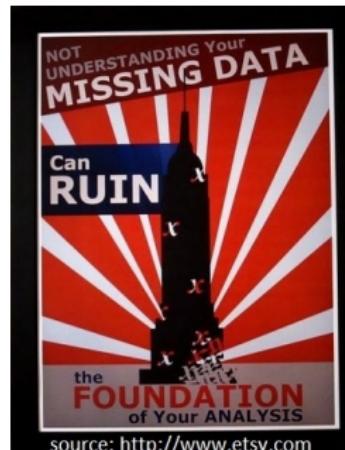


Contribution to missing values & principal components methods

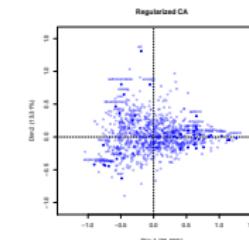
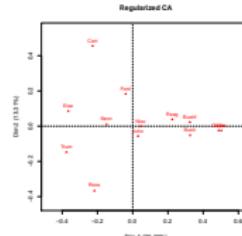
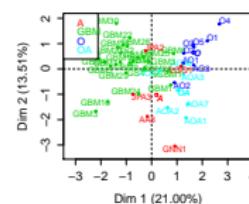
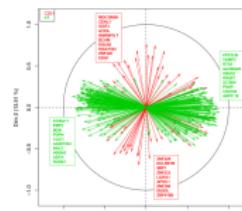
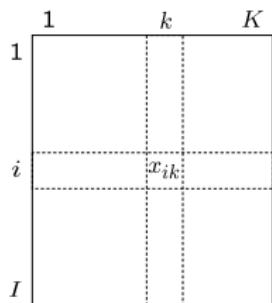
Julie Josse

Soutenance d'habilitation à diriger des recherches
Orsay University, 30 August 2016

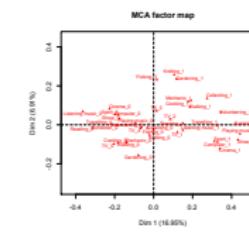
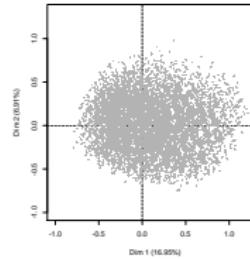


Research activities

Principal Component Analysis
 \Rightarrow continuous variables



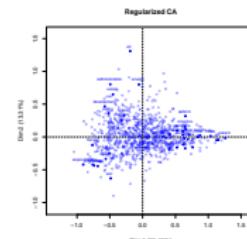
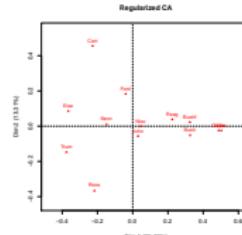
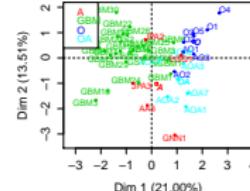
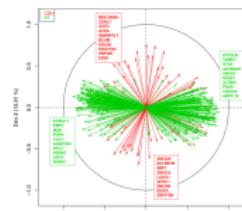
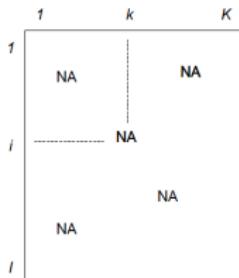
Correspondence Analysis
 \Rightarrow contingency table



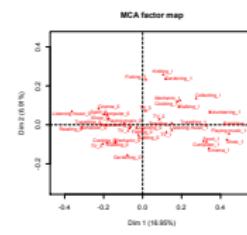
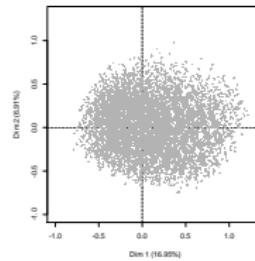
Multiple Correspondence Analysis
 \Rightarrow categorical variables

Research activities

Principal Component Analysis
 ⇒ continuous variables



Correspondence Analysis
 ⇒ contingency table



Multiple Correspondence Analysis
 ⇒ categorical variables

Research activities

⇒ Missing values

- PC with missing values
- Use PC methods for single and multiple imputation

⇒ Complete case

- Low rank matrix estimation (point estimates/ variability)
- Understanding the connections between PC and models

