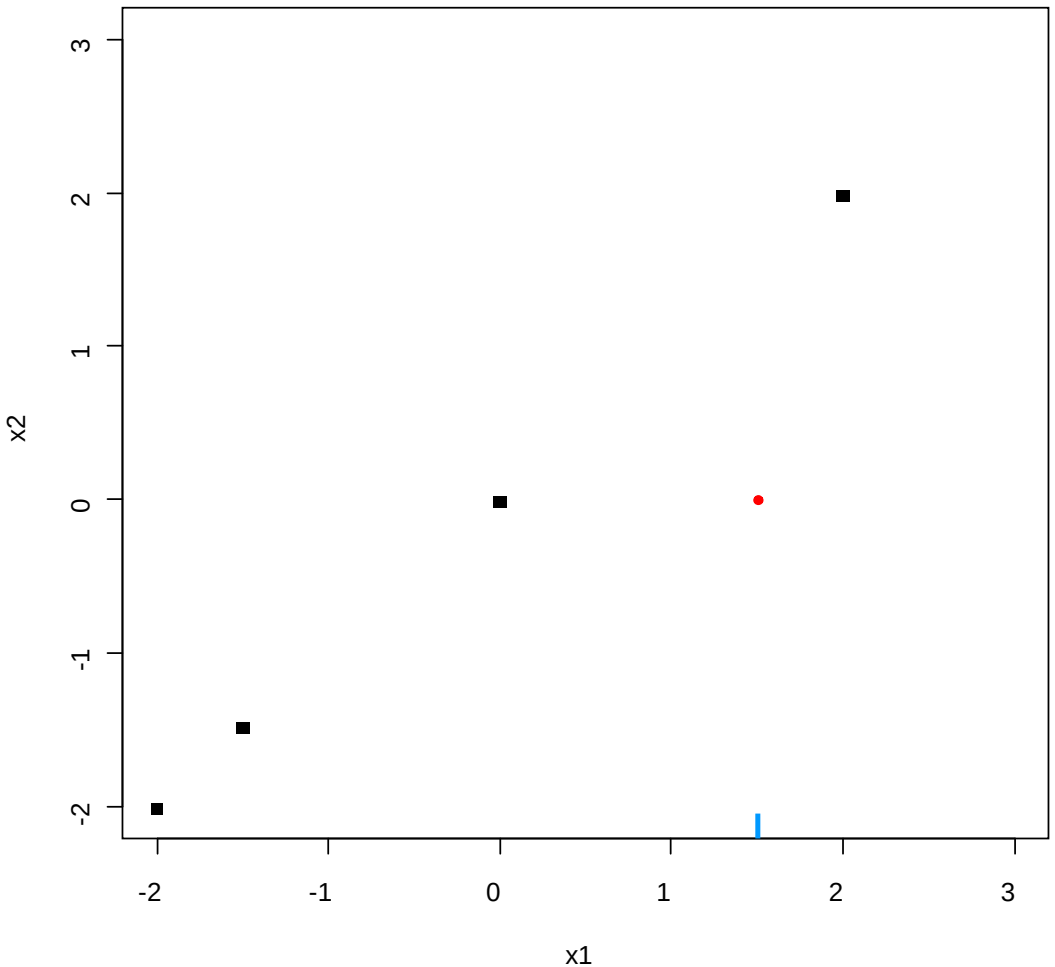
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98



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



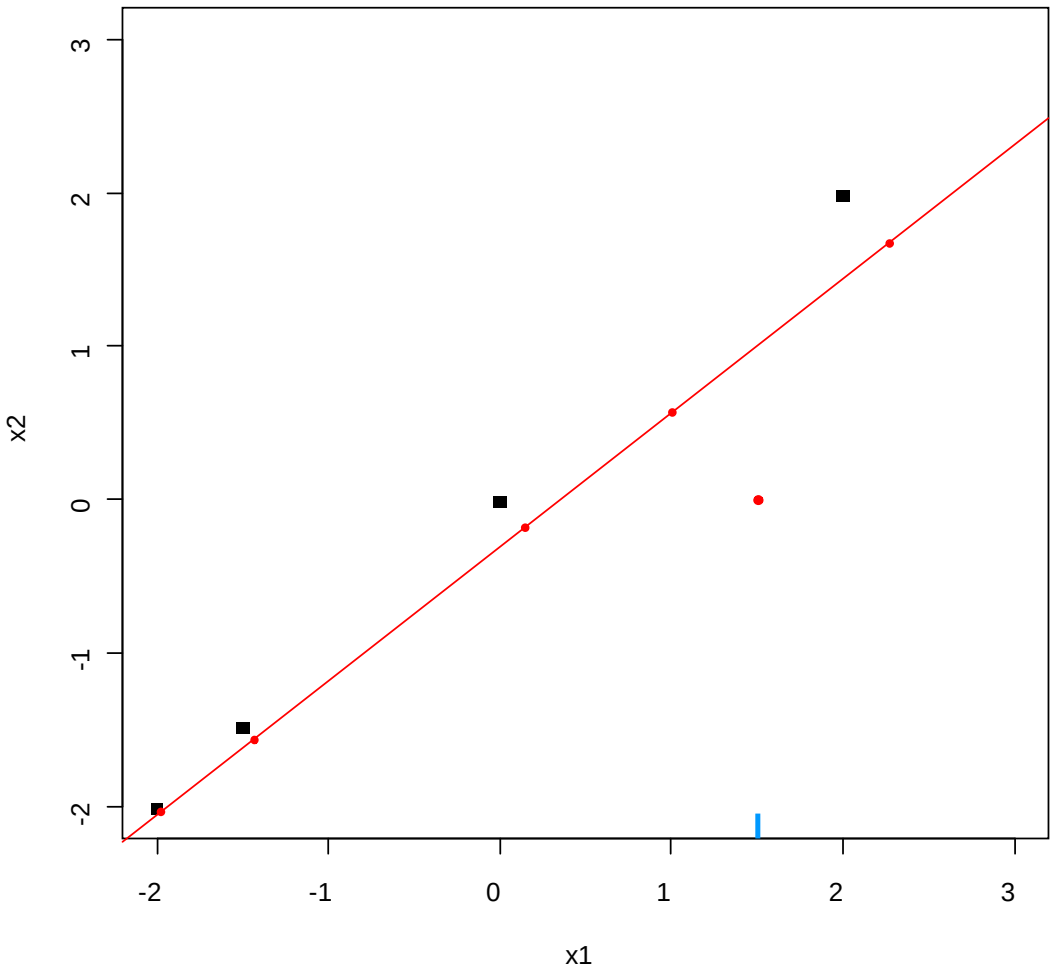
Initialization $\ell = 0$: X^0 (mean imputation)

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



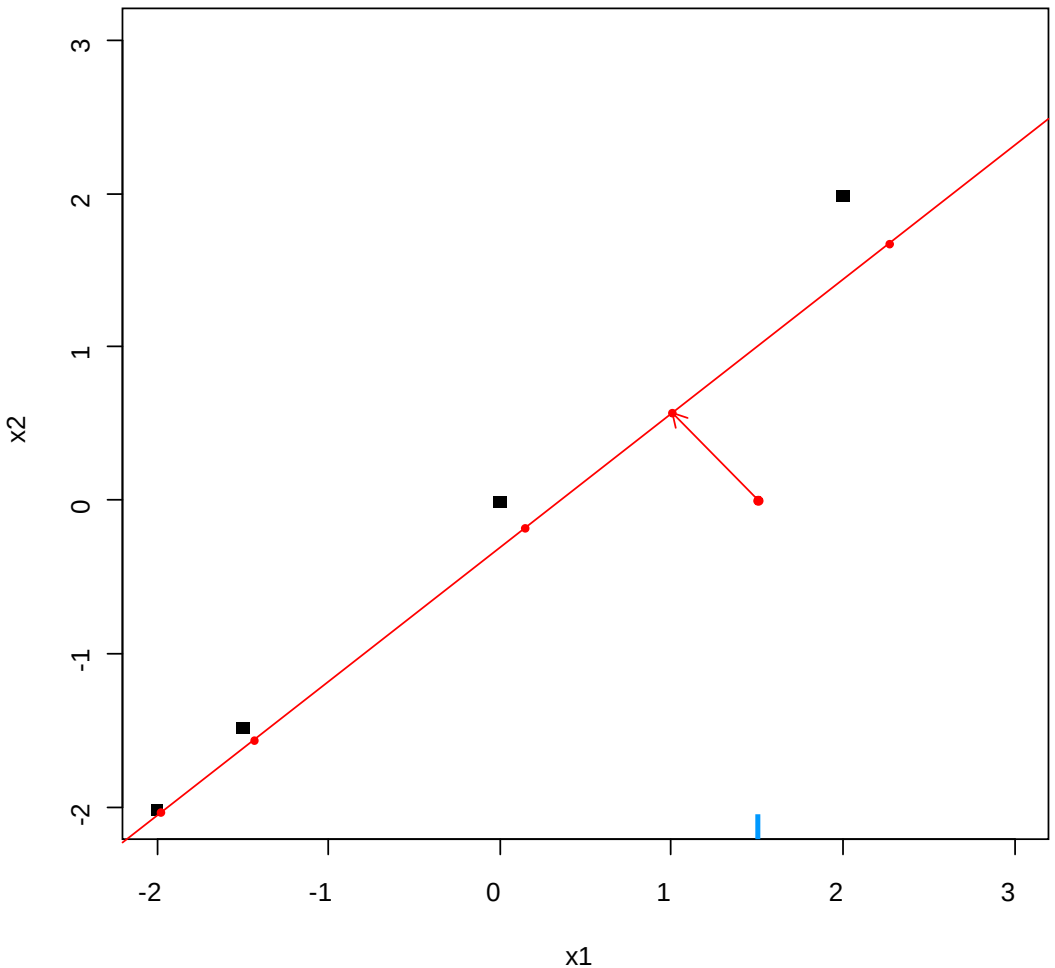
PCA on the completed data set $\rightarrow (U^\ell, \Lambda^\ell, D^\ell)$;

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

$\widehat{x1}$	$\widehat{x2}$
-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 $\hat{\mu}^{\ell} = U^{\ell} D^{\ell} V^{\ell\prime}$

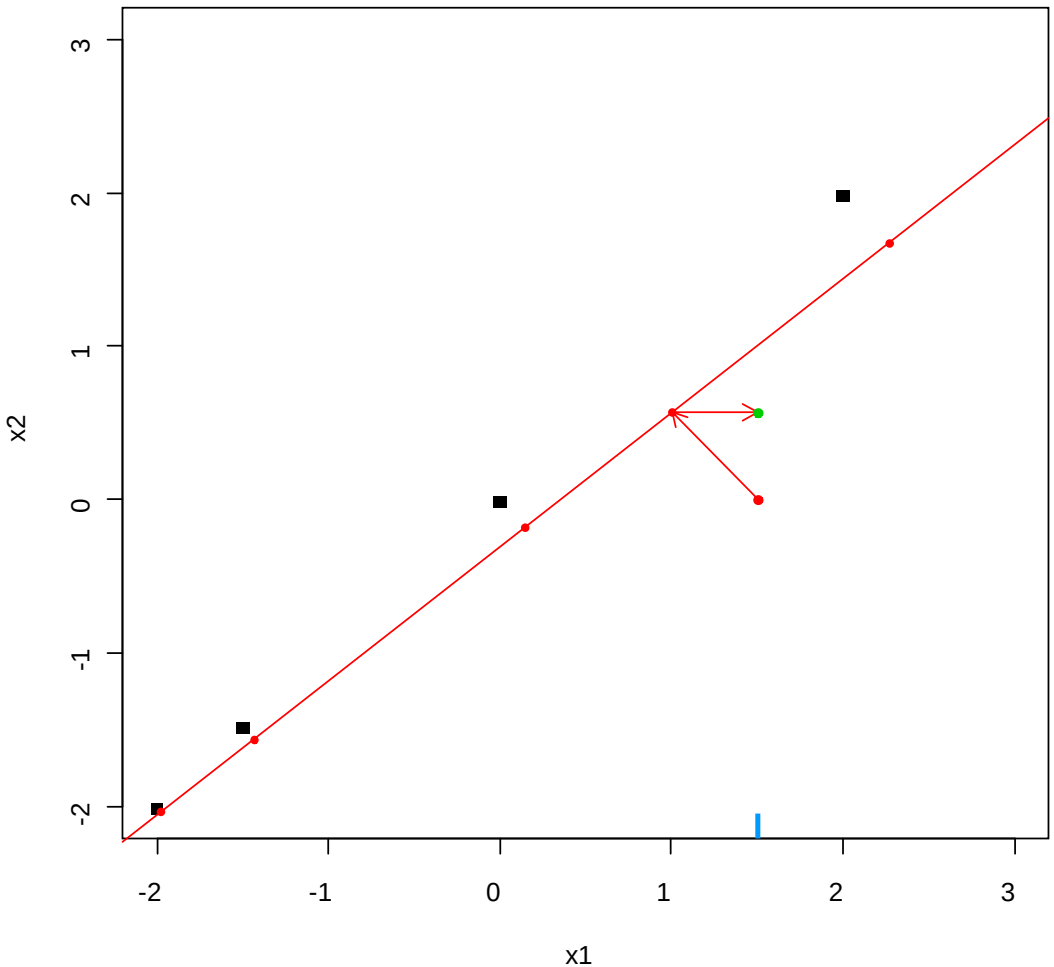
Iterative PCA

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

$\widehat{x1}$	$\widehat{x2}$
-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



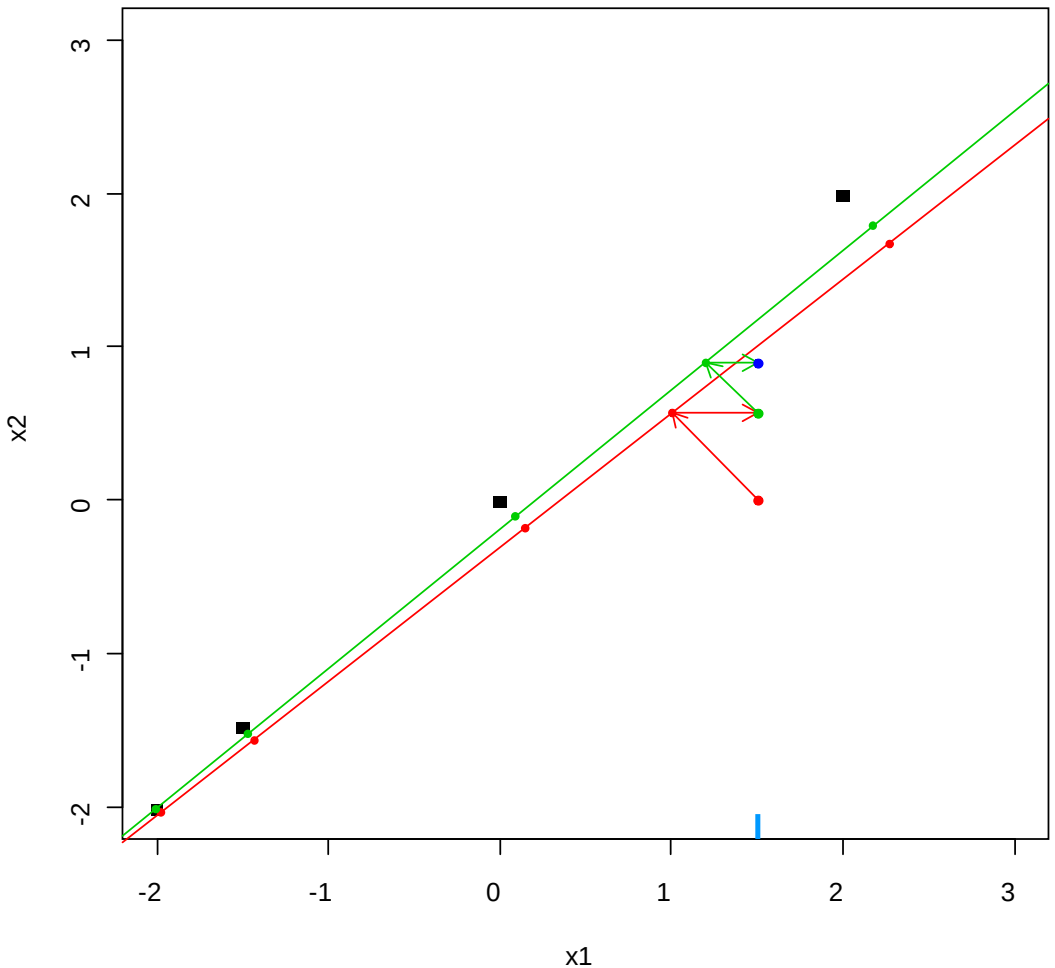
The new imputed dataset is $\hat{X}^\ell = W \odot X + (\mathbf{1} - W) \odot \hat{\mu}^\ell$

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



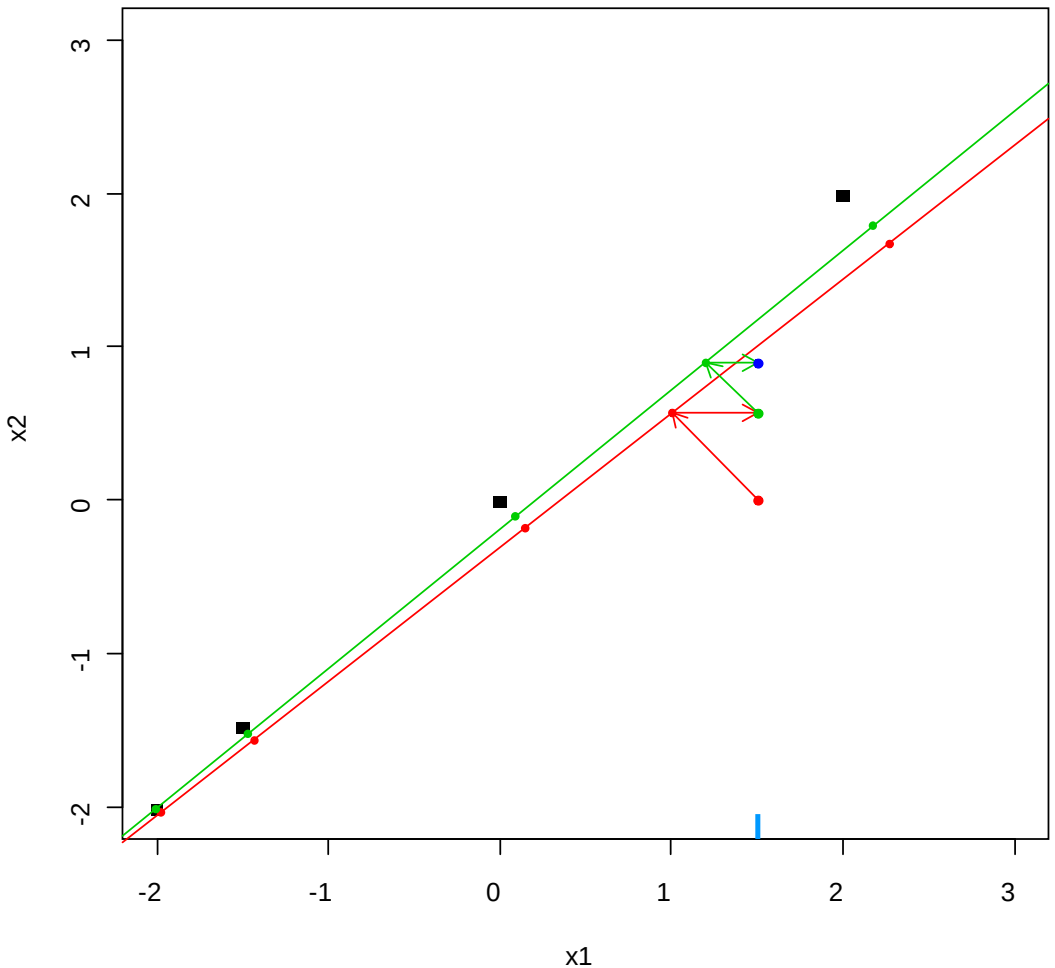
Iterative PCA

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



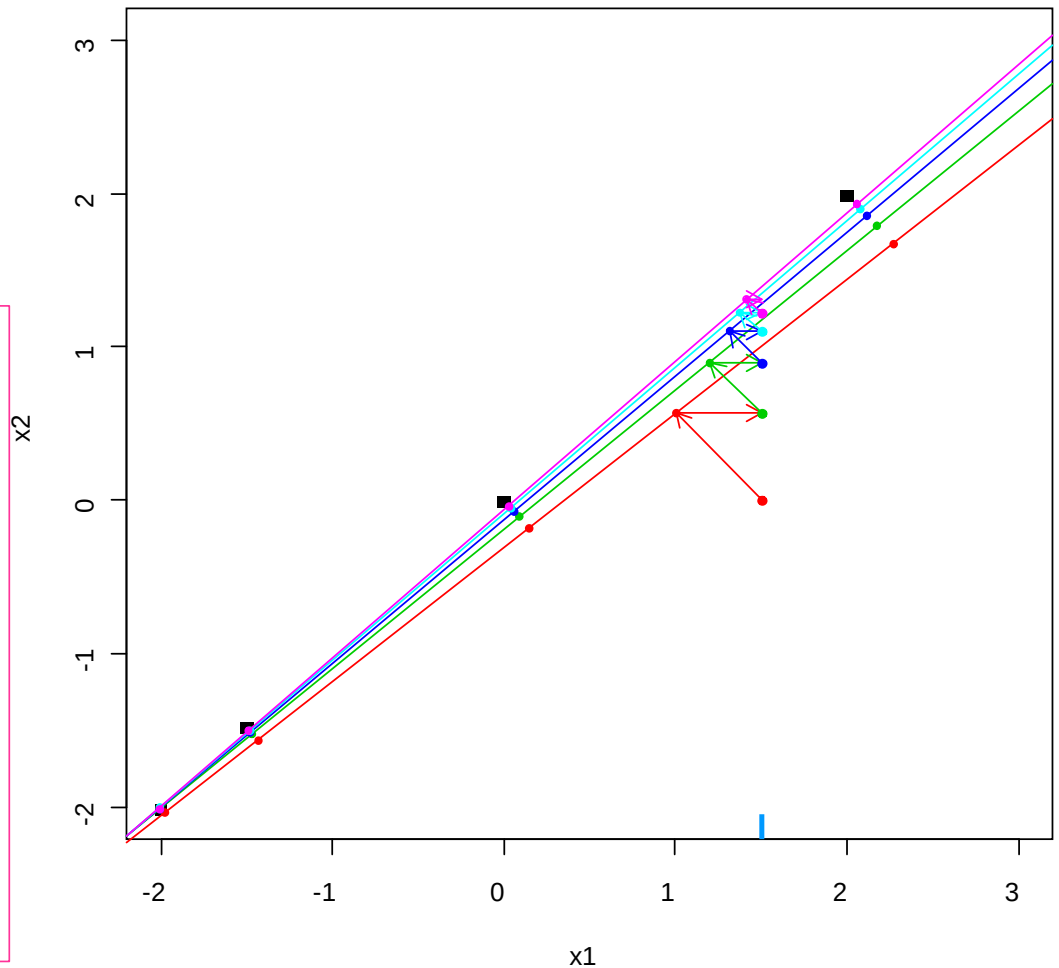
Iterative PCA

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

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

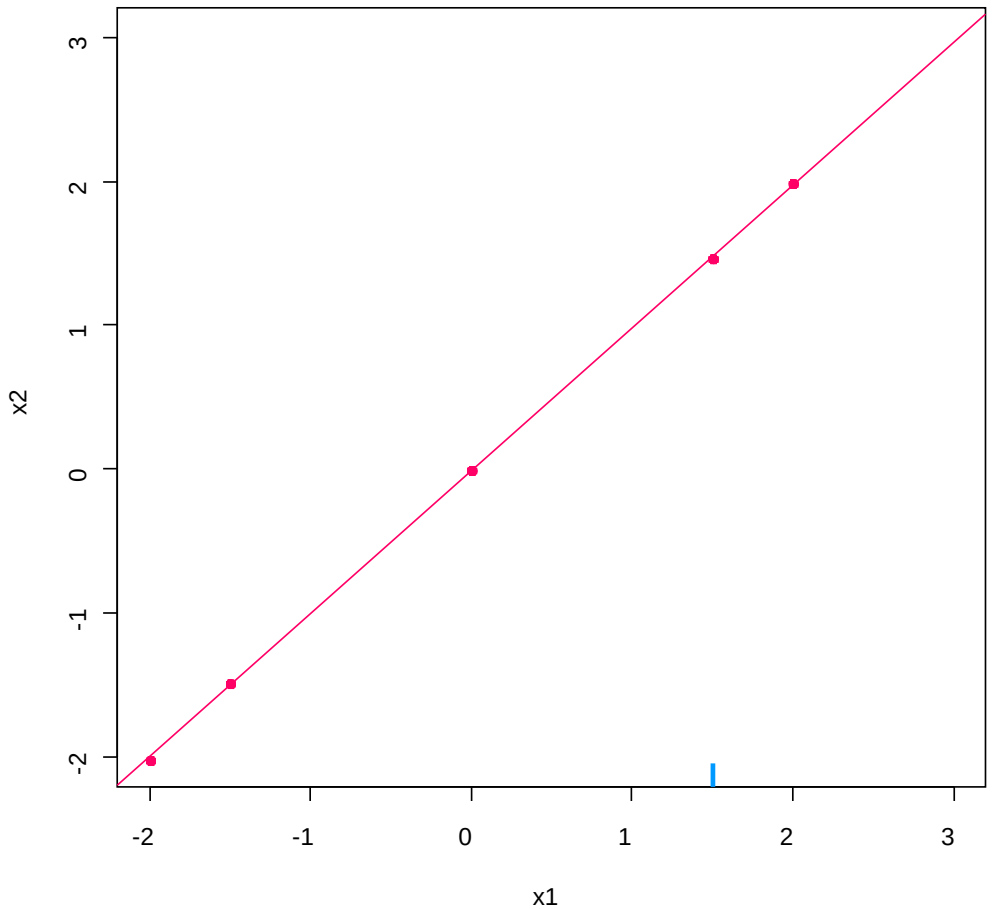
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



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) $(\hat{\mu}^\ell)^k = U^\ell D^\ell V^{\ell\prime}$ $\qquad X^\ell = W \odot X + (1 - W) \odot \hat{\mu}^\ell$
3. steps of estimation and imputation are repeated

$\Rightarrow$ $\hat{\mu}$ from incomplete data: EM algorithm

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

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

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

$\Rightarrow$ Completed data: good imputation (matrix completion, Netflix)
Reduction of variability (imputation by UDV')

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

Soft thresholding 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 $\hat{\mu}^{\text{PCA}} = \sum_{l=1}^{k} d_l u_l v_l^\top$ is replaced by

a "shrunk" impute $(\hat{\mu}^{\text{Soft}})_\lambda = \sum_{l=1}^{\min(n,p)} (d_l - \lambda)_+ u_l v_l^\top$

$$\text{argmin}_\mu \left\{ \|W \odot (X - \mu)\|_2^2 + \lambda \|\mu\|_* \right\}$$

SoftImpute for large matrices. T. Hastie, R. Mazumber, 2015, Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. *JMLR*.

Selecting regularization parameters

$(\hat{\mu}^{\text{Soft}})_\lambda = \sum_{l=1}^{\min(n,p)} (d_l - \lambda)_+ u_l v_l^\top \quad \Rightarrow$ Selecting λ?

Complete: $X = \mu + \varepsilon \; \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \quad \text{MSE}(\lambda) = \mathbb{E}\|\mu - \hat{\mu}_\lambda\|^2$

Stein Unbiased Risk Estimate (Candès *et al.*, 2012) (σ^2 known)
$\text{SURE} = -np\sigma^2 + \sum_k^{\min(n,p)} \min(\lambda^2, d_k) + 2\sigma^2 \text{div}(\hat{\mu}_\lambda)$

Selecting regularization parameters

$(\hat{\mu}^{\text{Soft}})_\lambda = \sum_{l=1}^{\min(n,p)} (d_l - \lambda)_+ u_l v_l^\top \quad \Rightarrow$ Selecting λ?

Complete: $X = \mu + \varepsilon \ \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \quad \text{MSE}(\lambda) = \mathbb{E}\|\mu - \hat{\mu}_\lambda\|^2$

Stein Unbiased Risk Estimate (Candès *et al.*, 2012) (σ^2 known)
$\text{SURE} = -np\sigma^2 + \text{RSS} + 2\sigma^2 \text{div}(\hat{\mu}_\lambda)$

Selecting regularization parameters

$(\hat{\mu}^{\text{Soft}})_\lambda = \sum_{l=1}^{\min(n,p)} (d_l - \lambda)_+ u_l v_l^\top \quad \Rightarrow$ Selecting λ?

Complete: $X = \mu + \varepsilon \; \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \quad \text{MSE}(\lambda) = \mathbb{E}\|\mu - \hat{\mu}_\lambda\|^2$

Stein Unbiased Risk Estimate (Candès *et al.*, 2012) (σ^2 known)
$\text{SURE} = -np\sigma^2 + \text{RSS} + 2\sigma^2 \text{div}(\hat{\mu}_\lambda)$

$\text{GSURE} = \frac{\text{RSS}}{(1 - \text{div}(\hat{\mu}_\lambda)/(np))^2}$ (σ^2 unknown)

Selecting regularization parameters

$(\hat{\mu}^{\text{Soft}})_\lambda = \sum_{l=1}^{\min(n,p)} (d_l - \lambda)_+ u_l v_l^\top \quad \Rightarrow$ Selecting λ?

Complete: $X = \mu + \varepsilon \; \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \quad \text{MSE}(\lambda) = \mathbb{E}\|\mu - \hat{\mu}_\lambda\|^2$

Stein Unbiased Risk Estimate (Candès *et al.*, 2012) (σ^2 known)
$\text{SURE} = -np\sigma^2 + \text{RSS} + 2\sigma^2 \text{div}(\hat{\mu}_\lambda)$

$\text{GSURE} = \frac{\text{RSS}}{(1 - \text{div}(\hat{\mu}_\lambda)/(np))^2}$ (σ^2 unknown)

Incomplete:

$$\text{SURE}^{\text{miss}} = -(np - |NA|)\sigma^2 + \sum_{ij \in obs} (X_{ij} - (\hat{\mu}_{ij})_\lambda^{\text{miss}})^2 + 2\sigma^2 \text{div}^{\text{miss}}(\hat{\mu}_\lambda^{\text{miss}})$$

$$\text{div}^{\text{miss}}(\hat{\mu}_\lambda^{\text{miss}}) = \sum_{ij \in obs} \frac{\hat{\mu}_\lambda^{\text{miss}}(X_{ij} + \delta) - \hat{\mu}_\lambda^{\text{miss}}(X_{ij})}{\delta}$$

Low rank matrix estimation, non linear shrinkage

$$X = \mu + \varepsilon, \text{ with } \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \; \hat{\mu}^{\text{shrink}} = \sum_{l=1}^{\min\{n,p\}} \psi(d_l) U_l V_l^\top$$

- Gavish & Donoho (2014). Asymptotic $n = n_p$, $p \to \infty$, $n_p/p \to \beta$, $0 < \beta \leq 1$

$$\psi(d_l) = \frac{1}{d_l} \sqrt{\left(d_l^2 - (\beta - 1)n\sigma^2\right)^2 - 4\beta n\sigma^4} \cdot 1\left(l \geq (1 + \sqrt{\beta})n\sigma^2\right)$$

- Verbank, Josse & Husson (2013) Asymptotic n, p fixed, $\sigma \to 0$

$$\hat{\mu}_{ij} = \sum_{l=1}^{k} \left(\frac{d_l^2 - \hat{\sigma}^2}{d_l^2}\right) d_l U_{il} V_{jl} = \sum_{l=1}^{k} \left(d_l - \frac{\hat{\sigma}^2}{d_l}\right) U_{il} V_{jl}$$

- Josse & Wager (2015). Bootstrap approach.
- Josse & Sardy (2014). Adaptive estimator $\hat{\mu}_{\lambda,\gamma}$. Finite sample.

$$\psi(d_l) = d_l \max\left(1 - \frac{\lambda^\gamma}{d_l^\gamma}, 0\right) \text{argmin}_\mu \left\{\|X - \mu\|_2^2 + \lambda^\gamma \|\mu\|_{*,w}, w_l = 1/d_l^{\gamma-1}\right\}$$

Properties

$\Rightarrow$ Iterative SVD algorithms to impute data

$\Rightarrow$ Very good quality of imputation. Popular in machine learning community with recommandation systems (Netflix: 99% missing).

Model makes sense: data = structure of rank k + noise

$\Rightarrow$ Different noise regime

- low noise: iterative PCA (tuning k: cross-validation, GCV)
- moderate noise: iterative regularized PCA (non-linear transformation, tuning σ)
- high noise (SNR low, k large): soft thresholding (tuning λ, σ)

$\Rightarrow$ Adaptive estimator is very flexible (selection with SURE)

$\Rightarrow$ Implemented in R packages denoiseR (Josse, Wager, Sardy) and missMDA (Josse, Husson)

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	

.....

Imputed Paris Hospitals data

Traumabase: 15000 patients/ 250 variables

	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

Imputed Paris Hospitals data

Traumabase: 15000 patients/ 250 variables

	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

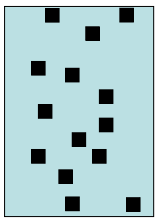
	SpO2	Temperature	Lactates	Hb	Glasgow.....
1	46	61	289.07	33	14
2	2	72	464.00	16	14
3	2	65	416.00	19	7
4	2	74	130.00	36	6
6	2	65	285.91	50	6
7	2	73	244.99	49	6
9	2	83	196.00	65	6
10	2	76	262.44	43	6
11	2	73	84.00	48	5

Imputation based on forests (Stekhoven, Buhlmann, 2011) cannot extrapolate but handle nonlinear relationships.

Outline

Multiple imputation (Rubin, 1987)

Iterative (reg) PCA: single impute with $\hat{\mu}^{\text{shrink}} = \sum_l \psi(d_l) U_l V_l^\top$

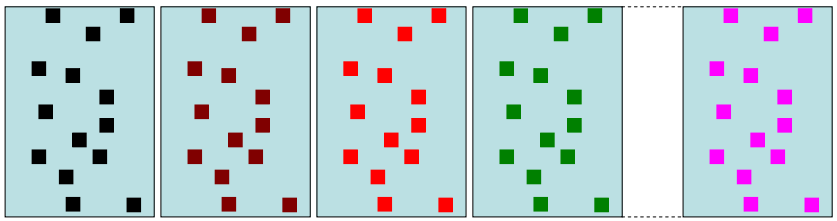


Regression $\hat{\beta}$. Underestimation of std errors: $\widehat{Var}\left(\hat{\beta}\right)$ too small.

$\Rightarrow$ A unique value can't reflect the uncertainty of prediction.

Multiple imputation (Rubin, 1987)

Iterative (reg) PCA: single impute with $\hat{\mu}^{\text{shrink}} = \sum_l \psi(d_l) U_l V_l^\top$



Regression $\hat{\beta}$. Underestimation of std errors: $\widehat{Var}\left(\hat{\beta}\right)$ too small.

$\Rightarrow$ A unique value can't reflect the uncertainty of prediction.

$\Rightarrow$ Multiple imputation: (Ex, one cell: pred 4.88 empirical interval [3.98 ; 5.89])

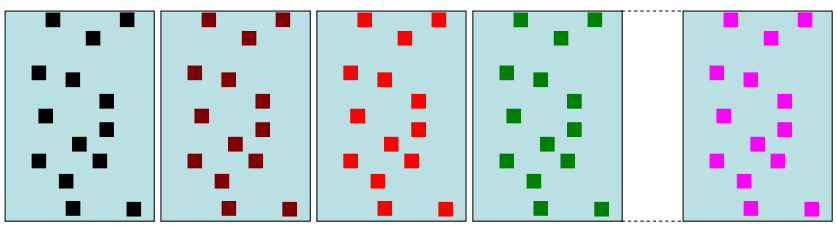
- B plausibles values for each missing entry.
- Perform the analysis on each imputed data $\hat{\beta}_b, \widehat{Var}\left(\hat{\beta}_b\right)$
- Combine: $\hat{\beta} = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b$ $T = \frac{1}{B}\sum_b \widehat{Var}\left(\hat{\beta}_b\right) + \frac{1}{B-1}\sum_b\left(\hat{\beta}_b - \hat{\beta}\right)^2$

PCA multiple imputation (proper)

Iterative (reg) PCA: single impute with $\hat{\mu}^{\text{shrink}} = \sum_l \psi(d_l) U_l V_l^\top$

Model $X = \mu + \varepsilon \; \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

Noise: $b = 1, ..., B$ missing X_{ij}^b drawn from $\mathcal{N}(\hat{\mu}_{ij}, \hat{\sigma}^2)$



PCA multiple imputation (proper)

Iterative (reg) PCA: single impute with $\hat{\mu}^{\text{shrink}} = \sum_l \psi(d_l) U_l V_l^\top$

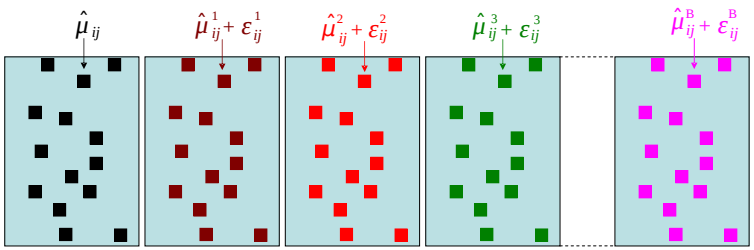
Model $X = \mu + \varepsilon \; \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

1) Variability of parameters: $\hat{\mu}^1 = U^1 D^1 V'^1, .., \hat{\mu}^B = U^B D^B V^{B\prime}$

$\Rightarrow$ nonparametric bootstrap:

- B samples (rows with replacement) with missing: $X^1, ..., X^B$
- Iterative reg PCA on each bootstrap sample: $\hat{\mu}^1, .., \hat{\mu}^B$

2) Noise: $b = 1, ..., B$ missing X_{ij}^b drawn from $\mathcal{N}((\hat{\mu}_{ij})^b, (\hat{\sigma}^2)^b)$



PCA multiple imputation (proper)

Iterative (reg) PCA: single impute with $\hat{\mu}^{\text{shrink}} = \sum_l \psi(d_l) U_l V_l^\top$

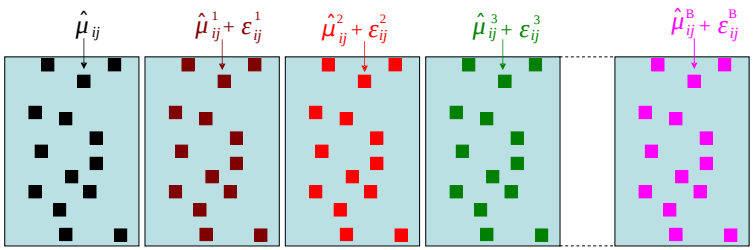
Model $X = \mu + \varepsilon$ $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

1) Variability of parameters: $\hat{\mu}^1 = U^1 D^1 V'^1, .., \hat{\mu}^B = U^B D^B V^{B\prime}$

$\Rightarrow$ nonparametric bootstrap:

- B samples (rows with replacement) with missing: $X^1, ..., X^B$
- Iterative reg PCA on each bootstrap sample: $\hat{\mu}^1, .., \hat{\mu}^B$

2) Noise: $b = 1, ..., B$ missing X_{ij}^b drawn from $\mathcal{N}((\hat{\mu}_{ij})^b, (\hat{\sigma}^2)^b)$



Husson, Josse, Wager (2014). Complete case: variability of PCA estimates: asymptotic, residuals bootstrap, rows bootstrap, a cell-wise jackknife

Audigier, Husson & Josse (2015). Multiple imputation for continuous variables using Bayesian PCA. (data augmentation)

Competitors to get B imputed data

$\Rightarrow$ Joint model: $X_{i.} \sim \mathcal{N}(\mu, \Sigma)$

1. Variability of parameters. Bootstrap rows: X^1, ... , X^B
   EM algorithm: $(\hat{\mu}^1, \hat{\Sigma}^1)$, ... , $(\hat{\mu}^B, \hat{\Sigma}^B)$
2. Noise. Imputation: X_{ij}^b drawn from $\mathcal{N}\left(\hat{\mu}^b, \hat{\Sigma}^b\right)$

$\Rightarrow$ Conditional modeling: one model/variable

1. For a variable j

   2.2 Fit a regression model X_j^{obs} on X_{-j} and impute X_j^{miss} with stochastic regression $\mathcal{N}\left(X_{-j}\hat{\beta}_{-j}, \hat{\sigma}^2_{-j}\right)$

2. Cycling through variables

With continuous variables and a regression/variable: $\mathcal{N}(\mu, \Sigma)$

Competitors to get B imputed data

$\Rightarrow$ Joint model: $X_{i.} \sim \mathcal{N}(\mu, \Sigma)$

1. Variability of parameters. Bootstrap rows: X^1, ... , X^B
   EM algorithm: $(\hat{\mu}^1, \hat{\Sigma}^1)$, ... , $(\hat{\mu}^B, \hat{\Sigma}^B)$
2. Noise. Imputation: X_{ij}^b drawn from $\mathcal{N}\left(\hat{\mu}^b, \hat{\Sigma}^b\right)$

$\Rightarrow$ Conditional modeling: one model/variable

1. For a variable j
   - 2.1 $(\beta_{-j}, \sigma_{-j})$ drawn from a bootstrap (or a posterior distribution)
   - 2.2 Fit a regression model X_j^{obs} on X_{-j} and impute X_j^{miss} with stochastic regression $\mathcal{N}\left(X_{-j}\hat{\beta}_{-j}, \hat{\sigma}^2_{-j}\right)$
2. Cycling through variables - Repeat B times

With continuous variables and a regression/variable: $\mathcal{N}(\mu, \Sigma)$

Competitors to get B imputed data

$\Rightarrow$ Joint model: $X_{i.} \sim \mathcal{N}(\mu, \Sigma)$

1. Variability of parameters. Bootstrap rows: X^1, ... , X^B
   EM algorithm: $(\hat{\mu}^1, \hat{\Sigma}^1)$, ... , $(\hat{\mu}^B, \hat{\Sigma}^B)$
2. Noise. Imputation: X_{ij}^b drawn from $\mathcal{N}\left(\hat{\mu}^b, \hat{\Sigma}^b\right)$

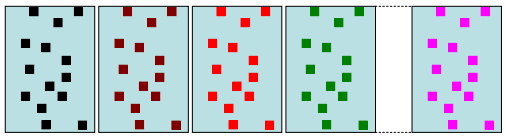
$\Rightarrow$ Conditional modeling: one model/variable

1. For a variable j
   - 2.1 $(\beta_{-j}, \sigma_{-j})$ drawn from a bootstrap (or a posterior distribution)
   - 2.2 Fit a regression model X_j^{obs} on X_{-j} and impute X_j^{miss} with stochastic regression $\mathcal{N}\left(X_{-j}\hat{\beta}_{-j}, \hat{\sigma}^2_{-j}\right)$
2. Cycling through variables - Repeat B times

With continuous variables and a regression/variable: $\mathcal{N}(\mu, \Sigma)$
More flexible? Tedious? (incompatible)

Properties

- Imputation model: B multiple imputed sets with PCA, JM, CM



- Analysis model: $\hat{\theta}_b, \widehat{Var}\left(\hat{\theta}_b\right)$ - combine Rubin's rules: $\hat{\theta}$, T

A regression on each imputed data set

$\Rightarrow$ Aim: inference with missing values

$\Rightarrow$ Good estimates of θ (bias) and coverage ≈ 0.95 (CI width OK): variability due to missing values is taken into account
reflects well the distribution of the data

$\Rightarrow$ PCA: small - large n/p; strong - weak relation; low-high % NA

Outline

Multiple Correspondence Analysis (MCA)

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

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

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: $F_l = \arg\max_{F_l \in \mathbb{R}^n} \frac{1}{m}\sum_{j=1}^{m} \eta^2(F_l, X_j)$

$$\eta^2(F_1, X_j) = \frac{\sum_{c=1}^{C_j}(F_{.c} - F_{..})^2}{\sum_i \sum_c (F_{ic})^2}$$

Benzecri, 1973 :*"All in all, doing a data analysis, in good mathematics, reduces to computing eigenvectors; all the science (or the art) of it is in finding the right matrix to diagonalize"*

Regularized iterative 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 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 MCA

- Initialization: imputation of the indicator matrix (proportions)
- Iterate until convergence
  1. estimation: MCA on the completed data $\rightarrow U, D, V$

	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 MCA

- Initialization: imputation of the indicator matrix (proportions)
- Iterate until convergence
  1. estimation: MCA on the completed data $\rightarrow U, D, V$
  2. imputation with the fitted matrix $\hat{\mu} = U_k D_k V_k'$

	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 MCA

- Initialization: imputation of the indicator matrix (proportions)
- Iterate until convergence
  1. estimation: MCA on the completed data $\rightarrow U, D, V$
  2. imputation with the fitted matrix $\hat{\mu} = U_k D_k V_k'$
  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 MCA

- Initialization: imputation of the indicator matrix (proportions)
- Iterate until convergence
  1. estimation: MCA on the completed data $\rightarrow U, D, V$
  2. imputation with the fitted matrix $\hat{\mu} = U_k D_k V_k'$
  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 MCA

- Initialization: imputation of the indicator matrix (proportions)
- Iterate until convergence
  1. estimation: MCA on the completed data $\rightarrow U, D, V$
  2. imputation with the fitted matrix $\hat{\mu} = U_k D_k V_k'$
  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 MCA

- Initialization: imputation of the indicator matrix (proportions)
- Iterate until convergence
  1. estimation: MCA on the completed data $\rightarrow U, D, V$
  2. imputation with the fitted matrix $\hat{\mu} = U_k D_k V_k'$
  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

Extended mixed data (Audigier, Husson & Josse, 2014) - weighted SVD

Multiple imputation with MCA

1. Variability of the parameters: B sets $(U_{n\times k}, D_{k\times k}, V_{J\times k}^{\top})$ using a non-parametric bootstrap

$\hat{X}_1$

1	0	...	1	0	0
1	0	...	1	0	0
1	0	...			
			0.01	0.80	0.19
			0	0	1
0.25	0.75				
0	1		0	0	1

$\hat{X}_2$

1	0	...	1	0	0
1	0	...	1	0	0
1	0	...			
			0.60	0.2	0.20
			0	0	1
0.26	0.74				
0	1		0	0	1

...