⇒ Transfer

- Selecting "tuning" parameters
- R packages: FactoMineR, missMDA, denoiseR

Outline

- ① New practices in visualization with PC methods
 - Singular values shrinkage
 - Bootstrap framework
- ② Inference with missing values using PC methods
 - Single and multiple imputation with PCA
 - Imputation with MCA
- ③ Models for PC methods
 - A multilogit bilinear model for MCA

Outline

1 Low-rank matrix estimation

- Singular values shrinkage
- Bootstrap framework

2 Inference with missing values using PC methods

- Single and multiple imputation with PCA
- Imputation with MCA

3 A multilogit bilinear model for MCA

Low-rank matrix estimation

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

⇒ Gaussian noise model: $X_{n \times p} = \mu_{n \times p} + \varepsilon_{n \times p}$, $\varepsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$
Image, recommendation system...

Low-rank matrix estimation

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

⇒ Gaussian noise model: $X_{n \times p} = \mu_{n \times p} + \varepsilon_{n \times p}$, $\varepsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$
Image, recommendation system...

⇒ Classical solution: truncated SVD $\hat{\mu}_k = \sum_{l=1}^k u_l d_l v_l^\top$

Hard thresholding: $d_l \cdot 1(l \leq k) = d_l \cdot 1(d_l \geq \lambda)$

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

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

Low-rank matrix estimation

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

⇒ Gaussian noise model: $X_{n \times p} = \mu_{n \times p} + \varepsilon_{n \times p}$, $\varepsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$
Image, recommendation system...

⇒ Classical solution: truncated SVD $\hat{\mu}_k = \sum_{l=1}^k u_l d_l v_l^\top$

Hard thresholding: $d_l \cdot 1(l \leq k) = d_l \cdot 1(d_l \geq \lambda)$

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

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

Not the best in MSE to recover μ !

Shrinking - thresholding SVD

$$\text{Soft thresholding: } \psi(d_I) = d_I \max\left(1 - \frac{\lambda}{d_I}, 0\right)$$

$$\hat{\mu}_\lambda = \sum_{I=1}^{\min\{n, p\}} u_I \psi(d_I) v_I^\top$$

$$\operatorname{argmin}_\mu \left\{ \|X - \mu\|_2^2 + \lambda \|\mu\|_* \right\}$$

\Rightarrow Selecting λ ? Risk: $\text{MSE}(\lambda) = \mathbb{E}\|\mu - \hat{\mu}_\lambda\|_2^2$

Stein Unbiased Risk Estimate (Candès et al., 2012) (σ^2 is known)

$$\text{SURE} = -np\sigma^2 + \sum_I^{\min(n,p)} \min(\lambda^2, d_I) + 2\sigma^2 \text{div}(\hat{\mu}_\lambda)$$

Shrinking - thresholding SVD

Soft thresholding: $\psi(d_l) = d_l \max\left(1 - \frac{\lambda}{d_l}, 0\right)$

$$\hat{\mu}_\lambda = \sum_{l=1}^{\min\{n, p\}} u_l \psi(d_l) v_l^\top$$

$$\operatorname{argmin}_\mu \left\{ \|X - \mu\|_2^2 + \lambda \|\mu\|_* \right\}$$

\Rightarrow Selecting λ ? Risk: $\text{MSE}(\lambda) = \mathbb{E}\|\mu - \hat{\mu}_\lambda\|_2^2$

Stein Unbiased Risk Estimate (Candès *et al.*, 2012) (σ^2 is known)

$$\text{SURE} = -np\sigma^2 + \text{RSS} + 2\sigma^2 \operatorname{div}(\hat{\mu}_\lambda)$$

Shrinking - thresholding SVD

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

\Rightarrow Non linear shrinkage: $\hat{\mu}^{\text{shrink}} = \sum_{l=1}^{\min\{n, p\}} u_l \psi(d_l) v_l^\top$

- Shabalin & Nobel (2013); Gavish & Donoho (2014). Asymptotic $n = n_p$ and $p \rightarrow \infty$, $n_p/p \rightarrow \beta$, $0 < \beta \leq 1$

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

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

$$\psi(d_l) = d_l \left(\frac{d_l^2 - \sigma^2}{d_l^2} \right) \cdot 1(l \leq k) \frac{\text{signal variance}}{\text{total variance}}$$

\Rightarrow Efron & Morris (1972) - empirical Bayes (Gaussian prior; PPCA)

Shrinking - thresholding SVD

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

\Rightarrow Non linear shrinkage: $\hat{\mu}^{\text{shrink}} = \sum_{l=1}^{\min\{n, p\}} u_l \psi(d_l) v_l^\top$

- Shabalin & Nobel (2013); Gavish & Donoho (2014). Asymptotic $n = n_p$ and $p \rightarrow \infty$, $n_p/p \rightarrow \beta$, $0 < \beta \leq 1$

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

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

$$\psi(d_l) = d_l \left(\frac{d_l^2 - \sigma^2}{d_l^2} \right) \cdot 1(l \leq k) = \left(d_l - \frac{\sigma^2}{d_l} \right) \cdot 1(l \leq k)$$

\Rightarrow Efron & Morris (1972) - empirical Bayes (Gaussian prior; PPCA)

Shrinking - thresholding SVD

- Josse & Sardy (2014). Adaptive estimator. Finite sample.

$$\psi(d_I) = d_I \max \left(1 - \frac{\lambda^\gamma}{d_I^\gamma}, 0 \right)$$

$$\operatorname{argmin}_\mu \left\{ \|X - \mu\|_2^2 + \lambda^\gamma \|\mu\|_{*,w}, \text{ with } w_I = 1/d_I^{\gamma-1} \right\}$$

⇒ Selecting (λ, γ) ? Data dependent:

$$\text{SURE} = -np\sigma^2 + \sum_{I=1}^{\min(n,p)} d_I^2 \min \left(\frac{\lambda^{2\gamma}}{d_I^{2\gamma}}, 1 \right) + 2\sigma^2 \operatorname{div}(\hat{\mu}_{\tau,\gamma})$$

$$\text{GSURE} = \frac{\text{RSS}}{(1 - \operatorname{div}(\hat{\mu}_{\tau,\gamma})/(np))^2} \quad (\sigma^2 \text{ unknown})$$

SVD via an autoencoder

Gaussian model: $X_{n \times p} = \mu_{n \times p} + \varepsilon_{n \times p}$, with $\varepsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$

⇒ Classical formulation:

- $\hat{\mu}_k = \operatorname{argmin}_{\mu} \left\{ \|X - \mu\|_2^2 : \operatorname{rank}(\mu) \leq k \right\}$

⇒ Another formulation: autoencoder (Boulard & Kamp, 1988)

- $\hat{\mu}_k = XB_k$, where $B_k = \operatorname{argmin}_B \left\{ \|X - XB\|_2^2 : \operatorname{rank}(B) \leq k \right\}$

SVD via an autoencoder

Gaussian model: $X_{n \times p} = \mu_{n \times p} + \varepsilon_{n \times p}$, with $\varepsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$

\Rightarrow Classical formulation:

- $\hat{\mu}_k = \operatorname{argmin}_{\mu} \left\{ \|X - \mu\|_2^2 : \operatorname{rank}(\mu) \leq k \right\}$

\Rightarrow Another formulation: autoencoder (Boulard & Kamp, 1988)

- $\hat{\mu}_k = XB_k$, where $B_k = \operatorname{argmin}_B \left\{ \|X - XB\|_2^2 : \operatorname{rank}(B) \leq k \right\}$

$$\text{TSVD: } \hat{\mu}_k = \sum_{l=1}^k u_l d_l v_l^\top = \sum_{l=1}^k f_l v_l^\top$$

$$F_{n \times k} = XV(V'V)^{-1}$$

$$V'_{k \times p} = (F'F)^{-1}F'X$$

$$\hat{\mu}_k = FV'$$

$$= XV(V'V)^{-1}V' = XPV_V$$

Parametric bootstrap

⇒ Autoencoder: compress X

$$\hat{\mu}_k = XB_k, \text{ where } B_k = \operatorname{argmin}_B \left\{ \|X - XB\|_2^2 : \operatorname{rank}(B) \leq k \right\}$$

⇒ Aim: recover μ

$$\hat{\mu}_k^* = XB_k^*, B_k^* = \operatorname{argmin}_B \left\{ \mathbb{E}_{X \sim \mathcal{L}(\mu)} \left[\| \mu - XB \|^2_2 \right] : \operatorname{rank}(B) \leq k \right\}$$

Parametric bootstrap

⇒ Autoencoder: compress X

$$\hat{\mu}_k = XB_k, \text{ where } B_k = \operatorname{argmin}_B \left\{ \|X - XB\|_2^2 : \operatorname{rank}(B) \leq k \right\}$$