$\hat{X}_B$

1	0	...	1	0
1	0	...	1	0
1	0	...		
			0.11	0.74
			0	0
0.20	0.80			
0	1		0	0

2. Categories drawn from multinomial disribution using the values in $\left(\hat{X}_b\right)_{1\leq b\leq B}$

y	...	Attack
y	...	Attack
y	...	
		Suicide
	...	Accident
n		
n	...	S

y	...	Attack
y	...	Attack
y	...	
		Attack
	...	Accident
n		
n	...	B

...

y	...	Attack
y	...	Attack
y	...	
		Suicide
	...	Accident
n		
n	...	Suicide

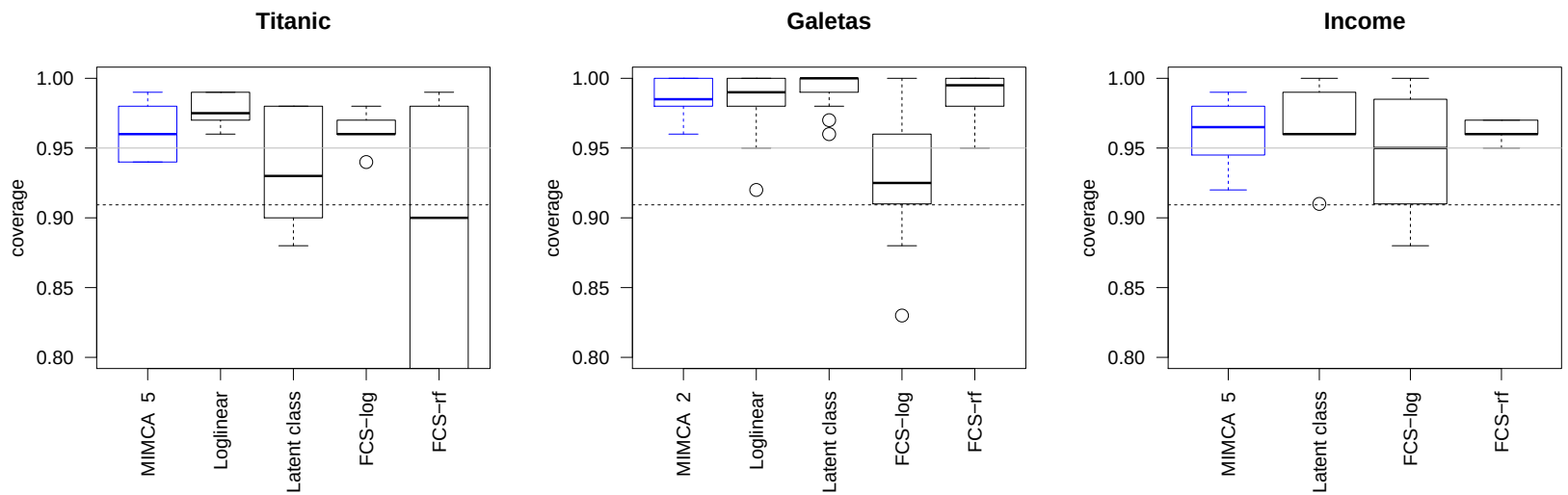
Audigier, Husson & Josse (2016)

Multiple imputation for categorical data

⇒ Joint modeling:

- Log-linear model (Schafer, 1997): pb many levels
- Latent class models (Vermunt, 2014) - nonparametric Bayesian (Si & Reiter, 2014, Murray & Reiter, 2016)

⇒ Conditional model: logistic, multinomial logit, random forests



Coverage of coefficients (2000 - 4 - 4) (1000 - 4 -11)(6000 - 14 - 9)

Results - Inference

⇒ MIMCA provides valid inference (ex. logistic reg with missing)

- reduce the dimensionality
- applied to data of various size (many levels, rare levels)

Time (seconds)	Titanic	Galetas	Income
rows-variables-levels	(2000 - 4 - 4)	(1000 - 4 -11)	(6000 - 14 - 9)
MIMCA	2.750	8.972	**58.729**
Loglinear	0.740	4.597	NA
Nonparametric bayes	10.854	17.414	143.652
Cond logistic	4.781	38.016	881.188
Cond forests	265.771	112.987	6329.514

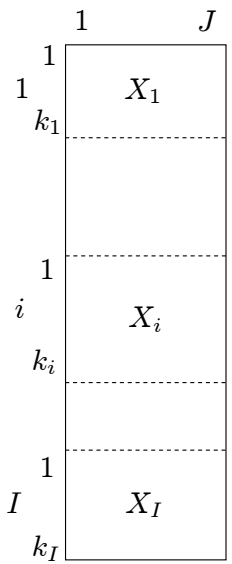
⇒ Model based on relationships between pairs of variables (CA on burt)

⇒ Require to select tuning parameters with missing values (k)

Take home message - On going work

Principal component methods powerful for single & multiple imputation of continuous & categorical data: need a small number of parameters and capture the similarities between rows and relationship between variables.

- Statistical Science issue
- MI for mixed data: extend multilogit model (MCA) (Fithian & J., 2015)
- Selecting variables
- Multilevel/Distributed computation (medical data) (Narasimhan & Robin)



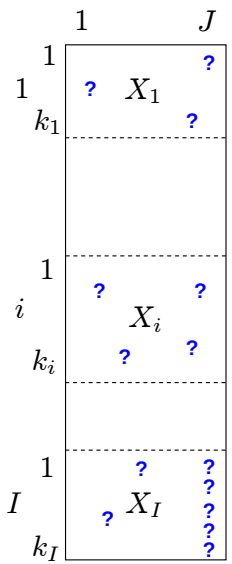
$$X_{i_{(k_i \times J)}} = 1_{k_i} m' + 1_{k_i} F_i^{b'} V^{b'} + F_i^w V^{w'} + E_i$$

1. Between part: Weighted SVD on WX_m; X_m $(I \times J)$ the variables means per group, W $(I \times I)$ $w_{ii} = \sqrt{k_i}$
2. Within part: SVD on the centered data per group X^w $(K \times J)$, $K = \sum_i k_i$

Take home message - On going work

Principal component methods powerful for single & multiple imputation of continuous & categorical data: need a small number of parameters and capture the similarities between rows and relationship between variables.

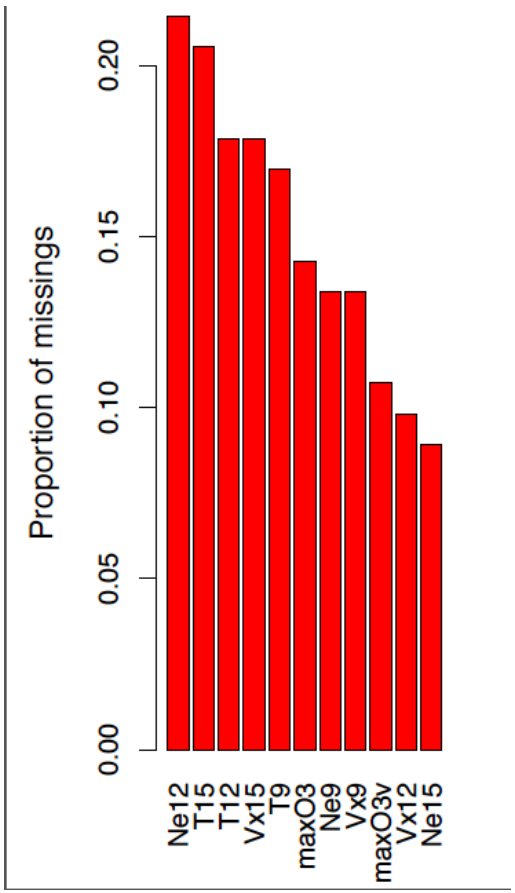
- Statistical Science issue
- MI for mixed data: extend multilogit model (MCA) (Fithian & J., 2015)
- Selecting variables
- Multilevel/Distributed computation (medical data) (Narasimhan & Robin)

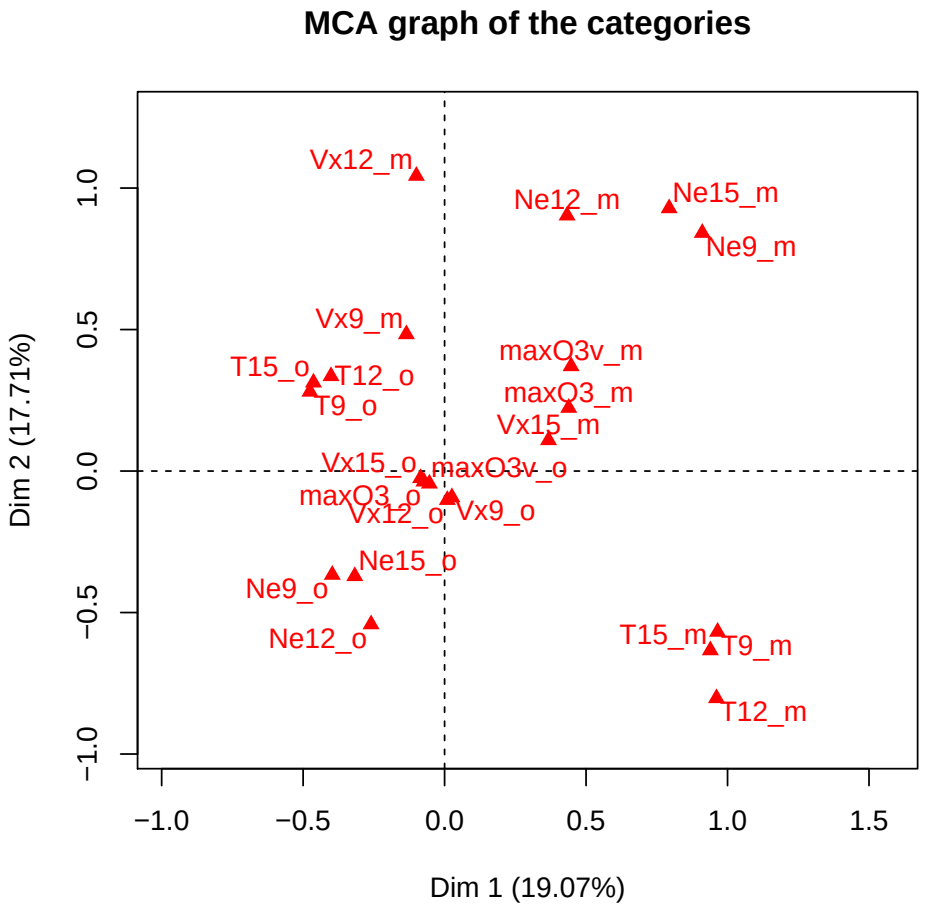


$$X_{i_{(k_i \times J)}} = 1_{k_i} m' + 1_{k_i} F_i^{b'} V^{b'} + F_i^w V^{w'} + E_i$$

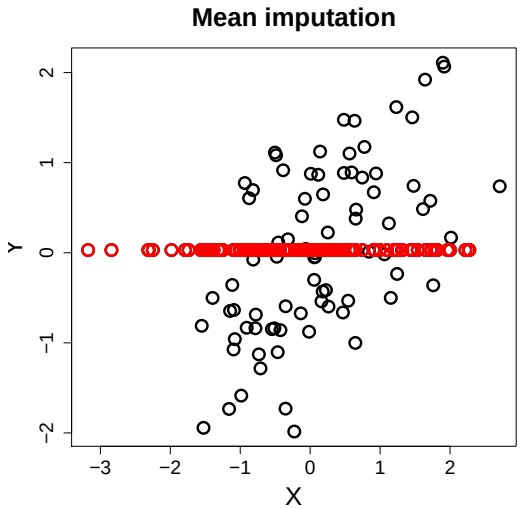
1. Between part: Weighted SVD on WX_m; X_m $(I \times J)$ the variables means per group, W $(I \times I)$ $w_{ii} = \sqrt{k_i}$
2. Within part: SVD on the centered data per group X^w $(K \times J)$, $K = \sum_i k_i$

Visualization with Multiple Correspondence Analysis





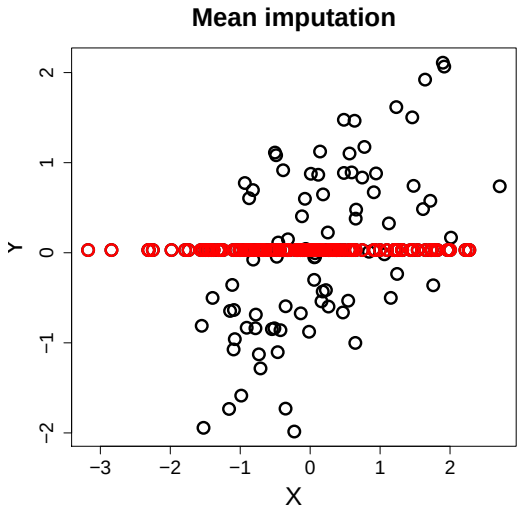
Single imputation methods

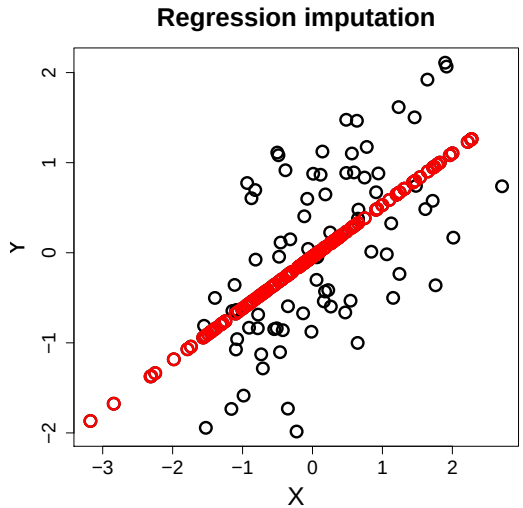


$\mu_y = 0$	0.01
$\sigma_y = 1$	0.5
$\rho = 0.6$	0.30
$CI\mu_y 95\%$	39.4

The idea of imputation is both seductive and dangerous *(Dempster and Rubin, 1983)*

Single imputation methods

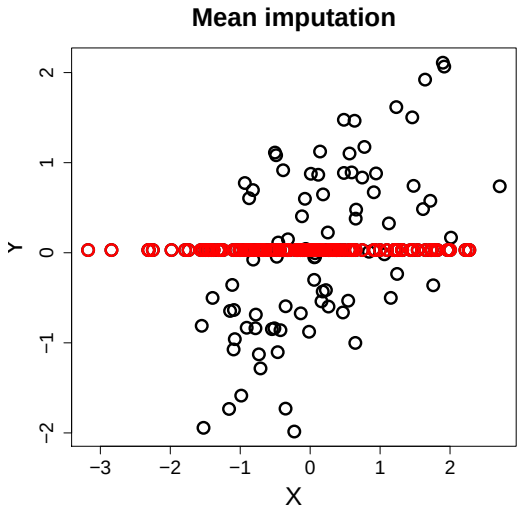


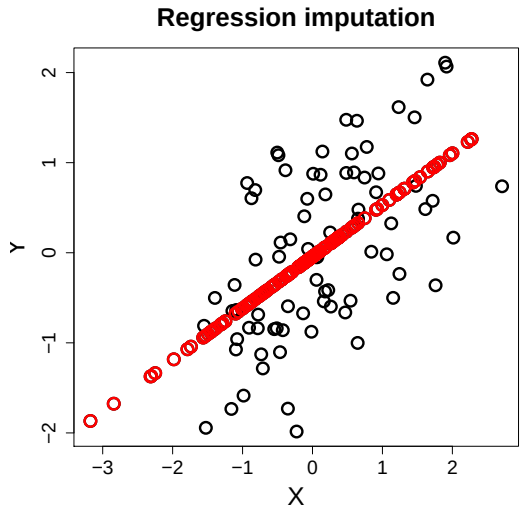


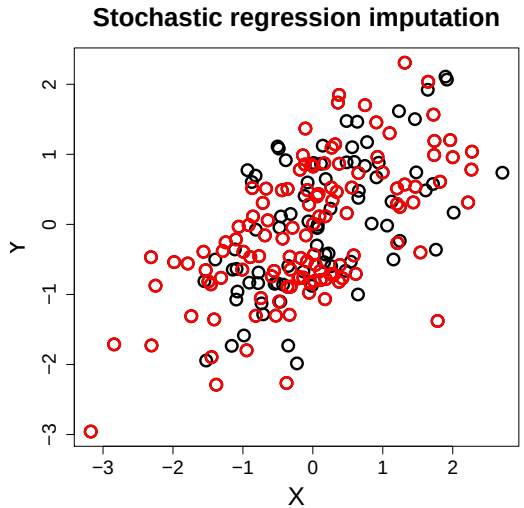
$\mu_y = 0$	0.01	0.01
$\sigma_y = 1$	0.5	0.72
$\rho = 0.6$	0.30	0.78
$CI\mu_y 95\%$	39.4	61.6

The idea of imputation is both seductive and dangerous *(Dempster and Rubin, 1983)*

Single imputation methods



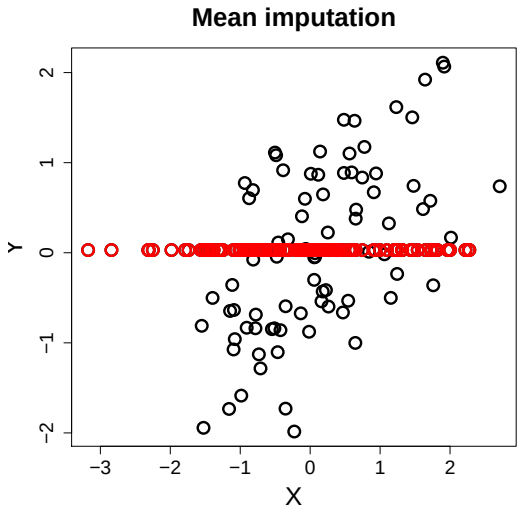


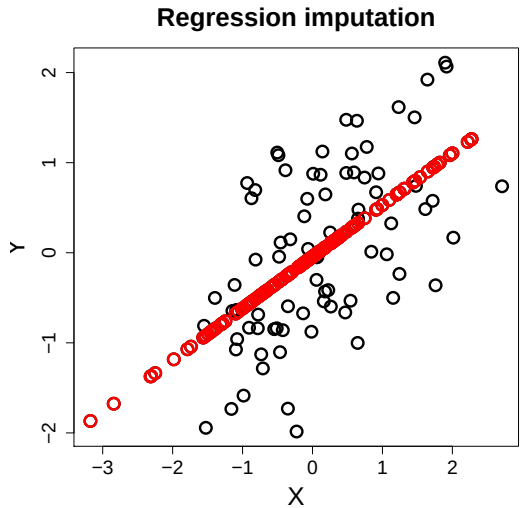


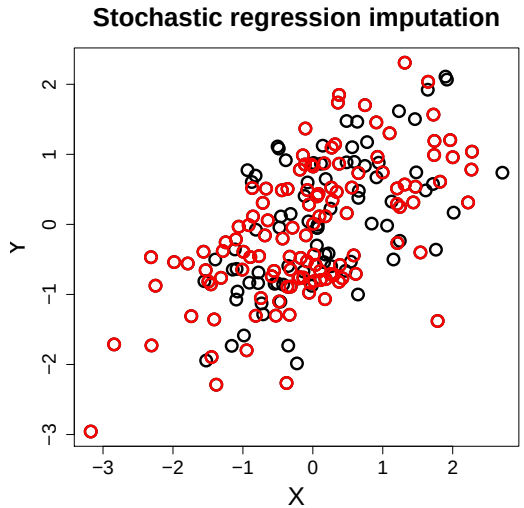
$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59
$CI\mu_y 95\%$	39.4	61.6	70.8

The idea of imputation is both seductive and dangerous *(Dempster and Rubin, 1983)*

Single imputation methods





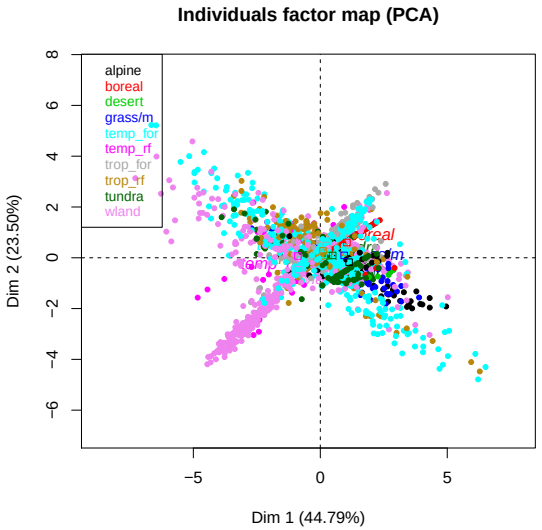


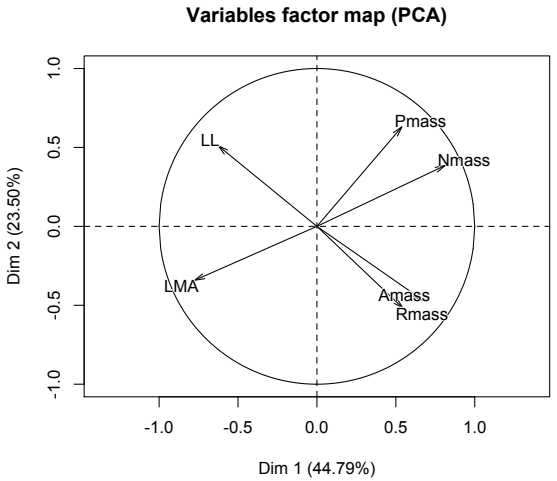
$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59
$CI\mu_y 95\%$	39.4	61.6	70.8

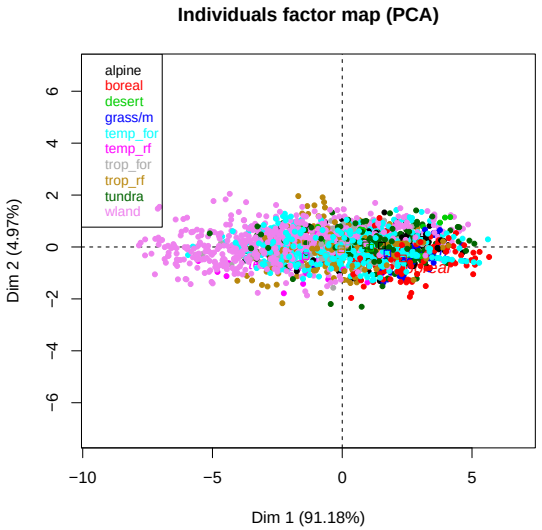
The idea of imputation is both seductive and dangerous *(Dempster and Rubin, 1983)*

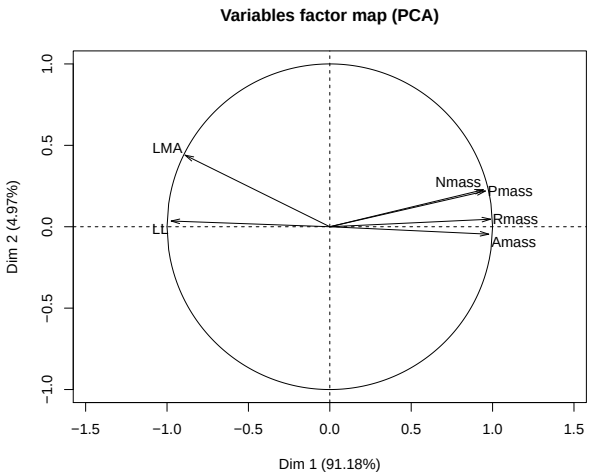
$\Rightarrow$ Standard errors of the parameters ($\hat{\sigma}_{\hat{\mu}_y}$) calculated from the imputed data set are underestimated

Dealing with missing values









Wright IJ, et al. (2004). The worldwide leaf economics spectrum. *Nature*,
69 000 species - LMA (leaf mass per area), LL (leaf lifespan), Amass (photosynthetic assimilation), Nmass (leaf nitrogen), Pmass (leaf phosphorus), Rmass (dark respiration rate) $\Rightarrow$ See codeMissUSER2016

Estimation of σ^2

$\Rightarrow$ Number of independent parameters

$$\hat{\sigma}^2 = \frac{RSS}{\text{ddl}} = \frac{n \sum_{l=k+1}^{p} d_l}{np - p - nk - pk + k^2 + k} \qquad (X_{n\times p}; F_{n\times k}; V_{p\times k})$$

$\Rightarrow$ Trace of the PCA projection matrix

2 projection matrices: $\|X_{n\times p} - F_{n\times k} V'_{k\times p}\|^2$

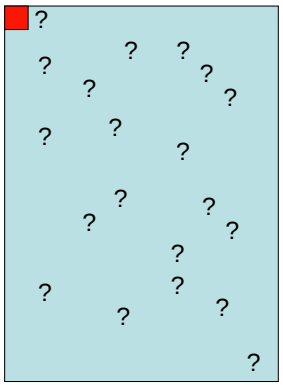
$$\begin{cases} V' = (F'F)^{-1}F'X & \Rightarrow P_F = F(F'F)^{-1}F' \\ F = XV(V'V)^{-1} & \Rightarrow P_V = V(V'V)^{-1}V' \end{cases}$$

$$\hat{\mu}^k = FV' = XP_V = P_F X$$

$$\text{vec}(\hat{\mu}) = P\text{vec}(X) \qquad P_{np\times np} = (P'_V \otimes \mathbb{I}_n) + (\mathbb{I}'_p \otimes P_F) - (P'_V \otimes P_F)$$

Pazman & Denis, 2002; Candes & Tao, 2009

Cross-validation to select S

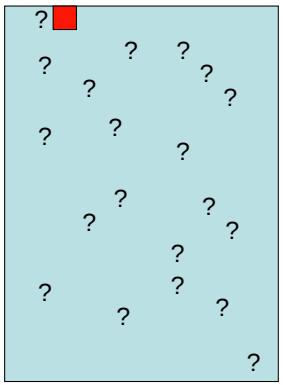


$\Rightarrow$ EM-CV (Bro *et al.* 2008)

$\Rightarrow \text{MSEP}_S = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (X_{ij} - (\hat{\mu}_{ij}^{S})^{-ij})^2$

$\Rightarrow$ Computational costly

Cross-validation to select S

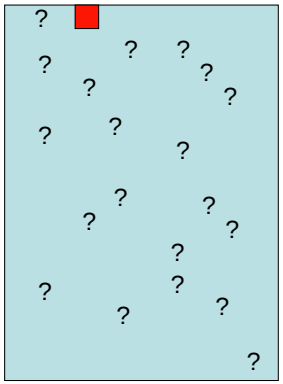


$\Rightarrow$ EM-CV (Bro *et al.* 2008)

$\Rightarrow \text{MSEP}_S = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (X_{ij} - (\hat{\mu}_{ij}^{S})^{-ij})^2$

$\Rightarrow$ Computational costly

Cross-validation to select S

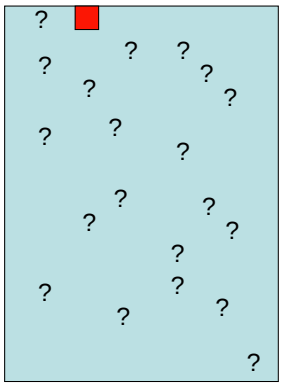


$\Rightarrow$ EM-CV (Bro *et al.* 2008)

$\Rightarrow \text{MSEP}_S = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (X_{ij} - (\hat{\mu}_{ij}^{S})^{-ij})^2$

$\Rightarrow$ Computational costly

Cross-validation to select S



$\Rightarrow$ EM-CV (Bro *et al.* 2008)

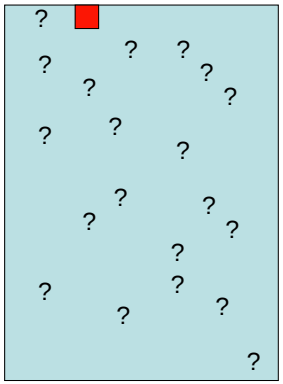
$\Rightarrow$ $\text{MSEP}_S = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(X_{ij} - (\hat{\mu}_{ij}^{S})^{-ij})^2$

$\Rightarrow$ Computational costly

$\Rightarrow$ In regression $\hat{y} = Py$ (Craven & Whaba, 1979)

$$\hat{y}_i^{-i} - y_i = \frac{\hat{y}_i - y_i}{1 - P_{i,i}}$$

Cross-validation to select S



$\Rightarrow$ EM-CV (Bro *et al.* 2008)
$\Rightarrow$ $\text{MSEP}_S = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(X_{ij} - (\hat{\mu}_{ij}^{S})^{-ij})^2$
$\Rightarrow$ Computational costly

$\Rightarrow$ In regression $\hat{y} = Py$ (Craven & Whaba, 1979)

$$\hat{y}_i^{-i} - y_i = \frac{\hat{y}_i - y_i}{1 - P_{i,i}}$$

$\Rightarrow$ Aim: write PCA as $\hat{\mu}^{(S)} = PX$

$$(\hat{\mu}_{ij}^{S})^{-ij} - x_{ij} \simeq \frac{(\hat{\mu}_{ij}^{S}) - X_{ij}}{1 - P_{ij,ij}}$$

PCA as a smoother

2 projection matrices: $\|X_{n\times p} - F_{n\times S} V'_{S\times p}\|_2^2$

$$\begin{cases} V' = (F'F)^{-1}F'X & \Rightarrow P_F = F(F'F)^{-1}F' \\ F = XV(V'V)^{-1} & \Rightarrow P_V = V(V'V)^{-1}V' \end{cases}$$

$$\hat{\mu}^S = FV' = XP_V = P_F X$$

$$\text{vec}(\hat{\mu}^{(S)}) = P^{(S)}\text{vec}(X) \quad P^{(S)}_{np\times np} = (P'_V \otimes \mathbb{I}_n) + (\mathbb{I}'_p \otimes P_F) - (P'_V \otimes P_F)$$

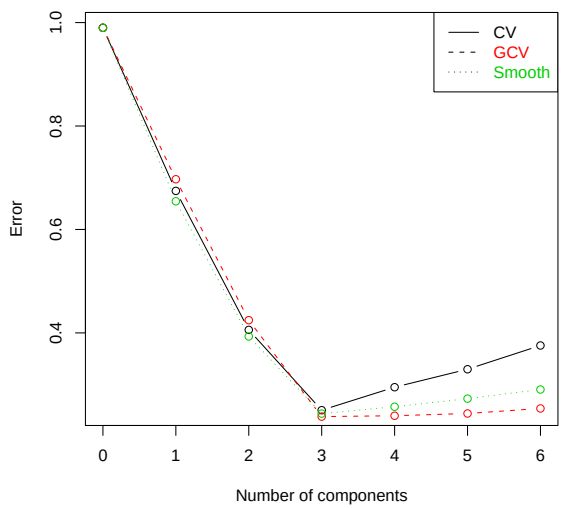
Pazman & Denis, 2002; Candes & Tao, 2009

$\Rightarrow$ Number of independent parameters:

$$\hat{\sigma}^2 = \frac{RSS}{\text{tr}\left(\mathbb{I}_{np} - P^{(S)}\right)} = \frac{n\sum_{s=S+1}^{min(n,p)} \lambda_s}{np - (nS + pS - S^2)}$$

Cross-validation approximations

```
> nb <- estim_ncp(don)
> nb$criterion
        0         1         2         3         4         5
1.2884873 0.8069719 0.6400517 0.7045074 2.2257738 3.0274337
```



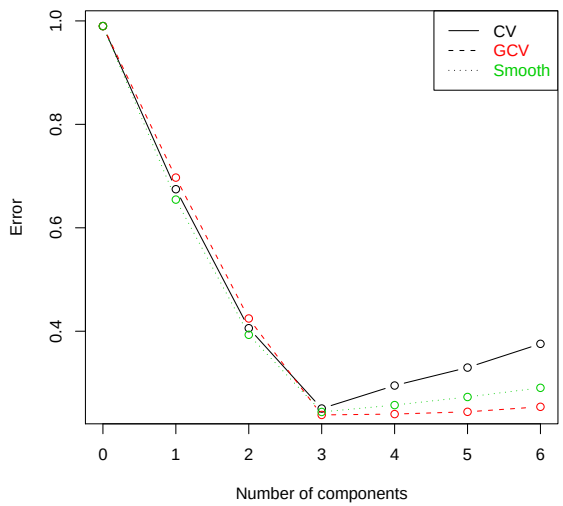
$$CV_S = \frac{1}{np}\sum_{i,j}\left(X_{ij} - (\hat{\mu}_{ij}^S)^{-ij}\right)^2$$

$$ACV_S = \frac{1}{np}\sum_{i,j}\left(\frac{X_{ij} - (\hat{\mu}_{ij}^S)}{1 - P_{ij,ij}}\right)^2$$

Josse, J. & Husson, F. Selecting the number of components in PCA using cross-validation approximations. *Computational Statistics and Data Analysis*.

Cross-validation approximations

```
> nb <- estim_ncp(don)
> nb$criterion
        0         1         2         3         4         5
1.2884873 0.8069719 0.6400517 0.7045074 2.2257738 3.0274337
```



$$\mathrm{CV}_S = \frac{1}{np}\sum_{i,j}\left(X_{ij} - (\hat{\mu}_{ij}^S)^{-ij}\right)^2$$

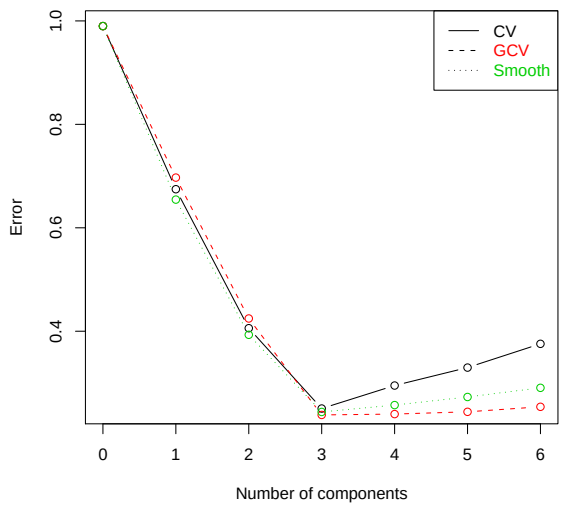
$$\mathrm{ACV}_S = \frac{1}{np}\sum_{i,j}\left(\frac{X_{ij} - (\hat{\mu}_{ij}^S)}{1 - P_{ij,ij}}\right)^2$$

$$\mathrm{GCV}_S = \frac{1}{np} \times \frac{\sum_{i,j}(X_{ij} - \hat{\mu}_{ij}^S))^2}{(1 - \mathrm{tr}(P^{(S)})/np)^2}$$

Josse, J. & Husson, F. Selecting the number of components in PCA using cross-validation approximations. *Computational Statistics and Data Analysis*.

Cross-validation approximations

```
> nb <- estim_ncp(don)
> nb$criterion
        0         1         2         3         4         5
1.2884873 0.8069719 0.6400517 0.7045074 2.2257738 3.0274337
```



$$\text{CV}_S = \frac{1}{np}\sum_{i,j}\left(X_{ij} - (\hat{\mu}_{ij}^S)^{-ij}\right)^2$$

$$\text{ACV}_S = \frac{1}{np}\sum_{i,j}\left(\frac{X_{ij} - (\hat{\mu}_{ij}^S)}{1 - P_{ij,ij}}\right)^2$$

$$\text{GCV}_S = \frac{np\sum_{i,j}(X_{ij} - \hat{\mu}_{ij}^S))^2}{(np - \text{tr}(P^{(S)}))^2}$$

$$\text{GCV NA}_S = \frac{np\|W * (X - \hat{\mu}^S))\|_2^2}{(np - |NA| - (nS + pS - S^2))}$$

Josse, J. & Husson, F. Selecting the number of components in PCA using cross-validation approximations. *Computational Statistics and Data Analysis*.

Iterative Random Forests imputation

1. Initial imputation: mean imputation
2. Fit a RF X_1^{obs} on X_{-1} and then predict X_1^{miss}
   Fit a RF X_2^{obs} on X_{-2} and then predict X_2^{miss}
   ...
   cycling through variables
3. Repeat until convergence

$\Rightarrow$ Conditional modeling based on RF

- number of trees: 100
- number of variables randomly selected at each node $\sqrt{p}$
- number of iterations: 4-5

$\Rightarrow$ Good for complex relationships between variables

Random Forests versus PCA

	Feat1	Feat2	Feat3	Feat4	Feat5...
C1	1	1	1	1	1
C2	1	1	1	1	1
C3	2	2	2	2	2
C4	2	2	2	2	2
C5	3	3	3	3	3
C6	3	3	3	3	3
C7	4	4	4	4	4
C8	4	4	4	4	4
C9	5	5	5	5	5
C10	5	5	5	5	5
C11	6	6	6	6	6
C12	6	6	6	6	6
C13	7	7	7	7	7
C14	7	7	7	7	7
Igor	8	NA	NA	8	8
Frank	8	NA	NA	8	8
Bertrand	9	NA	NA	9	9
Alex	9	NA	NA	9	9
Yohann	10	NA	NA	10	10
Jean	10	NA	NA	10	10

Random Forests versus PCA

	Feat1	Feat2	Feat3	Feat4	Feat5
C1	1	1.0	1.00	1	1
C2	1	1.0	1.00	1	1
C3	2	2.0	2.00	2	2
C4	2	2.0	2.00	2	2
C5	3	3.0	3.00	3	3
C6	3	3.0	3.00	3	3
C7	4	4.0	4.00	4	4
C8	4	4.0	4.00	4	4
C9	5	5.0	5.00	5	5
C10	5	5.0	5.00	5	5
C11	6	6.0	6.00	6	6
C12	6	6.0	6.00	6	6
C13	7	7.0	7.00	7	7
C14	7	7.0	7.00	7	7
Igor	8	6.87	6.87	8	8
Frank	8	6.87	6.87	8	8
Bertrand	9	6.87	6.87	9	9
Alex	9	6.87	6.87	9	9
Yohann	10	6.87	6.87	10	10
Jean	10	6.87	6.87	10	10

	Feat1	Feat2	Feat3	Feat4	Feat5
C1	1	1	1	1	1
C2	1	1	1	1	1
C3	2	2	2	2	2
C4	2	2	2	2	2
C5	3	3	3	3	3
C6	3	3	3	3	3
C7	4	4	4	4	4
C8	4	4	4	4	4
C9	5	5	5	5	5
C10	5	5	5	5	5
C11	6	6	6	6	6
C12	6	6	6	6	6
C13	7	7	7	7	7
C14	7	7	7	7	7
Igor	8	8	8	8	8
Frank	8	8	8	8	8
Bertrand	9	9	9	9	9
Alex	9	9	9	9	9
Yohann	10	10	10	10	10
Jean	10	10	10	10	10

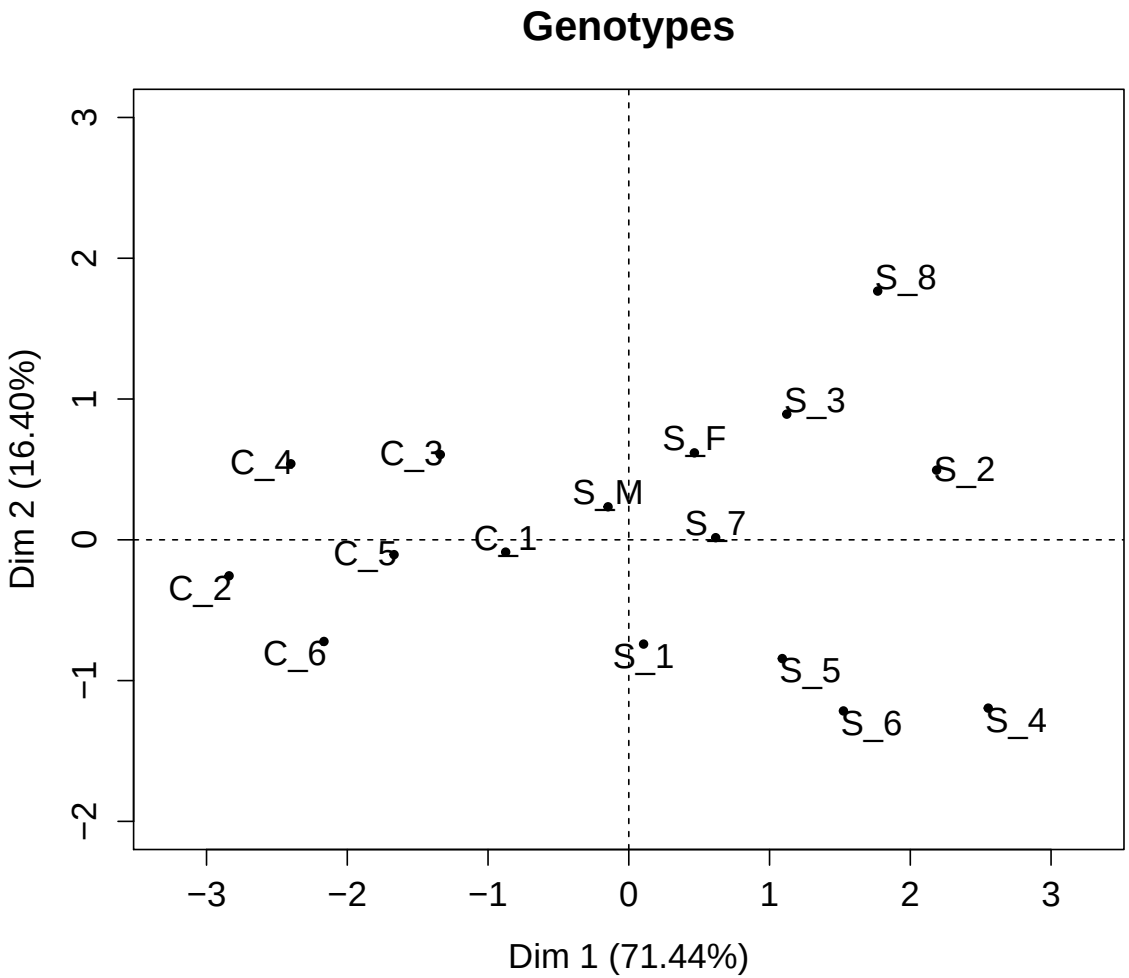
$\Rightarrow$ with Random Forests $\Rightarrow$ with PCA

(Stekhoven, Buhlmann, 2011 - Bartlett, Carpenter, 2014)

$\Rightarrow$ Non linear relationship well handled by forests

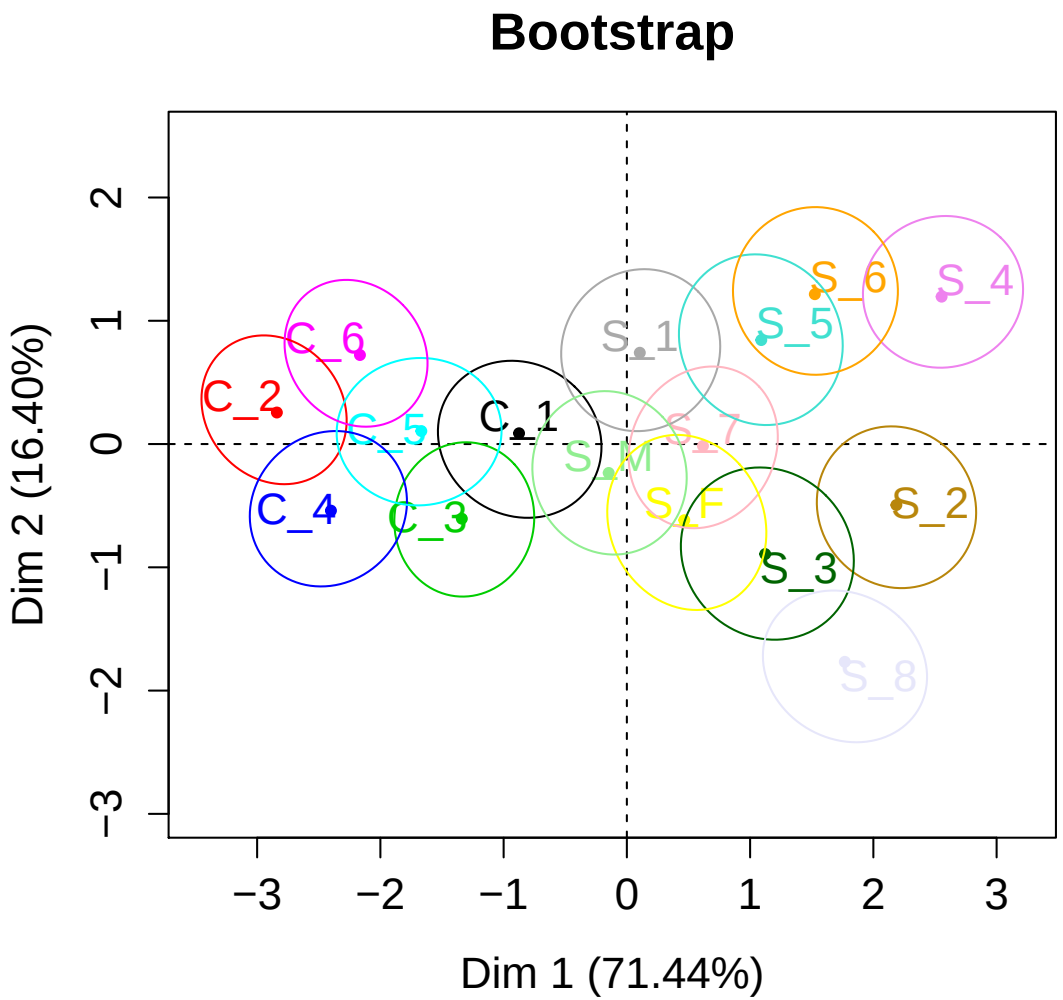
Outline

Inference in PCA



Josse, Husson & Wager (2015)

Inference in PCA



Josse, Husson & Wager (2015)

Confidence ellipses

Fixed-effects model $X_{n\times p} = \mu + \varepsilon$ $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with μ of rk k

MLE = LS $\hat{\mu}_k = U_{n\times k} D_{k\times k} V'_{k\times p}$.

Rq: Population data. (PPCA: random effects)

$\Rightarrow$ Asymptotic $\sigma^2 \to 0$

Confidence ellipses

Fixed-effects model $X_{n\times p} = \mu + \varepsilon$ $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with μ of rk k

MLE = LS $\hat{\mu}_k = U_{n\times k} D_{k\times k} V'_{k\times p}$.

Rq: Population data. (PPCA: random effects)

$\Rightarrow$ Asymptotic $\sigma^2 \to 0$ (Denis, Gower, 1994; Papadopoulo & Louraki 2000)

$$\mathbb{E}(\hat{\mu}_{ij}^k) = \mu_{ij} \qquad \mathbb{V}\left(\hat{\mu}_{ij}^k\right) = \sigma^2 P_{ij,ij}$$

$$\text{vec}(\hat{\mu}_k) = P\text{vec}(X) \; P_{np\times np} = (P'_V \otimes \mathbb{I}_n) + (\mathbb{I}'_p \otimes P_F) - (P'_V \otimes P_F)$$

2 projections: $\|X - FV'\|^2$ $\qquad$ $\hat{\mu}_k = FV' = XP_V = P_F X$

$$\begin{cases} V' = (F'F)^{-1}F'X & \Rightarrow P_F = F(F'F)^{-1}F' \\ F = XV(V'V)^{-1} & \Rightarrow P_V = V(V'V)^{-1}V' \end{cases}$$

Confidence ellipses

Fixed-effects model $X_{n\times p} = \mu + \varepsilon \ \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with μ of rk k

MLE = LS $\hat{\mu}_k = U_{n\times k} D_{k\times k} V'_{k\times p}$.