⇒ Aim: recover μ

$$\hat{\mu}_k^* = XB_k^*, B_k^* = \operatorname{argmin}_B \left\{ \mathbb{E}_{X \sim \mathcal{L}(\mu)} [\|\mu - XB\|_2^2] : \operatorname{rank}(B) \leq k \right\}$$

⇒ Parametric bootstrap: plug μ by X

$$\hat{\mu}_k^{\text{stable}} = X\hat{B}_k, \hat{B}_k = \operatorname{argmin}_B \left\{ \mathbb{E}_{\tilde{X} \sim \mathcal{L}(X)} [\|X - \tilde{X}B\|_2^2] : \operatorname{rank}(B) \leq k \right\}$$

$$\text{Model } X = \mu + \varepsilon \Rightarrow \tilde{X} = X + \varepsilon$$

Stable Autoencoder (Josse & Wager, 2016)

⇒ Bootstrap = Feature noising

$$\begin{aligned}\hat{\mu}_k^{\text{stable}} &= X\hat{B}_k, \hat{B}_k &= \operatorname{argmin}_B \left\{ \mathbb{E}_{\tilde{X} \sim \mathcal{L}(X)} \left[\|X - \tilde{X}B\|_2^2 \right] : \operatorname{rk}B \leq k \right\} \\ \hat{B}_k &= \operatorname{argmin}_B \left\{ \mathbb{E}_{\varepsilon} \left[\|X - (X + \varepsilon)B\|_2^2 \right] : \operatorname{rk}B \leq k \right\}\end{aligned}$$

Stable Autoencoder (Josse & Wager, 2016)

\Rightarrow Bootstrap = Feature noising

$$\begin{aligned}\hat{\mu}_k^{\text{stable}} &= X\hat{B}_k, \hat{B}_k &= \operatorname{argmin}_B \left\{ \mathbb{E}_{\tilde{X} \sim \mathcal{L}(X)} \left[\|X - \tilde{X}B\|_2^2 \right] : \operatorname{rk} B \leq k \right\} \\ \hat{B}_k &= \operatorname{argmin}_B \left\{ \mathbb{E}_{\varepsilon} \left[\|X - (X + \varepsilon)B\|_2^2 \right] : \operatorname{rk} B \leq k \right\}\end{aligned}$$

Theorem

\Rightarrow Equivalent to a ridge autoencoder problem

$$\hat{\mu}_{\lambda,k}^{\text{stable}} = X\hat{B}_\lambda, \quad \hat{B}_\lambda = \operatorname{argmin}_B \left\{ \|X - XB\|_2^2 + \lambda \|B\|_2^2 \right\} : \operatorname{rk} B \leq k$$

Stable Autoencoder (Josse & Wager, 2016)

\Rightarrow Bootstrap = Feature noising

$$\begin{aligned}\hat{\mu}_k^{\text{stable}} &= X\hat{B}_k, \hat{B}_k &= \operatorname{argmin}_B \left\{ \mathbb{E}_{\tilde{X} \sim \mathcal{L}(X)} \left[\|X - \tilde{X}B\|_2^2 \right] : \operatorname{rk} B \leq k \right\} \\ \hat{B}_k &= \operatorname{argmin}_B \left\{ \mathbb{E}_{\varepsilon} \left[\|X - (X + \varepsilon)B\|_2^2 \right] : \operatorname{rk} B \leq k \right\}\end{aligned}$$

Theorem

\Rightarrow Equivalent to a ridge autoencoder problem

$$\hat{\mu}_{\lambda,k}^{\text{stable}} = X\hat{B}_\lambda, \quad \hat{B}_\lambda = \operatorname{argmin}_B \left\{ \|X - XB\|_2^2 + \lambda \|B\|_2^2 \right\} : \operatorname{rk} B \leq k$$

\Rightarrow Solution: singular-value shrinkage estimator

$$\hat{\mu}_{\lambda,k}^{\text{stable}} = \sum_{l=1}^k u_l \frac{d_l}{1 + \lambda/d_l^2} v_l^\top \text{ with } \lambda = n\sigma^2$$

$$\hat{\mu}_{\lambda,k}^{\text{stable}} = X\hat{B}_\lambda, \quad \hat{B}_\lambda = (X^\top X + S)^{-1} X^\top X, \quad S = \operatorname{diag}(\lambda)$$

Stable Autoencoder (Josse & Wager, 2016)

\Rightarrow Bootstrap = Feature noising

$$\begin{aligned}\hat{\mu}_k^{\text{stable}} &= X\hat{B}_k, \hat{B}_k &= \operatorname{argmin}_B \left\{ \mathbb{E}_{\tilde{X} \sim \mathcal{L}(X)} \left[\|X - \tilde{X}B\|_2^2 \right] : \operatorname{rk} B \leq k \right\} \\ \hat{B}_k &= \operatorname{argmin}_B \left\{ \mathbb{E}_{\varepsilon} \left[\|X - (X + \varepsilon)B\|_2^2 \right] : \operatorname{rk} B \leq k \right\}\end{aligned}$$

Theorem

\Rightarrow Equivalent to a ridge autoencoder problem

$$\hat{\mu}_{\lambda,k}^{\text{stable}} = X\hat{B}_\lambda, \quad \hat{B}_\lambda = \operatorname{argmin}_B \left\{ \|X - XB\|_2^2 + \lambda \|B\|_2^2 \right\} : \operatorname{rk} B \leq k$$

\Rightarrow Solution: singular-value shrinkage estimator

$$\hat{\mu}_{\lambda,k}^{\text{stable}} = \sum_{l=1}^k u_l \frac{d_l}{1 + \lambda/d_l^2} v_l^\top \text{ with } \lambda = n\sigma^2$$

$$\hat{\mu}_{\lambda,k}^{\text{stable}} = X\hat{B}_\lambda, \quad \hat{B}_\lambda = (X^\top X + S)^{-1} X^\top X, \quad S = \operatorname{diag}(\lambda)$$

Feature noising and regularization in regression

Bishop (1995). Training with noise is equivalent to Tikhonov regularization. *Neural computation.*

$$\hat{\beta} = \operatorname{argmin}_{\beta} \left\{ \mathbb{E}_{\varepsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)} [\|Y - (X + \varepsilon)\beta\|_2^2] \right\}$$

Many noisy copies of the data and average out the auxiliary noise is equivalent to ridge regularization with $\lambda = n\sigma^2$:

$$\hat{\beta}_\lambda^{(R)} = \operatorname{argmin}_{\beta} \left\{ \|Y - X\beta\| + \lambda \|\beta\|_2^2 \right\}$$

⇒ Control overfitting by artificially corrupting the training data

Iterative Stable Autoencoder

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

Iterative algorithm. Init: $\hat{\mu} = X$, $\hat{B}_\lambda = (X^\top X + S)^{-1} X^\top X$

$$① \hat{\mu} = X\hat{B}$$

$$② \hat{B} = (\hat{\mu}^\top \hat{\mu} + S)^{-1} \hat{\mu}^\top \hat{\mu} \quad S = \operatorname{diag}(n\sigma^2)$$

Iterative Stable Autoencoder

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

Iterative algorithm. Init: $\hat{\mu} = X$, $\hat{B}_\lambda = (X^\top X + S)^{-1} X^\top X$

$$① \quad \hat{\mu} = X\hat{B}$$

$$② \quad \hat{B} = (\hat{\mu}^\top \hat{\mu} + S)^{-1} \hat{\mu}^\top \hat{\mu} \quad S = \operatorname{diag}(n\sigma^2)$$

A centered parametric bootstrap

$$\hat{\mu}_k^* = XB_k^*, B_k^* = \operatorname{argmin}_B \left\{ \mathbb{E}_{X \sim \mathcal{L}(\mu)} \left[\left\| \mu - XB \right\|_2^2 \right] : \operatorname{rk} B \leq k \right\}$$

$\Rightarrow X$ as a proxy for μ

$$\hat{\mu}_k^{\text{stb}} = X\hat{B}_k, \hat{B}_k = \operatorname{argmin}_B \left\{ \mathbb{E}_{\tilde{X} \sim \mathcal{L}(X)} \left[\left\| X - \tilde{X}B \right\|_2^2 \right] : \operatorname{rk} B \leq k \right\}$$

$\Rightarrow \hat{\mu} = X\hat{B}$ as a proxy for μ

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

Iterative Stable Autoencoder

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

Iterative algorithm. Init: $\hat{\mu} = X$, $\hat{B}_\lambda = (X^\top X + S)^{-1} X^\top X$

$$① \quad \hat{\mu} = X\hat{B}$$

$$② \quad \hat{B} = (\hat{\mu}^\top \hat{\mu} + S)^{-1} \hat{\mu}^\top \hat{\mu} \quad S = \operatorname{diag}(n\sigma^2)$$