Rq: Population data. (PPCA: random effects)

$\Rightarrow$ Asymptotic $\sigma^2 \to 0$ (Denis, Gower, 1994; Papadopoulo & Louraki 2000)

$$\mathbb{E}(\hat{\mu}_{ij}^k) = \mu_{ij} \qquad \mathbb{V}\left(\hat{\mu}_{ij}^k\right) = \sigma^2 P_{ij,ij}$$

$\text{vec}(\hat{\mu}_k) = P\text{vec}(X)$ $P_{np\times np} = (P'_V \otimes \mathbb{I}_n) + (\mathbb{I}'_p \otimes P_F) - (P'_V \otimes P_F)$

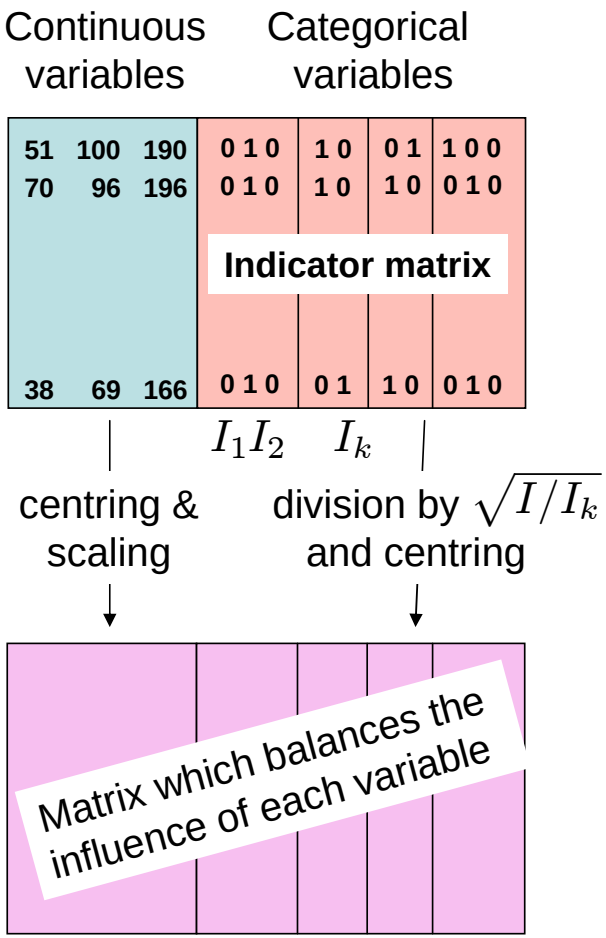
2 projections: $\|X - FV'\|^2$ $\qquad$ $\hat{\mu}_k = FV' = XP_V = P_F X$

$$\begin{cases} V' = (F'F)^{-1}F'X & \Rightarrow P_F = F(F'F)^{-1}F' \\ F = XV(V'V)^{-1} & \Rightarrow P_V = V(V'V)^{-1}V' \end{cases}$$

$\Rightarrow$ Draw $(\hat{\mu}^1, ..., \hat{\mu}^B)$

Principal component method for mixed data (complete)

Factorial Analysis Mixed Data FAMD (Escofier, 1979), PCAMIX (Kiers, 1991)



A PCA is performed on the weighted matrix: SVD $(X, D_\Sigma^{-1}, \frac{1}{I} I_I)$, with X the matrix with the continuous variables and the indicator matrix, D_Σ, the diagonal matrix with the standard deviation and the weights $\sqrt{(I_S/I)}$.

Properties of FAMD (complete)

Benzecri, 1973 :*"All in all, doing a data analysis, in good mathematics, is simply searching eigenvectors; all the science (or the art) of it is just to find the right matrix to diagonalize"*

- The distance between observations is:

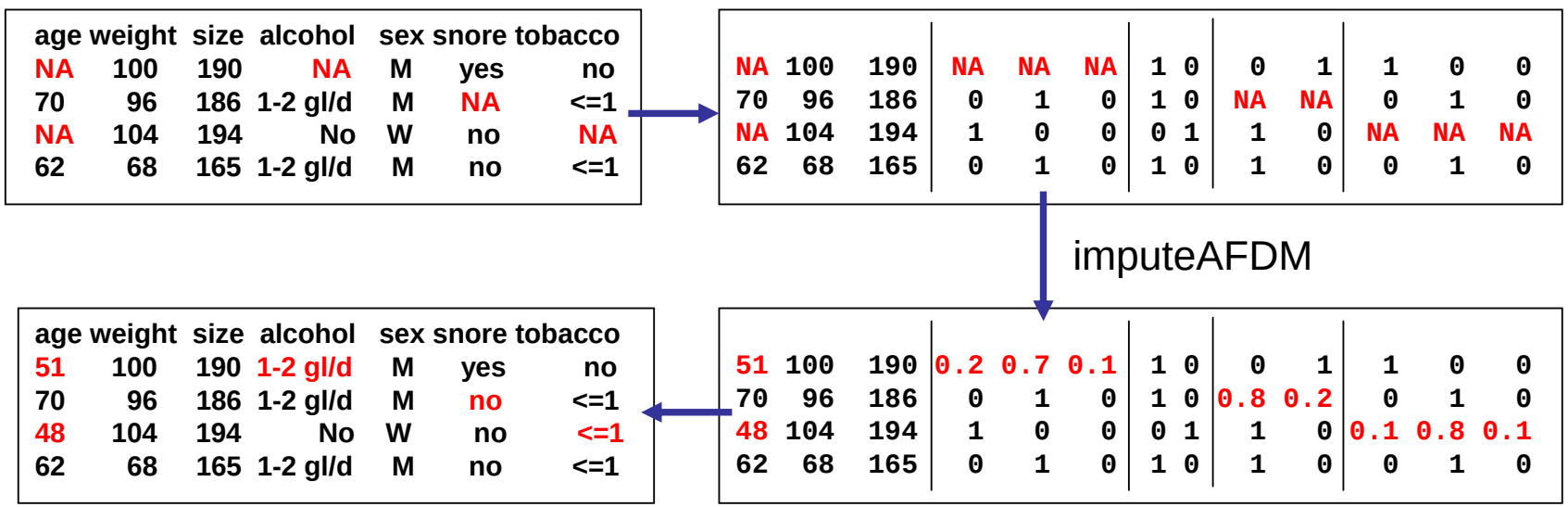
$$d^2(i,l) = \sum_{k=1}^{K_{cont}} \frac{1}{\sigma_S}(x_{ik} - x_{lk})^2 + \sum_{q=1}^{Q} \sum_{k=1}^{K_q} \frac{1}{l_{k_q}}(x_{iq} - x_{lq})^2$$

- The principal component F_s maximises:

$$\sum_{k=1}^{K_{cont}} r^2(F_s, v_S) + \sum_{q=1}^{Q_{cat}} \eta^2(F_s, v_q)$$

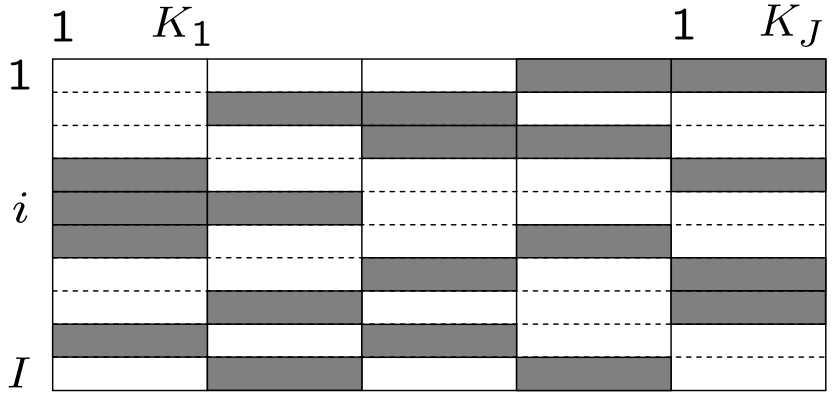
Iterative FAMD algorithm

1. Initialization: imputation mean (continuous) and proportion (dummy)
2. Iterate until convergence
   - (a) estimation: FAMD on the completed data $\Rightarrow U, \Lambda, V$
   - (b) imputation of the missing values with the fitted matrix $\hat{X} = U_S \Lambda_S^{1/2} V_S'$
   - (c) means, standard deviations and column margins are updated



$\Rightarrow$ Imputed values can be seen as degrees of membership

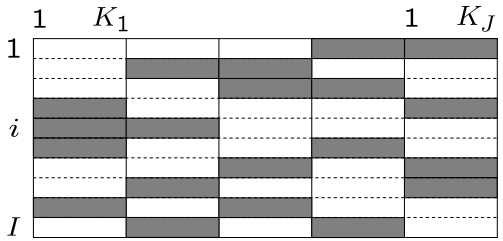
Multi-blocks data set



L'OREAL: 100 000 women in different countries - 300 questions

- Self-assessment questionnaire: life style, skin and hair characteristics, care and consumer habits
- Clinical assessments by a dermatologist: facial skin complexion, wrinkles, scalp dryness, greasiness
- Hair assessments by a hair dresser: abundance, volume, breakage, curliness
- Skin and Hair photographs and measurements: sebum quantity, etc.

Multi-blocks data set



- Sensory analysis: products described by people and by physico-chemical measurements

  (each judge can't taste more than 8 products: Planned missing products per judge, experimental design: BIB)

- Biology. DNA/RNA (samples without expression data)

Continuous / categorical / contingency sets of variables

$\Rightarrow$ Missing rows per subtable

$\Rightarrow$ Regularized iterative Multiple Factor Analysis (Husson & Josse, 2013)

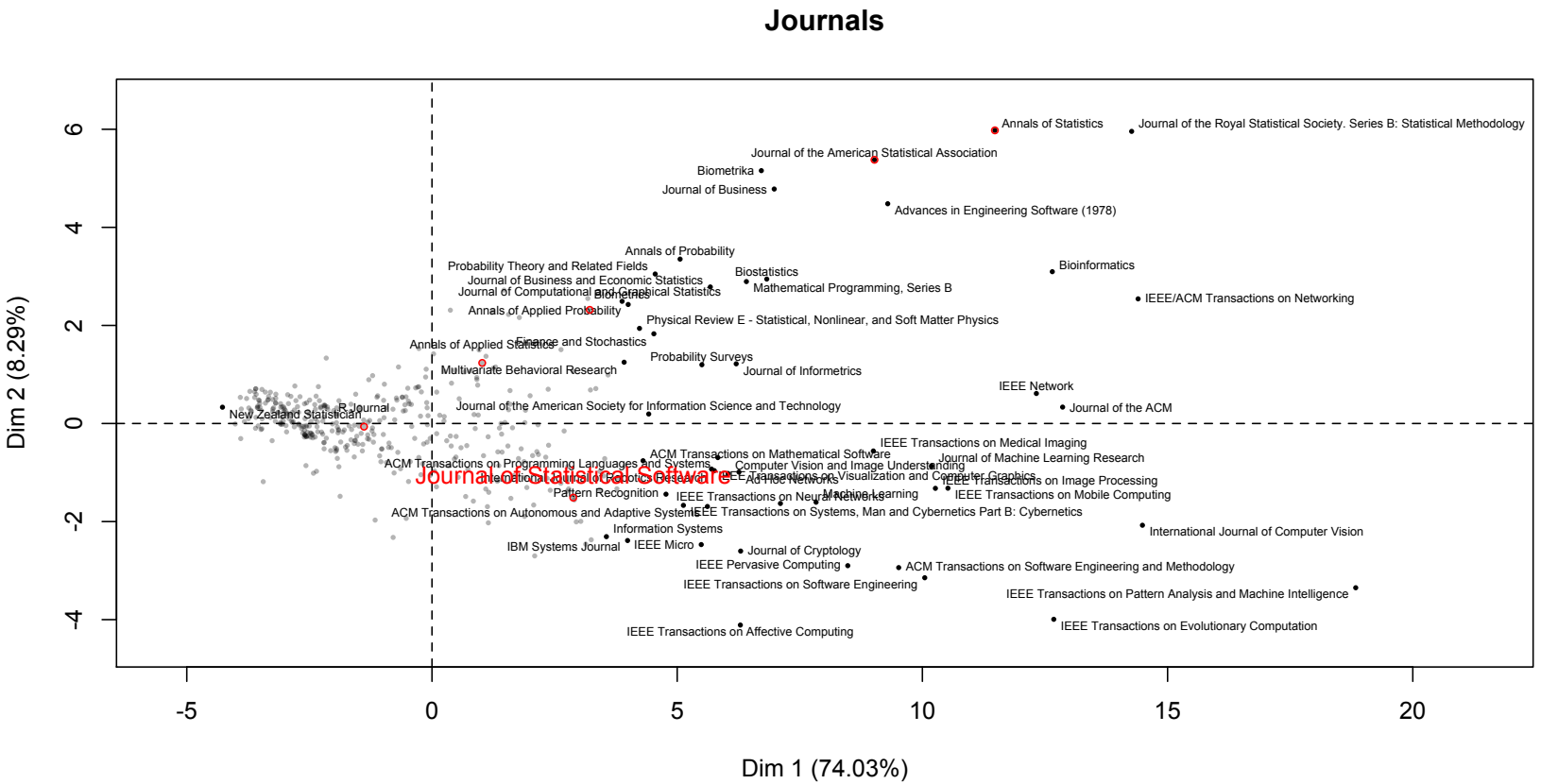
Journal impact factors

journalmetrics.com provides 27000 journals/ 15 years of metrics.

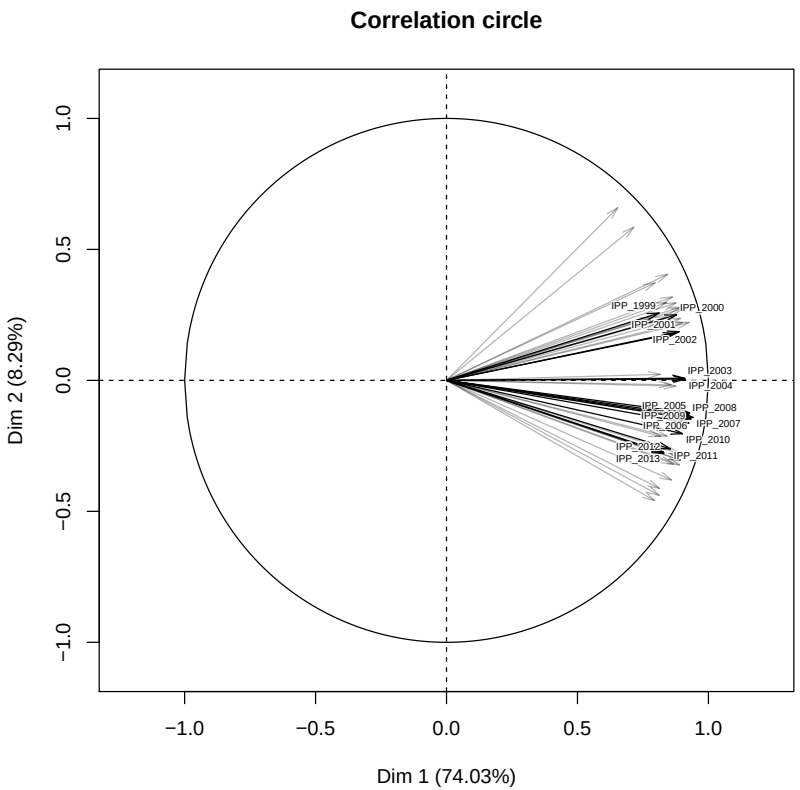
443 journals (Computer Science, Statistics, Probability and Mathematics). 45 metrics, some may be NA, 15 years by 3 types of measures:

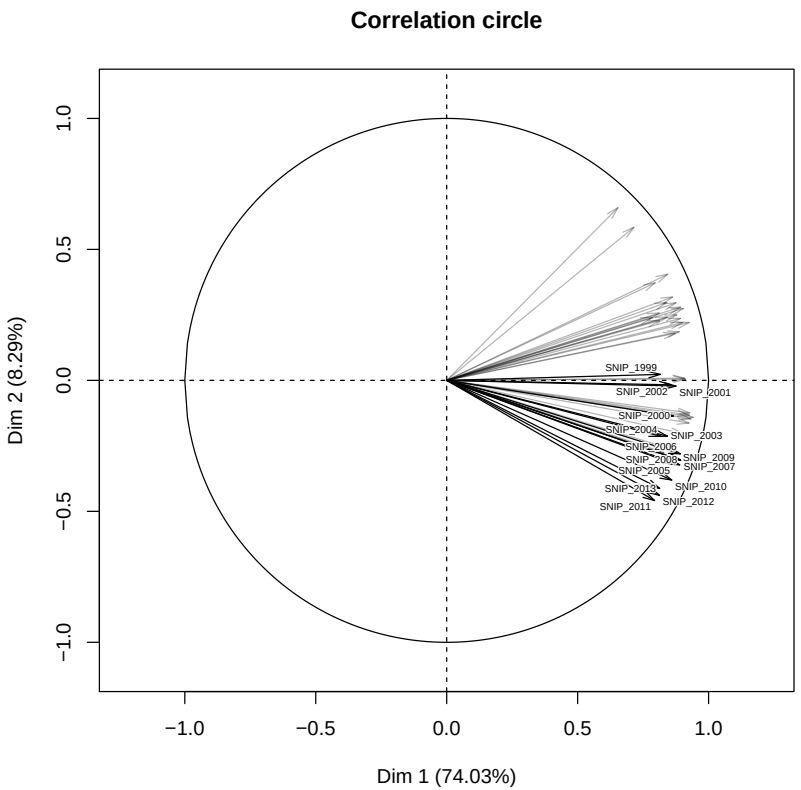
- IPP - Impact Per Publication (like the ISI impact factor but for 3 (rather than 2) years.
- SNIP - Source Normalized Impact Per Paper: Tries to weight by the number of citations per subject field to adjust for different citation cultures.
- SJR - SCImago Journal Rank: Tries to capture average prestige per publication.

MFA with missing values

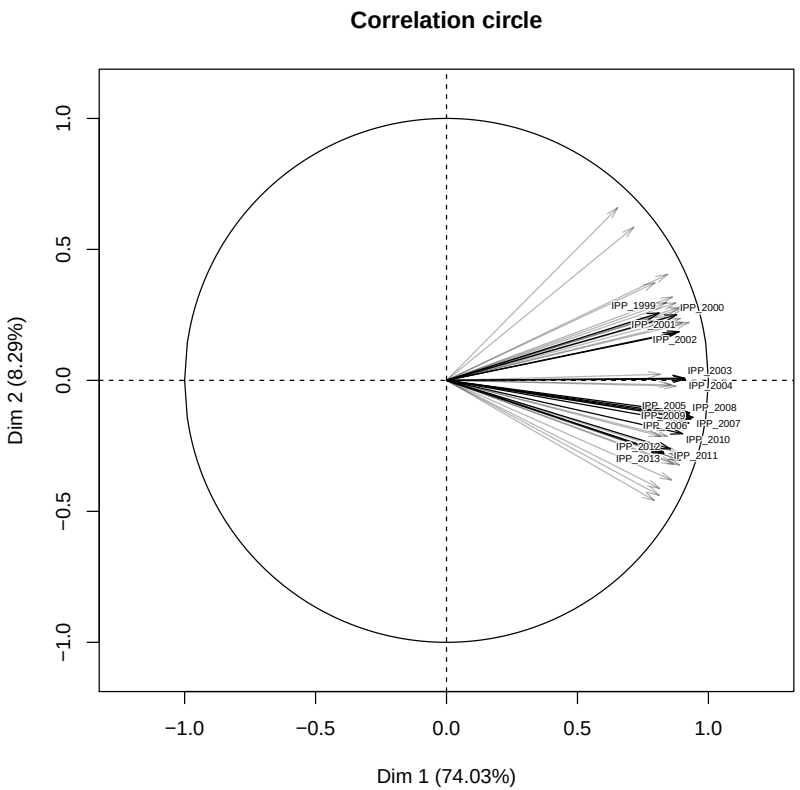


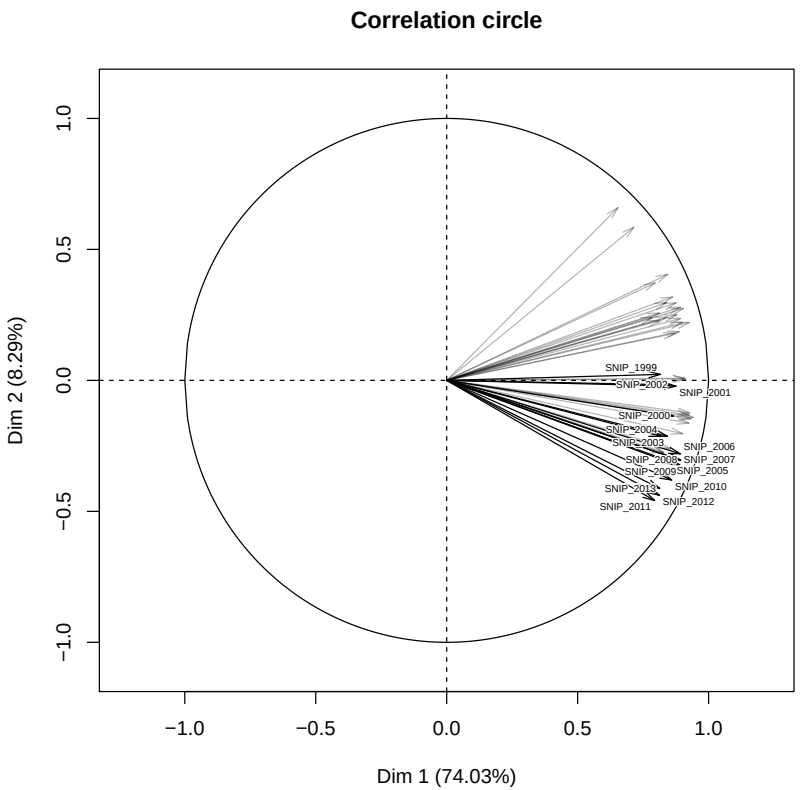
MFA with missing values





MFA with missing values

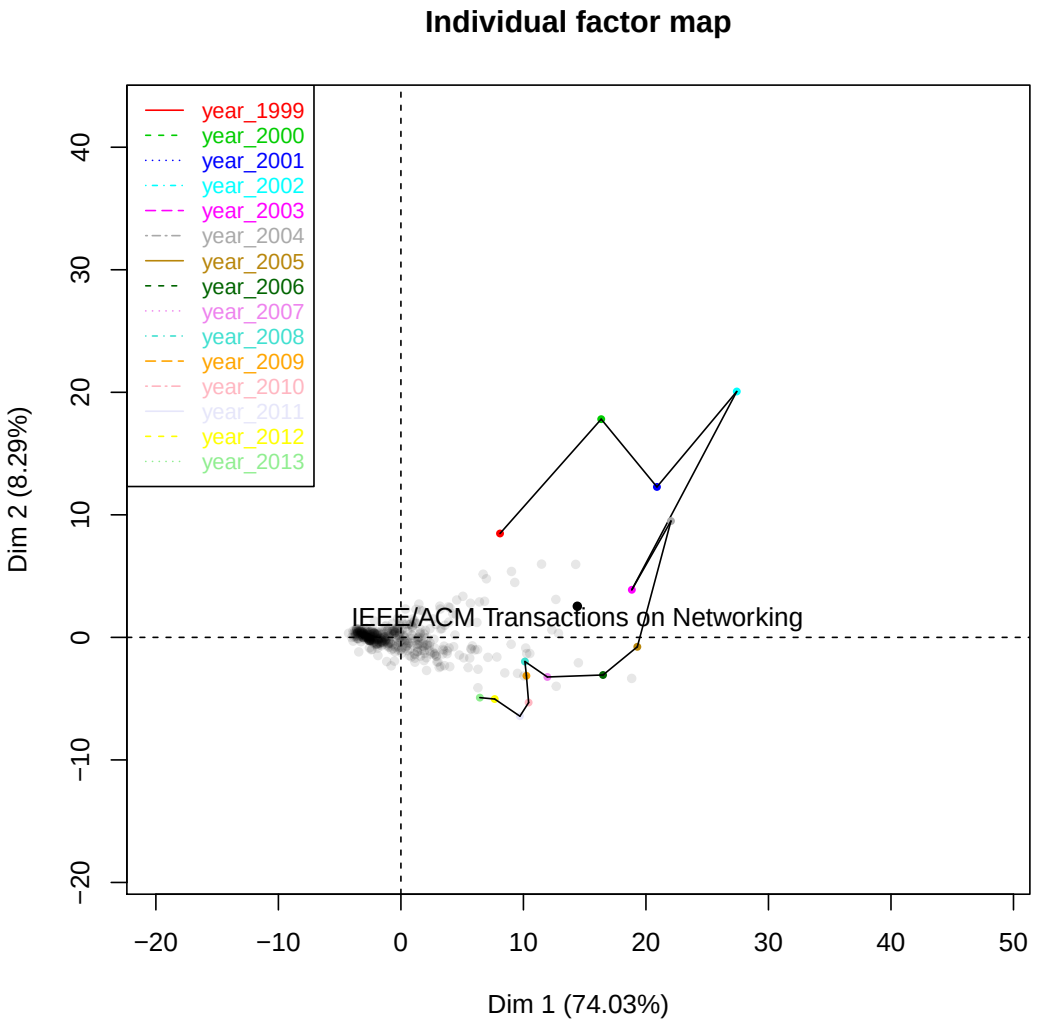




MFA with missing values

Rows: 47000 journals / Groups: 15 years of data/ Variables: 3 scores each year. Many missing...

ACM Transactions on Networking trajectory.pdf



Confidence ellipses

$\Rightarrow$ Bootstrap

$\Rightarrow$ Rows bootstrap (Holmes, 1985, Timmerman *et al*, 2007)

- random sample from a population - sampling variance
- confidence areas around the position of the variables

$\Rightarrow$ Residuals bootstrap

- full population data - $X = \mu + \varepsilon$ - noise fluctuation
- confidence areas around the individuals and the variables

Confidence ellipses

$\Rightarrow$ Bootstrap

$\Rightarrow$ Rows bootstrap (Holmes, 1985, Timmerman *et al*, 2007)

- random sample from a population - sampling variance
- confidence areas around the position of the variables

$\Rightarrow$ Residuals bootstrap

- full population data - $X = \mu + \varepsilon$ - noise fluctuation
- confidence areas around the individuals and the variables

1. residuals $\hat{\varepsilon} = X - \hat{\mu}^k$. Bootstrap or draw from $\mathcal{N}(0, \hat{\sigma}^2) : \varepsilon^b$
2. $X^b = \hat{\mu}^k + \varepsilon^b$
3. PCA on $X^b \rightarrow (\hat{\mu}^1 = U^1 D^1 V^{1'}, ..., \hat{\mu}^B = U^B D^B V^{B'})$

Confidence ellipses

$\Rightarrow$ Jackknife

Confidence ellipses

$\Rightarrow$ Cell-wise Jackknife

$\hat{\mu}^{(k)^{(-ij)}}$ the estimator from the matrix without the cell (ij)

$$\hat{\mu}^{(k)^{(ij)}}_{jackk} = \hat{\mu}^k + \sqrt{np}\left(\hat{\mu}^{(k)^{(-ij)}} - \hat{\mu}^k\right)$$

Represent pseudo-values: $(\hat{\mu}^1, ..., \hat{\mu}^{np})$
Requires a method to deal with missing values - costly

Confidence ellipses

$\Rightarrow$ Cell-wise Jackknife

$\hat{\mu}^{(k)^{(-ij)}}$ the estimator from the matrix without the cell (ij)

$$\hat{\mu}^{(k)^{(ij)}}_{jackk} = \hat{\mu}^k + \sqrt{np}\left(\hat{\mu}^{(k)^{(-ij)}} - \hat{\mu}^k\right)$$

Represent pseudo-values: $(\hat{\mu}^1, ..., \hat{\mu}^{np})$
Requires a method to deal with missing values - costly

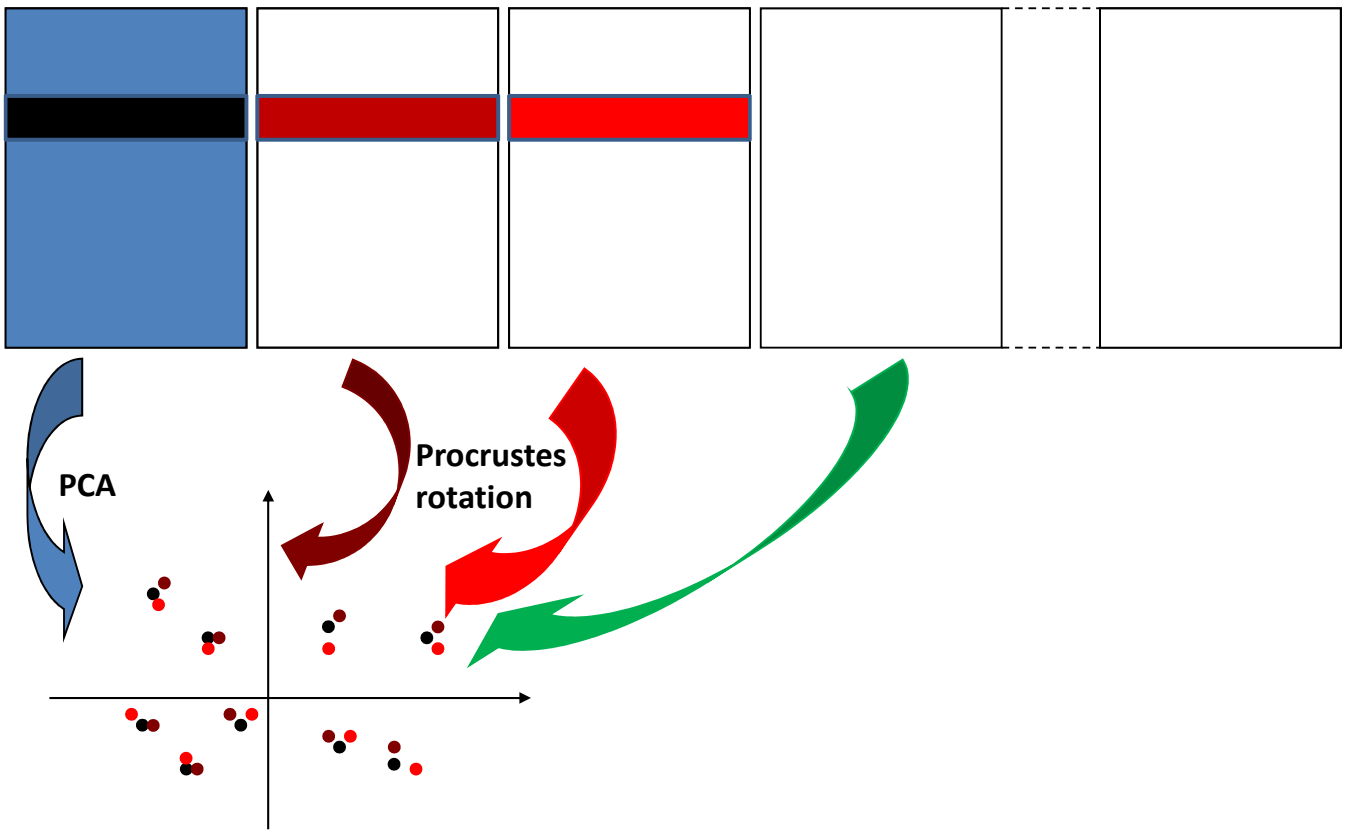
$\Rightarrow$ Approximate Jackknife (Craven & Wahba, 1979 lemma)

In regression $\hat{y} = Py$: $\hat{y}_i^{-i} - y_i = \frac{\hat{y}_i - y_i}{1 - P_{i,i}}$
In PCA $\text{vec}(\hat{\mu}_k) = P\text{vec}(X)$

$$x_{ij} - \hat{\mu}_{ij}^{(k)^{(-ij)}} \simeq \frac{x_{ij} - \hat{\mu}_{ij}^k}{1 - P_{ij,ij}}$$

Visualizing confidence areas

$\Rightarrow$ Spread of $(\hat{\mu}^1, ..., \hat{\mu}^B)$ gives the variability of PCA $\hat{\mu}^k_{n\times p}$

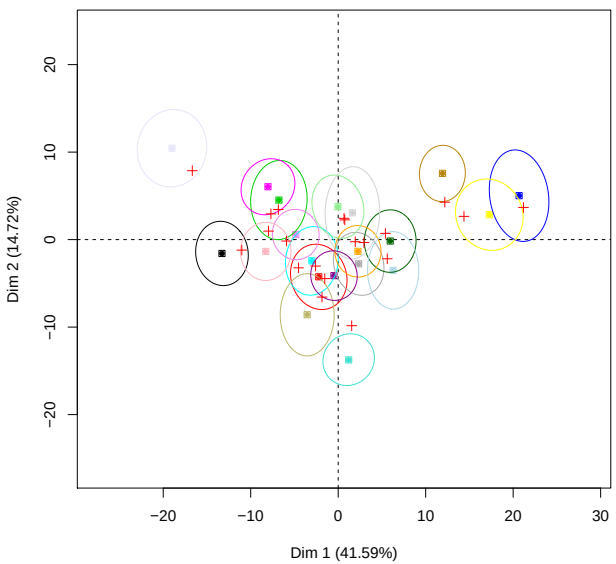


Simulations

Model: $X = \mu + \varepsilon,\ \varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma^2\right)$

- n/p: 100/20, 50/50 and 20/100
- k: 2, 4 - the ratio d_1/d_2: 4, 1 - SNR: 4, 1, 0.8

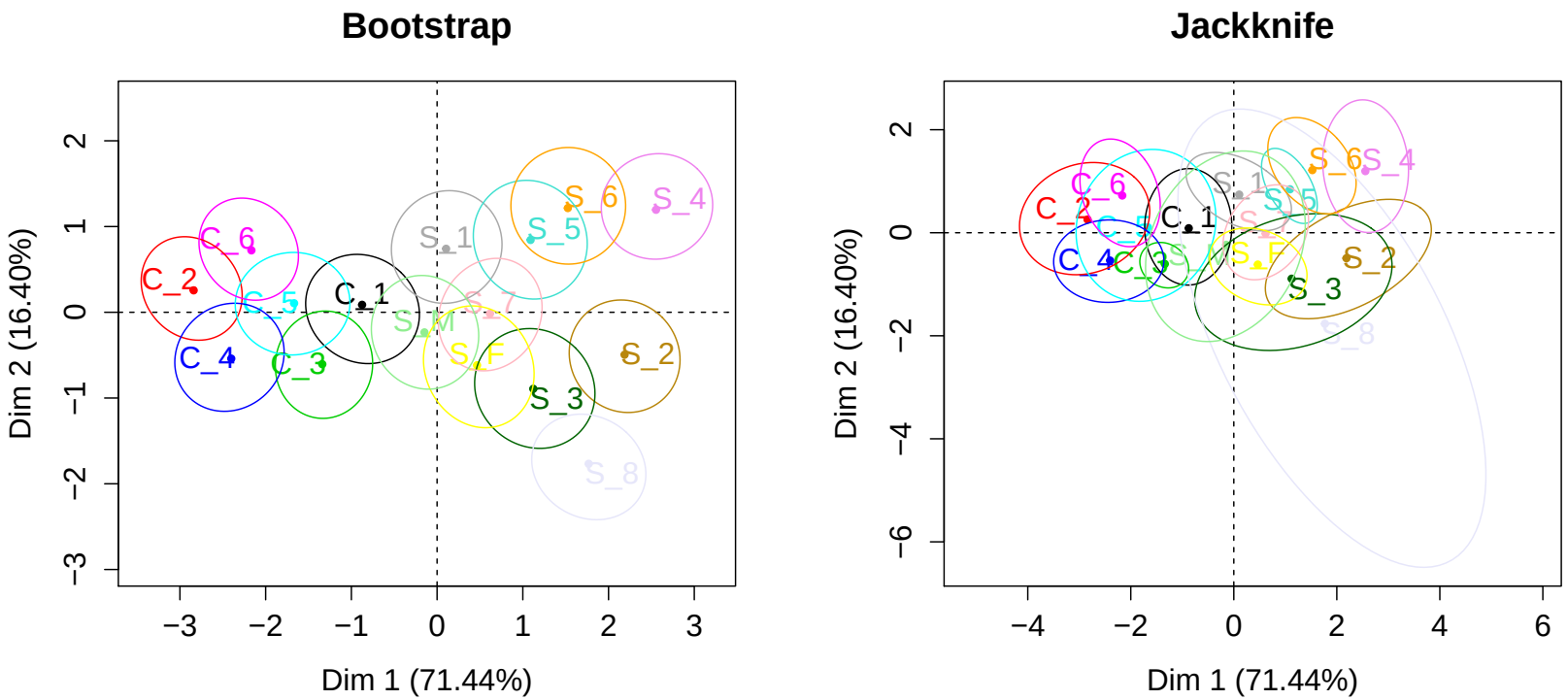
Coverage properties



$\Rightarrow$ Different regimes:

- High SNR - n large: Asymptotic $\approx$ Bootstrap
- Small SNR: Jackknifes

Inference in PCA graphs



- Simulation based on the data
- Incorporating the uncertainty of k!
- Variance of other estimators

Bayesian treatment of SVD

$$X = \mu + \varepsilon = UDV' + \varepsilon$$

$\Rightarrow$ Overparametrization - priors and posteriors meeting constraints?

- **Hoff 2009** uniform prior for U and V special cases of von Mises-Fisher distributions (Stiefel manifold); $(d_l)_{l=1...k} \sim \mathcal{N}(0, s_\lambda^2)$
- conditional posterior of U and V are also VMF
- draw from the posterior of μ. Point estimate $\hat{\mu}$, small MSE

$\Rightarrow$ Point estimate $\hat{\mu}$ (shrinked) - uncertainty $(\mu^1, ..., \mu^B)$, also on k.

Bayesian treatment of SVD

$$X = \mu + \varepsilon = UDV' + \varepsilon$$

$\Rightarrow$ Overparametrization - priors and posteriors meeting constraints?

- **Hoff 2009** uniform prior for U and V special cases of von Mises-Fisher distributions (Stiefel manifold); $(d_l)_{l=1...k} \sim \mathcal{N}(0, s_\lambda^2)$
- conditional posterior of U and V are also VMF
- draw from the posterior of μ. Point estimate $\hat{\mu}$, small MSE

- **Josse *et.* 2013** $u_{il} \sim \mathcal{N}(0,1)$ $v_{jl} \sim \mathcal{N}(0,1)$ $(d_l) \sim \mathcal{N}(0, s_\lambda^2)$
- posteriors do not meet the constraints
- posterior for μ: draw $\mu^1, ..., \mu^B \Rightarrow$ postprocessing step SVD on each matrix $\mu^1 = U^1 D^1 V^1, ..., \mu^B = U^B D^B V^B$

$\Rightarrow$ Point estimate $\hat{\mu}$ (shrinked) - uncertainty $(\mu^1, ..., \mu^B)$, also on k.

Random effect models: Probabilistic PCA

$\Rightarrow$ A specific factor analysis model (Roweis, 1998, Tipping & Bishop, 1999)

$$x_{ij} = (Z_{n\times k} B'_{p\times k})_{ij} + \varepsilon_{ij},\ z_i \sim \mathcal{N}(0, \mathbb{I}_k),\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_p)$$

- Conditional independence : $x_i|z_i \sim \mathcal{N}(Bz_i,\ \sigma^2 \mathbb{I}_p)$
- Distribution *i.i.d*: $x_i \sim \mathcal{N}(0, \Sigma = BB' + \sigma^2 \mathbb{I}_p)$

- Max Lik.: $\hat{\sigma}^2 = \frac{1}{p-k} \sum_{l=k+1}^{p} \lambda_l \qquad \hat{B} = V(D - \sigma^2 \mathbb{I}_k)^{\frac{1}{2}}$

$\Rightarrow$ BLUP $\mathbb{E}(z_i|x_i)$: $\hat{Z} = X\hat{B}(\hat{B}'\hat{B} + \sigma^2 \mathbb{I}_k)^{-1}$

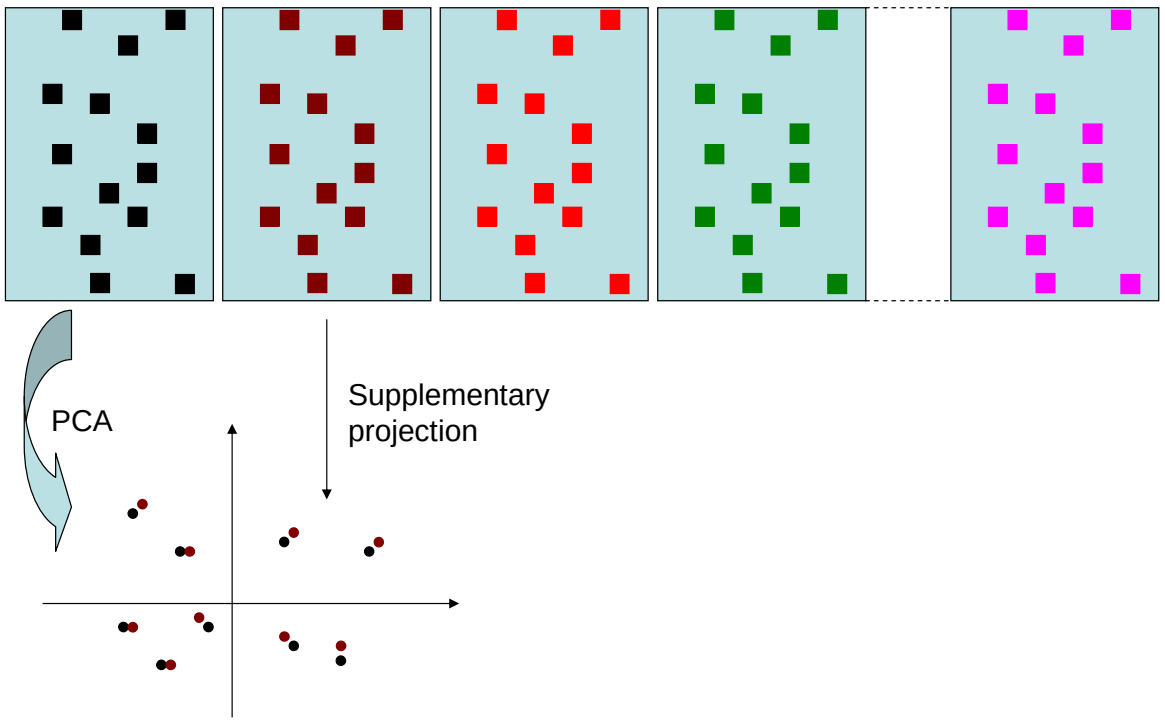
$$\hat{\mu}^{\text{ppca}} = \hat{Z}\hat{B}' \qquad \hat{\mu}_{ij}^{\text{ppca}} = \sum_{l=1}^{k} \left(d_l - \frac{\sigma^2}{d_l} \right) u_{il} v_{jl}$$

$\Rightarrow$ Bayesian SVD with *a priori* on U: Regularized PCA

Supplementary projection

Imputed by Regularized PCA/ B imputed data sets from MIPCA

Same observed values (blue)/ different predictions for missing values



Regularized iterative PCA

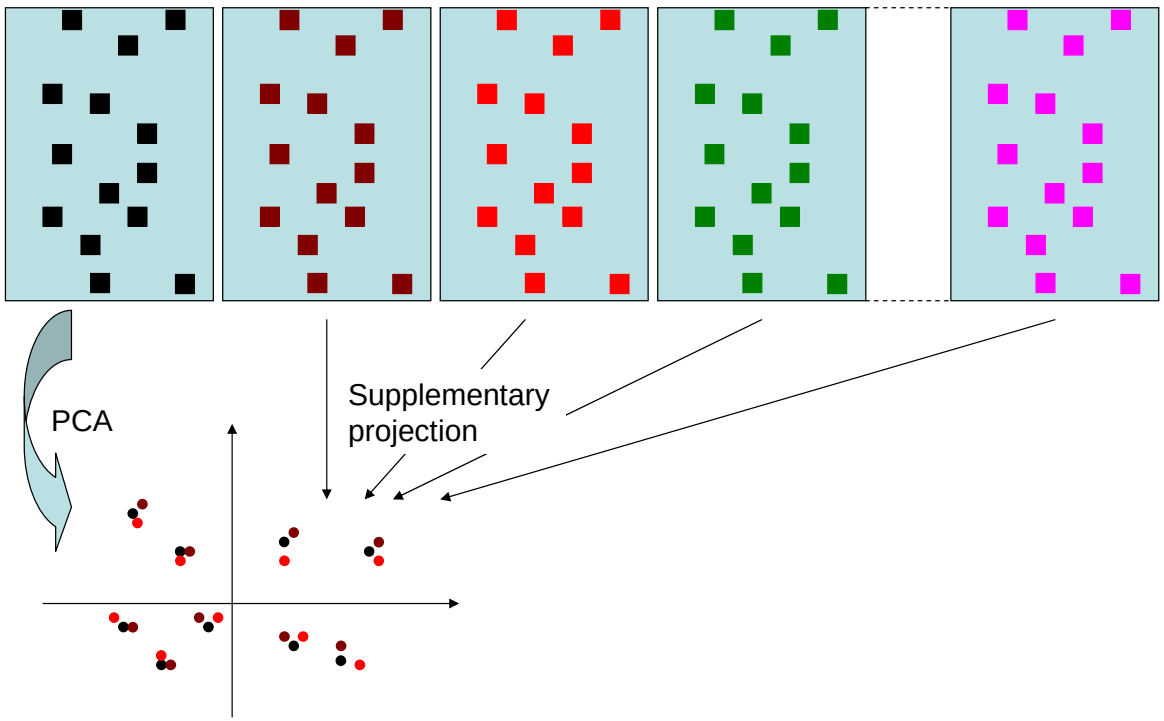
$\Rightarrow$ reference configuration

$\Rightarrow$ Individuals' position (and variables) with other predictions

Supplementary projection

Imputed by Regularized PCA/ B imputed data sets from MIPCA

Same observed values (blue)/ different predictions for missing values



Regularized iterative PCA

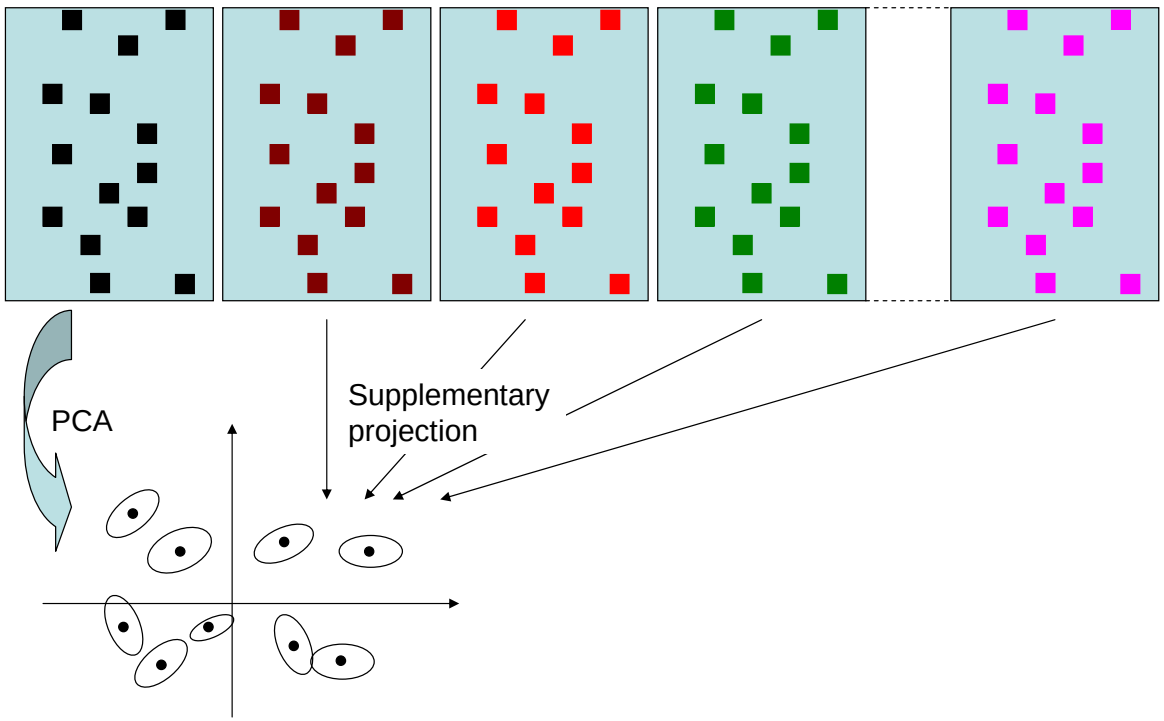
$\Rightarrow$ reference configuration

$\Rightarrow$ Individuals' position (and variables) with other predictions

Supplementary projection

Imputed by Regularized PCA/ B imputed data sets from MIPCA

Same observed values (blue)/ different predictions for missing values

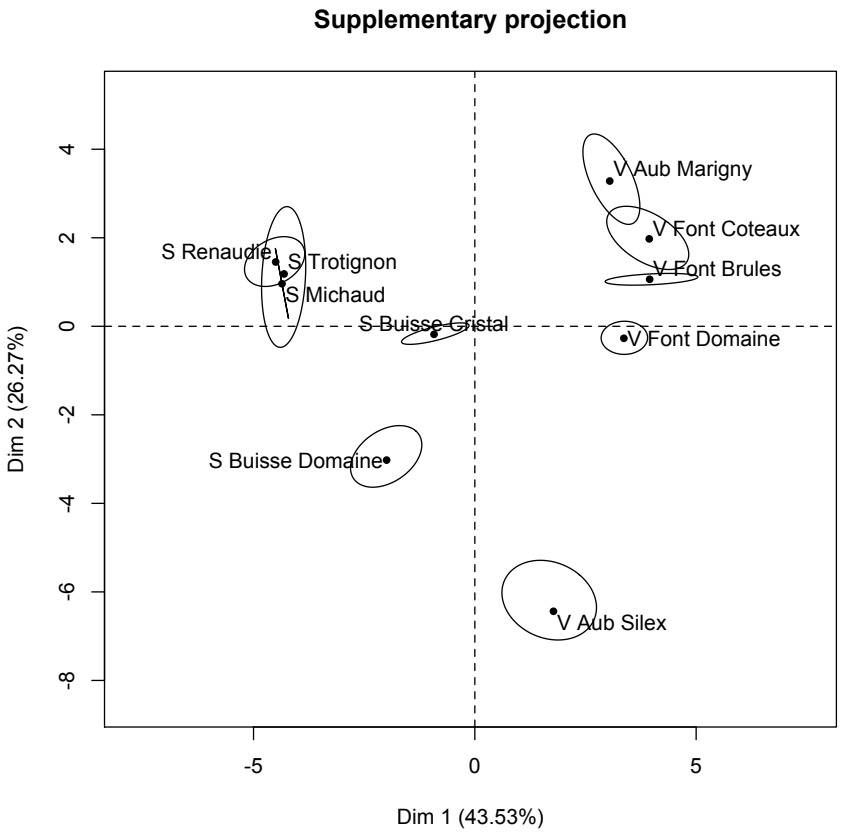


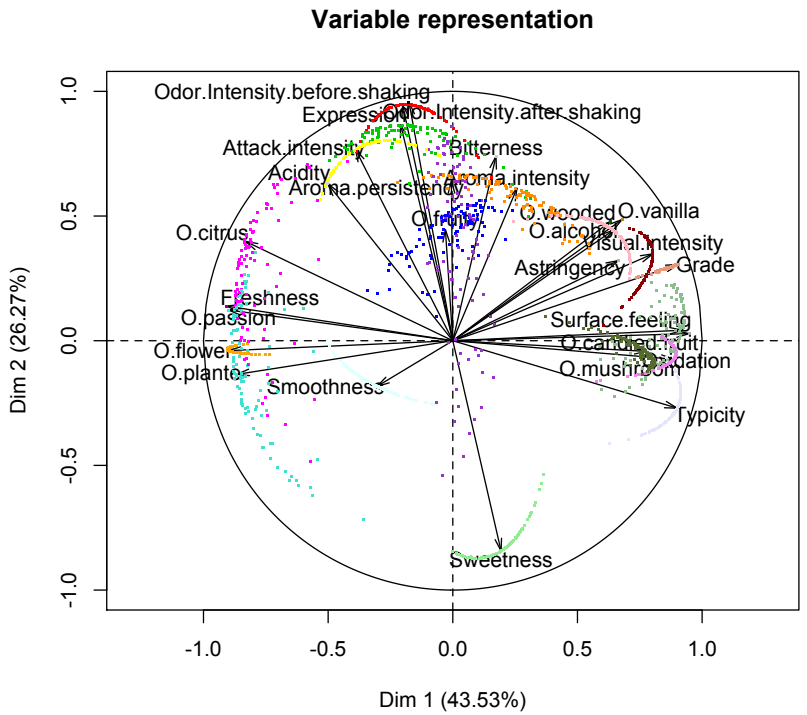
Regularized iterative PCA

$\Rightarrow$ reference configuration

$\Rightarrow$ Individuals' position (and variables) with other predictions

Multiple imputation: visualization of variance of prediction





⇒ Does single imputation make sense?