Theorem

ISA converges to a low rank solution

Proposition

In the isotropic Gaussian case ISA converges to

$$\hat{\mu}^{\text{ISA}} = \sum_{l=1}^{\min\{n, p\}} u_l \psi(d_l) v_l^\top, \quad \psi(d) = \begin{cases} \frac{1}{2} \left(d + \sqrt{d^2 - 4n\sigma^2} \right) & \text{for } d^2 \geq 4n\sigma^2, \\ 0 & \text{else.} \end{cases}$$

⇒ Connection with optimal shrinkage under operator norm loss

Properties

$$X = \mu + \varepsilon$$

⇒ Different noise regimes of predilection:

- low noise: truncated SVD
- moderate noise: Asympt, SA, ISA (nonlinear transformation)
- high noise (SNR low, k large): Soft thresholding (collapses in other regimes)

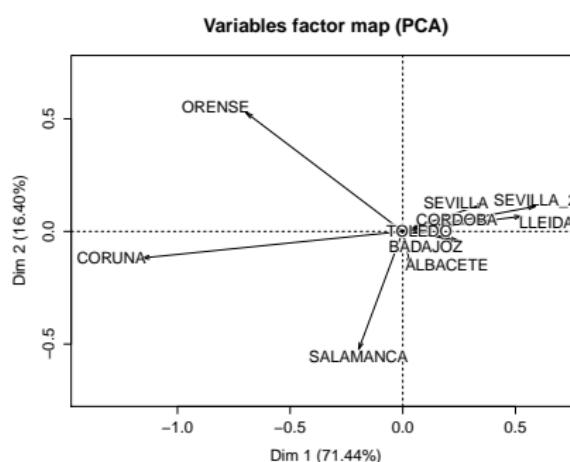
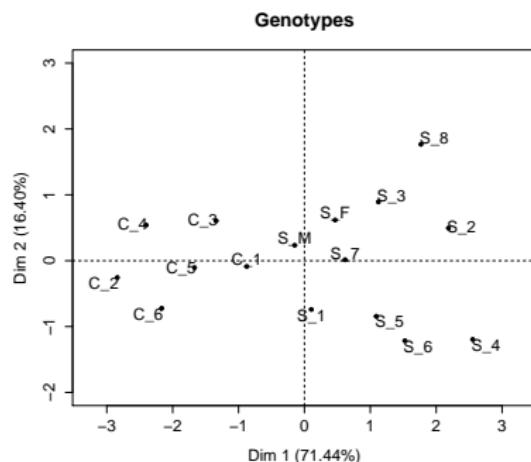
⇒ Adaptive estimator is very flexible! (selection (λ, γ) with SURE)

⇒ ISA performs well in MSE (other than Gaussian noise)

Principal component analysis

⇒ Do the genotypes have different performances in different environments?

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



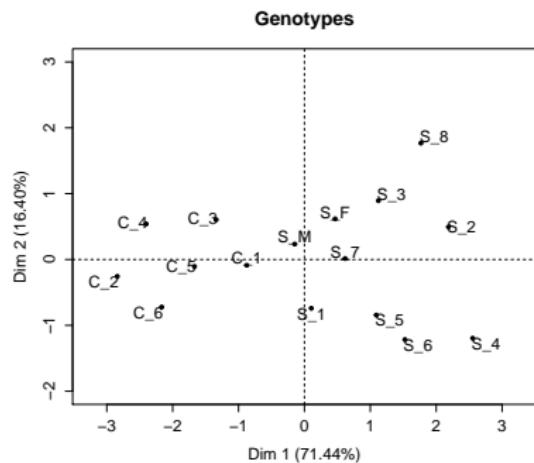
Scores $F = UD^{1/2}$ and variables coordinates $VD^{1/2}$

⇒ Separates Complete (C) and Substituted (S)

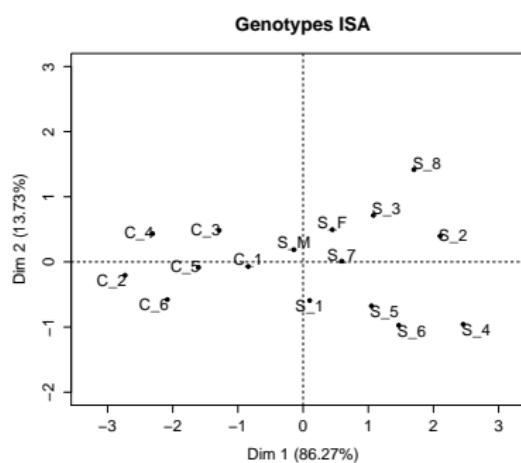
Exploratory multivariate analysis

$\Rightarrow X = \mu + \varepsilon$ fixed effects: interest in rows (not *i.i.d*) and columns