The log-bilinear model & CA

$\Rightarrow$ The saturated log-linear models (Christensen, 1990; Agresti, 2013).

$$\log \mu_{ij} = \alpha_i + \beta_j + \Gamma_{ij}$$

$\Rightarrow$ The RC association model (Goodman, 1985; Gower, 2011; GAMMI)

$$\log \mu_{ij} = \alpha_i + \beta_j + \sum_{k=1}^{K} d_k u_{ik} v_{jk}$$

Estimation: iterative weighted least squares, steps of GLM.

The log-bilinear model & CA

$\Rightarrow$ The saturated log-linear models (Christensen, 1990; Agresti, 2013).

$$\log \mu_{ij} = \alpha_i + \beta_j + \Gamma_{ij}$$

$\Rightarrow$ The RC association model (Goodman, 1985; Gower, 2011; GAMMI)

$$\log \mu_{ij} = \alpha_i + \beta_j + \sum_{k=1}^{K} d_k u_{ik} v_{jk}$$

Estimation: iterative weighted least squares, steps of GLM.

$\Rightarrow$ CA (Greenacre,1984) texts corpus, spectral clustering on graphs:

$$z_{ij} = \frac{x_{ij}/N - r_i c_j}{\sqrt{r_i c_j}} \text{ i.e. } Z = D_r^{-1/2}(X/N - rc^T)D_c^{-1/2}$$

if $\mathbf{X}$ adjacency matrix, $\mathbf{Z}$ is symmetric normalized graph Laplacian

CA approximates the log-bilinear model

$\Rightarrow$ SVD $Z = \widetilde{U}_K D_K \widetilde{V}_K^T$. Standard row and col coord $U_K = D_r^{-1/2}\widetilde{U}_K$, $V_K = D_c^{-1/2}\widetilde{V}_K$. If the low-rank approx is good:

$$U_K D_K V_K^T \approx D_r^{-1/2} Z D_c^{-1/2} = D_r^{-1}(X/N - rc^T)D_c^{-1} \qquad (1)$$

By "solving for X" in (**??**), we get the *reconstruction formula*:

$$\widehat{X}/N = rc^T + D_r(U_K D_K V_K^T)D_c \text{ i.e. } \frac{\hat{x}_{ij}}{N} = r_i c_j \left(1 + \sum_{k=1}^{K} d_k u_{ik} v_{jc}\right) \qquad (2)$$

CA approximates the log-bilinear model

$\Rightarrow$ SVD $Z = \widetilde{U}_K D_K \widetilde{V}_K^T$. Standard row and col coord $U_K = D_r^{-1/2}\widetilde{U}_K$, $V_K = D_c^{-1/2}\widetilde{V}_K$. If the low-rank approx is good:

$$U_K D_K V_K^T \approx D_r^{-1/2} Z D_c^{-1/2} = D_r^{-1}(X/N - rc^T)D_c^{-1} \qquad (1)$$

By "solving for X" in (**??**), we get the *reconstruction formula*:

$$\widehat{X}/N = rc^T + D_r(U_K D_K V_K^T)D_c \text{ i.e. } \frac{\hat{x}_{ij}}{N} = r_i c_j \left(1 + \sum_{k=1}^{K} d_k u_{ik} v_{jc}\right) \qquad (2)$$

$\Rightarrow$ Connection (Escofier, 1982) : when $\sum_{k=1}^{K} d_k u_{ik} v_{jk} << 1$, eq. (**??**) is:

$$\log(\hat{x}_{ij}) \approx \log(N) + \log(r_i) + \log(c_j) + \sum_{k=1}^{K} d_k u_{ik} v_{jk}$$

Outline

Alcohol data

INPES (Santé publique France)

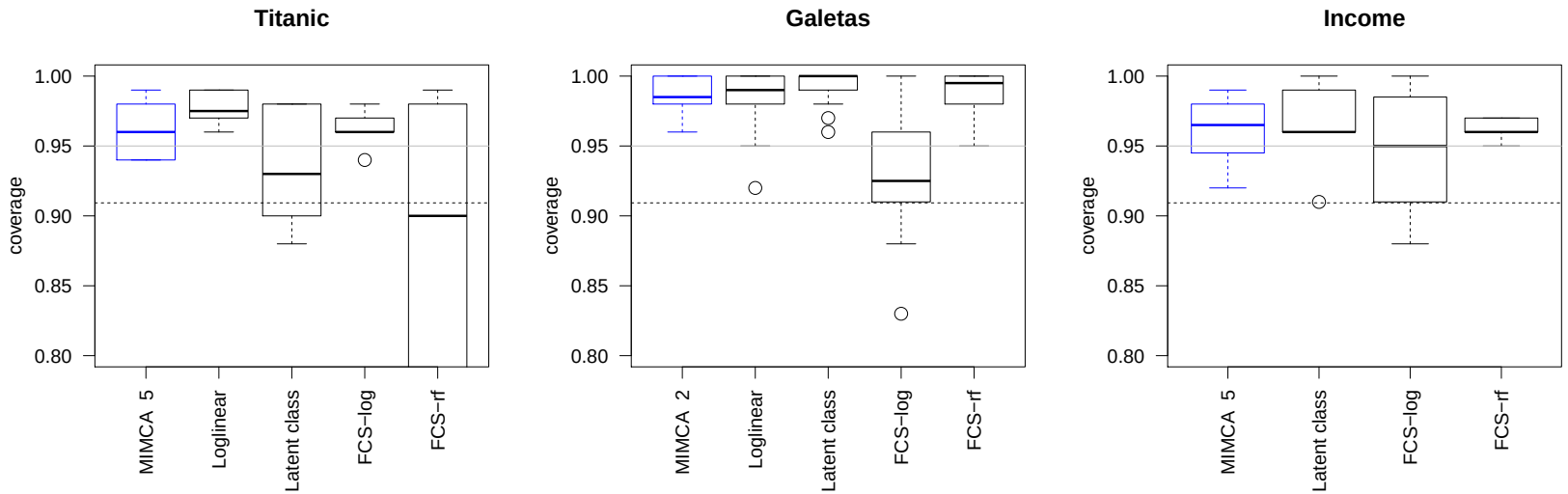
```
region                    sex        age           year          edu          drunk         alcohol      gla
Ile de France      :8120  F:29776   18_25: 6920   2005:27907   E1:12684  0     :44237   <1/m :12889   0
Rhone Alpes        :5421  M:23165   26_34: 9401   2010:25034   E2:23521  1-2   : 4952   0    : 6133   0-2
Provence Alpes     :4116            35_44:10899                E3:6563   10-19:  839   1-2/m: 7583   10-
Nord Pas de Calais :3819            45_54: 9505                E4:10100  20-29:  212   1-2/w: 9526   3-4
Pays de Loire      :3152            55_64: 9503                NA:73     3-5   : 1908   3-4/w: 6815   5-6
Bretagne           :3038            65_+ : 6713                          30+   :  404   5-6/w: 3402   7-9
(Other)            :25275                                                6-9   :  389   7/w  : 6593

binge                  Pbsleep                 Tabac
<2/m:10323             Never:20605             Frequent  : 9176
0   :34345             Often: 10172            Never     :39080
1/m : 6018             Rare :22134             Occasional: 4588
1/w : 1800             NA:  30                 NA:  97
7/w :  374
NA  :   81
```

```
   region sexe   age year edu drunk alcohol glasses binge
1 Rhône Alpes    M 45_54 2005   1     0       0     0-2     0
2 Rhône Alpes    M 45_54 2005   2     0       0     0-2     0
3 Rhône Alpes    M 55_64 2005   2     0       0     0-2     0
4 Rhône Alpes    M 18_25 2005   3     0       0     0-2     0
5 Rhône Alpes    M 18_25 2005   2     0       0     0-2     0
6 Rhône Alpes    M 26_34 2005   2     0       0     0-2     0
```

Simulations

- Quantities of interest: $\theta =$ parameters of a logistic model
- 200 simulations from real data sets
  - the real data set is considered as a population
  - drawn one sample from the data set
  - generate 20% of missing values
  - multiple imputation using $M = 5$ imputed data
- Criteria
  - bias
  - CI width, coverage

Results - Inference



	Titanic	Galetas	Income
Number of variables	4	4	14
Number of categories	≤ 4	≤ 11	≤ 9

MI using the loglinear model

- Hypothesis $X = (x_{ijk})_{i,j,k}$:
  $X|\theta \sim \mathcal{M}(n, \theta)$ where:

$$log(\theta_{ijk}) = \lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_S^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

1. Variability of the parameters
   - prior on θ : $\theta|\theta \in \Theta \sim \mathcal{D}(\alpha)$
   - posterior: $\theta|x, \theta \in \Theta \sim \mathcal{D}(\alpha')$
   - Data Augmentation (M.A. Tanner, W.H. Wong, 1987)
2. Imputation according to the loglinear model using the set of M parameters

- Implemented: `R` package `cat` (J.L. Schafer)

MI using a DPMPM model (Si and Reiter, 2013)

- Hypothesis: $\mathbb{P}(X = (x_1, \ldots, x_K); \theta) = \sum_{\ell=1}^{L} \left(\theta_\ell \prod_{k=1}^{K} \theta_{x_k}^{(\ell)} \right)$

1. Variability of the parameters:
   - a hierarchic prior on θ:
     $\alpha \sim \mathcal{G}(.25, .25)$ $\zeta_\ell \sim \mathcal{B}(1, \alpha)$ $\theta_\ell = \zeta_\ell \prod_{g<\ell}(1 - \zeta_g)$ for ℓ in $1, \ldots, \infty$
   - posterior on θ: untractable
     $\rightarrow$ Gibbs sampler and Data Augmentation
2. Imputation according to the mixture model using the set of M parameters

- Implemented: `R` package `mi` (Gelman *et al.*)

Maximizing the likelihood

Data: $A = [A_1, \ldots, A_m]$

Model: $\pi_{ijc} = \frac{\exp(\theta_{ij(c)})}{\sum_{c'=1}^{C_j} \exp(\theta_{ij(c')})}$ $\theta_{ij}(c) = \beta_j(c) + \sum_{l=1}^{k} d_l u_{il} v_{jl}(c)$

Estimation: MLE but overfitting problems $\theta_{ij(c)} \to \infty$

Penalized likelihood: $-\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{c=1}^{C_j} A_{ic}^{j} \log(\pi_{ijc}) + \lambda \sum_{l=1}^{k^\star} d_l$

Algorithm: MM with quadratic majorization (Groenen & Josse, 2016)

$$f_{ij}(\theta_i) = -\sum_{c=1}^{C_j} A_{ic}^{j} \log(\pi_{ijc}) = -\sum_{c=1}^{C_j} A_{ic}^{j} \log\left(\frac{\exp(\theta_{ij(c)})}{\sum_{c'=1}^{C_j} \exp(\theta_{ij(c')})}\right)$$

With $\nabla f_{ij}(\theta_i) = A_{ij} - \pi_{ij}$ and $H = \nabla^2 f_{ij}(\theta_i) = (\mathrm{Diag}(\pi_{ij}) - \pi_{ij}\pi_{ij}')$

Outline

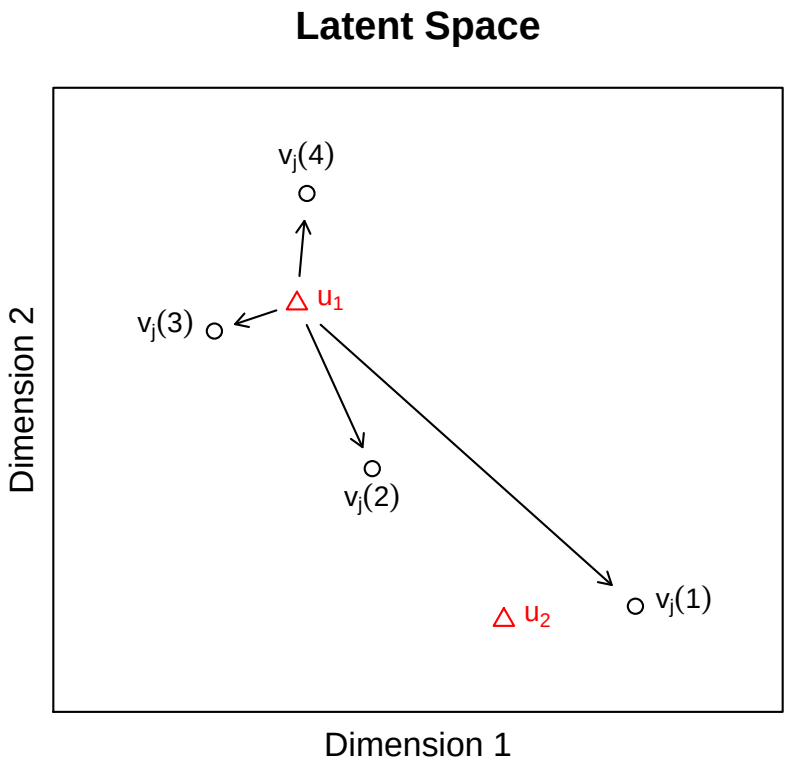
Multilogit-bilinear model

$$\mathbb{P}(x_{ij} = c) = \pi_{ijc} = \frac{e^{\theta_{ij}(c)}}{\sum_{c'=1}^{C_j} e^{\theta_{ij}(c')}},$$

$$\theta_{ij}(c) = \beta_j(c) + \Gamma_i^j(c) = \beta_j(c) + \sum_{l=1}^{k} d_l u_{il} v_{jl}(c)$$

$\tilde{\mathbf{v}}_j(c) = (\sqrt{d_1} v_{j1}(c), \sqrt{d_2} v_{j2}(c))$: question j, category c with coordinates one point/ C_j categories. The latent variables $\tilde{\mathbf{u}}_i = \mathbf{D}_2^{1/2} \mathbf{u}_i$

$$\mathbb{P}(x_{ij} = c) \propto \exp\left\{\tilde{\beta}_j(c) - \tfrac{1}{2}\|\tilde{\mathbf{v}}_j(c) - \tilde{\mathbf{u}}_i\|^2\right\}$$



Relationship with MCA

Data: $X_{n\times m}$ - $A = \left[A^1 | \cdots | A^m\right]$, $A^j \in \{0,1\}^{n\times C_j}$

Model: $\pi_{ijc} = \frac{e^{\beta_j(c)+\Gamma_i^j(c)}}{\sum_{c'=1}^{C_j} e^{\beta_j(c')+\Gamma_i^j(c')}}$ Param: $\zeta = \binom{\beta}{\text{vec}(\Gamma)}$; $\zeta_0 = \binom{log(p)}{0}$

Relationship with MCA

Data: $X_{n\times m}$ - $A = \left[A^1 | \cdots | A^m\right]$, $A^j \in \{0,1\}^{n\times C_j}$

Model: $\pi_{ijc} = \frac{e^{\beta_j(c)+\Gamma_i^j(c)}}{\sum_{c'=1}^{C_j} e^{\beta_j(c')+\Gamma_i^j(c')}}$ Param: $\zeta = \binom{\beta}{\text{vec}(\Gamma)}$; $\zeta_0 = \binom{log(p)}{0}$

$\Rightarrow$ Rationale: Taylor expand ℓ around the *independence model* ζ_0:

$$\tilde{\ell}(\beta,\Gamma) = \ell(\zeta_0) + \ell'(\zeta_0)^T(\zeta-\zeta_0) + \frac{1}{2}(\zeta-\zeta_0)^T\ell''(\zeta_0)(\zeta-\zeta_0)$$

$\tilde{\ell}(\beta,\Gamma)$ a quadratic function of its arguments, then maximizing the latter amounts to a generalized SVD $\Rightarrow$ MCA.

Relationship with MCA

Data: $X_{n\times m}$ - $A = \left[A^1 | \cdots | A^m\right]$, $A^j \in \{0,1\}^{n\times C_j}$

Model: $\pi_{ijc} = \frac{e^{\beta_j(c)+\Gamma_i^j(c)}}{\sum_{c'=1}^{C_j} e^{\beta_j(c')+\Gamma_i^j(c')}}$ Param: $\zeta = \binom{\beta}{\text{vec}(\Gamma)}$; $\zeta_0 = \binom{log(p)}{0}$

$\Rightarrow$ Rationale: Taylor expand ℓ around the *independence model* ζ_0:

$$\tilde{\ell}(\beta,\Gamma) = \ell(\zeta_0) + \ell'(\zeta_0)^T(\zeta-\zeta_0) + \frac{1}{2}(\zeta-\zeta_0)^T\ell''(\zeta_0)(\zeta-\zeta_0)$$

$\tilde{\ell}(\beta,\Gamma)$ a quadratic function of its arguments, then maximizing the latter amounts to a generalized SVD $\Rightarrow$ MCA.

$\Rightarrow$ The joint likelihood is $\prod_{i=1}^n \prod_{j=1}^m \prod_{c=1}^{C_j} \pi_{ijc}^{A_{ic}^j}$ (independence)

$\Rightarrow$ The log-likelihood for the MultiLogit Bilinear model is:

$$\ell = \sum_{i,j,c} A_{ic}^j \log(\pi_{ijk}) = \sum_{i,j,c} A_{ic}^j \log\left(\frac{\exp(\beta_j(c)+\Gamma_i^j(c))}{\sum_{c'=1}^{C_j}\exp(\beta_j(c')+\Gamma_i^j(c'))}\right)$$

Relationship with MCA

$$\ell(\beta, \Gamma; A) = \beta^j(A_i^j) + \Gamma_i^j(A_i^j) - \log\left(\sum_{c=1}^{C_j} e^{\beta^j(c)+\Gamma_i^j(c)}\right)$$

Compute $\frac{\partial \ell}{\partial \Gamma_i^j(c)}$ and $\frac{\partial \ell}{\partial \Gamma_i^j(c)\partial \Gamma_{i'}^{j'}(c')}$, assess at $(\beta_0 = \log(p), 0)$ to get:

$$\tilde{\ell}(\beta, \Gamma) \approx \langle \Gamma, A - \mathbb{1}p^T \rangle - \tfrac{1}{2}\|\Gamma D_p^{1/2}\|_F^2$$

Solution: rank k SVD of $(A - \mathbb{1}p^T)D_p^{-1/2} \Rightarrow$ SVD in MCA.

Theorem (Fithian & Josse, 2016): *The one-step likelihood estimate for the MultiLogit Bilinear model with rank constraint k, obtained by expanding around the independence model* $(\beta_0 = \log p, \Gamma_0 = 0)$*, is* $(\beta_0, \widehat{\mathbf{\Gamma}}_{MCA})$.

Lemma: Let $\mathbf{G} \in \mathbb{R}^{n\times n}, \mathbf{H}_1 \in \mathbb{R}^{n\times n}, \mathbf{H}_2 \in \mathbb{R}^{m\times m}$, with $\mathbf{H}_1, \mathbf{H}_2 \succ 0$.
$\text{argmax}_{\mathbf{\Gamma}:\, \text{rank}(\mathbf{\Gamma})\leq K} \langle \mathbf{\Gamma}, \mathbf{G}\rangle - \frac{1}{2}\|\mathbf{H}_1\mathbf{\Gamma}\mathbf{H}_2\|_F^2$

$\mathbf{\Gamma}^* = \mathbf{H}_1^{-1}\left[\text{SVD}_K(\mathbf{H}_1^{-1}\mathbf{G}\mathbf{H}_2^{-1})\right]\mathbf{H}_2^{-1}$

Results

- Low interaction: MCA and MLE agree
- Strong signal: MCA less appropriate?
- MCA estimates better than MLE in difficult settings! (small n, m, noisy): overfitting! penalized estimate improves on MCA

$\Rightarrow$ MCA a proxy to estimate the model's parameters

Results

- Low interaction: MCA and MLE agree
- Strong signal: MCA less appropriate?
- MCA estimates better than MLE in difficult settings! (small n, m, noisy): overfitting! penalized estimate improves on MCA

$\Rightarrow$ MCA a proxy to estimate the model's parameters

- Initialization of the algorithms by PC methods like MCA
- Regularized MCA: influenced by the way the problem is written
- Select k with model based criteria (asymptotic picture unclear)
- Solving log-linear models by iterative CA

$log(\theta_{ijk}) = \lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_S^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + ...$

Majorization function

Theorem
$g_{ij}(\theta_i, \theta_i^{(0)}) = f_{ij}(\theta_i^{(0)}) + (\theta_i - \theta_i^{(0)})'\nabla f_{ij}(\theta_i^{(0)}) + 1/4\|\theta_i - \theta_i^{(0)}\|^2$ *is a majorizing function of* $f_{ij}(\theta_i)$ *(using De Leeuw, 2005)*

$\frac{1}{2}\mathbf{I} - \nabla^2 f_{ij}(\theta_i)$ is positive semi-definite (largest eigenvalue of $\nabla^2 f_{ij}(\theta_i)$ is smaller than $1/2$)

$$H = \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & -\pi_1\pi_3 & \dots \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & -\pi_2\pi_3 & \dots \\ \dots & \dots & \pi_3(1-\pi_3) & \dots \\ \dots & \dots & \dots. & \dots \end{pmatrix}$$

Gerschgorin disks: the eigenvalue ϕ is always smaller than a diagonal element plus the sum of its absolute off-diagonal row (or col) values
$\phi \leq \pi_{ijc} - \pi_{ijc}^2 + \pi_{ijc}\sum_{\ell \neq c}\pi_{ij\ell}$
$\phi \leq \pi_{ijc} - \pi_{ijc}^2 + \pi_{ijc}\sum_{\ell=1}^{C_j}\pi_{ij\ell} - \pi_{ijc}^2$
$\phi \leq 2(\pi_{ijc} - \pi_{ijc}^2) = 2\pi_{ijc}(1-\pi_{ijc})$.

$2\pi_{ijc}(1-\pi_{ijc})$ reaches its maximum of $1/2$ at $\pi_{ijc} = 1/2$

Majorization function

Theorem
$g_{ij}(\theta_i, \theta_i^{(0)}) = f_{ij}(\theta_i^{(0)}) + (\theta_i - \theta_i^{(0)})'\nabla f_{ij}(\theta_i^{(0)}) + 1/4\|\theta_i - \theta_i^{(0)}\|^2$ *is a majorizing function of* $f_{ij}(\theta_i)$ *(using De Leeuw, 2005)*

$\frac{1}{2}\mathbf{I} - \nabla^2 f_{ij}(\theta_i)$ is positive semi-definite (largest eigenvalue of $\nabla^2 f_{ij}(\theta_i)$ is smaller than $1/2$)

Update parameters $\theta_{ij}(c) = \beta_j(c) + \sum_{l=1}^{k} d_k u_{il} v_{jl}(c)$

$$L(\beta, U, D, V) \leq \frac{1}{4}\sum_{i,j,c} A_{ic}^{j}(z_{ijc} - \theta_{ij(c)})^2 + \lambda\left(\sum_{l=1}^{k^\star} d_l\right) + c$$

- Update U and V: SVD of $Z = P\Phi Q'$, $U = P$, $V = Q'$
- Update D: $\sum_{l=1}^{k^\star}[(\phi_l - d_l)^2 + \lambda d_l]$ $d_l = \max(0, \phi_l - \lambda)$

Bayesian PCA complete case

$$\begin{aligned} X_{ij} &= \mu_{ij} + \varepsilon_{ij} \ , \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ &= \sum_{l=1}^{S} \mu_{ij}^{(l)} + \varepsilon_{ij} = \sum_{l=1}^{S} \sqrt{d_l} R_{il} Q_{jl} + \varepsilon_{ij} \end{aligned}$$

$\Rightarrow$ Priors on Q, D, R: overparametrization issue

• Hoff (2009): Uniform prior for Q and R von Mises-Fisher distributions (Stiefel manifold); $(d_l)_{l=1\ldots k} \sim \mathcal{N}\left(0, s_\lambda^2\right)$

• Josse & Denis (2013): disregarding the constraints; $Q_{il} \sim \mathcal{N}\left(0,1\right)$, $R_{jl} \sim \mathcal{N}\left(0,1\right)$, $(d_l)_{l=1\ldots k} \sim \mathcal{N}\left(0, s_\lambda^2\right)$

$\Rightarrow$ Priors on Q, D, R: prior on μ

Bayesian PCA complete case

Model: $X_{ij} = \sum_{l=1}^{S} \mu_{ij}^{(l)} + \varepsilon_{ij}$, $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

Prior: $\mu_{ij}^{(l)} \sim \mathcal{N}(0, \tau_l^2)$

Posterior: $\left(\mu_{ij}^{(l)} | X_{ij}^{(l)}\right) = \mathcal{N}(\Phi_l X_{ij}^{(l)}, \sigma^2 \Phi_l)$ with $\Phi_l = \frac{\tau_l^2}{\tau_l^2 + \sigma^2}$

Empirical Bayes: $\left(X_{ij}^{(s)}\right) \sim \mathcal{N}(0, \tau_l^2 + \sigma^2)$ max the likelihood τ_l^2:
$\hat{\tau_s}^2 = (\lambda_s - \hat{\sigma}^2)$

$$\hat{\Phi}_l = \frac{\lambda_s - \hat{\sigma}^2}{\lambda_s} = \frac{\text{signal variance}}{\text{total variance}}$$

Multiple imputation Bayes-PCA

Variability of the parameters, B plausible $(\hat{\mu}_{ij})^1, ..., (\hat{\mu}_{ij})^B$

$\Rightarrow$ Posterior distribution: Bayesian PCA

$$\left(\mu_{ij}^{(l)} | X_{ij}^{(l)}\right) = \mathcal{N}(\Phi_l X_{ij}^{(l)}, \sigma^2 \Phi_l)$$

$\Rightarrow$ Data Augmentation (Tanner and Wong, 1987)

(I) given μ and σ^2, imputing the missing values X_{ij} by a draw from the predictive distribution $\mathcal{N}(\mu_{ij}, \sigma^2)$

(P) drawing μ_{ij} from its posterior distribution

Properties

$\Rightarrow$ Joint model: $X_{i.} \sim \mathcal{N}(\mu, \Sigma)$

1. Bootstrap rows: $X^1, \ldots, X^B$
   EM algorithm: $(\hat{\mu}^1, \hat{\Sigma}^1), \ldots, (\hat{\mu}^B, \hat{\Sigma}^B)$
2. Imputation: X_{ij}^b drawn from $\mathcal{N}\left(\hat{\mu}^b, \hat{\Sigma}^b\right)$

$\Rightarrow$ Conditional modeling: one model/variable

1. For a variable j
   2.1 $(\beta^{-j}, \sigma^{-j})$ drawn from a bootstrap or a posterior distribution
   2.2 Imputation: stochastic regression X_{ij} from $\mathcal{N}\left(X_{-j}\beta^{-j}, \sigma^{-j}\right)$
2. Cycling through variables - Repeat B times

With continuous variables and a regression/variable: $\mathcal{N}(\mu, \Sigma)$

Properties

$\Rightarrow$ Joint model: $X_{i.} \sim \mathcal{N}(\mu, \Sigma)$

1. Bootstrap rows: $X^1, \ldots, X^B$
   EM algorithm: $(\hat{\mu}^1, \hat{\Sigma}^1), \ldots, (\hat{\mu}^B, \hat{\Sigma}^B)$
2. Imputation: X_{ij}^b drawn from $\mathcal{N}\left(\hat{\mu}^b, \hat{\Sigma}^b\right)$

$\Rightarrow$ Conditional modeling: one model/variable

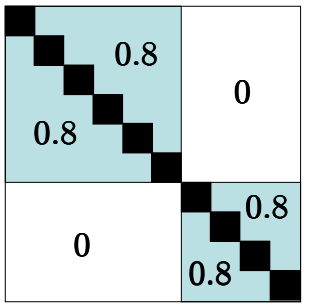
1. For a variable j
   2.1 $(\beta^{-j}, \sigma^{-j})$ drawn from a bootstrap or a posterior distribution
   2.2 Imputation: stochastic regression X_{ij} from $\mathcal{N}\left(X_{-j}\beta^{-j}, \sigma^{-j}\right)$
2. Cycling through variables - Repeat B times

With continuous variables and a regression/variable: $\mathcal{N}(\mu, \Sigma)$
More flexible? Tedious?

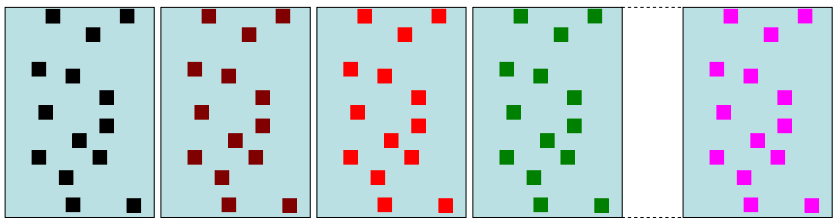
Simulations

The simulated data $\mathcal{N}(\mu, \Sigma)$

- 2 underlying dimensions (control k)
- n (30,200), p (6,60), ρ (0.3,0.8), %NA (10,30)



$\Rightarrow$ Imputation with $B = 100$ imputed tables with PCA, JM, CM



Estimate (analysis model): $\hat{\theta}_b, \widehat{Var}\left(\hat{\theta}_b\right)$: $\theta_1 = \mathbb{E}[Y], \theta_2 = \beta_1$

Rubin: $\hat{\theta} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$ $T = \frac{1}{B}\sum_b \widehat{Var}\left(\hat{\theta}_b\right) + \frac{1}{B-1}\sum_b\left(\hat{\theta}_b - \hat{\theta}\right)^2$

$\Rightarrow$ Bias, CI width, coverage - 1000 simulations

Results for the expectation

	parameters				confidence interval width			coverage		
	n	p	ρ	%	*Joint*	*Cond*	*MIPCA*	*Joint*	*Cond*	*MIPCA*
1	30	6	0.3	0.1	0.803	0.805	0.781	0.955	0.953	0.950
2	30	6	0.3	0.3		1.010	0.898		0.971	0.949
3	30	6	0.9	0.1	0.763	0.759	0.756	0.952	0.95	0.949
4	30	6	0.9	0.3		0.818	0.783		0.965	0.953
5	30	60	0.3	0.1			0.775			0.955
6	30	60	0.3	0.3			0.864			0.952
7	30	60	0.9	0.1			0.742			0.953
8	30	60	0.9	0.3			0.759			0.954
9	200	6	0.3	0.1	0.291	0.294	0.292	0.947	0.947	0.946
10	200	6	0.3	0.3	0.328	0.334	0.325	0.954	0.959	0.952
11	200	6	0.9	0.1	0.281	0.281	0.281	0.953	0.95	0.952
12	200	6	0.9	0.3	0.288	0.289	0.288	0.948	0.951	0.951
13	200	60	0.3	0.1		0.304	0.289		0.957	0.945
14	200	60	0.3	0.3		0.384	0.313		0.981	0.958
15	200	60	0.9	0.1		0.282	0.279		0.951	0.948
16	200	60	0.9	0.3		0.296	0.283		0.958	0.952

$\Rightarrow$ Good estimates of θ and coverage ≈ 0.95: variability due to missing is taken into account

$\Rightarrow$ PCA: small - large n/p; strong - weak relation; low-high % NA

Iterative Random Forests imputation

1. Initial imputation: mean imputation
2. Fit a RF X_1^{obs} on X_{-1} and then predict X_1^{miss}
   Fit a RF X_2^{obs} on X_{-2} and then predict X_2^{miss}
   ...
   cycling through variables
3. Repeat until convergence

$\Rightarrow$ Conditional modeling based on RF

- number of trees: 100
- number of variables randomly selected at each node $\sqrt{p}$
- number of iterations: 4-5

$\Rightarrow$ Good for complex relationships between variables

Random Forests versus PCA

	Feat1	Feat2	Feat3	Feat4	Feat5...
C1	1	1	1	1	1
C2	1	1	1	1	1
C3	2	2	2	2	2
C4	2	2	2	2	2
C5	3	3	3	3	3
C6	3	3	3	3	3
C7	4	4	4	4	4
C8	4	4	4	4	4
C9	5	5	5	5	5
C10	5	5	5	5	5
C11	6	6	6	6	6
C12	6	6	6	6	6
C13	7	7	7	7	7
C14	7	7	7	7	7
Igor	8	NA	NA	8	8
Frank	8	NA	NA	8	8
Bertrand	9	NA	NA	9	9
Alex	9	NA	NA	9	9
Yohann	10	NA	NA	10	10
Jean	10	NA	NA	10	10

Random Forests versus PCA

	Feat1	Feat2	Feat3	Feat4	Feat5
C1	1	1.0	1.00	1	1
C2	1	1.0	1.00	1	1
C3	2	2.0	2.00	2	2
C4	2	2.0	2.00	2	2
C5	3	3.0	3.00	3	3
C6	3	3.0	3.00	3	3
C7	4	4.0	4.00	4	4
C8	4	4.0	4.00	4	4
C9	5	5.0	5.00	5	5
C10	5	5.0	5.00	5	5
C11	6	6.0	6.00	6	6
C12	6	6.0	6.00	6	6
C13	7	7.0	7.00	7	7
C14	7	7.0	7.00	7	7
Igor	8	6.87	6.87	8	8
Frank	8	6.87	6.87	8	8
Bertrand	9	6.87	6.87	9	9
Alex	9	6.87	6.87	9	9
Yohann	10	6.87	6.87	10	10
Jean	10	6.87	6.87	10	10

	Feat1	Feat2	Feat3	Feat4	Feat5
C1	1	1	1	1	1
C2	1	1	1	1	1
C3	2	2	2	2	2
C4	2	2	2	2	2
C5	3	3	3	3	3
C6	3	3	3	3	3
C7	4	4	4	4	4
C8	4	4	4	4	4
C9	5	5	5	5	5
C10	5	5	5	5	5
C11	6	6	6	6	6
C12	6	6	6	6	6
C13	7	7	7	7	7
C14	7	7	7	7	7
Igor	8	8	8	8	8
Frank	8	8	8	8	8
Bertrand	9	9	9	9	9
Alex	9	9	9	9	9
Yohann	10	10	10	10	10
Jean	10	10	10	10	10

$\Rightarrow$ with Random Forests $\Rightarrow$ with PCA

(Stekhoven, Buhlmann, 2011 - Bartlett, Carpenter, 2014)

$\Rightarrow$ Non linear relationship well handled by forests

MI using the loglinear model

- Hypothesis $X = (x_{ijk})_{i,j,k}$:
  $X|\theta \sim \mathcal{M}(n, \theta)$ where:

$$log(\theta_{ijk}) = \lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_S^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

1. Variability of the parameters
   - prior on θ : $\theta|\theta \in \Theta \sim \mathcal{D}(\alpha)$
   - posterior: $\theta|x, \theta \in \Theta \sim \mathcal{D}(\alpha')$
   - Data Augmentation (M.A. Tanner, W.H. Wong, 1987)
2. Imputation according to the loglinear model using the set of M parameters

- Implemented: `R` package `cat` (J.L. Schafer)

MI using a DPMPM model (Si and Reiter, 2013)

- Hypothesis: $\mathbb{P}(X = (x_1, \ldots, x_K); \theta) = \sum_{\ell=1}^{L} \left(\theta_\ell \prod_{k=1}^{K} \theta_{x_k}^{(\ell)} \right)$

1. Variability of the parameters:
   - a hierarchic prior on θ:
     $\alpha \sim \mathcal{G}(.25, .25)$ $\zeta_\ell \sim \mathcal{B}(1, \alpha)$ $\theta_\ell = \zeta_\ell \prod_{g<\ell}(1 - \zeta_g)$ for ℓ in $1, \ldots, \infty$
   - posterior on θ: untractable
     $\rightarrow$ Gibbs sampler and Data Augmentation
2. Imputation according to the mixture model using the set of M parameters

- Implemented: `R` package `mi` (Gelman *et al.*)

Conclusion

Multiple imputation methods for continuous and categorical data using dimensionality reduction method

Properties:

- requires a small number of parameters
- captures the relationships between variables
- captures the similarities between individuals

From a practical point of view:

- can be applied on data sets of various dimensions
- provides correct inferences for analysis model based on relationships between pairs of variables
- requires to choose the number of dimensions S

Perspective:

- mixed data

To conclude

Take home message:

- Principal component methods powerful to impute large data sets with continuous, categorical variables: reduce the dimensionality and take into account the similarities between rows and relationship between variables.
- Single imputation aims to complete a dataset as best as possible (prediction). Multiple imputation aims to perform other statistical methods after and to estimate parameters and their variability taking into account the missing values uncertainty.
- R packages FactoMineR, MissMDA, denoiseR

*"**The idea of imputation is both seductive and dangerous**. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases."*

Outline

Feature Noising = Ridge = Shrinkage (Proof)

$$\begin{aligned}\mathbb{E}_\varepsilon\left[\|X-(X+\varepsilon)B\|_2^2\right] &= \|X-XB\|_2^2 + \mathbb{E}_\varepsilon\left[\|\varepsilon B\|_2^2\right] \\ &= \|X-XB\|_2^2 + \sum_{i,j,k} B_{ij}^2 \operatorname{Var}\left[\varepsilon_{jk}\right] \\ &= \|X-XB\|_2^2 + n\sigma^2\|B\|_2^2\end{aligned}$$

Let $X = UDV^\top$ be the SVD of X, $\lambda = n\sigma^2$

$$\widehat{B}_\lambda = V\widetilde{B}_\lambda V^T, \text{ where } \widetilde{B}_\lambda = \operatorname{argmin}_B\left\{\|D-DB\|_2^2 + \lambda\|B\|_2^2\right\}$$

$$\widetilde{B}_{ii} = \operatorname{argmin}_{B_{ii}}\left\{(1-B_{ii})^2 D_{ii}^2 + \lambda B_{ii}^2\right\} = \frac{D_{ii}^2}{\lambda + D_{ii}^2}$$

$$\hat{\mu}_\lambda^{(k)} = \sum_{i=1}^{n} U_{i\cdot}\frac{D_{ii}}{1+\lambda/D_{ii}^2}V_{i\cdot}^\top$$

Simulations $X_{200\times500} = \mu + \varepsilon$

k	SNR	Bootstrap		TSVD	Adaptive (λ, γ)	Asympt	Soft (λ)
		SA (σ, k)	ISA (σ)	(k)	(σ) - SURE	(σ)	(σ) - SURE
MSE							
10	4	**0.004**	**0.004**	**0.004**	**0.004**	**0.004**	0.008
100	4	**0.037**	**0.036**	**0.037**	**0.037**	**0.037**	0.045
10	2	**0.017**	**0.017**	**0.017**	**0.017**	**0.017**	0.033
100	2	**0.142**	**0.143**	0.152	**0.142**	0.146	0.156
10	1	**0.067**	**0.067**	0.072	**0.067**	**0.067**	0.116
100	1	0.511	0.775	0.733	**0.448**	0.600	**0.448**
10	0.5	0.277	**0.251**	0.321	**0.253**	**0.250**	0.353
100	0.5	1.600	1.000	3.164	**0.852**	0.961	**0.852**
Rank							
10	4		**10**		11	**10**	65
100	4		**100**		103	**100**	193
10	2		**10**		11	**10**	63
100	2		**100**		114	**100**	181
10	1		**10**		11	**10**	59
100	1		29.6		154	**64**	154
10	0.5		**10**		15	**10**	51
100	0.5		0		87	15	86

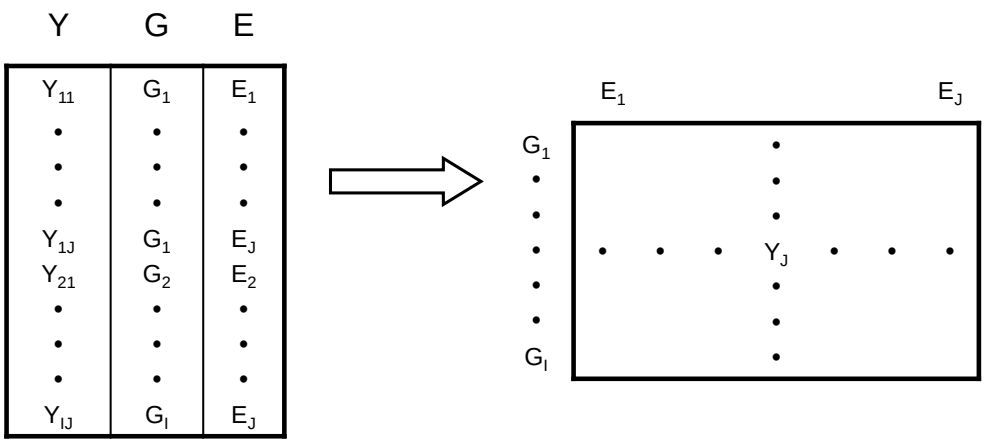
Genotypes - Environement data

16 triticales (6 complete - 8 substituted - 2 checks)
10 locations in Spain (heterogeneous pH)

$\Rightarrow$ Adaptation of the triticales to the Spanish soil

$\Rightarrow$ AMMI (Additive Main effects and Multiplicative Interaction)

$$y_{ij} = \mu + \alpha_i + \beta_j + \sum_{l=1}^{k} d_l u_{il} v_{jl} + \varepsilon_{ij}$$

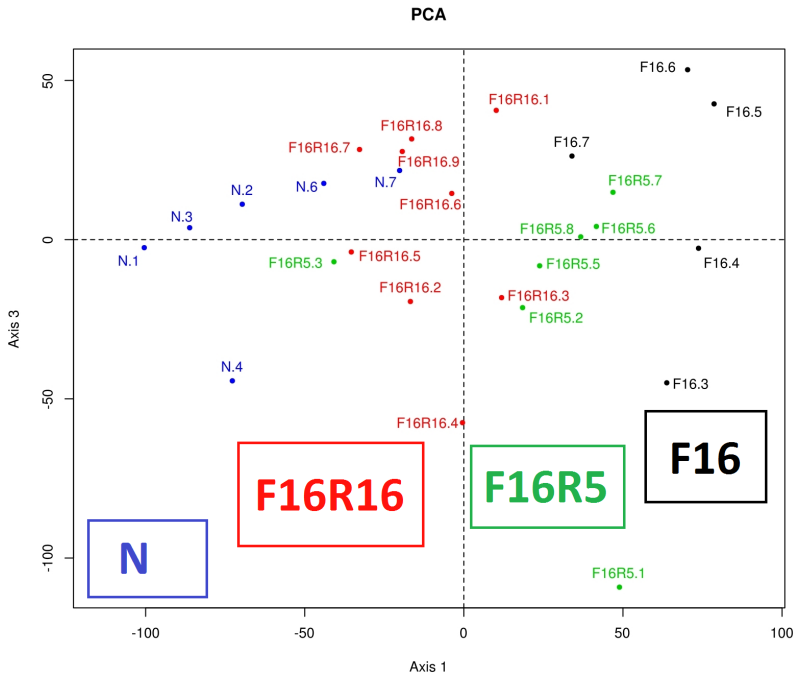


$\Rightarrow$ PCA on the residuals matrix of the 2-way anova

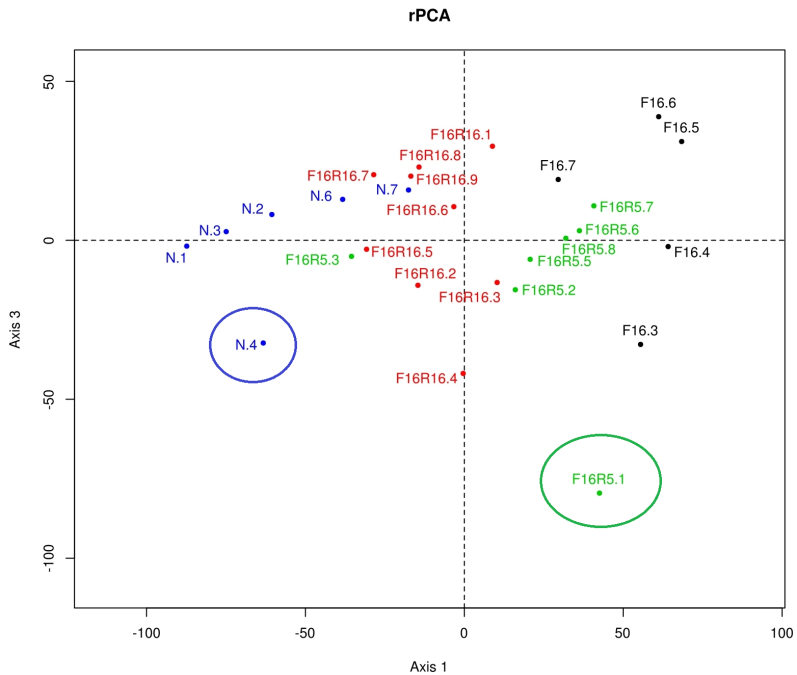
Expression data

- 27 chickens submitted to 4 nutritional statuses
- 12664 gene expressions
- $X = \mu + \varepsilon$. PCA $\hat{\mu}_k = \sum_{l=1}^{k} u_l \, d_l \, v_l^\top$

PCA



rPCA



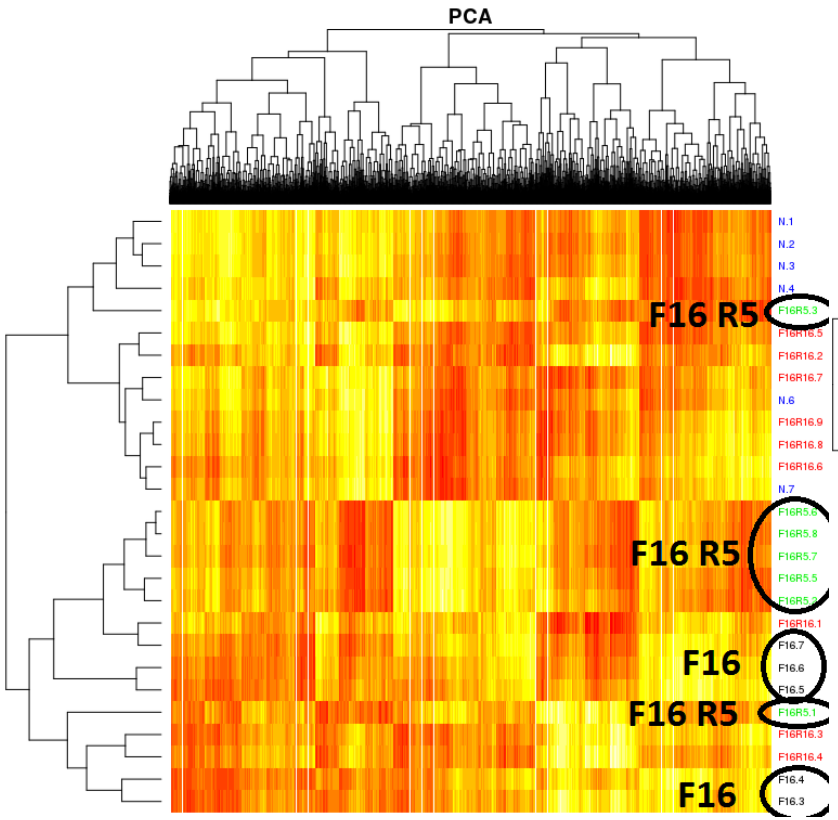
Scores $F = UD$

rPCA

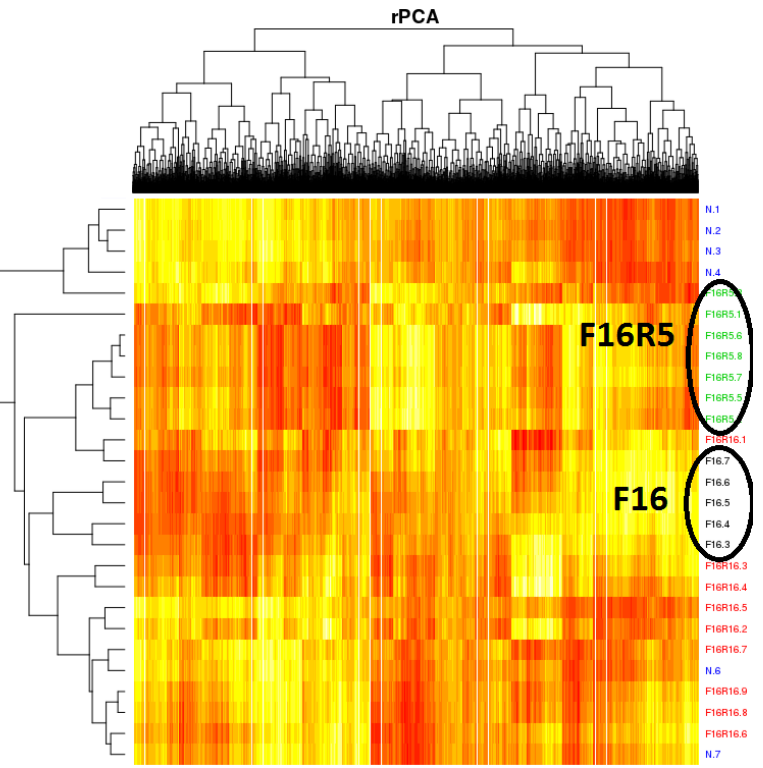
Expression data

Clustering on the denoised matrix $\hat{\mu}_k = \sum_{l=1}^{k} u_l \, d_l \, v_l^\top$

PCA



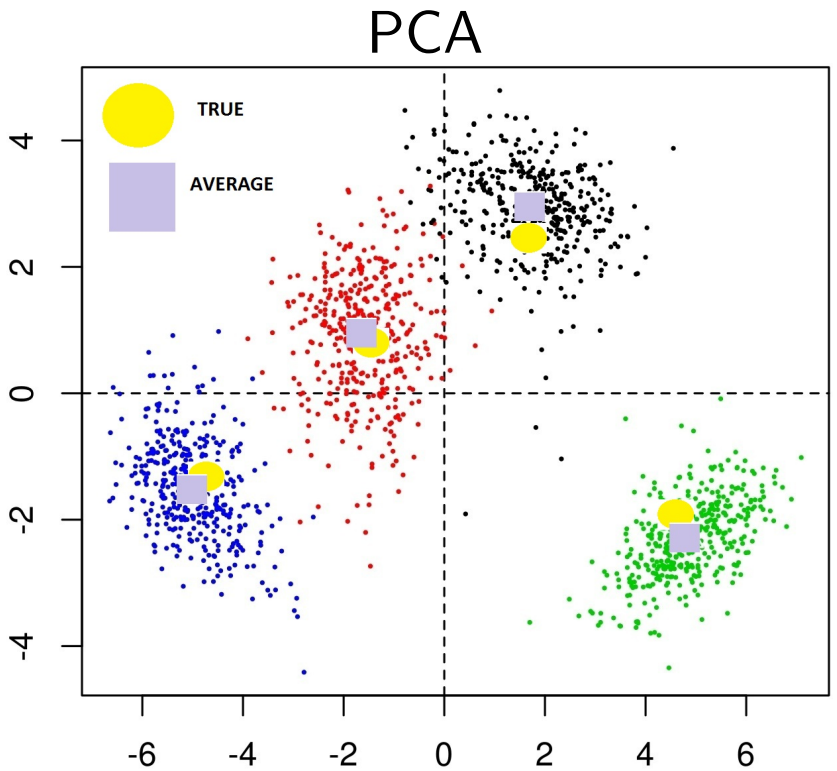
rPCA

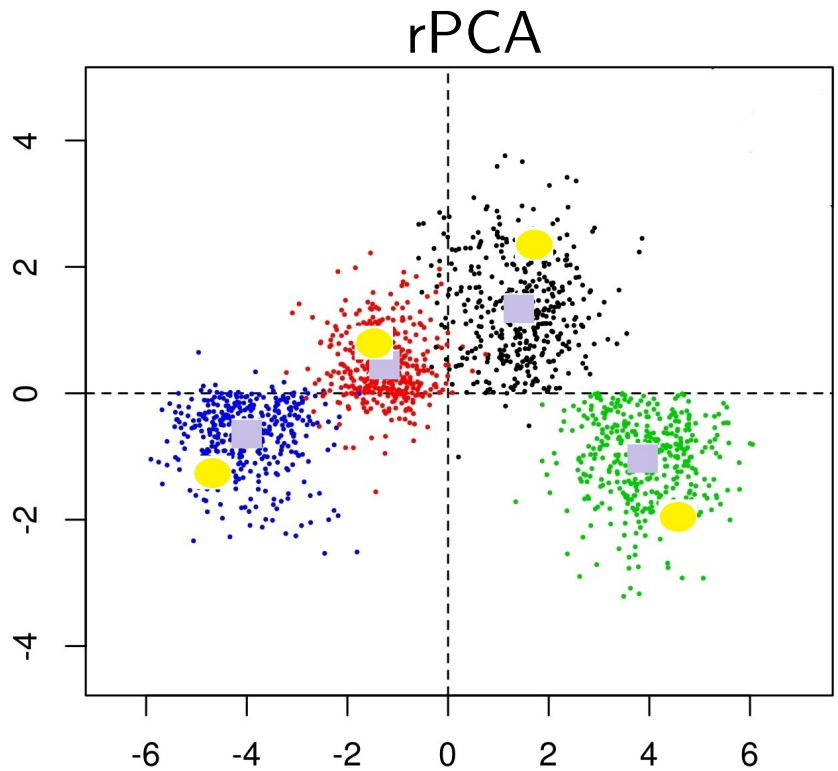


Bi-clustering from rPCA better find the 4 nutritional statuses!

Illustration of the regularization

$$X_{4\times10}^{sim} = \mu_{4\times10} + \varepsilon_{4\times10}^{sim} \qquad sim = 1, ..., 1000$$



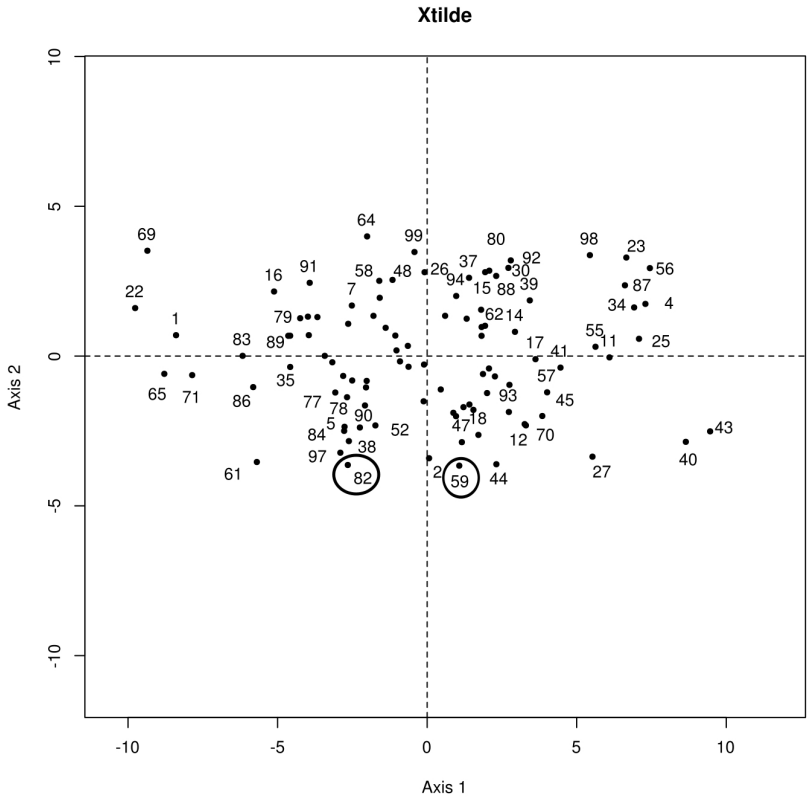


$\Rightarrow$ rPCA more biased less variable

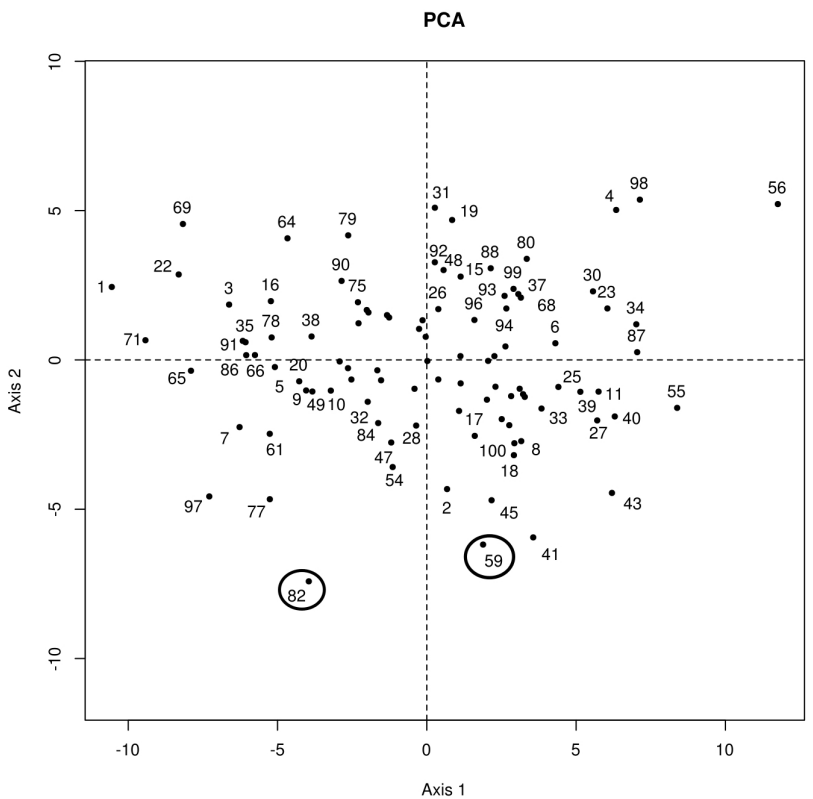
Rows representation

100 rows, 20 variables, 2 dimensions, SNR=0.8, $d_1/d_2 = 4$

True configuration



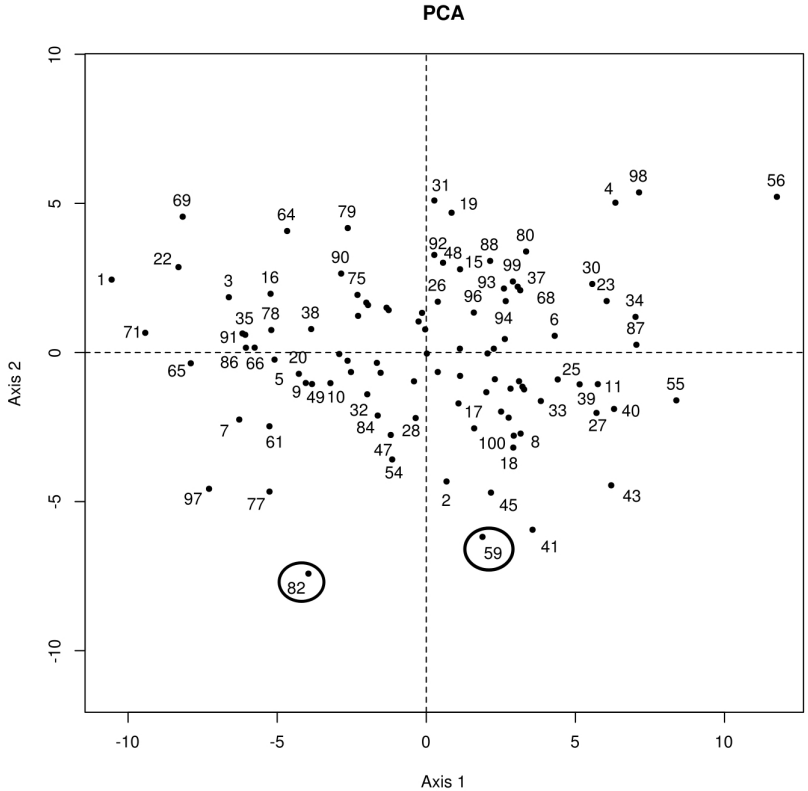
PCA



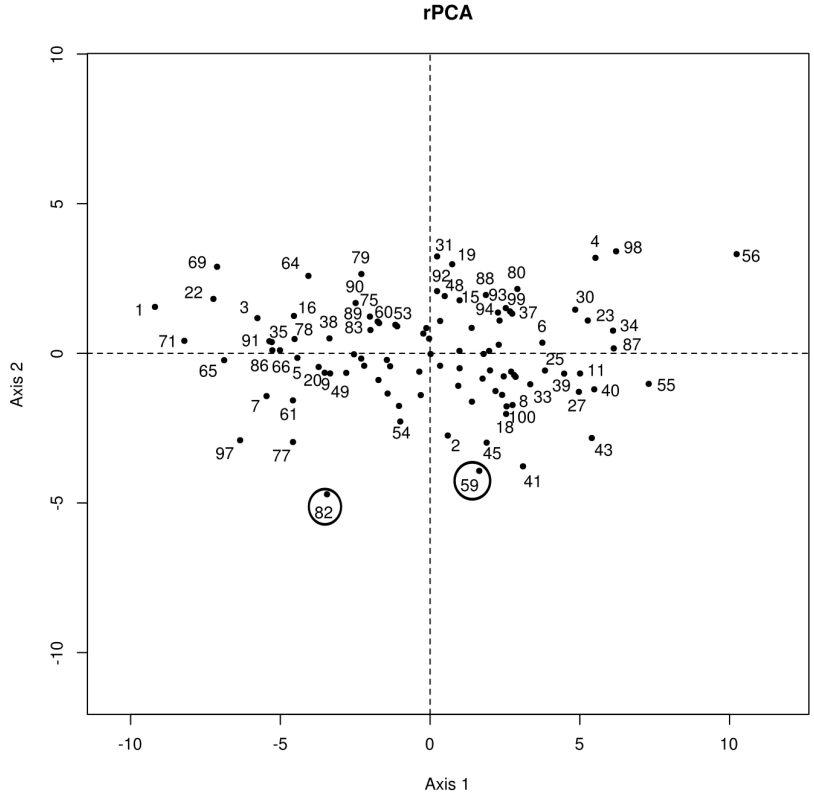
Rows representation

100 rows, 20 variables, 2 dimensions, SNR=0.8, $d_1/d_2 = 4$

PCA



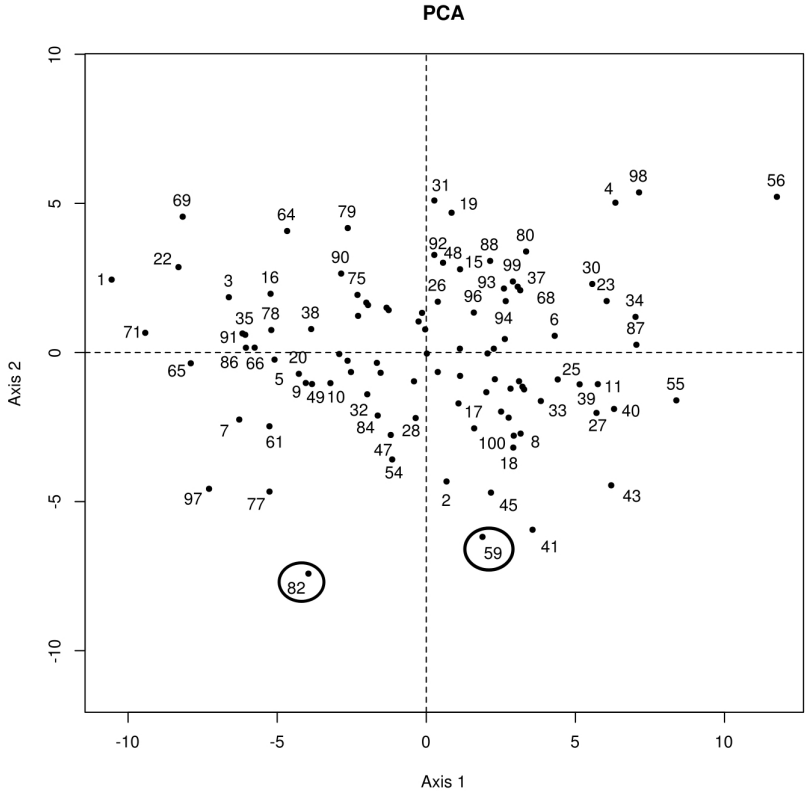
Non-linear



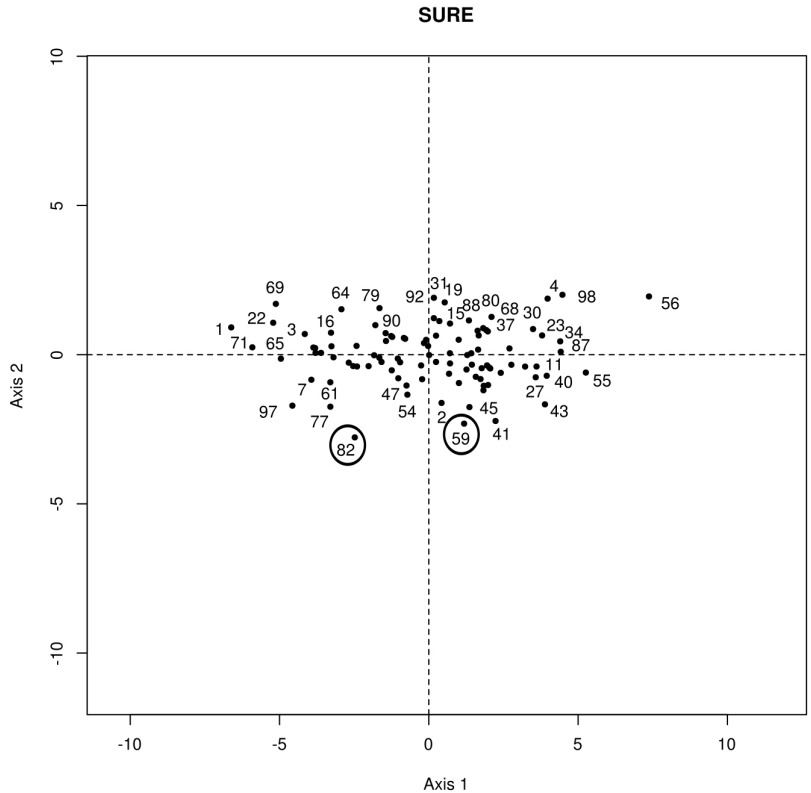
Rows representation

100 rows, 20 variables, 2 dimensions, SNR=0.8, $d_1/d_2 = 4$

PCA



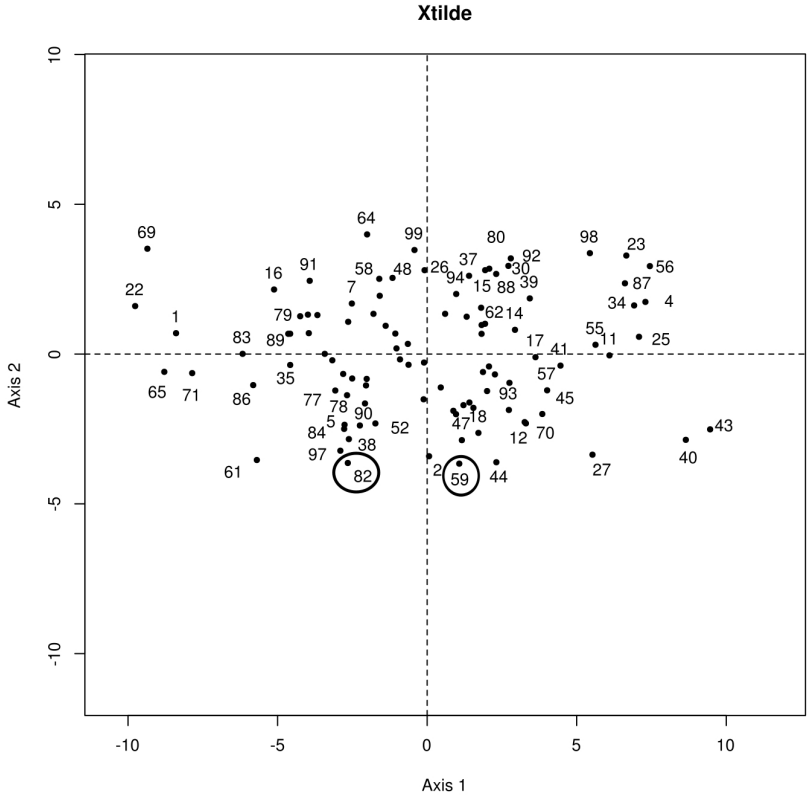
Soft



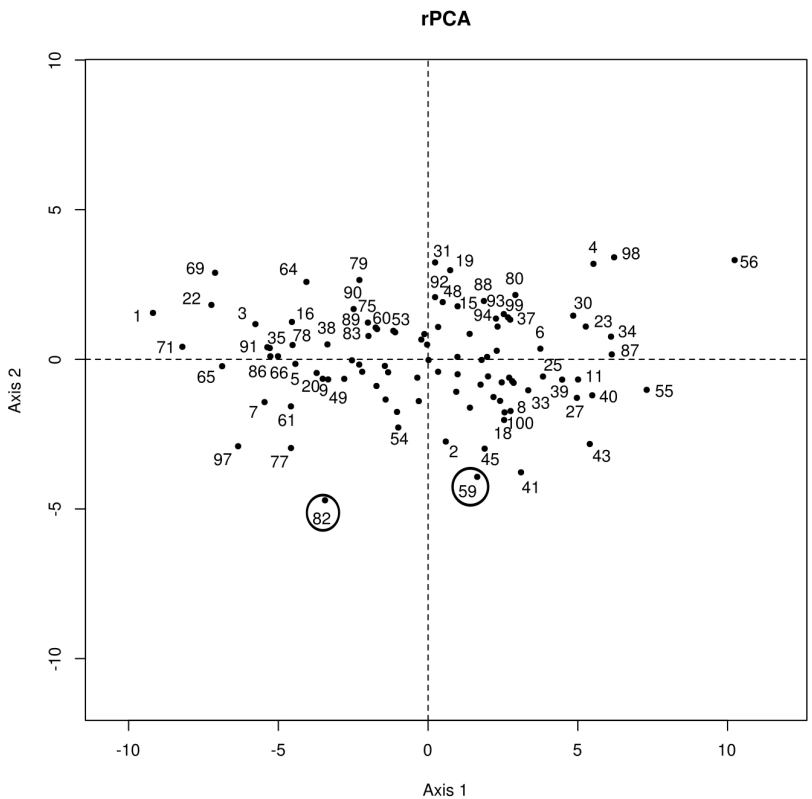
Rows representation

100 rows, 20 variables, 2 dimensions, SNR=0.8, $d_1/d_2 = 4$

True configuration



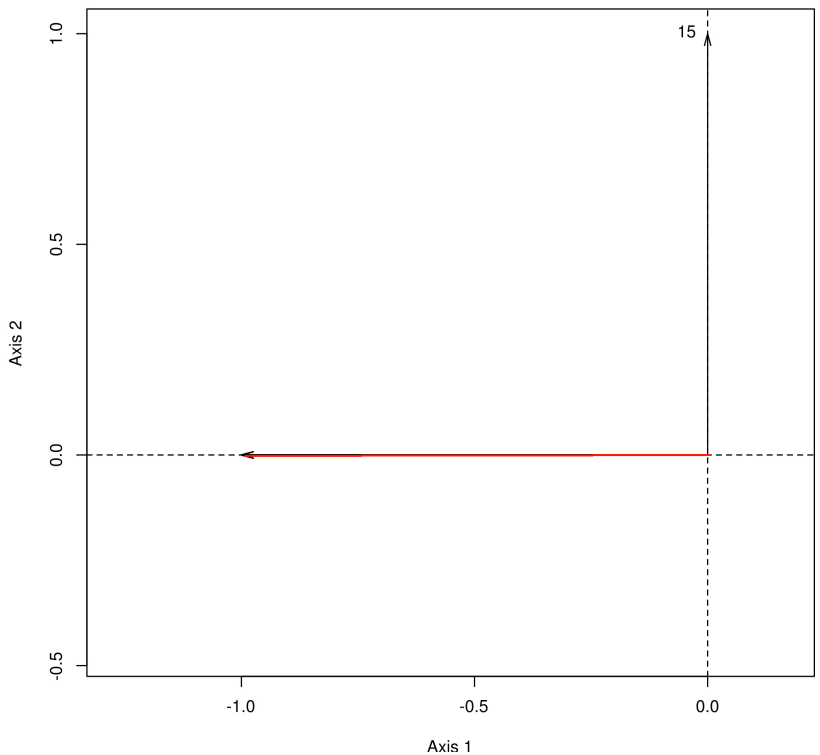
Non-linear



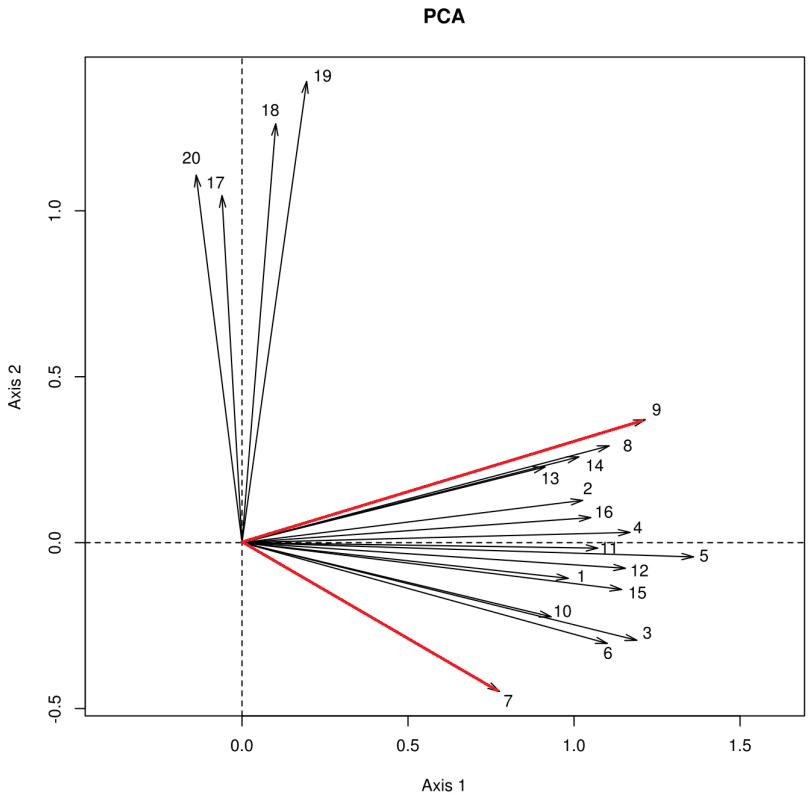
$\Rightarrow$ Improve the recovery of the inner-product matrix $\mu\mu'$

Variables representation

True configuration



PCA

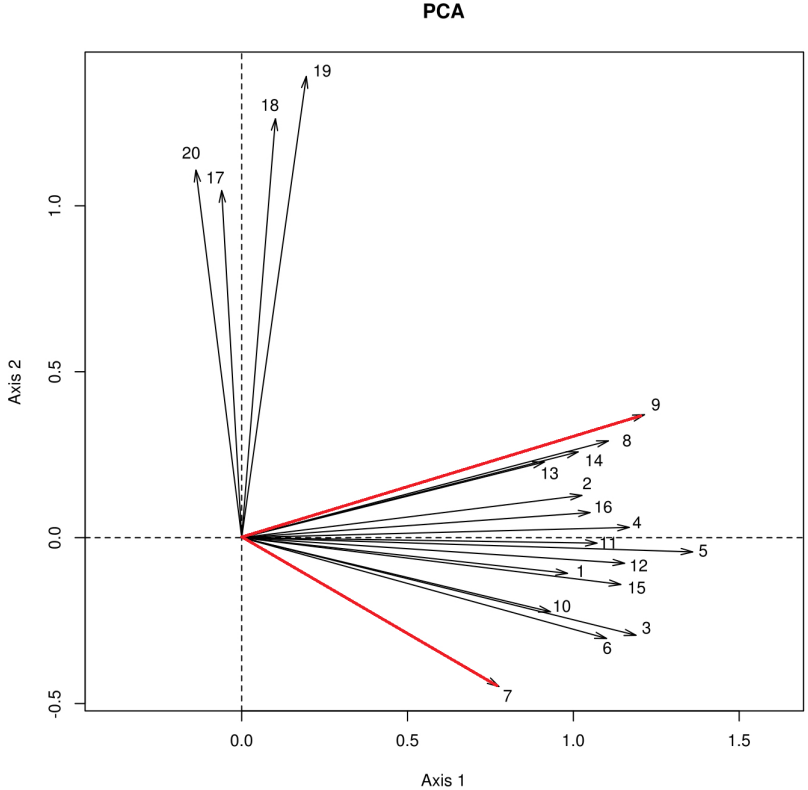


$\text{cor}(X_7, X_9) = 1$

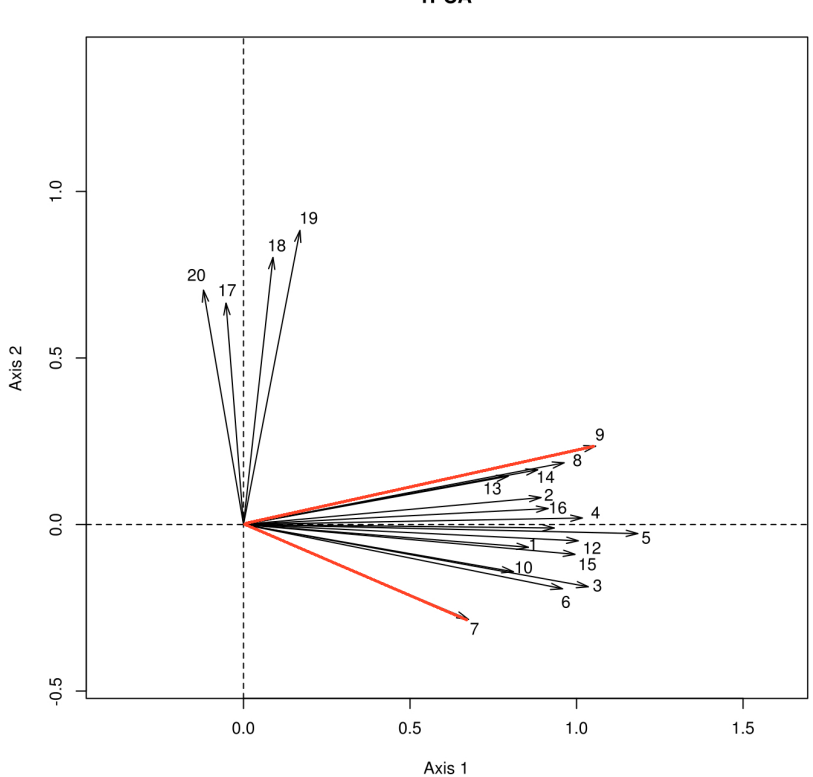
0.68

Variables representation

PCA



Non-linear

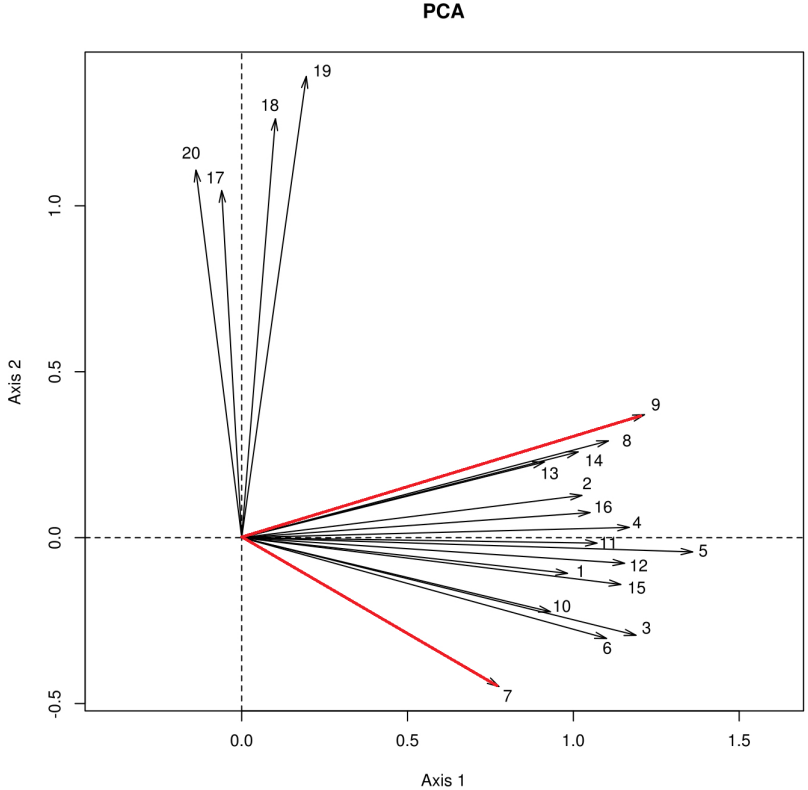


$\text{cor}(X_7, X_9) = 0.68$

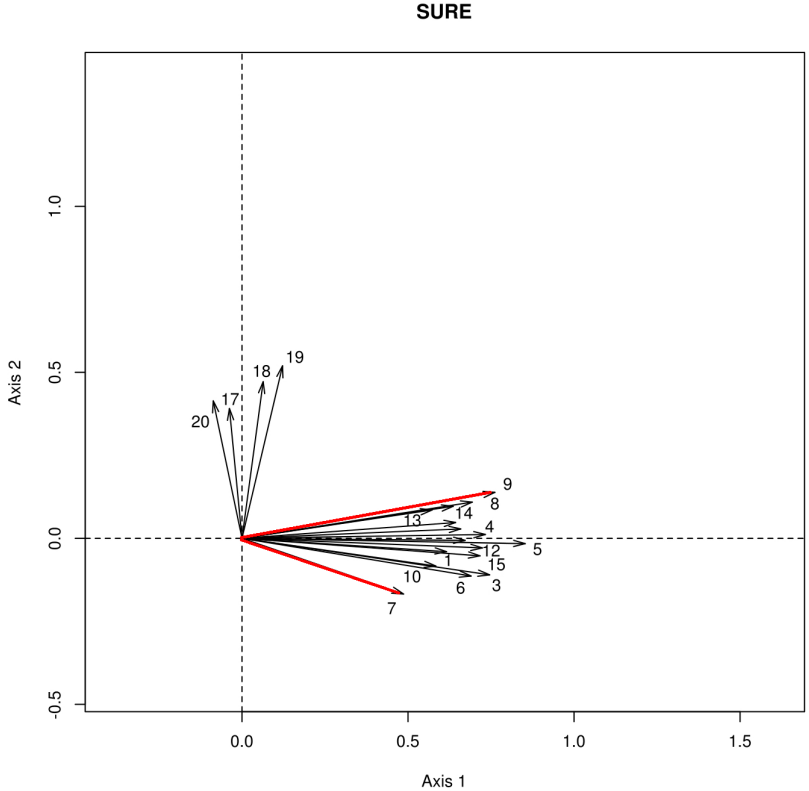
0.81

Variables representation

PCA



Soft

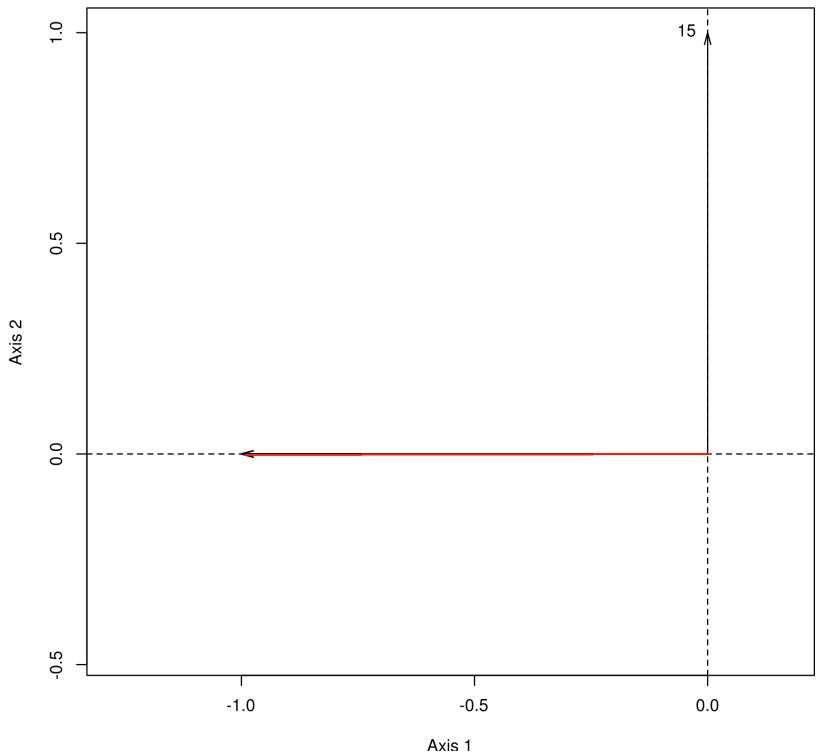


$\text{cor}(X_7, X_9) = 0.68$

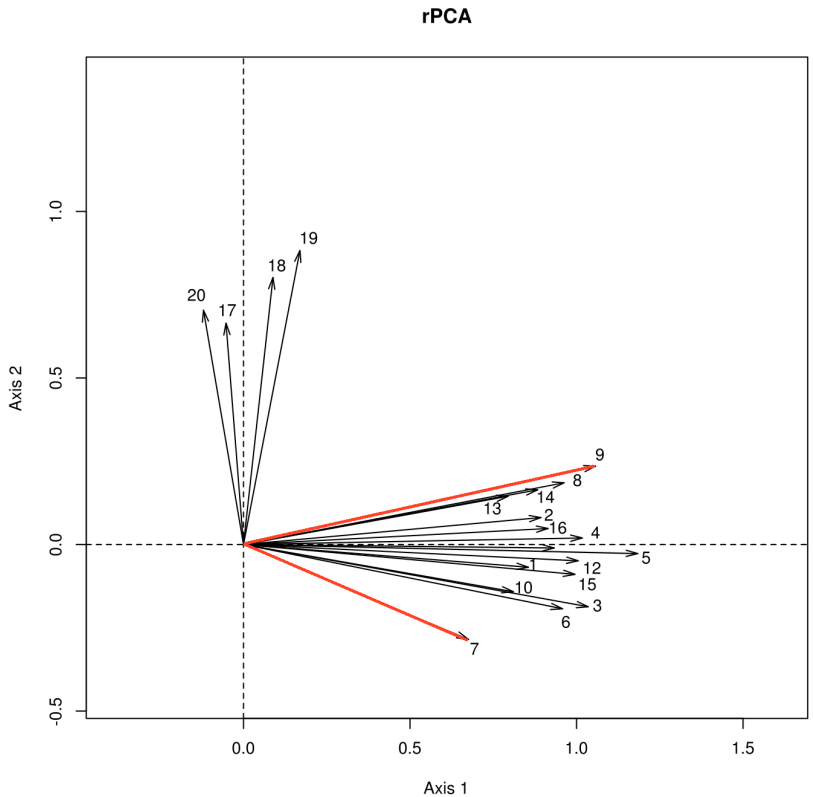
0.82

Variables representation

True configuration



Non-linear



$\text{cor}(X_7, X_9) = 1$ 0.81

$\Rightarrow$ Improve the recovery of the covariance matrix $\mu'\mu$

Drop-out training

Drop-out (Hinton *et al.* 2012) randomly omits subsets of features at each iteration of a training algorithm (improves neural networks)

Wager, *et al.* (2013). Dropout training as adaptive regularization.

- GLM: equivalence between noising schemes and regularization
- Full potential with drop-out noise: $X_{ij} = 0$ with prob δ and $X_{ij}/(1-\delta) \rightarrow$ nice penalty (ex. logistic reg - rare features)

Wager, *et al.* (2014). Altitude training: bounds for single layer.

- *"marathon runner who practices on altitude: once a classifier learns to perform well on training examples corrupted by dropout, it will do very well on the uncorrupted test"*

Count data

$\Rightarrow$ Model: $X \in \mathbb{R}^{n \times p} \sim \mathcal{L}(\mu)$ with $\mathbb{E}[X] = \mu$ of low-rank k

Poisson noise: $X_{ij} \sim \text{Poisson}(\mu_{ij})$

$\Rightarrow \hat{\mu}_k = \sum_{l=1}^{k} u_l \, d_l \, v_l^\top$

Count data

$\Rightarrow$ Model: $X \in \mathbb{R}^{n \times p} \sim \mathcal{L}(\mu)$ with $\mathbb{E}[X] = \mu$ of low-rank k

Poisson noise: $X_{ij} \sim \text{Poisson}(\mu_{ij})$

$\Rightarrow \hat{\mu}_k = \sum_{l=1}^{k} u_l \, d_l \, v_l^\top$

$\Rightarrow$ Bootstap estimator $=$ noising: $\widetilde{X}_{ij} \sim \text{Poisson}(X_{ij})$

$$\hat{\mu}^{\text{boot}} = X\widehat{B} \quad \widehat{B} \quad = \quad \text{argmin}_B \left\{ \mathbb{E}_{\widetilde{X} \sim \mathcal{L}(X)} \left[\left\| X - \widetilde{X} B \right\|_2^2 \right] \right\}$$

$\Rightarrow$ Estimator robust to subsampling the obs used to build X

Bootstrap estimators

$\Rightarrow$ Feature noising $=$ regularization

$$\widehat{B} = \text{argmin}_B \left\{ \|X - XB\|_2^2 + \left\| S^{\frac{1}{2}} B \right\|_2^2 \right\} \quad S_{jj} = \sum_{i=1}^{n} \text{Var}_{\widetilde{X} \sim \mathcal{L}(X)} \left[\widetilde{X}_{ij} \right]$$

$\hat{\mu} = X(X^\top X + S)^{-1} X^\top X$, S diagonal with row-sums of X

$\Rightarrow$ New estimator $\hat{\mu}$ that does not reduce to singular value shrinkage: new singular vectors!

$\Rightarrow$ ISA estimator - iterative algorithm

1. $\hat{\mu} = X\hat{B}$
2. $\hat{B} = (\hat{\mu}^\top \hat{\mu} + S)^{-1} \hat{\mu}^\top \hat{\mu}$

Perfume data set

⇒ 12 luxury perfumes described by 39 words - N=1075

	floral	fruity	strong	soft	light	...
Angel	2	11	18	3	1	...
Aromatics Elixir	2	3	29	2	0	...
Chanel 5	5	0	19	3	1	...
Cinéma	14	14	3	12	9	...
Coco Mademoiselle	10	10	6	10	7	...
.....	.	.	.	.	.	

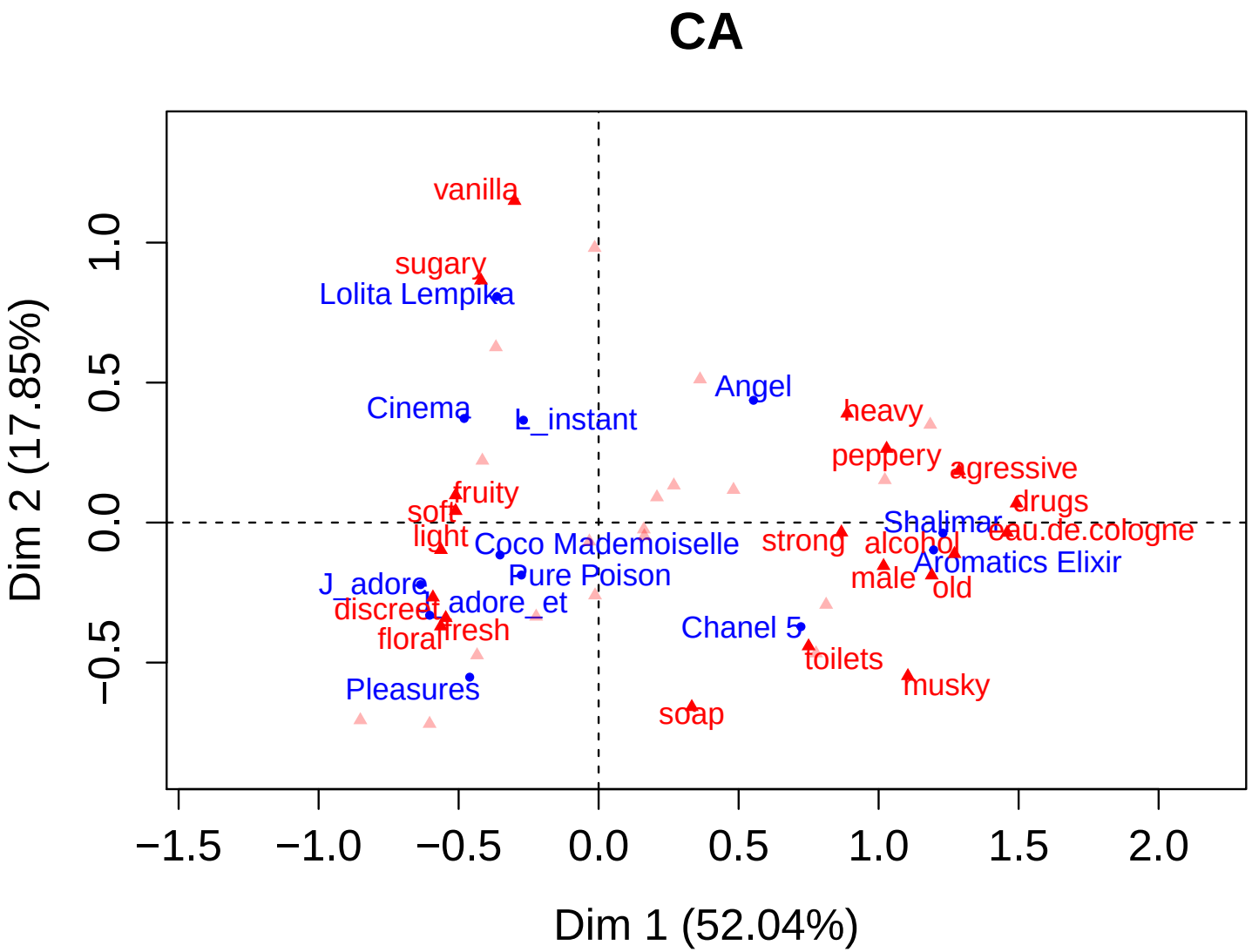
⇒ Correspondence Analysis

$$M = R^{-\frac{1}{2}} \left(X - \frac{1}{N} r c^\top \right) C^{-\frac{1}{2}}, \quad \text{where } R = \text{diag}\,(r),\ C = \text{diag}\,(c)$$

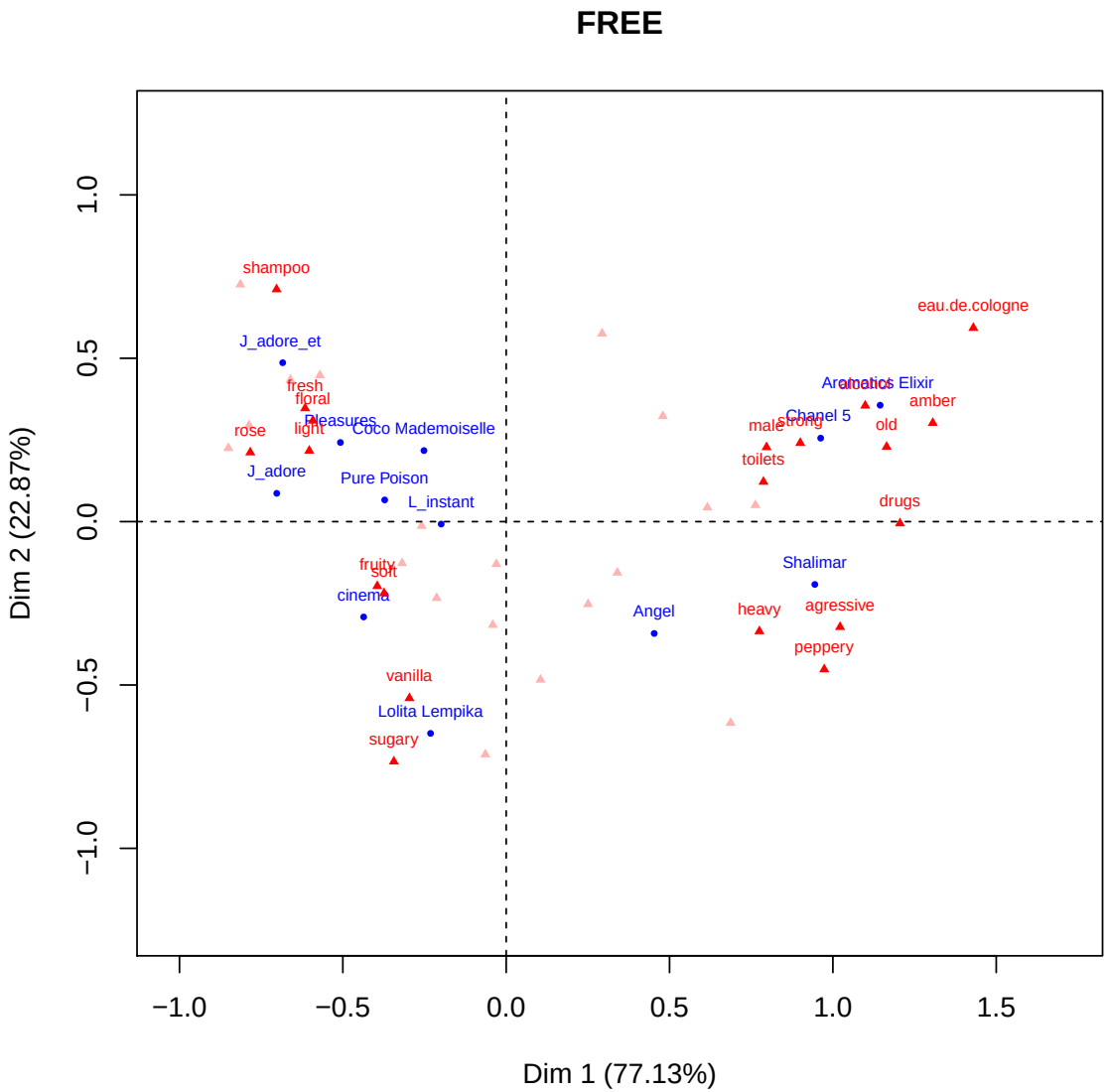
If the low-rank approx is good:

$$\hat{\mu}_k^{CA} = R^{\frac{1}{2}}\, \widehat{M}_k\, C^{\frac{1}{2}} + \frac{1}{N} r c^\top. \tag{3}$$

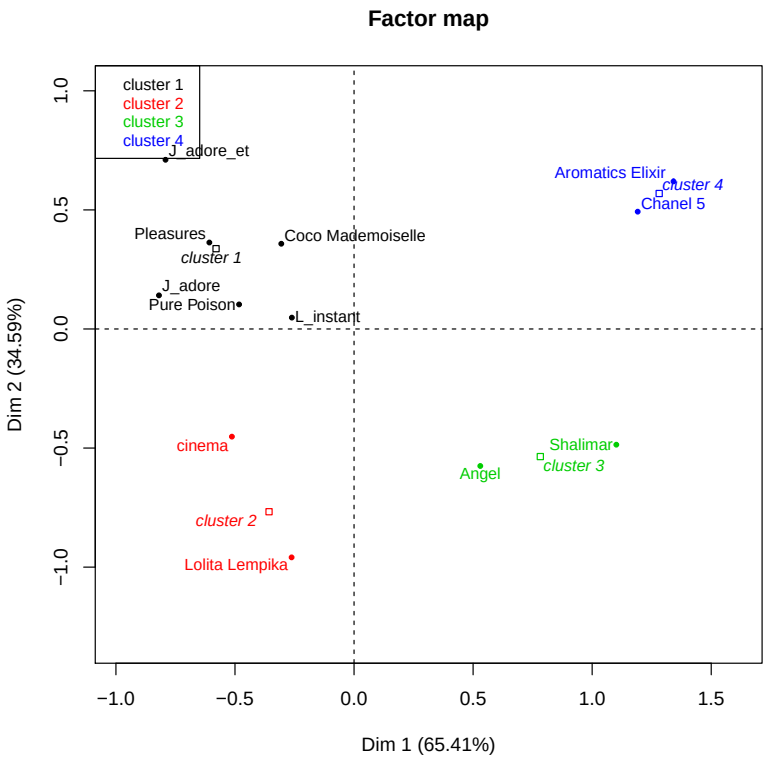
Perfume data set

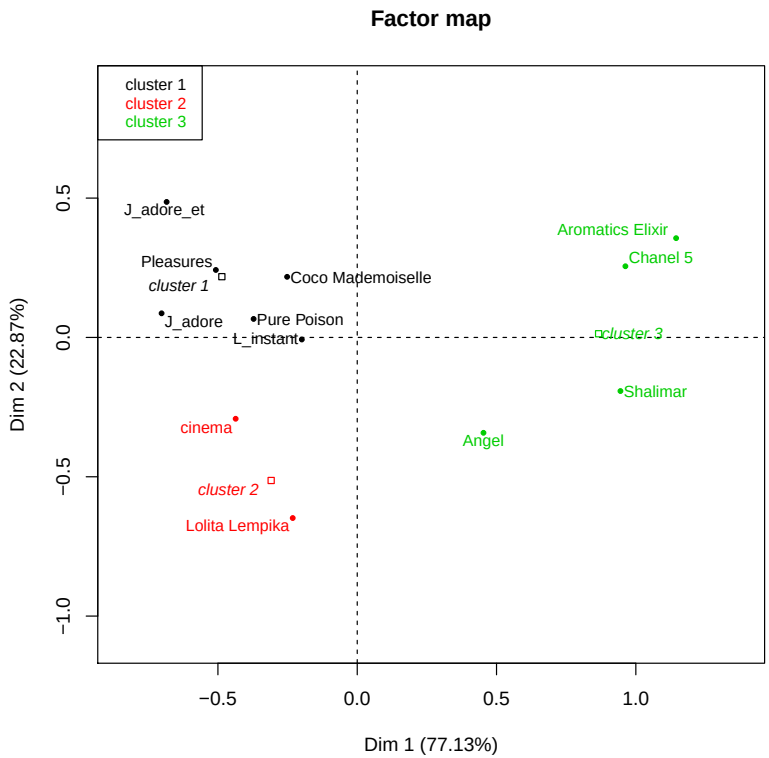


Perfume data set



Regularized Correspondence Analysis



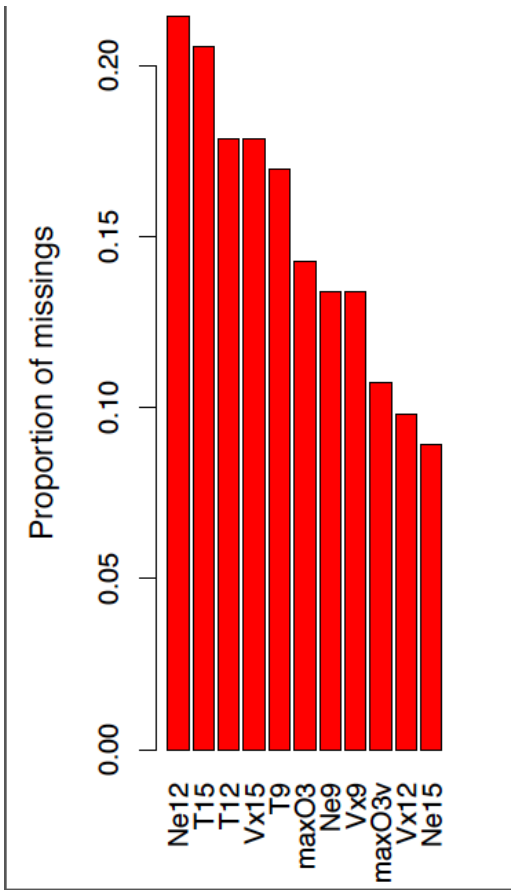


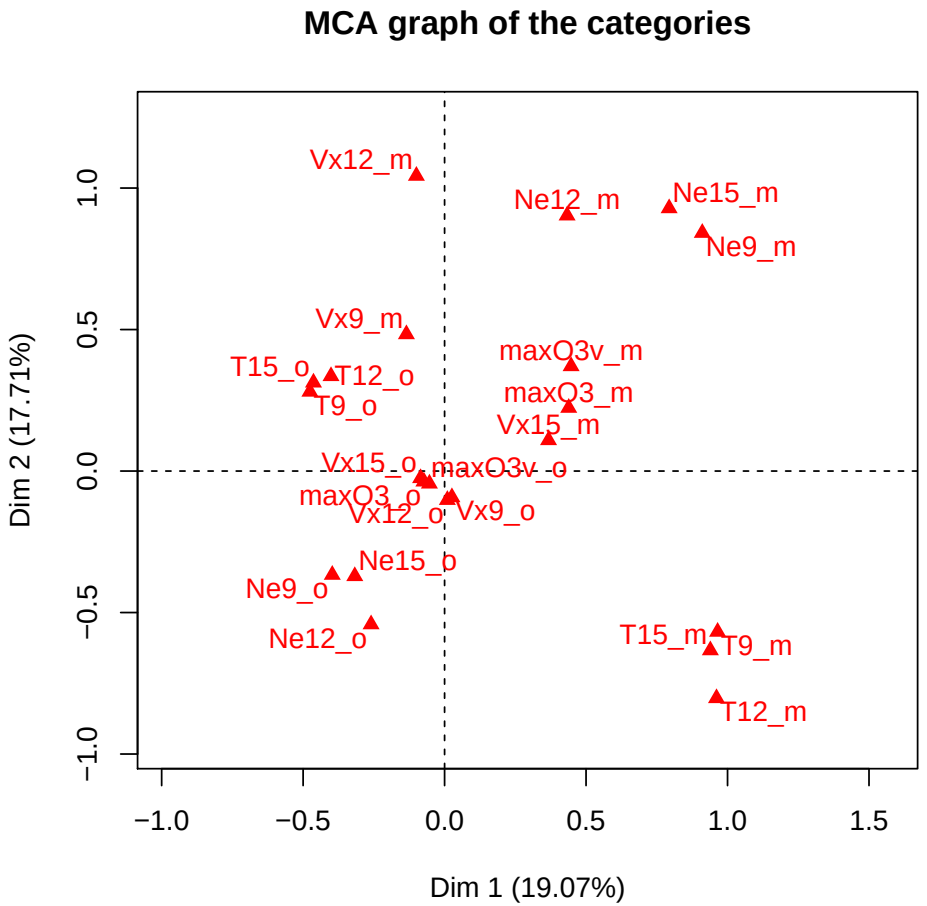
Outline

Ozone data set

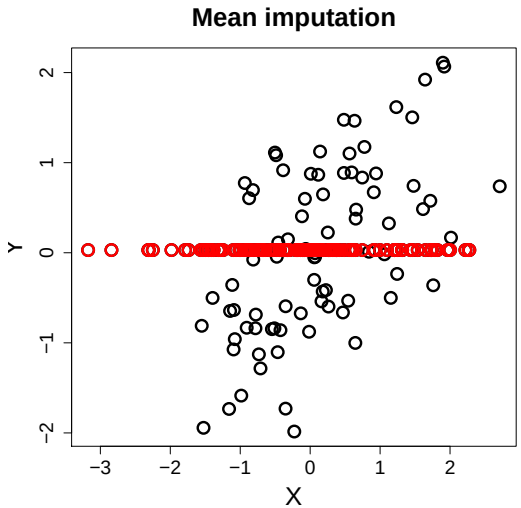
	O3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	O3v
0601	NA	15.6	18.5	18.4	4	4	8	NA	-1.7101	-0.6946	84
0602	82	17	18.4	17.7	5	5	7	NA	NA	NA	87
0603	92	NA	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82
0604	114	16.2	NA	NA	1	1	0	NA	NA	NA	92
0605	94	17.4	20.5	NA	8	8	7	-0.5	NA	-4.3301	114
0606	80	17.7	NA	18.3	NA	NA	NA	-5.6382	-5	-6	94
0607	NA	16.8	15.6	14.9	7	8	8	-4.3301	-1.8794	-3.7588	80
0610	79	14.9	17.5	18.9	5	5	4	0	-1.0419	-1.3892	NA
0611	101	NA	19.6	21.4	2	4	4	-0.766	NA	-2.2981	79
0612	NA	18.3	21.9	22.9	5	6	8	1.2856	-2.2981	-3.9392	101
0613	101	17.3	19.3	20.2	NA	NA	NA	-1.5	-1.5	-0.8682	NA
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
0919	NA	14.8	16.3	15.9	7	7	7	-4.3301	-6.0622	-5.1962	42
0920	71	15.5	18	17.4	7	7	6	-3.9392	-3.0642	0	NA
0921	96	NA	NA	NA	3	3	3	NA	NA	NA	71
0922	98	NA	NA	NA	2	2	2	4	5	4.3301	96
0923	92	14.7	17.6	18.2	1	4	6	5.1962	5.1423	3.5	98
0924	NA	13.3	17.7	17.7	NA	NA	NA	-0.9397	-0.766	-0.5	92
0925	84	13.3	17.7	17.8	3	5	6	0	-1	-1.2856	NA
0927	NA	16.2	20.8	22.1	6	5	5	-0.6946	-2	-1.3681	71
0928	99	16.9	23	22.6	NA	4	7	1.5	0.8682	0.8682	NA
0929	NA	16.9	19.8	22.1	6	5	3	-4	-3.7588	-4	99
0930	70	15.7	18.6	20.7	NA	NA	NA	0	-1.0419	-4	NA

Visualization with Multiple Correspondence Analysis





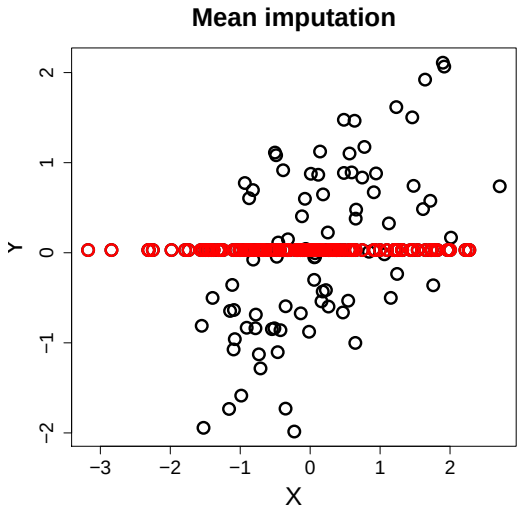
Single imputation methods

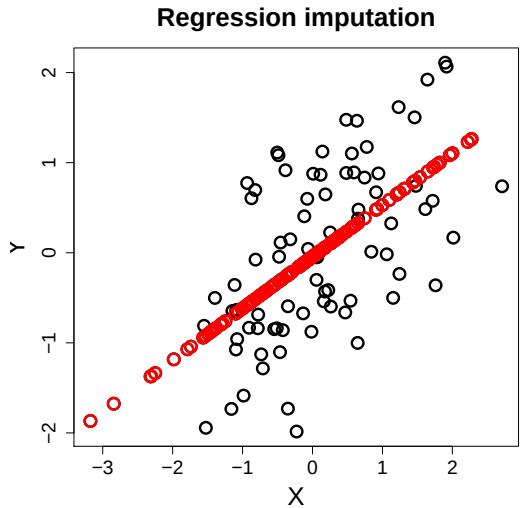


$\mu_y = 0$	0.01
$\sigma_y = 1$	0.5
$\rho = 0.6$	0.30
$CI\mu_y 95\%$	39.4

The idea of imputation is both seductive and dangerous *(Dempster and Rubin, 1983)*

Single imputation methods

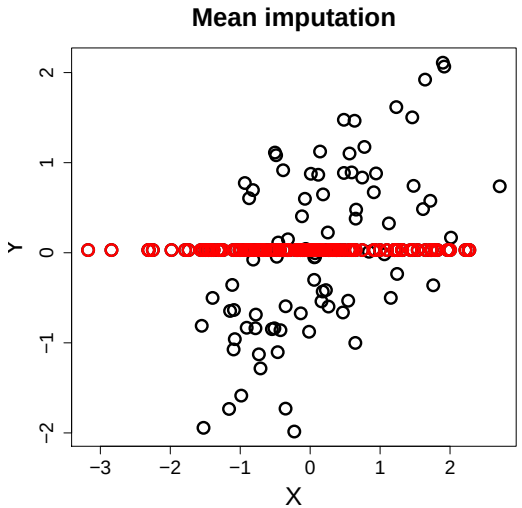


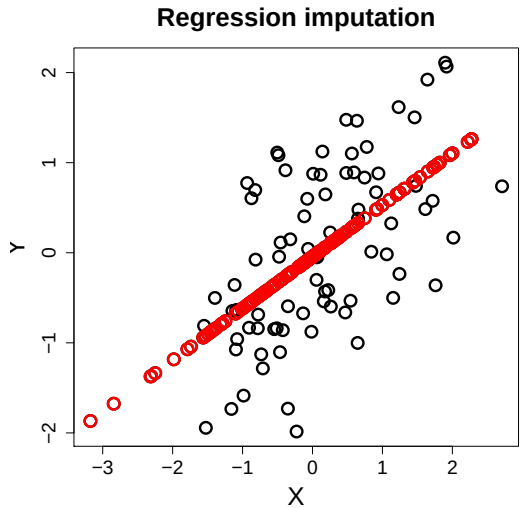


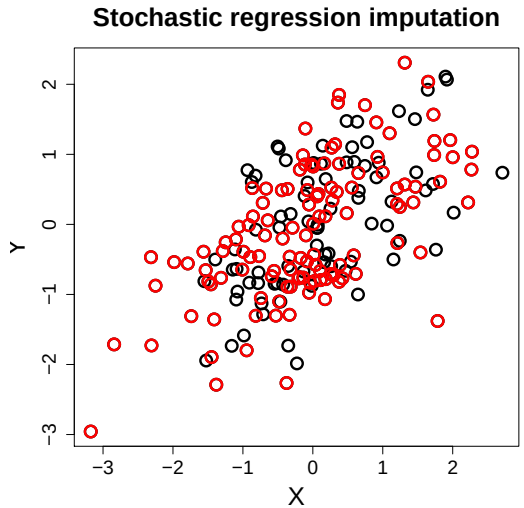
$\mu_y = 0$	0.01	0.01
$\sigma_y = 1$	0.5	0.72
$\rho = 0.6$	0.30	0.78
$CI\mu_y 95\%$	39.4	61.6

The idea of imputation is both seductive and dangerous *(Dempster and Rubin, 1983)*

Single imputation methods



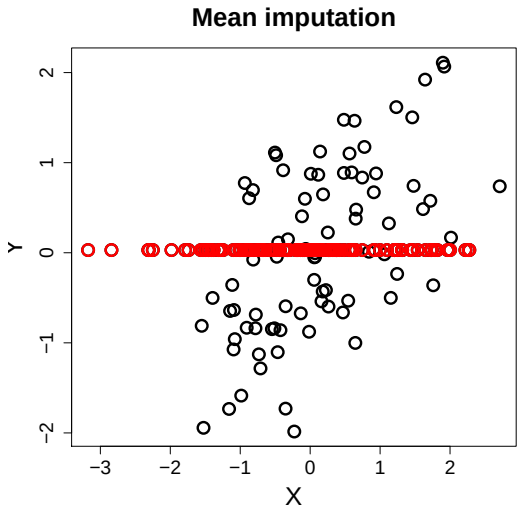


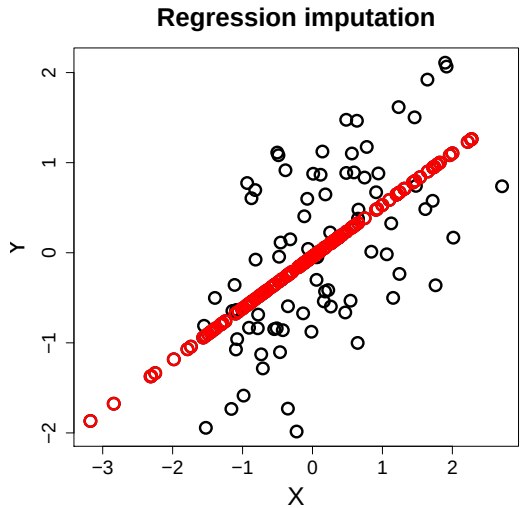


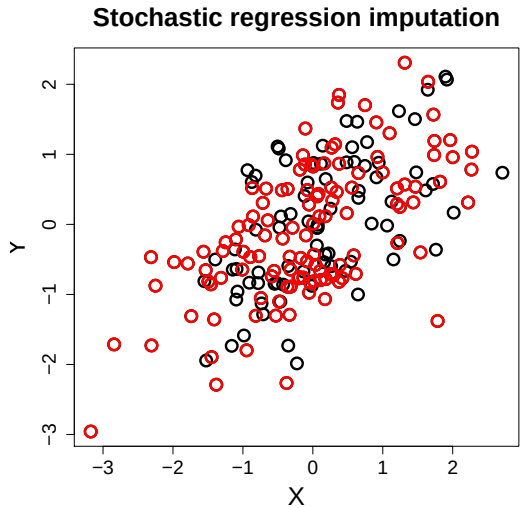
$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59
$CI\mu_y 95\%$	39.4	61.6	70.8

The idea of imputation is both seductive and dangerous *(Dempster and Rubin, 1983)*

Single imputation methods







$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59
$CI\mu_y 95\%$	39.4	61.6	70.8

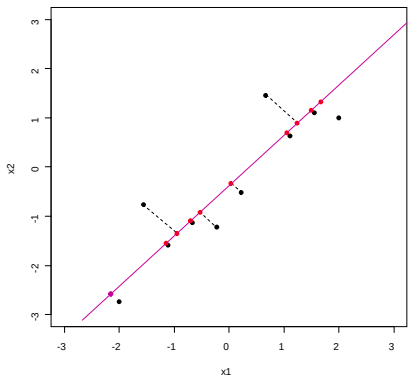
The idea of imputation is both seductive and dangerous *(Dempster and Rubin, 1983)*

$\Rightarrow$ Standard errors of the parameters ($\hat{\sigma}_{\hat{\mu}_y}$) calculated from the imputed data set are underestimated

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



$\Rightarrow$ Minimizes distance between observations and their projection

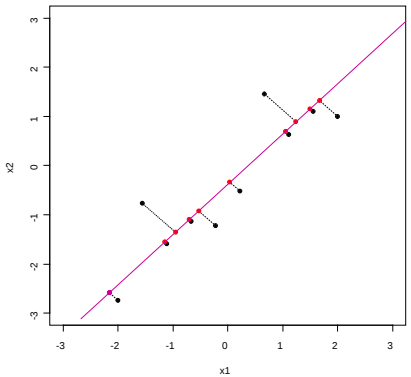
PCA reconstruction

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 $S < p$:

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

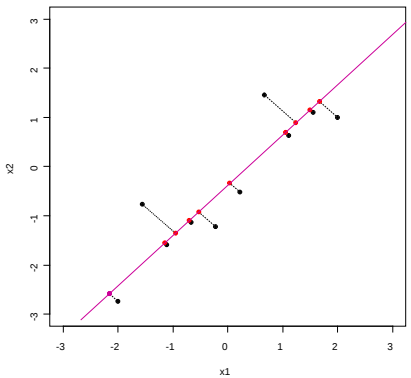
PCA reconstruction

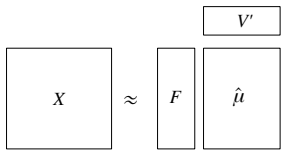
X

-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 $S < p$:

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

SVD X: $\hat{\mu}^{\text{PCA}} = U_{n\times S}\Lambda^{\frac{1}{2}}_{S\times S}V'_{p\times S}$ $F = U\Lambda^{\frac{1}{2}}$ PC - scores

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

Missing values in PCA

$\Rightarrow$ PCA: least squares

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

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

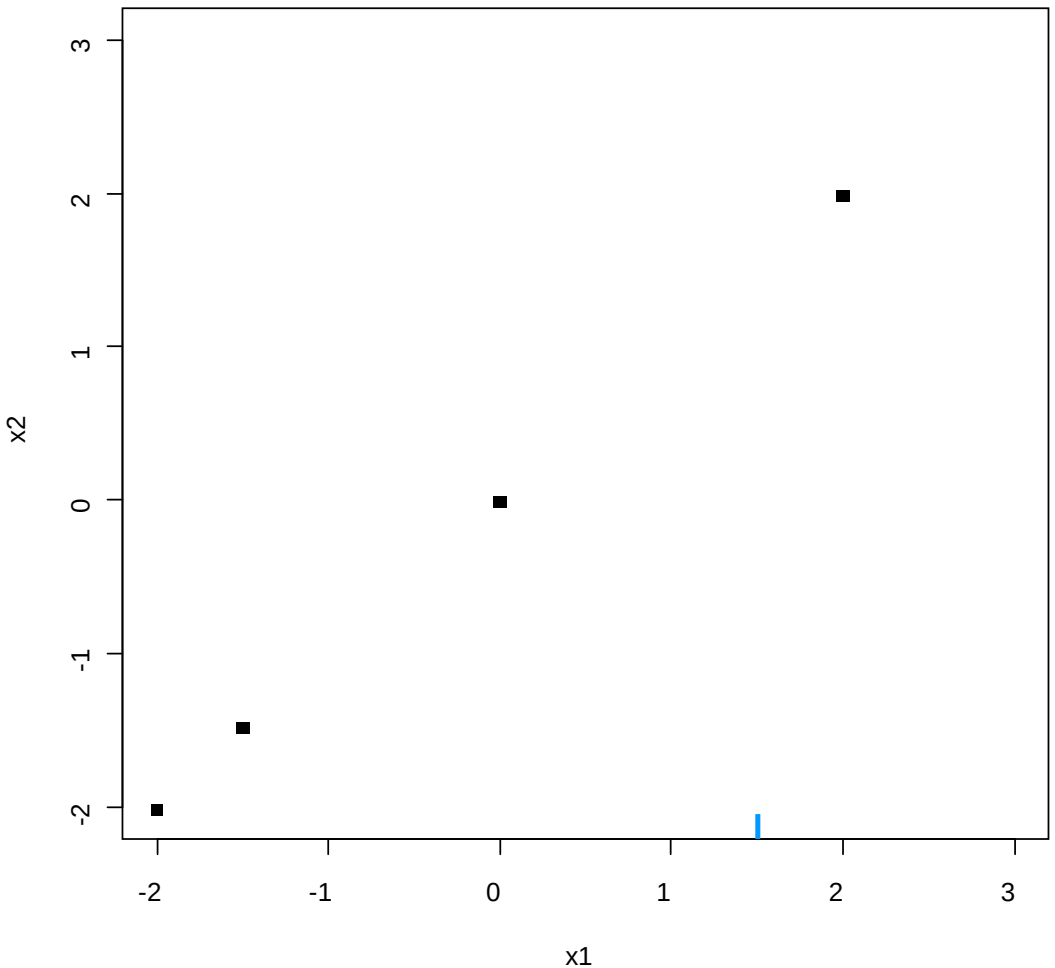
$$\text{argmin}_{\mu} \left\{ \| W_{n \times p} * (X - \mu) \|_2^2 : \text{rank}\,(\mu) \leq S \right\}$$

with $W_{ij} = 0$ if X_{ij} is missing, $W_{ij} = 1$ otherwise

Many algorithms: weighted alternating least squares (Gabriel & Zamir, 1979); iterative PCA (Kiers, 1997)

Iterative PCA

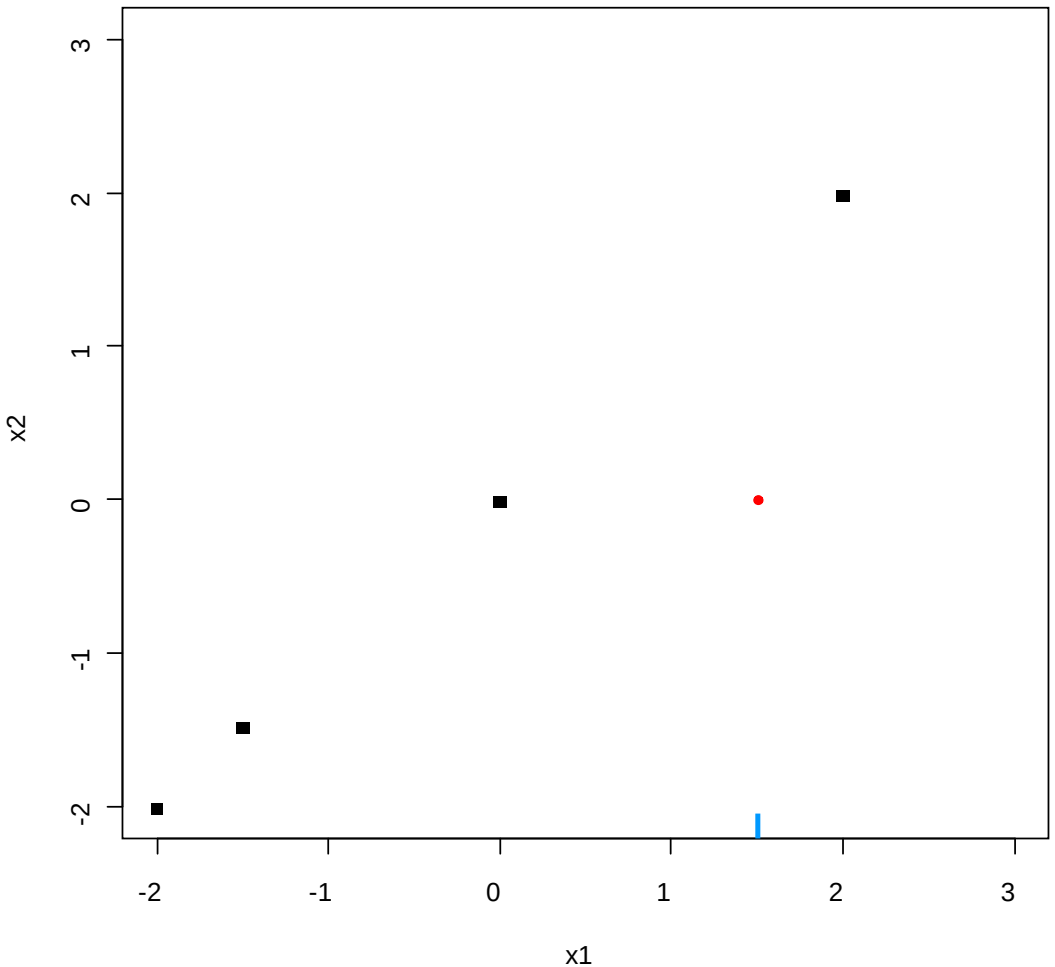
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98



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



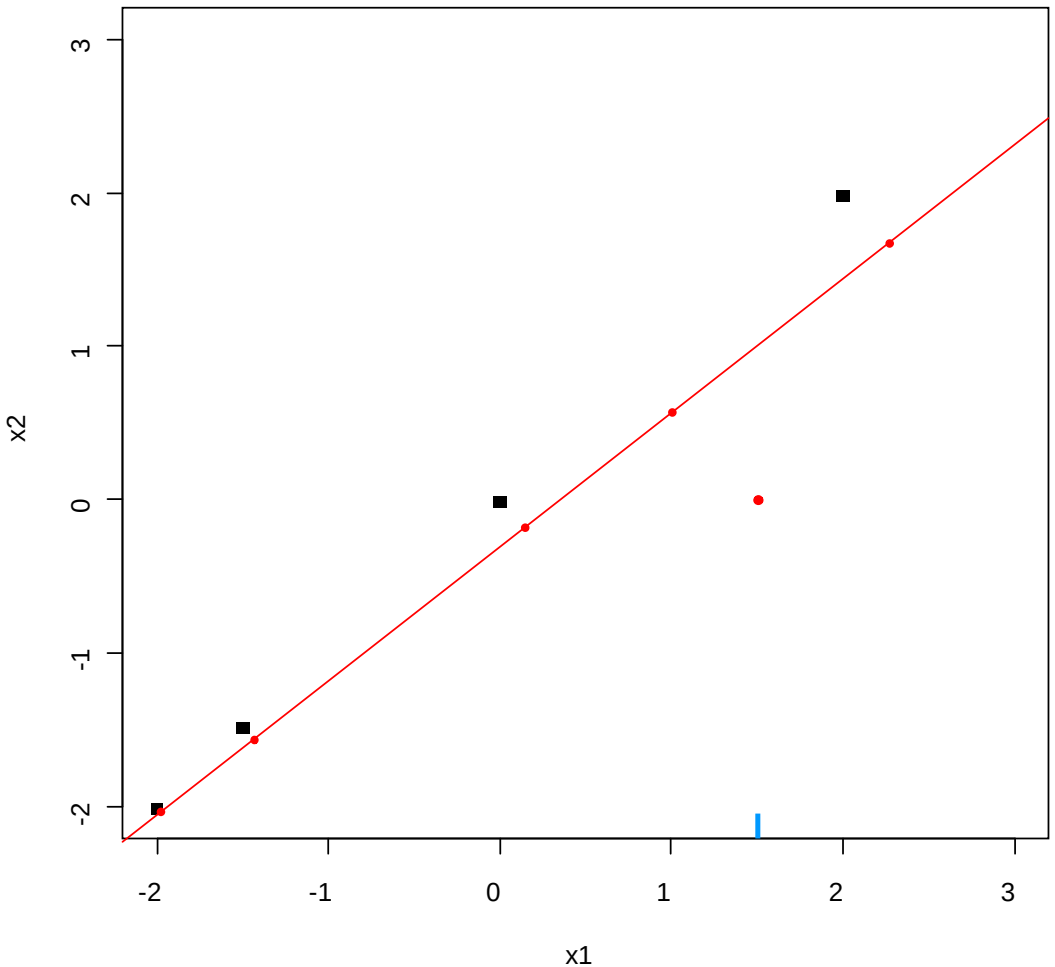
Initialization $\ell = 0$: X^0 (mean imputation)

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



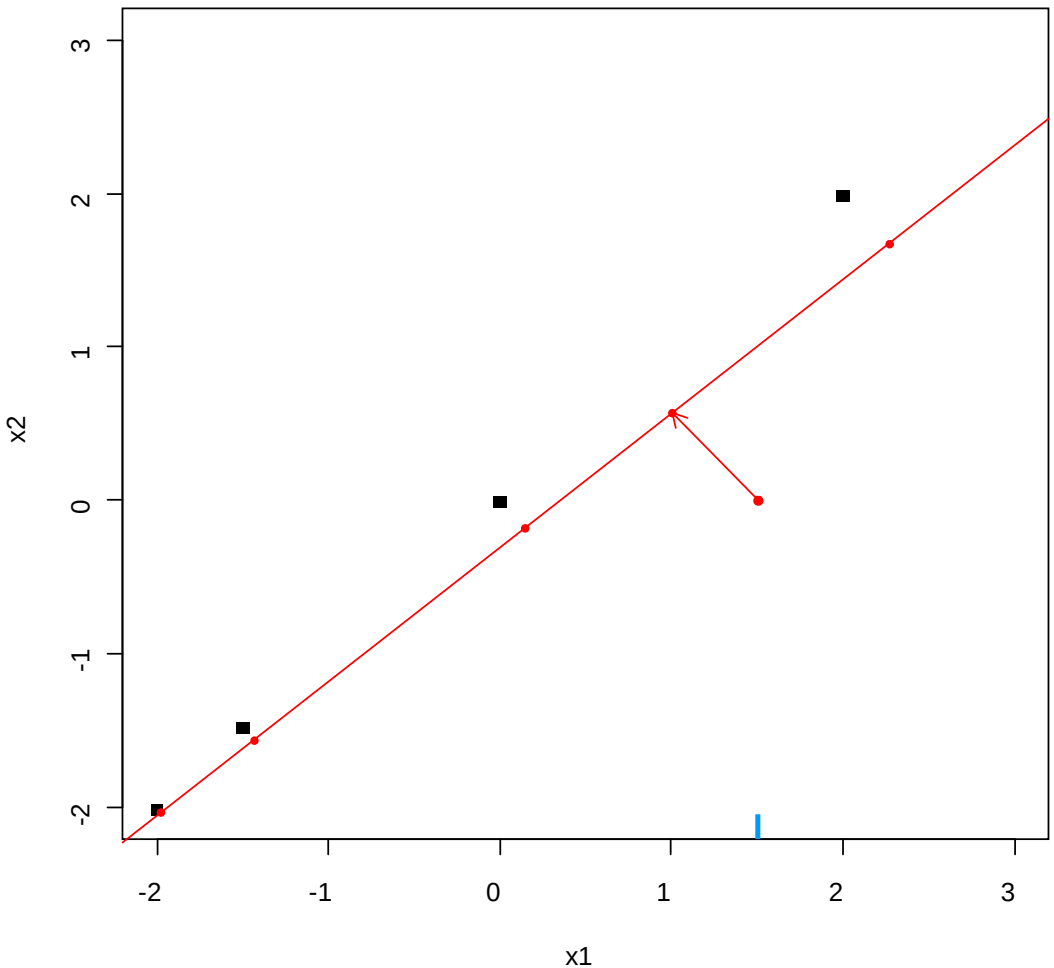
PCA on the completed data set $\rightarrow (U^\ell, \Lambda^\ell, V^\ell)$;

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 $\hat{\mu}^{\ell} = U^{\ell}\Lambda^{1/2^{\ell}}V^{\ell\prime}$

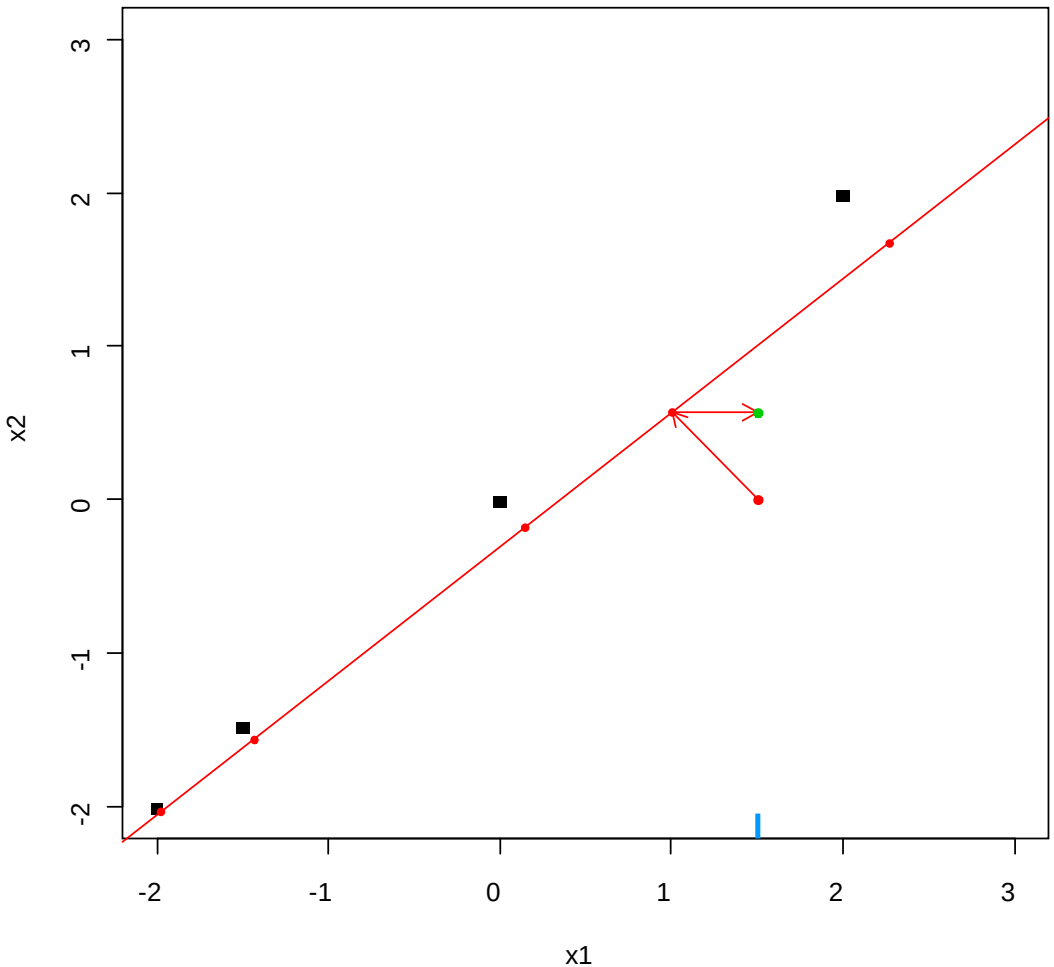
Iterative PCA

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

$\widehat{x1}$	$\widehat{x2}$
-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



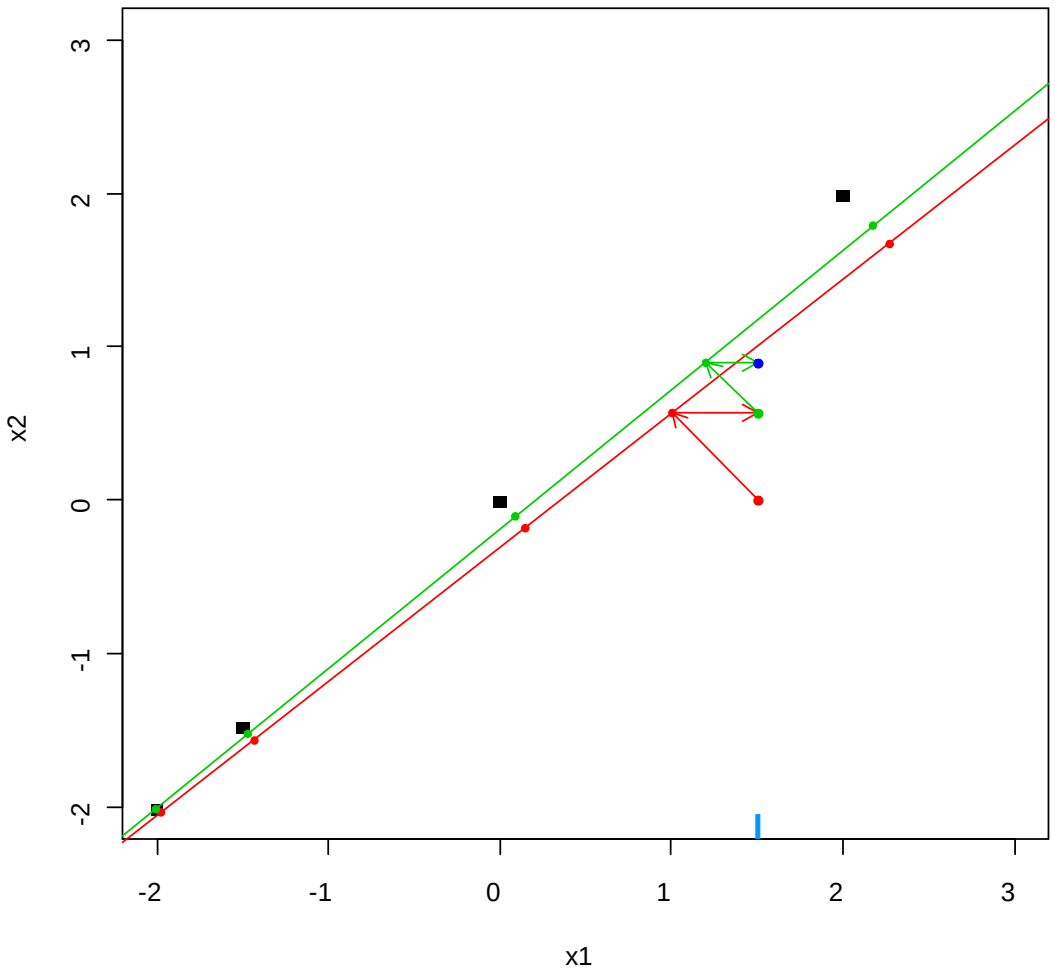
The new imputed dataset is $\hat{X}^{\ell} = W \odot X + (\mathbf{1} - W) \odot \hat{\mu}^{\ell}$

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



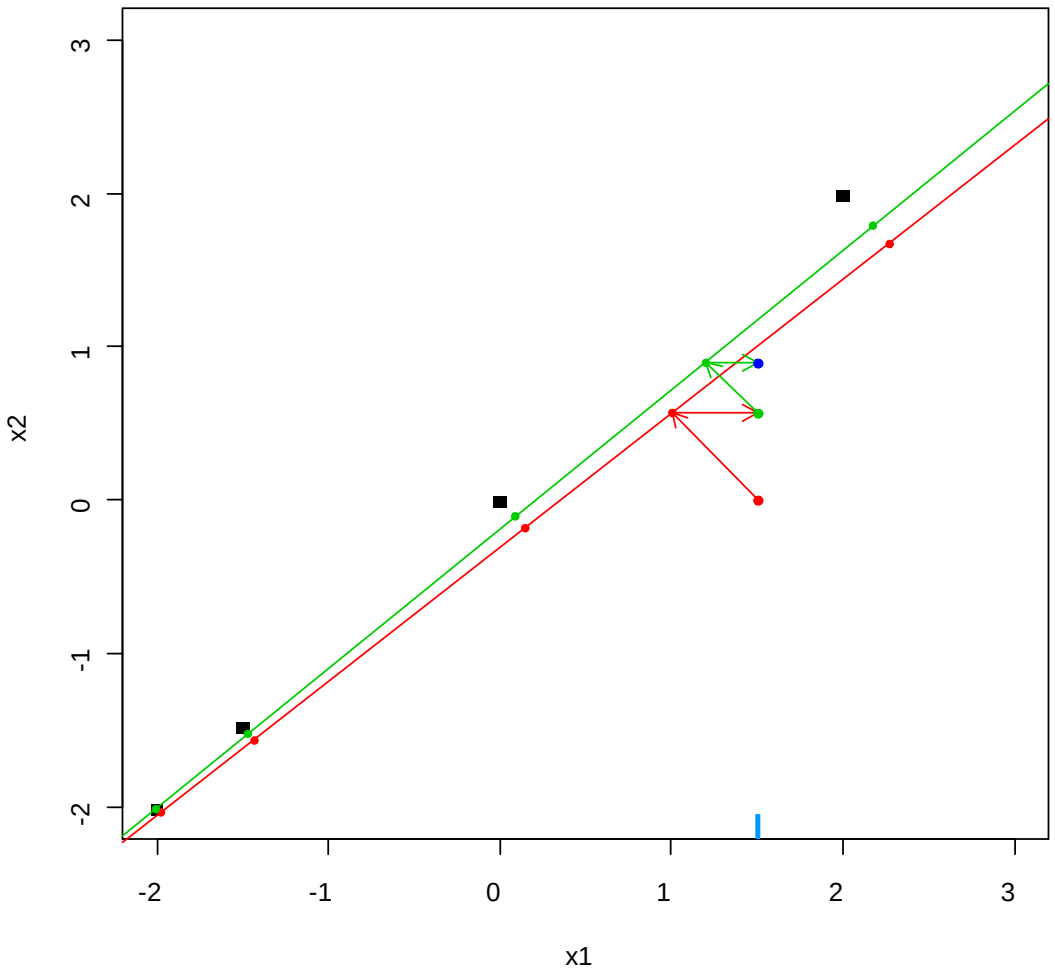
Iterative PCA

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



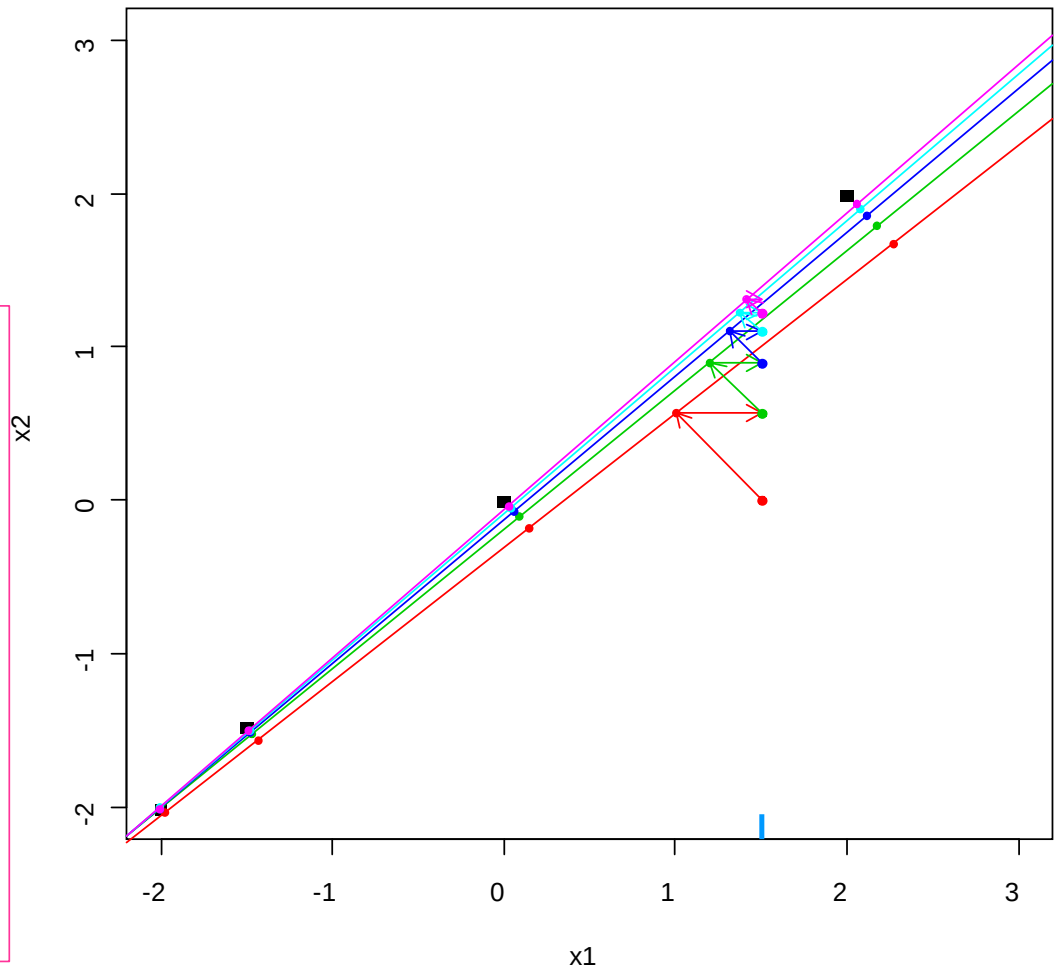
Iterative PCA

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

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

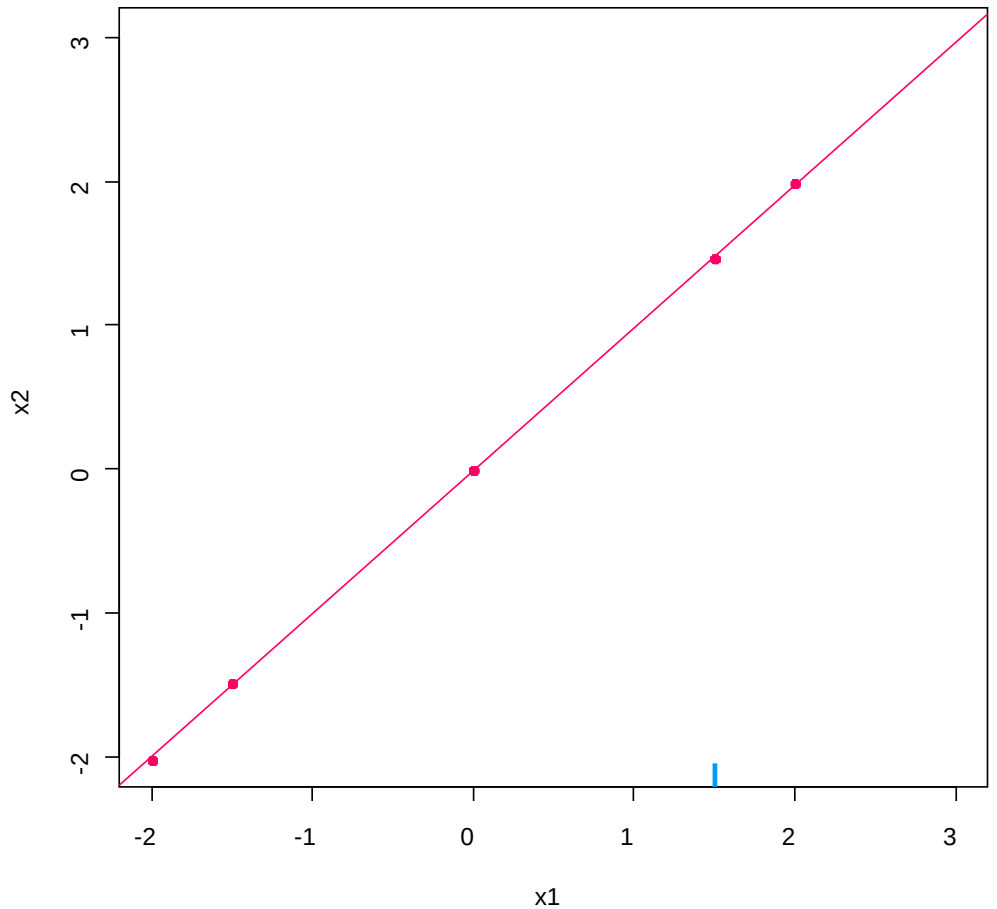
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



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 $\rightarrow (U^\ell, \Lambda^\ell, V^\ell)$

Missing values imputed with the fitted matrix $\hat{\mu}^\ell = U^\ell \Lambda^{1/2^\ell} V^{\ell\prime}$

Outline

Models for categorical variables

- Log-linear models - gold standard (Christensen, 1990; Agresti, 2013)
  $\Rightarrow$ Pb with high dimensional data.

- Latent variables models:
  - categoricals: latent class models (Goodman, 1974) - unsupervised clustering for one latent variable. Nonparametric Bayesian extensions (Dunson, 2009, 2012)
  - continuous: latent-trait models (Lazar, 1968) - item response theory (psychology & education, Van der Linden, 1997)
    - fixed: often one latent variable, difficulties to estimate.
    - random: Gaussian distribution on the latent variables (Moustaki, 2000; Sanchez, 2013)

  $\Rightarrow$ Binary data: Collins, Dasgupta, & Schapire (2001), Buntine (2002), Hoff (2009), De Leeuw (2006), Li & Tao (2013)

Relationship with MCA

Lemma

Let $\mathbf{G} \in \mathbb{R}^{n\times n}, \mathbf{H}_1 \in \mathbb{R}^{n\times n}, \mathbf{H}_2 \in \mathbb{R}^{m\times m}$, *with* $\mathbf{H}_1, \mathbf{H}_2 \succ 0$.

$$\text{argmax}_{\mathbf{\Gamma}:\,\text{rank}(\mathbf{\Gamma})\le K} \langle \mathbf{\Gamma}, \mathbf{G}\rangle - \frac{1}{2}\|\mathbf{H}_1\mathbf{\Gamma}\mathbf{H}_2\|_F^2$$

$$\mathbf{\Gamma}^* = \mathbf{H}_1^{-1}\left[\text{SVD}_K(\mathbf{H}_1^{-1}\mathbf{G}\mathbf{H}_2^{-1})\right]\mathbf{H}_2^{-1}$$

Thus, using Lemma **??**, the solution $\langle \mathbf{\Gamma}, \mathbf{A} - \mathbb{1}\mathbf{p}^T\rangle - \frac{1}{2}\|\mathbf{\Gamma}\mathbf{D}_p^{1/2}\|_F^2$ is given by the rank K SVD of $(\mathbf{A} - \mathbb{1}\mathbf{p}^T)\mathbf{D}_\mathbf{p}^{-1/2}$ which is precisely the SVD performed in MCA.

Theorem

The one-step likelihood estimate for the MultiLogit Bilinear model with rank constraint K, obtained by expanding around the independence model $(\beta_0 = \log \boldsymbol{p}, \mathbf{\Gamma}_0 = 0)$, is $(\beta_0, \widehat{\mathbf{\Gamma}}_{MCA})$.

Relationship with MCA

$$\ell(\beta, \mathbf{\Gamma}; \mathbf{A}) = \beta^j(A_i^j) + \Gamma_i^j(A_i^j) - \log\left(\sum_{c=1}^{C_j} e^{\beta^j(c)+\Gamma_i^j(c)}\right)$$

$$\frac{\partial \ell}{\partial \Gamma_i^j(c)} = 1_{x_{ij}=c} - \frac{e^{\beta^j(c)+\Gamma_i^j(c)}}{\sum_{c'=1}^{C_j} e^{\beta^j(c')+\Gamma_i^j(c')}} = A_{ic}^j - \pi_{ijc} \tag{4}$$

$$\frac{\partial \ell}{\partial \Gamma_i^j(c)\partial \Gamma_{i'}^{j'}(c')} = \begin{cases} \pi_{ijc}\pi_{ijc'} - \pi_{ijc}1_{c=c'} & j = j', i = i' \\ 0 & \text{o.w.} \end{cases} \tag{5}$$

Evaluating (**??**) at $\zeta_0 = (\beta_0 = \log(\mathbf{p}), 0)$ gives $A_{ic}^j - p^j(c)$ - idem (**??**)

$$\tilde{\ell}(\beta, \mathbf{\Gamma}) \approx \langle \mathbf{\Gamma}, \mathbf{A} - \mathbb{1}\mathbf{p}^T \rangle - \frac{1}{2}\|\mathbf{\Gamma}\mathbf{D}_p^{1/2}\|_F^2$$

Majorization

- Majorization (or MM) algorithms (De Leeuw & Heiser, 1977; Lange, 2004) use in each iteration a majorizing function $g(\theta, \theta_0)$.
- Current estimate θ_0 is called *supporting point*.
- Requirements:
  1. $f(\theta_0) = g(\theta_0, \theta_0)$.
  2. $f(\theta) \leq g(\theta, \theta_0)$.
- *Sandwich inequality*: $f(\theta^+) \leq g(\theta^+, \theta_0) \leq g(\theta_0, \theta_0) = f(\theta_0)$ with $\theta^+ = \operatorname{argmin} g(\theta, \theta_0)$
- Any majorization algorithm is guaranteed to descent.

Selecting λ

$\Rightarrow$ 2 steps : QUT to select the rank K - Shrinkage with CV

Rationale with Lasso:

- Lasso often used for screening (Buhlmann van de Geer, 2011)
- Selecting λ with CV or STEIN focuses on predictive properties
- Optimal threshold for prediction $\neq$ optimal for selecting var

$\Rightarrow$ Quantile Universal Threshold (Sardy, 2016) : select the threshold at the bulk edge of what a threshold should be under the null. Guaranteed variable screening with high proba. Be careful, biaised!

Quantile Universal Threshold

Ex PCA. $X = \mu + \varepsilon$, with $\varepsilon_{ij} \sim N(0, \sigma^2) \rightarrow \hat{\mu}_K = \sum_{k=1}^{K} u_k \, d_k \, v_k^\top$

Soft-threshold: $\text{argmin}_\mu \, \|X - \mu\|_2^2 + \lambda \|\mu\|_* \rightarrow d_k \max\left(1 - \frac{\lambda}{d_k}, 0\right)$

$\Rightarrow$ Selecting λ to have good estimation of the rank

1. Generate data under the null hypothesis of no signal, $\mu = 0$
2. Compute the first singular value d_1
3. Repeat 1000 times 1 and 2
4. Use the $(1 - \alpha)$-quantile of the distribution of d_1 as threshold

(Exact results Zanella 2009; Asymptotic results, random matrix theory Shabalin 2013, Paul 2007, Baik 2006...)

$\Rightarrow$ Suppose to know σ!

Selecting λ

$\Rightarrow$ 2 steps : QUT to select the rank K - Shrinkage with CV

Model: $\pi_{ijc} = \frac{\exp(\beta_j(c)+\sum_{k=1}^{K} d_k u_{ik} v_{jk}(c))}{\sum_{c'=1}^{C_j} \exp(\beta_j(c')+\sum_{k=1}^{K} d_k u_{ik} v_{jk}(c'))}$

Lik:$L(\beta, \mathbf{U}, \mathbf{D}, \mathbf{V}) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{c=1}^{C_j} A_{ic}^{j} \log(\pi_{ijc}) + \lambda \sum_{k=1}^{K^\star} d_k$

1. Generate under the null of no interaction and take λ the quantile of the distribution of d_1: good rank recovery
2. For a rank K_{QUT}, estimate λ with Cross-Validation to determine the amount of shrinkage (Lasso + LS)
   k-fold CV, λ with the best out-of-sample deviance is chosen.

Simulations

- n: 50, 100, 300
- m: 20, 100, 300 - 3 categories/variables
- Interaction K: 2, 6
- (d_1/d_2): 2, 1
- the strength of the interaction (low, strong).

$$\tilde{\mathbf{u}}_i \sim \mathcal{N}_K\left(0, \begin{pmatrix} d_1 & 0 \\ 0 & d_K \end{pmatrix}\right)$$

$$\tilde{\mathbf{v}}_j(c) \sim \mathcal{N}_K\left(0, \begin{pmatrix} d_1 & 0 \\ 0 & d_K \end{pmatrix}\right)$$

$$\theta_{ij}^c = -\frac{1}{2}\|\tilde{\mathbf{u}}_i - \tilde{\mathbf{v}}_j(c)\|^2$$

$$\mathbb{P}(x_{ij} = c) \propto e^{\theta_{ij}^c}$$

Simulations results

	n	p	rank	ratio	strength	model	MCA
1	50	20	2	1	0.1	0.044	**0.035**
2	50	20	2	1	1	0.020	0.045
3	50	20	2	2	0.1	0.048	**0.036**
4	50	20	2	2	1	**0.0206**	0.042
5	50	20	6	1	0.1	0.111	**0.064**
6	50	20	6	1	1	0.045	**0.026**
7	50	20	6	2	0.1	0.115 (0.028)	**0.071**
8	50	20	6	2	1	**0.032**	0.051
9	300	100	2	1	0.1	**0.005**	*0.006*
10	300	100	2	1	1	**0.004**	0.042
11	300	100	2	2	0.1	**0.0047**	*0.005*
12	300	100	2	2	1	**0.0037** (0.00369)	0.040
13	300	300	2	1	0.1	**0.003**	*0.004*
14	300	300	2	1	1	**0.002**	0.039
15	300	300	2	2	0.1	**0.003**	*0.004*
16	300	300	2	2	1	**0.002**	0.039
17	300	100	6	1	0.1	0.019	***0.015***
18	300	100	6	1	1	**0.011**	0.023
19	300	100	6	2	0.1	0.018 (0.010)	***0.017***
20	300	100	6	2	1	**0.010**	0.056
21	300	300	6	1	0.1	0.011	***0.008***
22	300	300	6	1	1	**0.006**	0.022
23	300	300	6	2	0.1	**0.009**	*0.012*
24	300	300	6	2	1	**0.006**	0.061

Simulations results

	n	p	rank	ratio	strength	model	MCA
1	50	20	2	1	0.1	0.044	**0.035**
2	50	20	2	1	1	0.020	0.045
3	50	20	2	2	0.1	0.048	**0.036**
4	50	20	2	2	1	**0.0206**	0.042
5	50	20	6	1	0.1	0.111	**0.064**
6	50	20	6	1	1	0.045	**0.026**
7	50	20	6	2	0.1	0.115 (0.028)	**0.071**
8	50	20	6	2	1	**0.032**	0.051
9	300	100	2	1	0.1	**0.005**	*0.006*
10	300	100	2	1	1	**0.004**	0.042
11	300	100	2	2	0.1	**0.0047**	*0.005*
12	300	100	2	2	1	**0.0037** (0.00369)	0.040
13	300	300	2	1	0.1	**0.003**	*0.004*
14	300	300	2	1	1	**0.002**	0.039
15	300	300	2	2	0.1	**0.003**	*0.004*
16	300	300	2	2	1	**0.002**	0.039
17	300	100	6	1	0.1	0.019	***0.015***
18	300	100	6	1	1	**0.011**	0.023
19	300	100	6	2	0.1	0.018 (0.010)	***0.017***
20	300	100	6	2	1	**0.010**	0.056
21	300	300	6	1	0.1	0.011	***0.008***
22	300	300	6	1	1	**0.006**	0.022
23	300	300	6	2	0.1	**0.009**	*0.012*

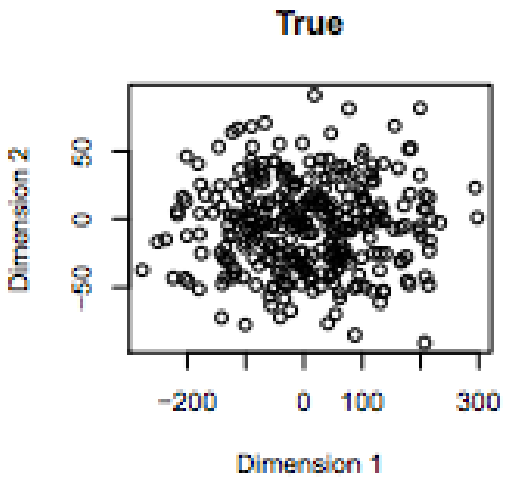
Simulations results

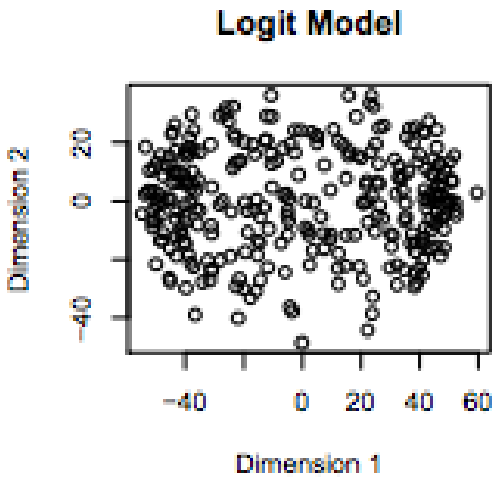
	n	p	rank	ratio	strength	model	MCA
1	50	20	2	1	0.1	0.044	**0.035**
2	50	20	2	1	1	0.020	0.045
3	50	20	2	2	0.1	0.048	**0.036**
4	50	20	2	2	1	**0.0206**	0.042
5	50	20	6	1	0.1	0.111	**0.064**
6	50	20	6	1	1	0.045	**0.026**
7	50	20	6	2	0.1	0.115 (0.028)	**0.071**
8	50	20	6	2	1	**0.032**	0.051
9	300	100	2	1	0.1	**0.005**	*0.006*
10	300	100	2	1	1	**0.004**	0.042
11	300	100	2	2	0.1	**0.0047**	*0.005*
12	300	100	2	2	1	**0.0037** (0.00369)	0.040
13	300	300	2	1	0.1	**0.003**	*0.004*
14	300	300	2	1	1	**0.002**	0.039
15	300	300	2	2	0.1	**0.003**	*0.004*
16	300	300	2	2	1	**0.002**	0.039
17	300	100	6	1	0.1	0.019	***0.015***
18	300	100	6	1	1	**0.011**	0.023
19	300	100	6	2	0.1	0.018 (0.010)	***0.017***
20	300	100	6	2	1	**0.010**	0.056
21	300	300	6	1	0.1	0.011	***0.008***
22	300	300	6	1	1	**0.006**	0.022
23	300	300	6	2	0.1	**0.009**	*0.012*
24	300	300	6	2	1	**0.006**	0.061

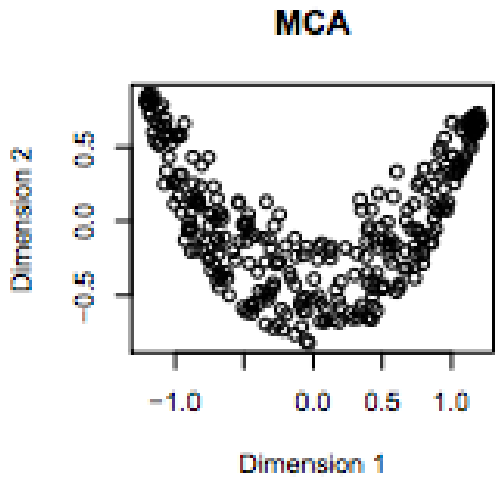
Simulations results

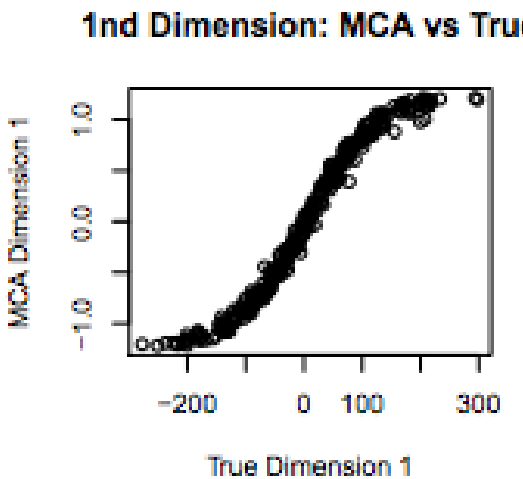
	n	p	rank	ratio	strength	model	MCA
1	50	20	2	1	0.1	0.044	**0.035**
2	50	20	2	1	1	0.020	0.045
3	50	20	2	2	0.1	0.048	**0.036**
4	50	20	2	2	1	**0.0206**	0.042
5	50	20	6	1	0.1	0.111	**0.064**
6	50	20	6	1	1	0.045	**0.026**
7	50	20	6	2	0.1	0.115 (0.028)	**0.071**
8	50	20	6	2	1	**0.032**	0.051
9	300	100	2	1	0.1	**0.005**	*0.006*
10	300	100	2	1	1	**0.004**	0.042
11	300	100	2	2	0.1	**0.0047**	*0.005*
12	300	100	2	2	1	**0.0037** (0.00369)	0.040
13	300	300	2	1	0.1	**0.003**	*0.004*
14	300	300	2	1	1	**0.002**	0.039
15	300	300	2	2	0.1	**0.003**	*0.004*
16	300	300	2	2	1	**0.002**	0.039
17	300	100	6	1	0.1	0.019	***0.015***
18	300	100	6	1	1	**0.011**	0.023
19	300	100	6	2	0.1	0.018 (0.010)	***0.017***
20	300	100	6	2	1	**0.010**	0.056
21	300	300	6	1	0.1	0.011	***0.008***
22	300	300	6	1	1	**0.006**	0.022
23	300	300	6	2	0.1	**0.009**	*0.012*
24	300	300	6	2	1	**0.006**	0.061

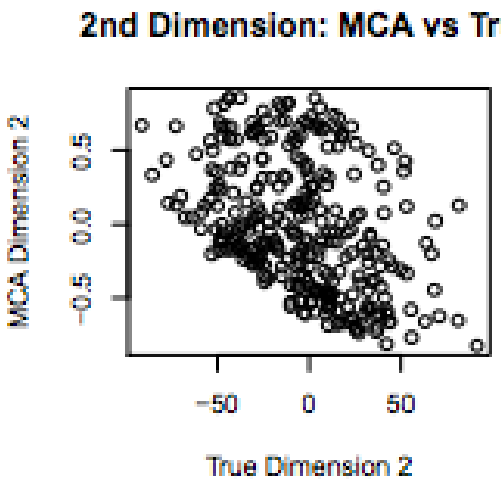
Simulation

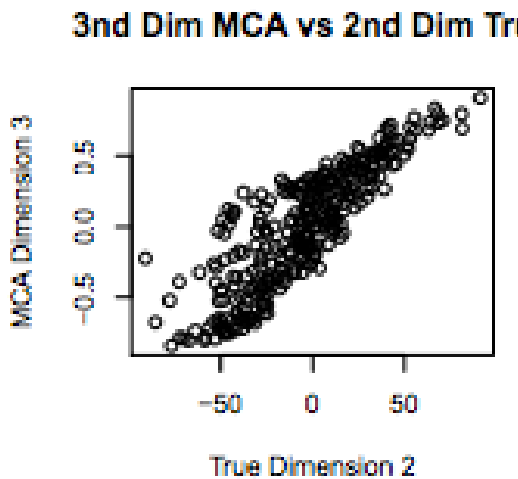

True
Dimension 2
Dimension 1
−200
0
100
300
−50
0
50


Logit Model
Dimension 2
Dimension 1
−40
0
20
40
60
−40
0
20

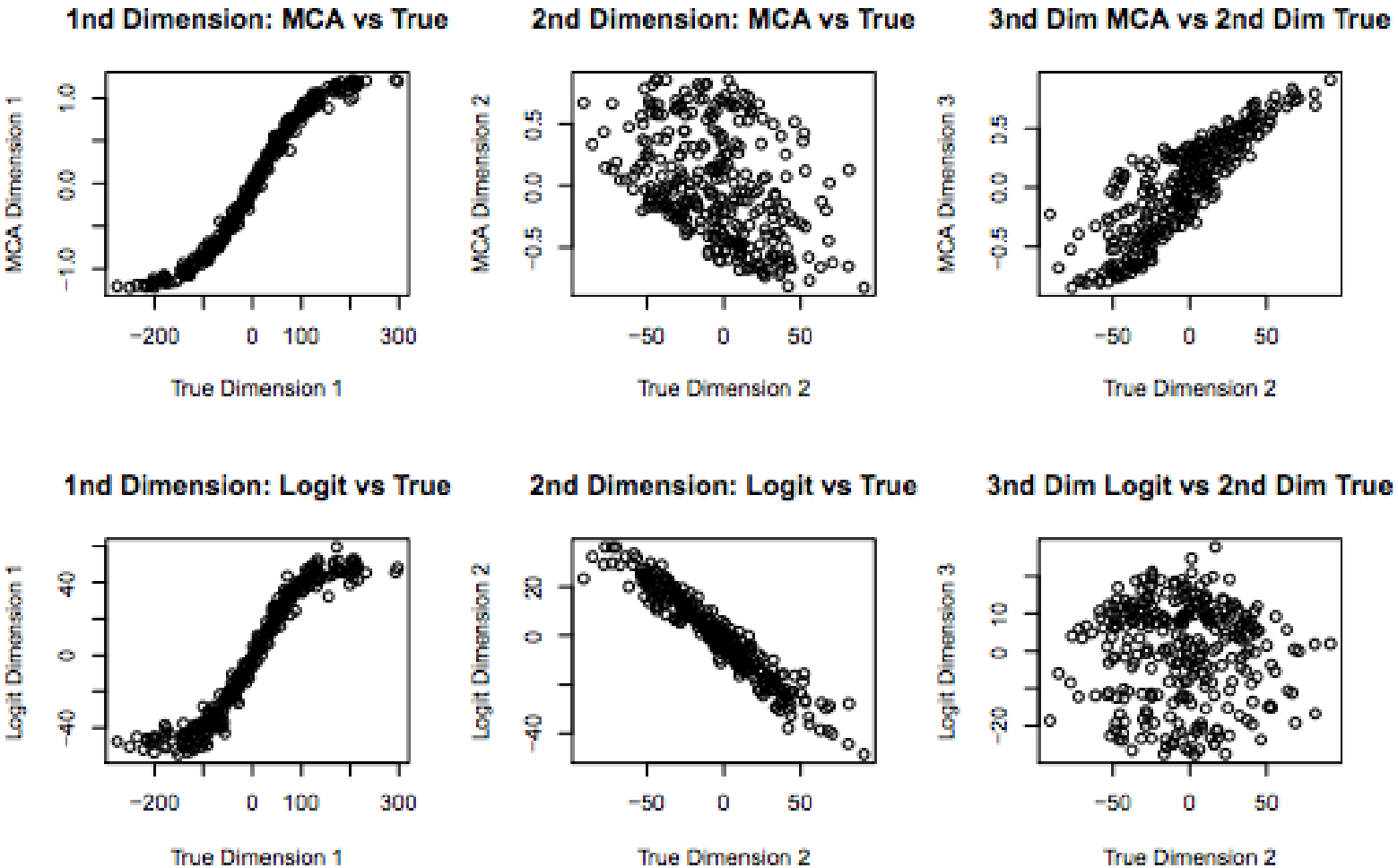

MCA
Dimension 2
Dimension 1
−1.0
0.0
0.5
1.0
−0.5
0.0
0.5


1nd Dimension: MCA vs True
MCA Dimension 1
True Dimension 1
−200
0
100
300
−1.0
0.0
1.0


2nd Dimension: MCA vs True
MCA Dimension 2
True Dimension 2
−50
0
50
−0.5
0.0
0.5

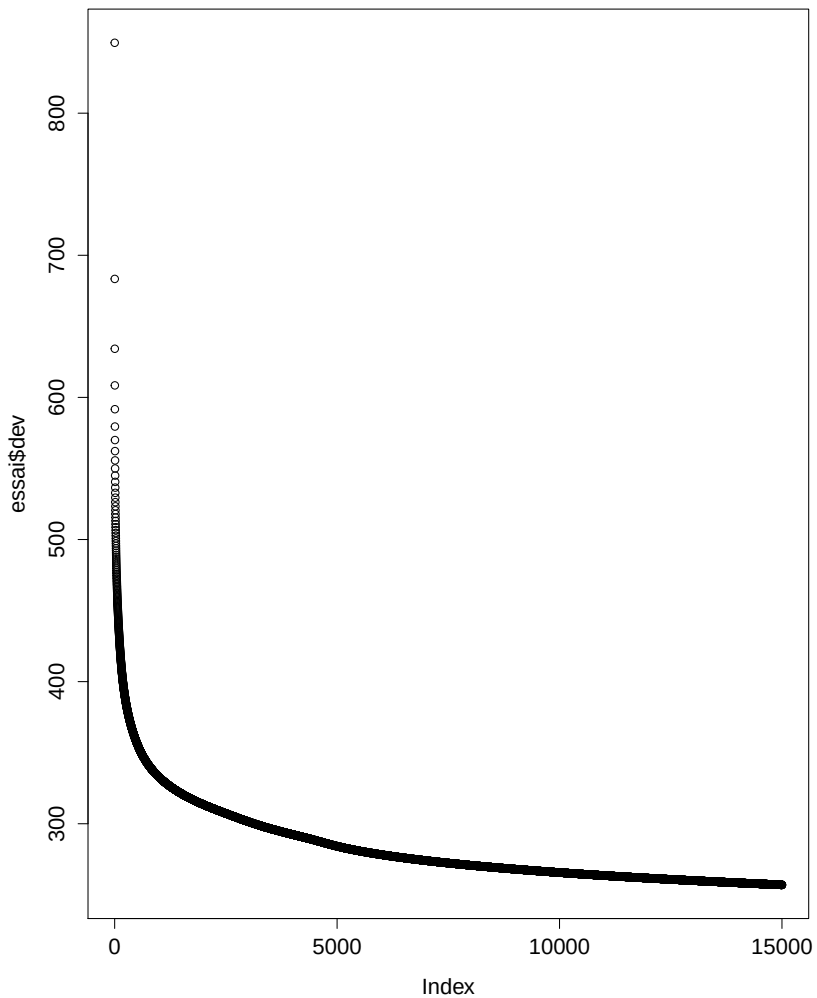

3nd Dim MCA vs 2nd Dim True
MCA Dimension 3
True Dimension 2
−50
0
50
−0.5
0.0
0.5

Simulation


1nd Dimension: MCA vs True
MCA Dimension 1
True Dimension 1
2nd Dimension: MCA vs True
MCA Dimension 2
True Dimension 2
3nd Dim MCA vs 2nd Dim True
MCA Dimension 3
True Dimension 2
1nd Dimension: Logit vs True
Logit Dimension 1
True Dimension 1
2nd Dimension: Logit vs True
Logit Dimension 2
True Dimension 2
3nd Dim Logit vs 2nd Dim True
Logit Dimension 3
True Dimension 2

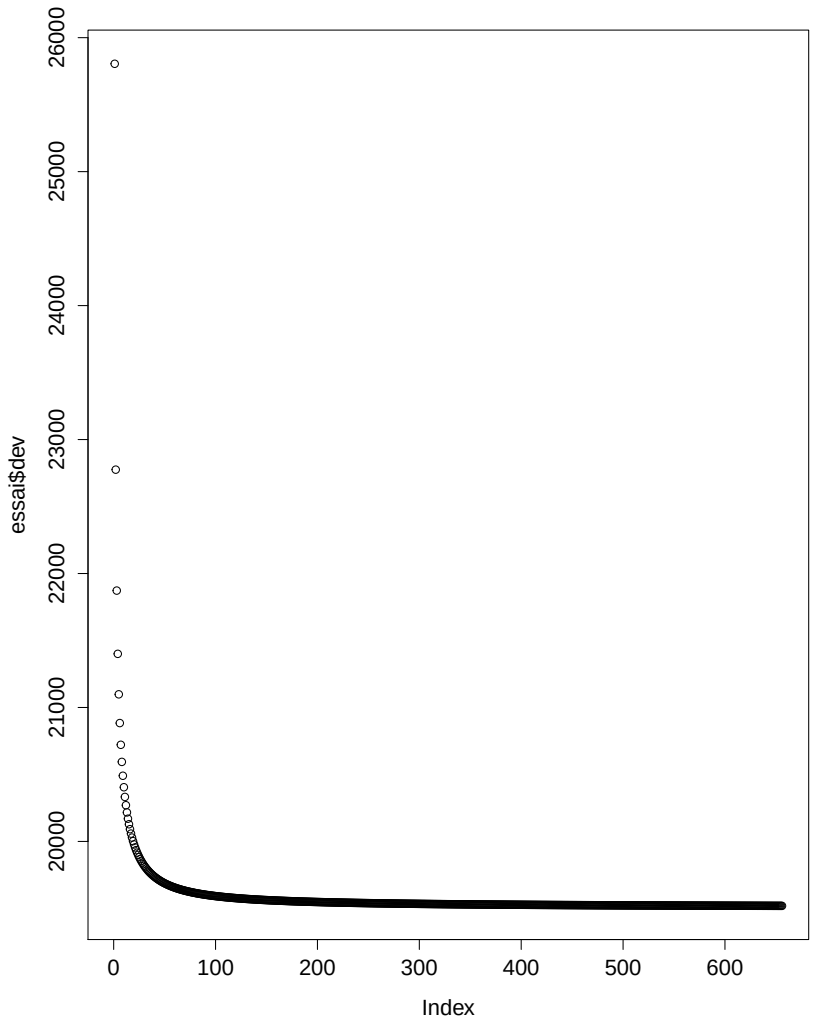
Overfitting

$\Rightarrow n = 50,\ p = 20,\ r = 6,\ strength = 0.1$



Overfitting

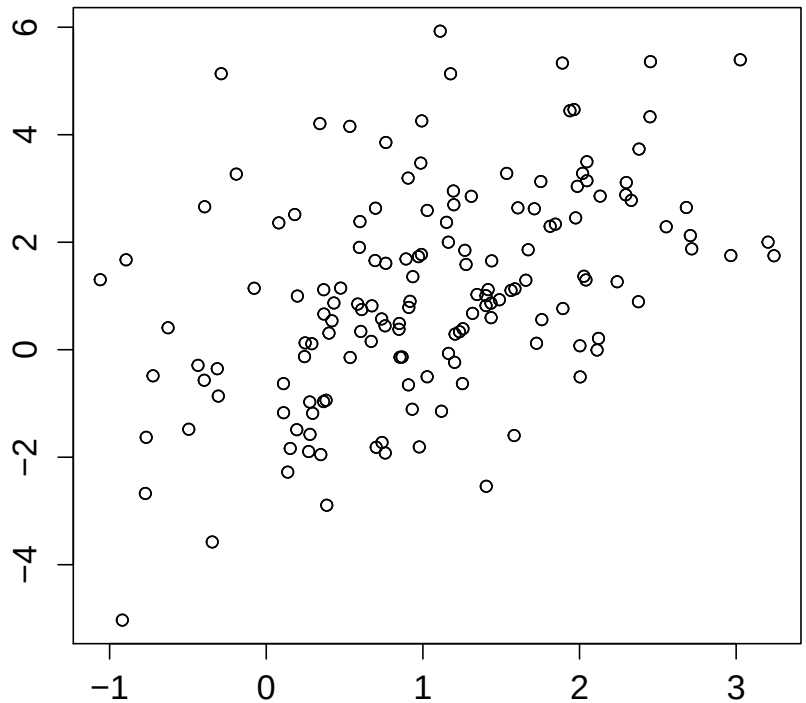
$\Rightarrow n = 50,\ p = 20,\ r = 6,\ strength = 0.1$



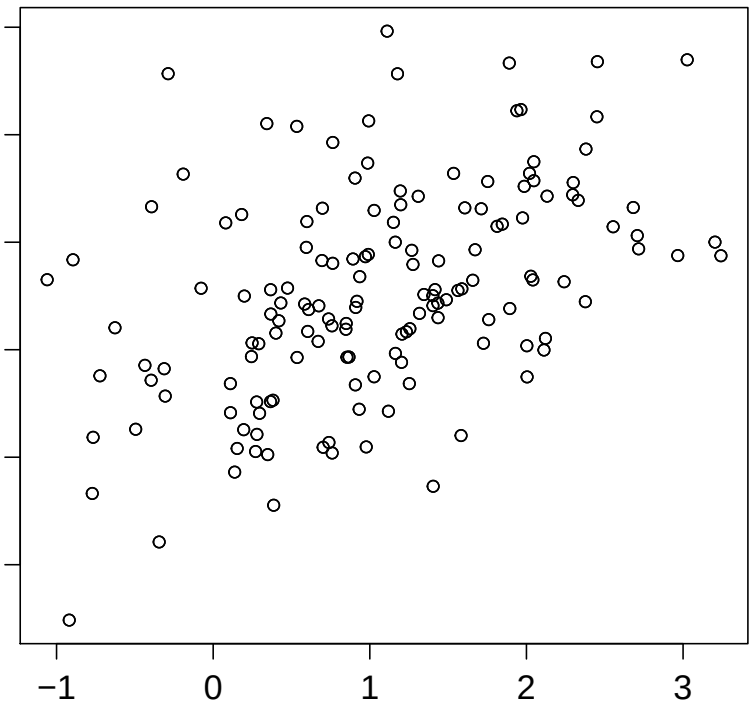
Imputation close to data and extrapolation

$$150 \text{ observasions of } X \sim \mathit{Normal}\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$$

Missing completely at random



Missing (not) at random

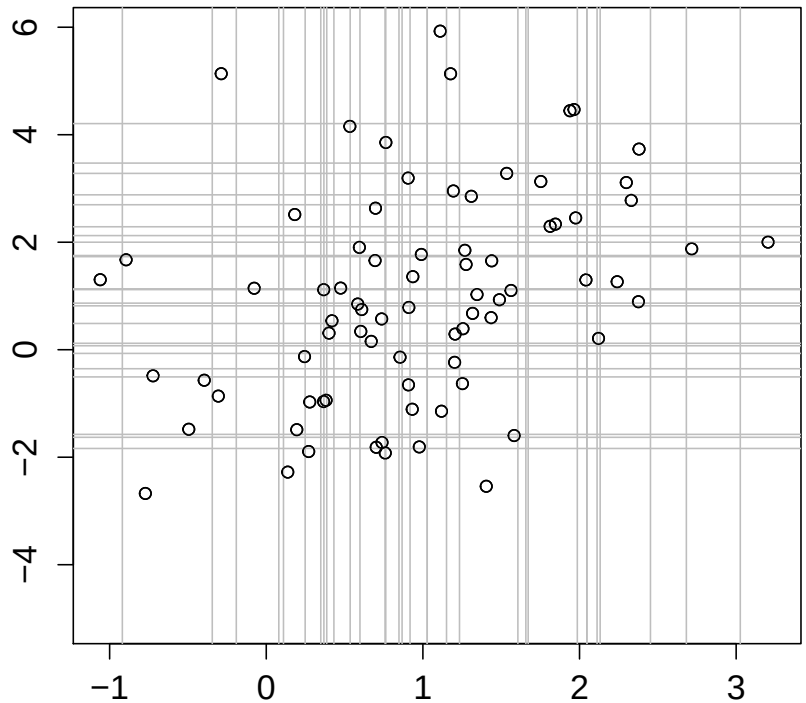


◆ – regression ● – zonoid depth ▲ – random forest

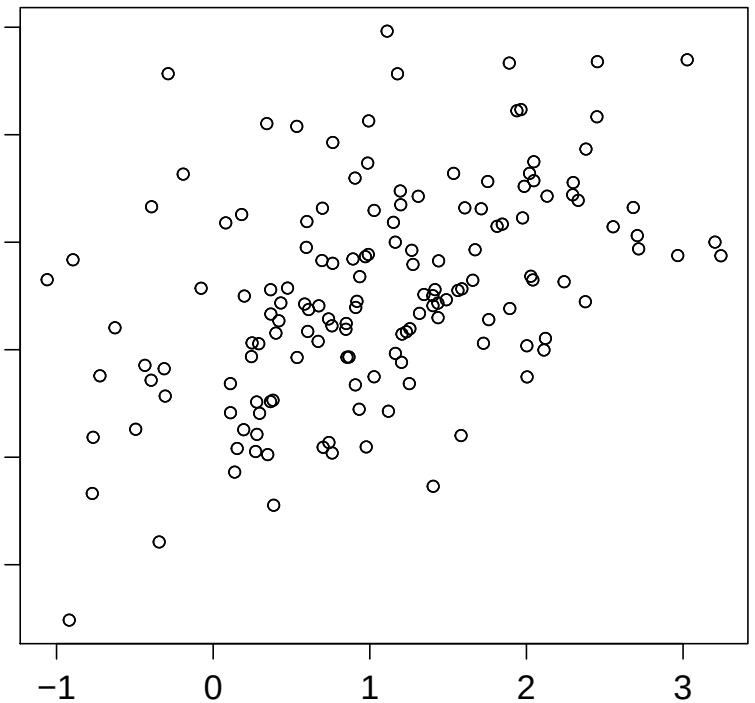
Imputation close to data and extrapolation

$$150 \text{ observasions of } X \sim \mathit{Normal}\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$$

Missing completely at random



Missing (not) at random

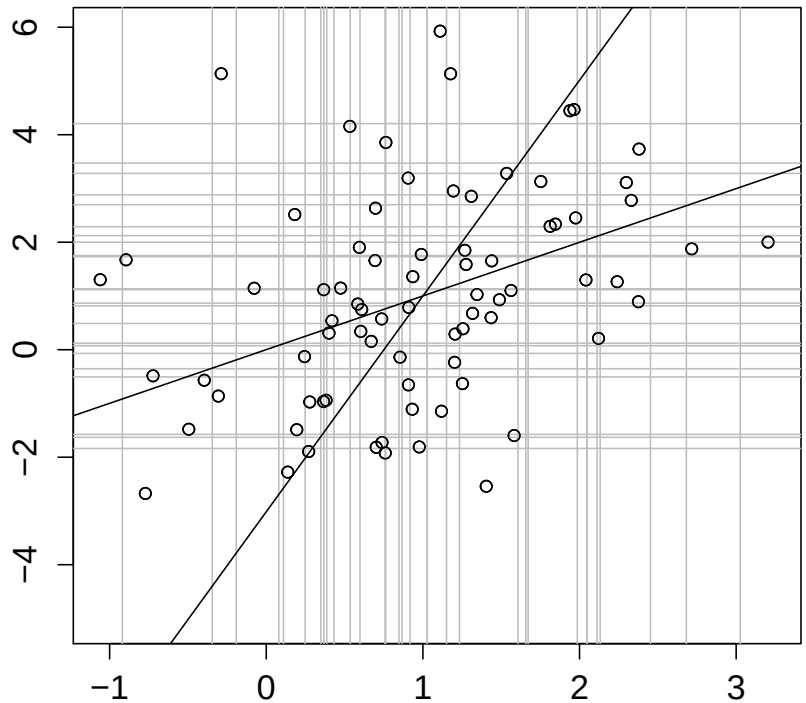


◆ – regression ● – zonoid depth ▲ – random forest

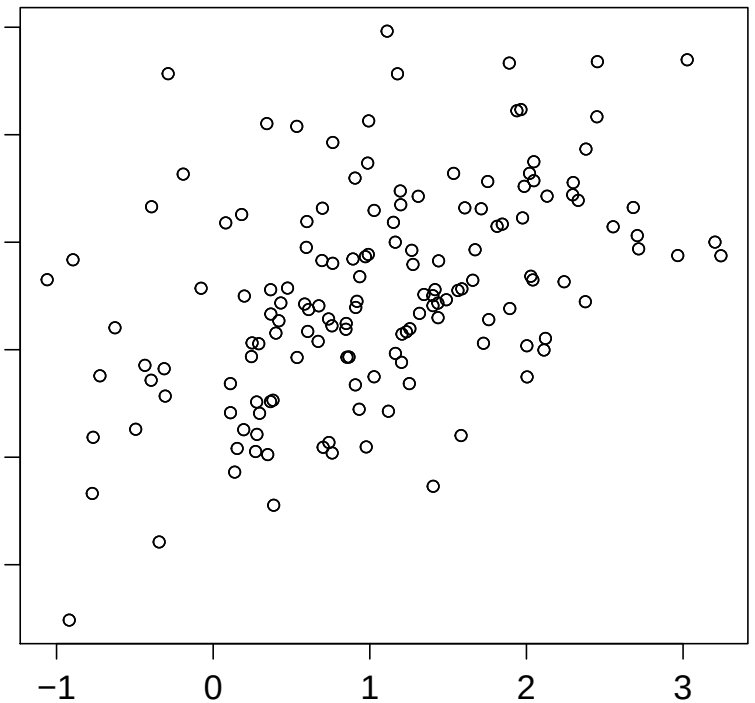
Imputation close to data and extrapolation

$$150 \text{ observasions of } X \sim \mathit{Normal}\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$$

Missing completely at random



Missing (not) at random

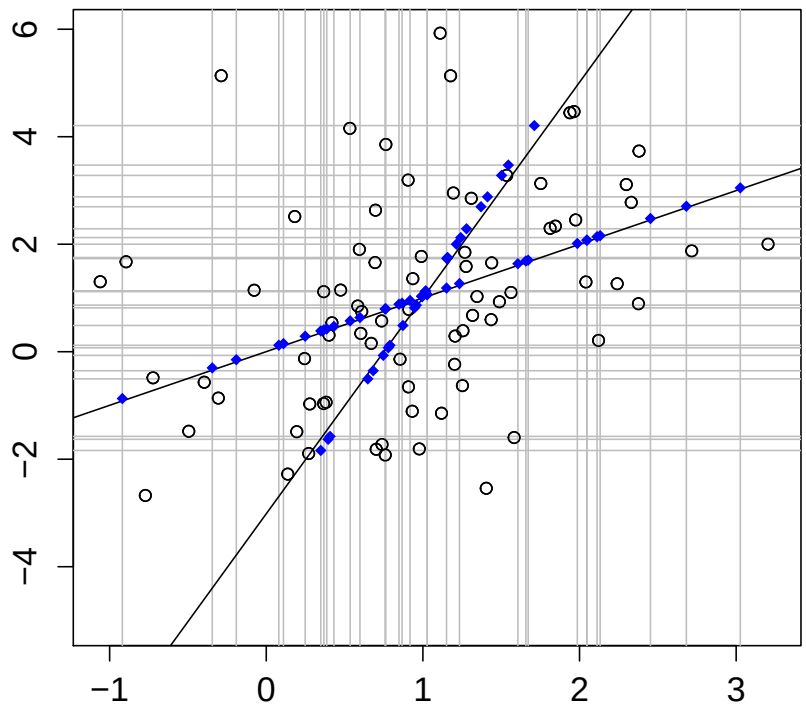


◆ – regression ● – zonoid depth ▲ – random forest

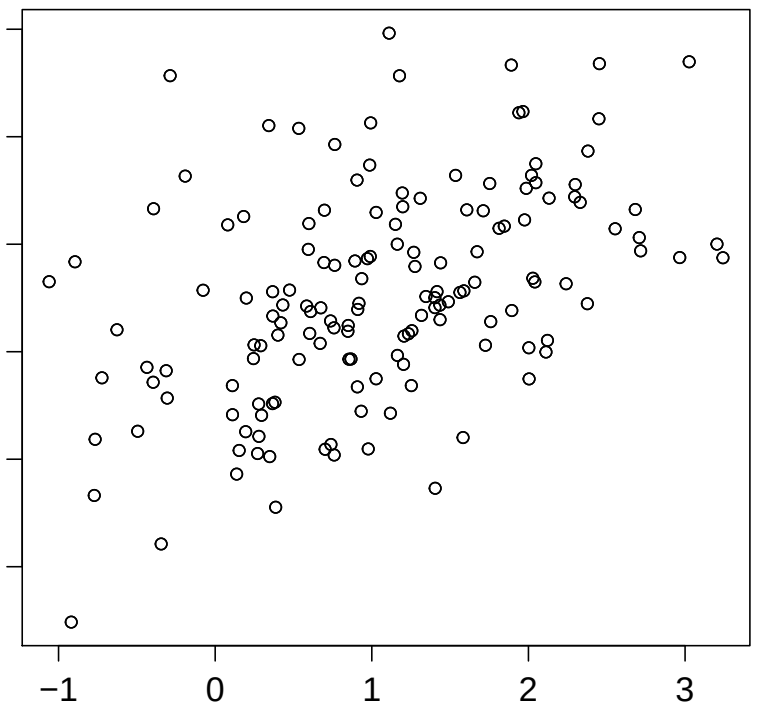
Imputation close to data and extrapolation

$$150 \text{ observasions of } X \sim \mathit{Normal}\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$$

Missing completely at random



Missing (not) at random

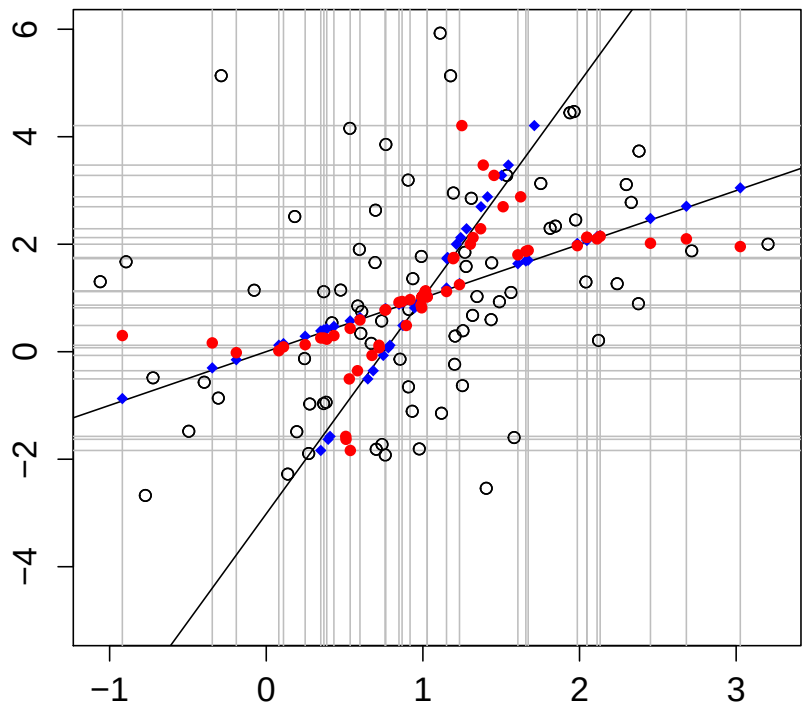


◆ – regression ● – zonoid depth ▲ – random forest

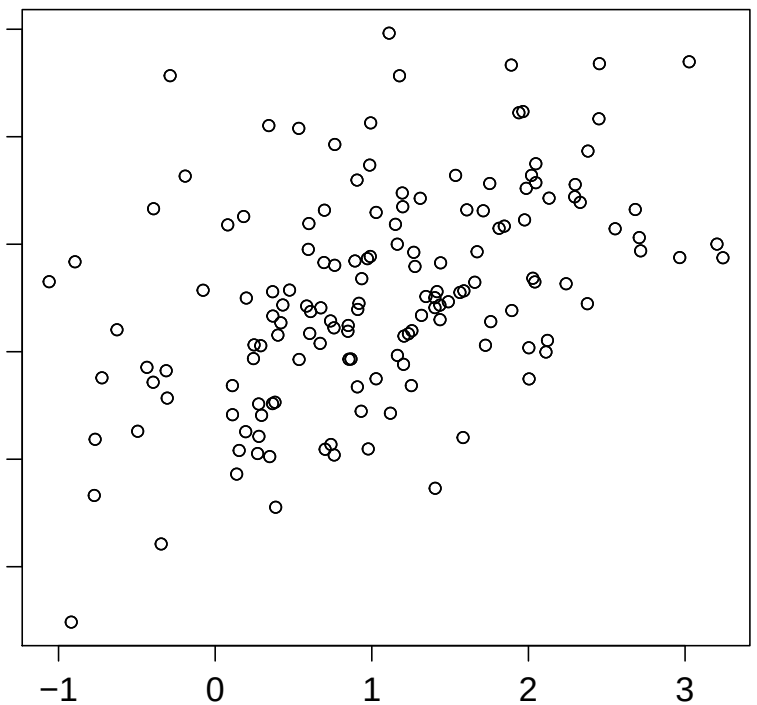
Imputation close to data and extrapolation

$$150 \text{ observasions of } X \sim \mathit{Normal}\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$$

Missing completely at random



Missing (not) at random

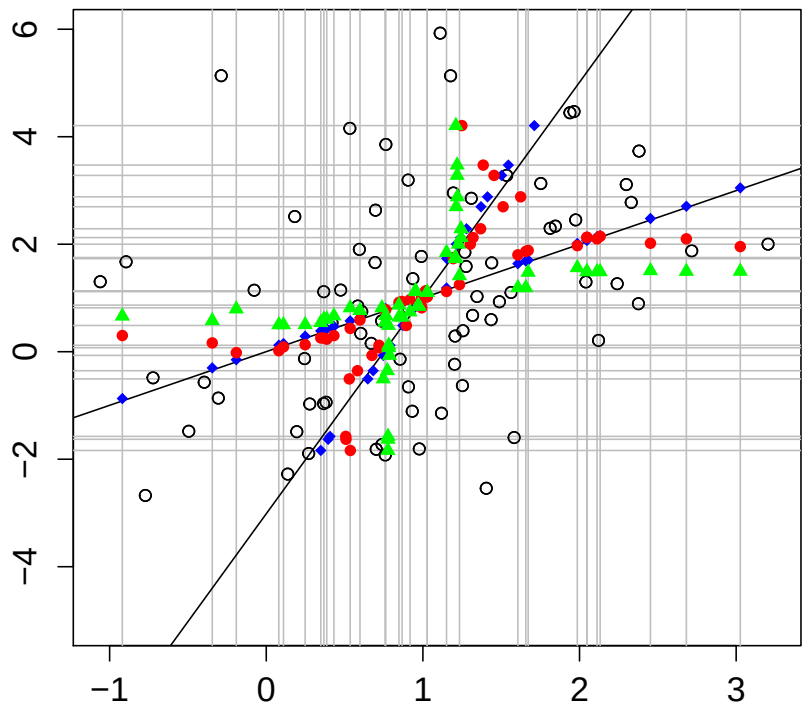


◆ – regression ● – zonoid depth ▲ – random forest

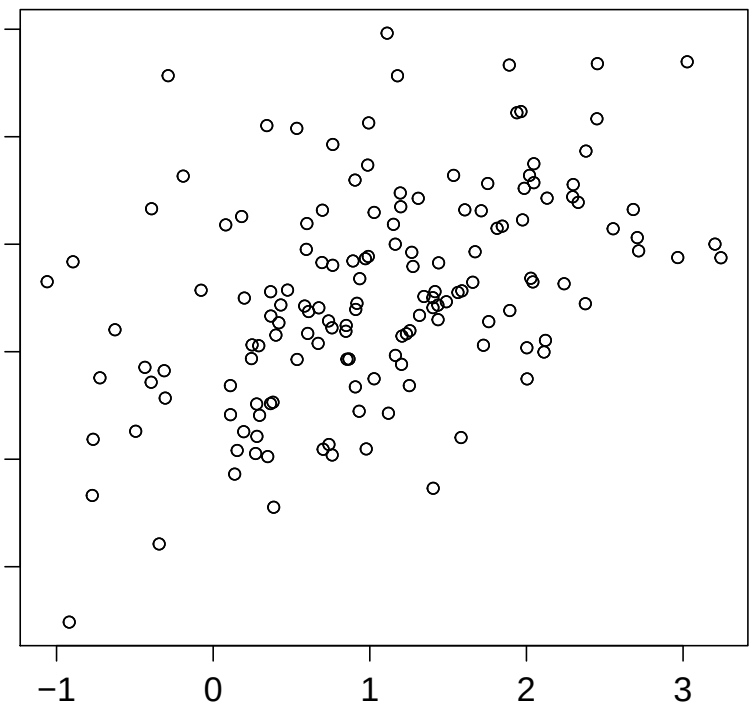
Imputation close to data and extrapolation

$$150 \text{ observasions of } X \sim \mathit{Normal}\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$$

Missing completely at random



Missing (not) at random

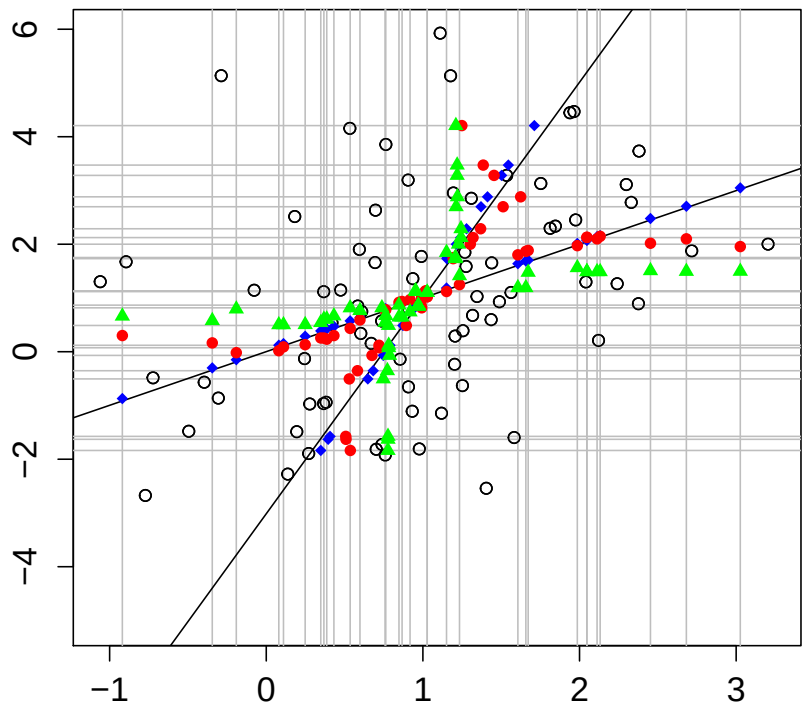


◆ – regression ● – zonoid depth ▲ – random forest

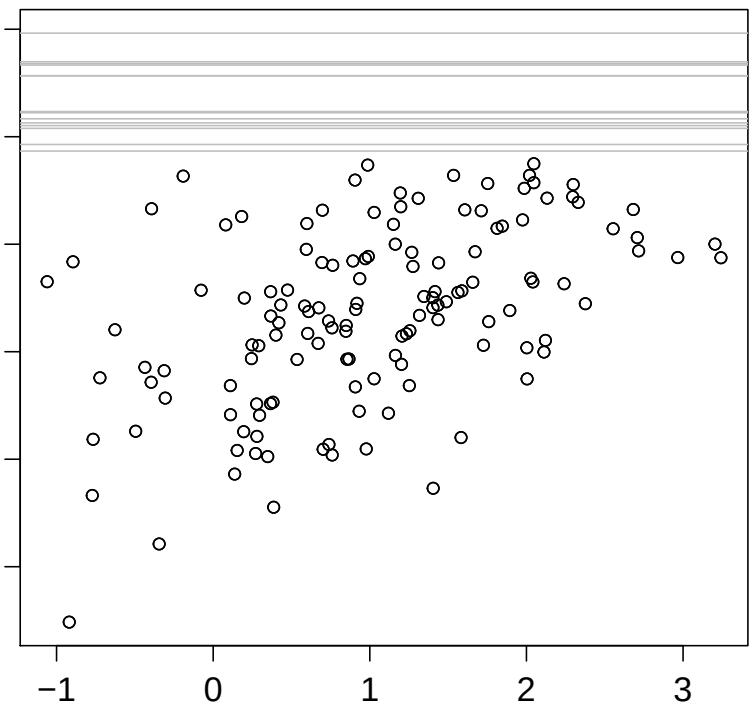
Imputation close to data and extrapolation

$$150 \text{ observasions of } X \sim \mathit{Normal}\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$$

Missing completely at random



Missing (not) at random

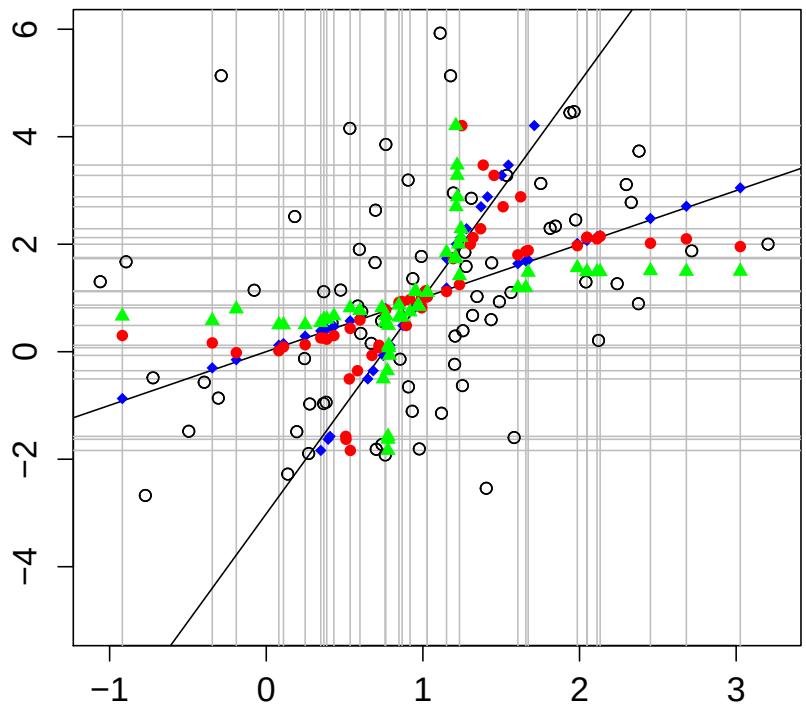


◆ – regression ● – zonoid depth ▲ – random forest

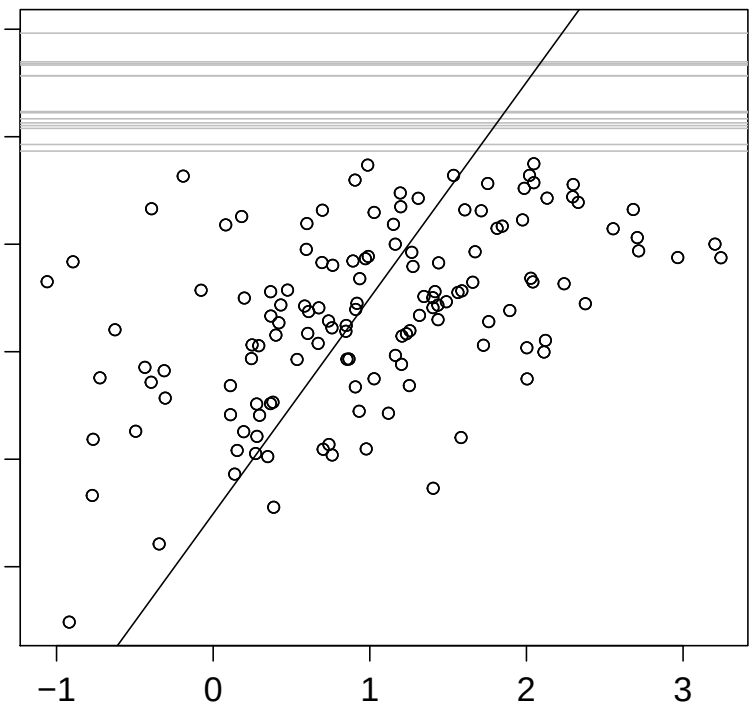
Imputation close to data and extrapolation

$$150 \text{ observasions of } X \sim \mathit{Normal}\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$$

Missing completely at random



Missing (not) at random

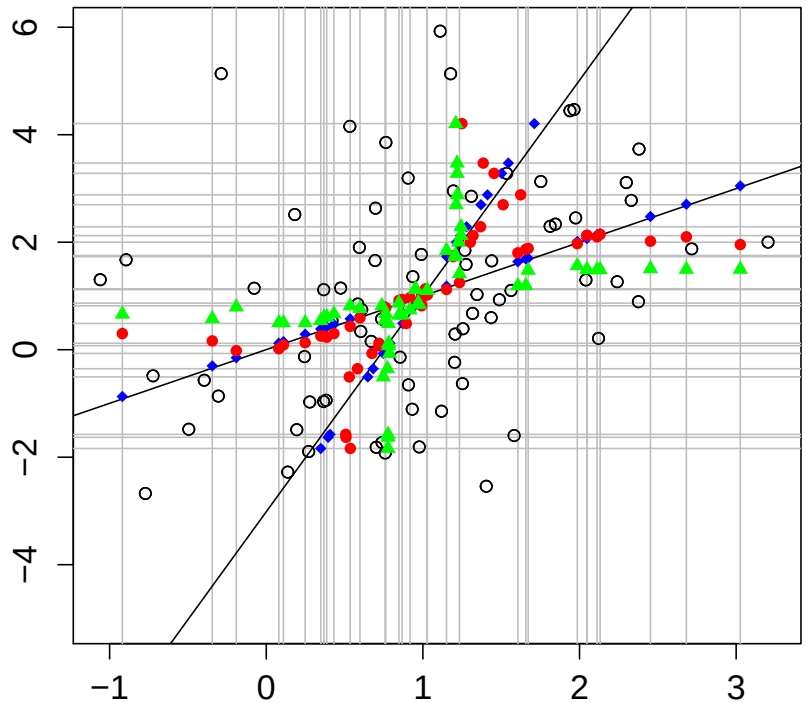


◆ – regression ● – zonoid depth ▲ – random forest

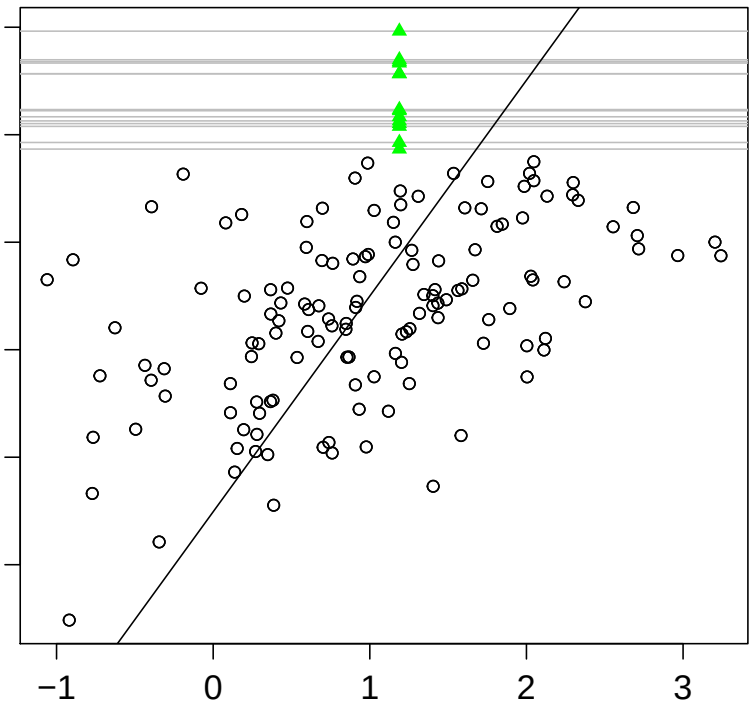
Imputation close to data and extrapolation

$$150 \text{ observasions of } X \sim \mathit{Normal}\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$$

Missing completely at random



Missing (not) at random

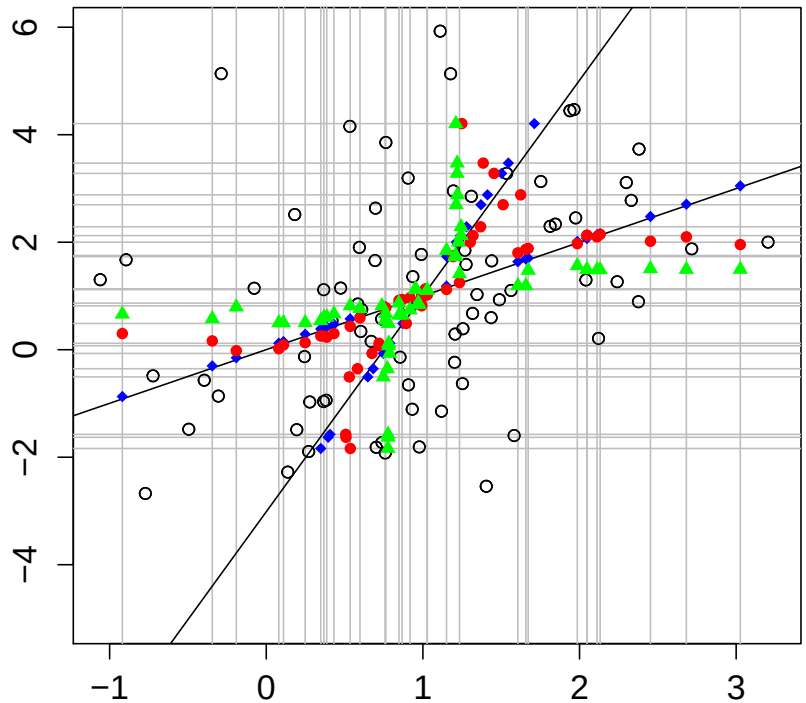


◆ – regression ● – zonoid depth ▲ – random forest

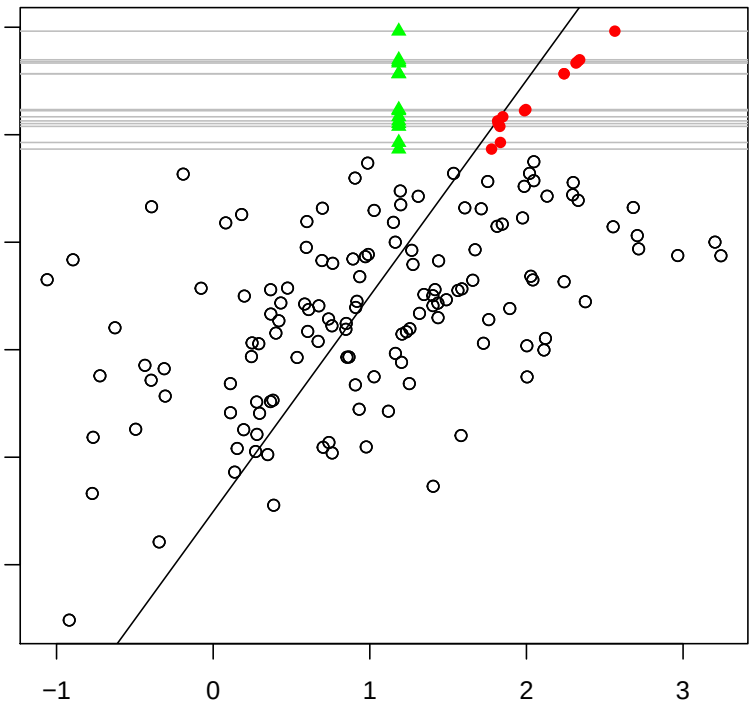
Imputation close to data and extrapolation

$$150 \text{ observasions of } X \sim \mathit{Normal}\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$$

Missing completely at random



Missing (not) at random

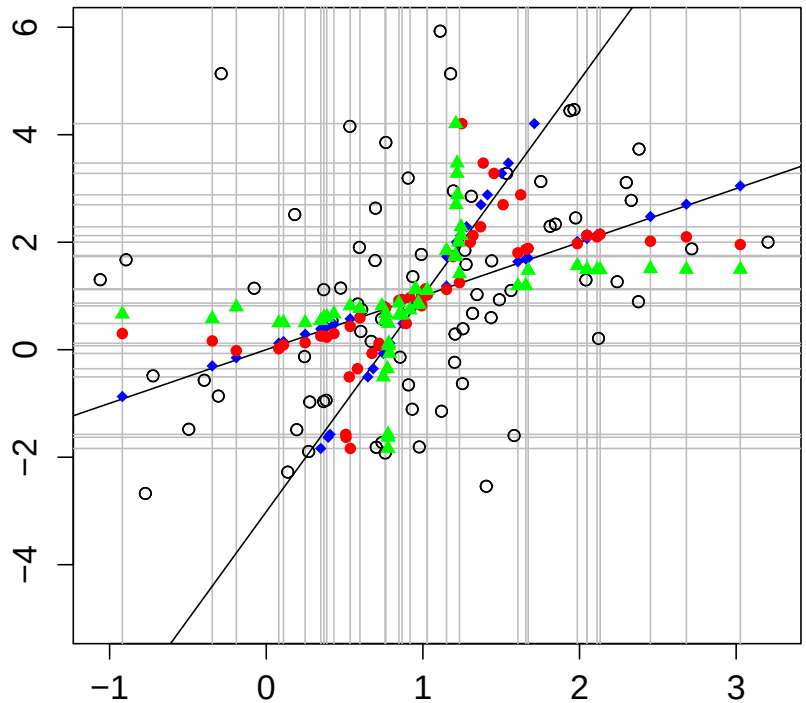


◆ – regression ● – zonoid depth ▲ – random forest

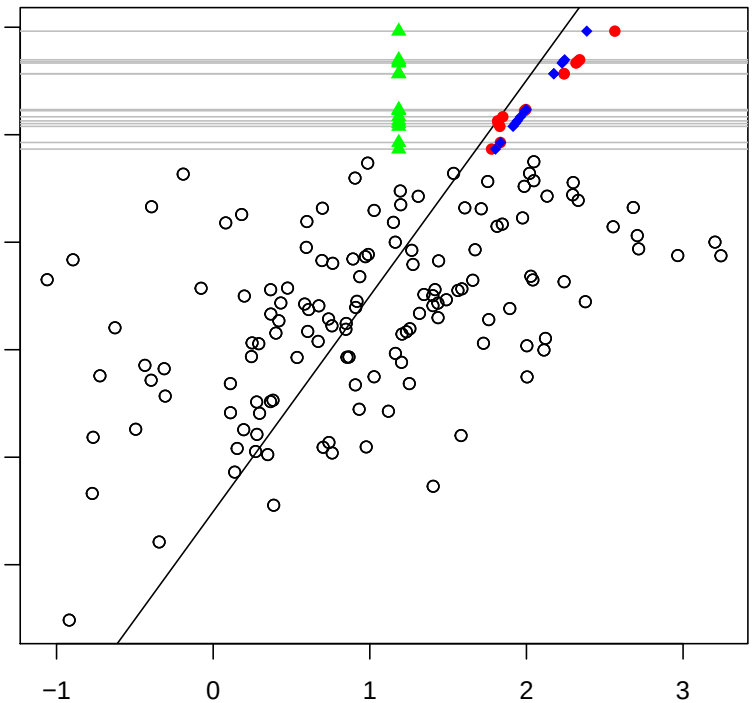
Imputation close to data and extrapolation

$$150 \text{ observasions of } X \sim \mathit{Normal}\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$$

Missing completely at random



Missing (not) at random

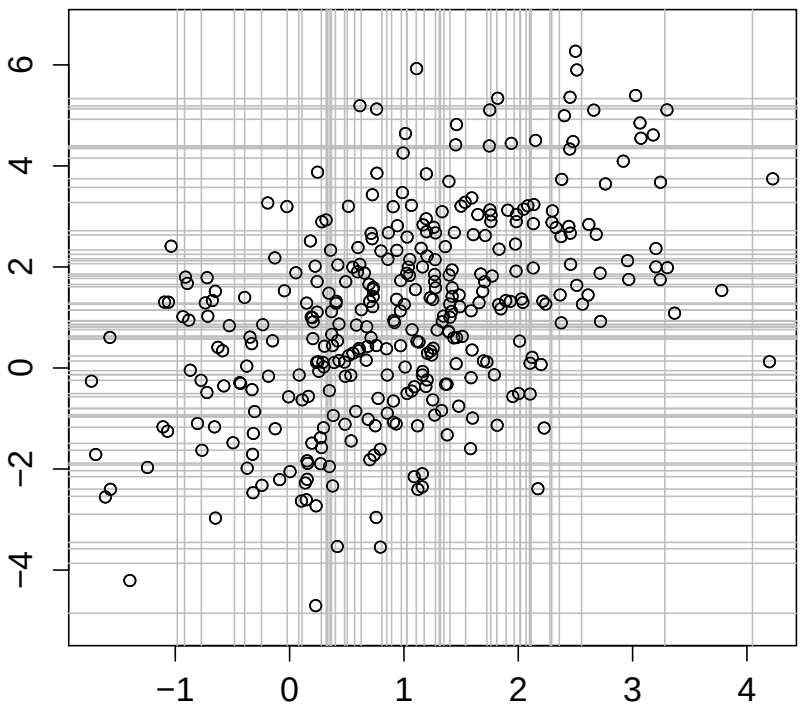


◆ – regression ● – zonoid depth ▲ – random forest

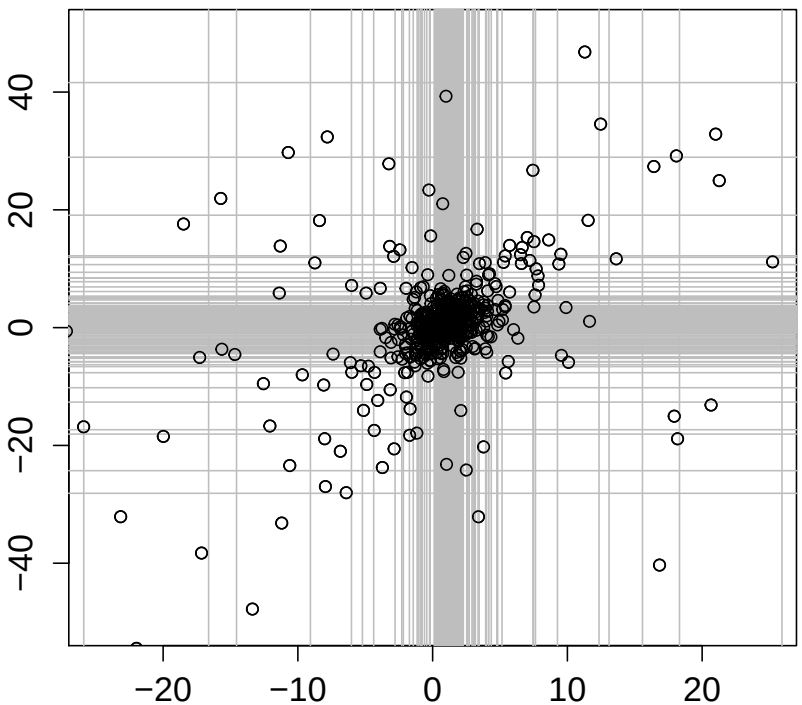
Robust imputation

15% (left) and 100% (right) from *Cauchy* $\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix} \right)$

Robust to outliers (MCAR)



Robust to disribution (MCAR)

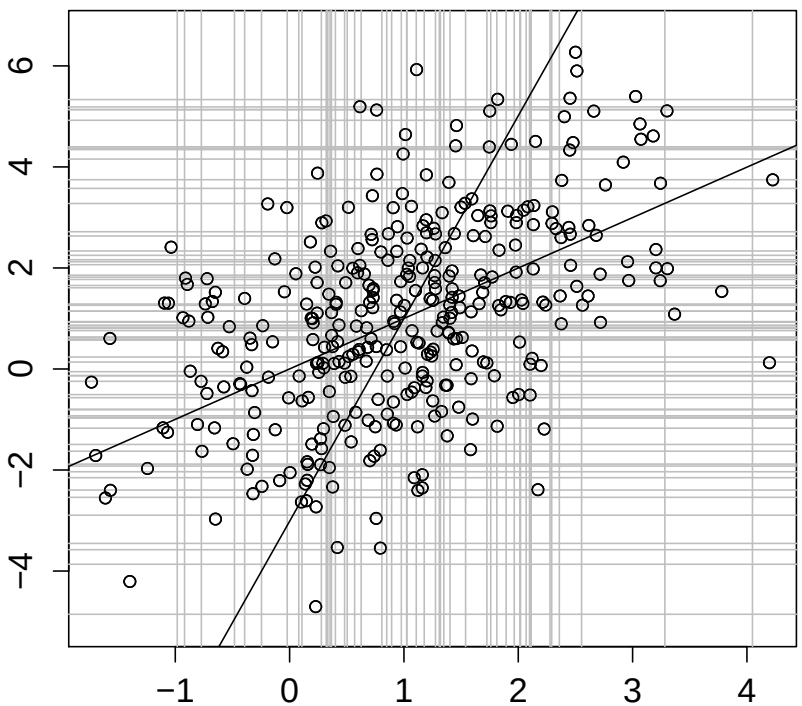


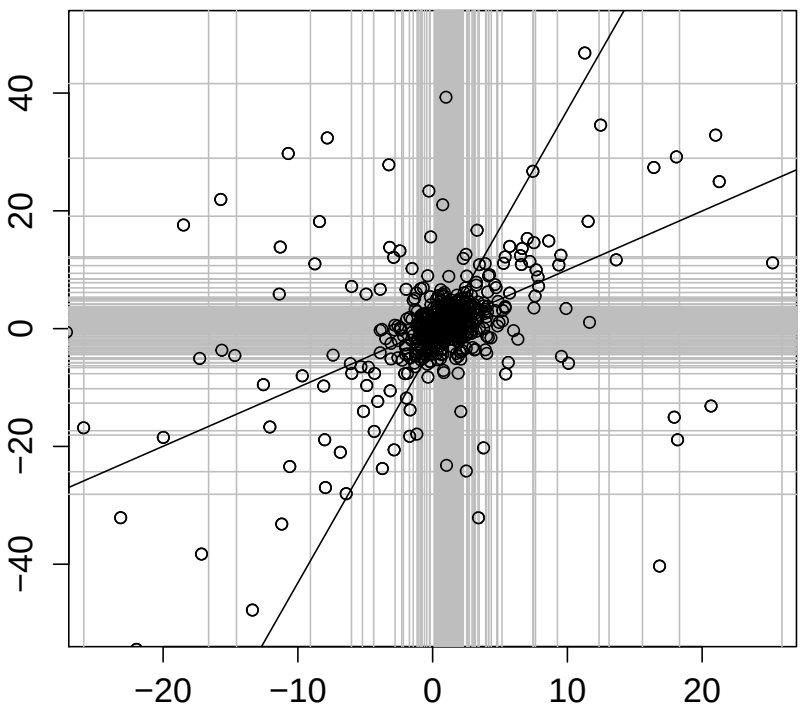
◆ – regression ● – Tukey depth

Robust imputation

15% (left) and 100% (right) from *Cauchy* $\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix} \right)$

Robust to outliers (MCAR) Robust to disribution (MCAR)



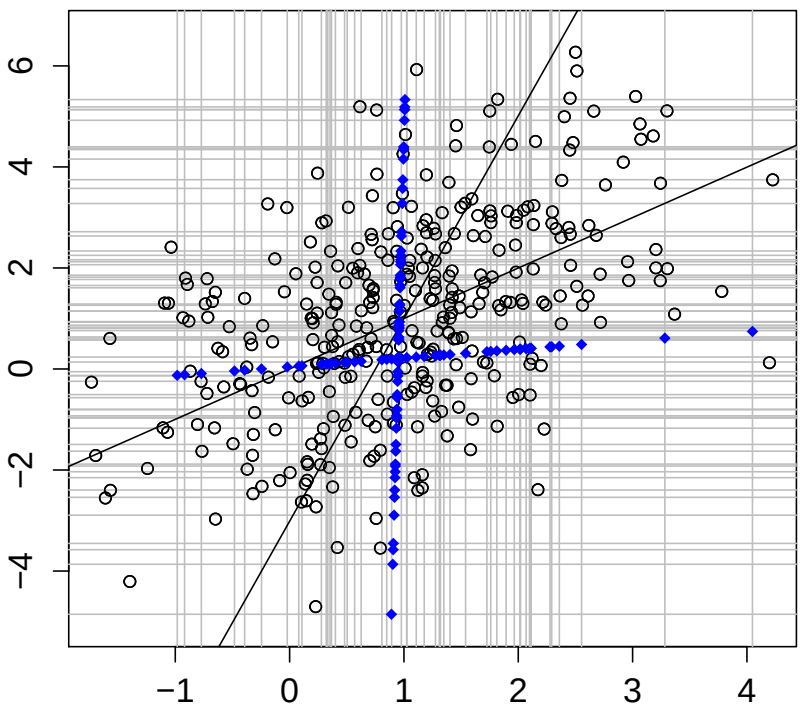


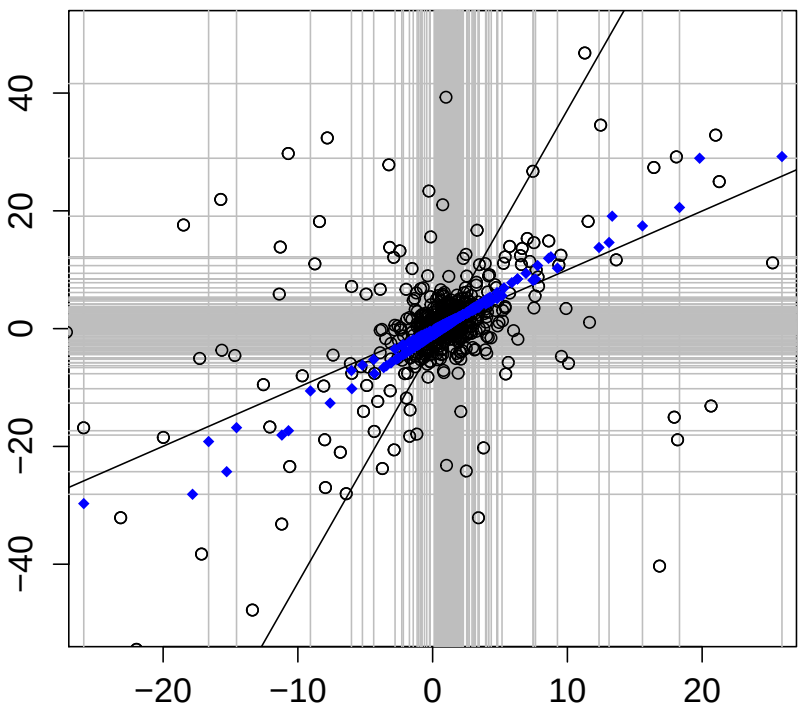
◆ – regression ● – Tukey depth

Robust imputation

15% (left) and 100% (right) from *Cauchy* $\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix} \right)$

Robust to outliers (MCAR) Robust to disribution (MCAR)



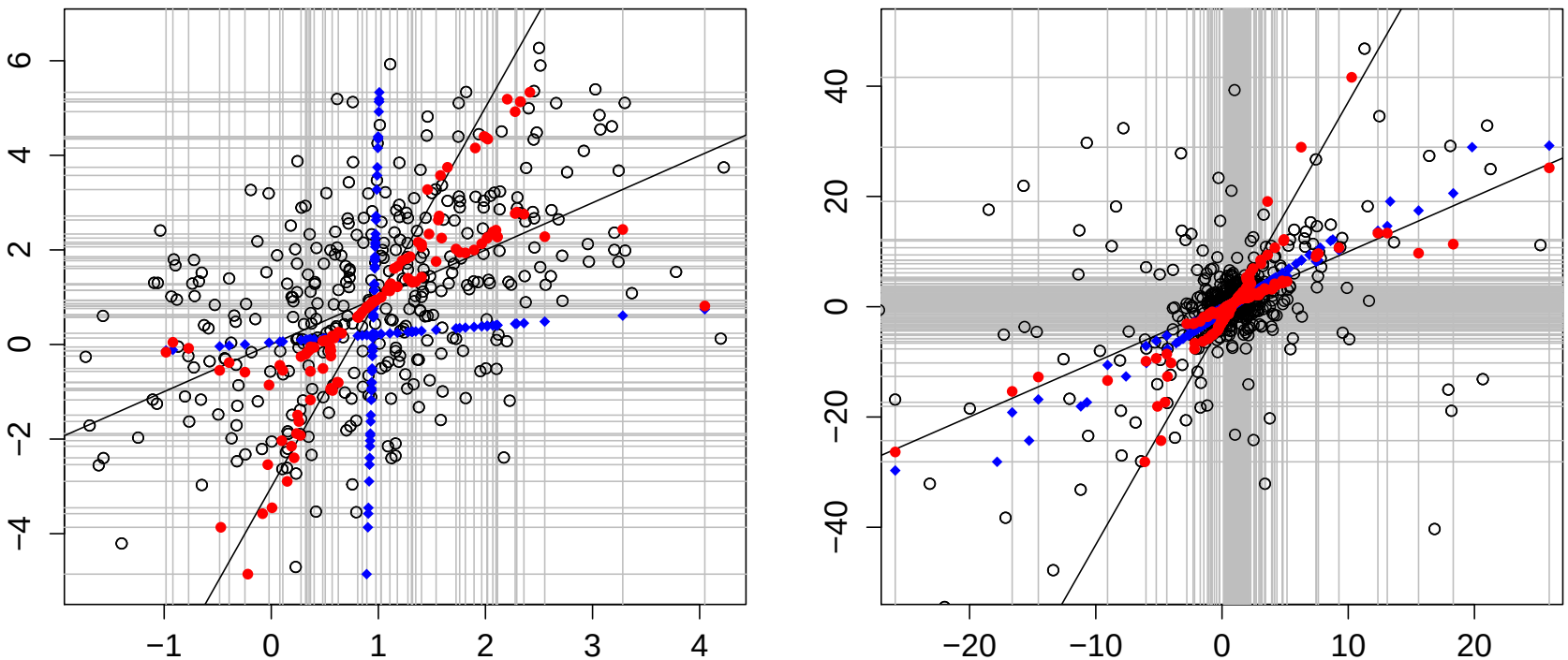


◆ – regression ● – Tukey depth

Robust imputation

15% (left) and 100% (right) from *Cauchy* $\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix} \right)$

Robust to outliers (MCAR) Robust to disribution (MCAR)



◆ – regression ● – Tukey depth