$\Rightarrow \mu$ better estimated with shrinked estimator $\hat{\mu}^{\text{ISA}}$ than $\hat{\mu}_k^{\text{TSVD}}$



PCA (TSVD)

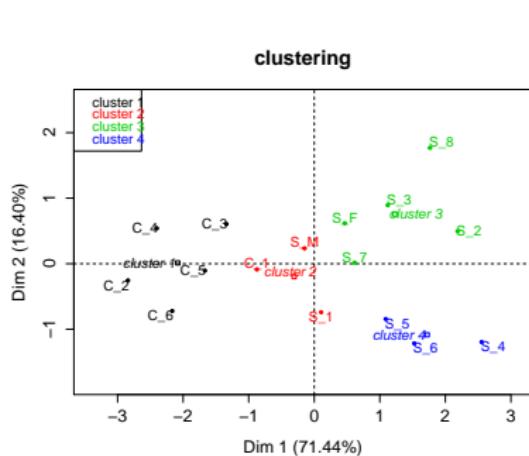


Regularized PCA (ISA)

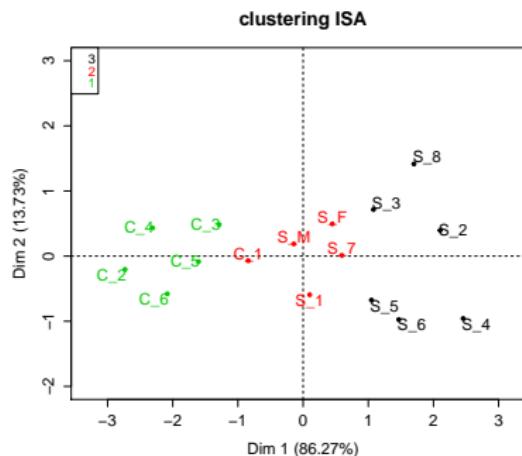
Exploratory multivariate analysis

$\Rightarrow X = \mu + \varepsilon$ fixed effects: interest in rows (not *i.i.d*) and columns

$\Rightarrow \mu$ better estimated with shrinked estimator $\hat{\mu}^{\text{ISA}}$ than $\hat{\mu}_k^{\text{TSVD}}$



PCA (TSVD)



Regularized PCA (ISA)

Discussion: low rank estimation

- Bayesian SVD model (constraints): $X = \mu + \varepsilon = UDV' + \varepsilon$

Hoff, 2009: Uniform priors U, V special cases of von Mises-Fisher distribution, Stiefel manifold.

$\Rightarrow \hat{\mu}$ (shrunk) - uncertainty $(\hat{\mu}^1, \dots, \hat{\mu}^B)$, also on k .

Future research:

- Count data $X_{ij} \sim \mathcal{P}(\mu_{ij}) \quad \log \mu_{ij} = \alpha_i + \beta_j + \sum_{l=1}^k d_l u_{il} v_{jl}$

$$\arg \min_{d,u,v} \left\{ - \sum_{i,j} \left[X_{ij} \sum_{l=1}^k d_l u_{il} v_{jl} - \exp \left(\sum_{l=1}^k d_l u_{il} v_{jl} \right) \right] + \sum_{l=1}^L \gamma_l d_l \right\}$$

\Rightarrow Optimal / data driven choice of weights γ_l ?

\Rightarrow Empirical Bayes: $\mu_{ij} \sim \text{Gamma}(\alpha, \theta_{ij})$; $\log \theta_{ij} = \sum_{l=1}^k d_l u_{il} v_{jl}$;
von Mises-Fisher Uniform priors

- Bayesian ISA?
- Bagging PCA to shrink - Adaptive with outliers

Outline

1 Low-rank matrix estimation

- Singular values shrinkage
- Bootstrap framework

2 Inference with missing values using PC methods

- Single and multiple imputation with PCA
- Imputation with MCA

3 A multilogit bilinear model for MCA

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.

⇒ Still an issue in the "big data" area



Data integration: data from different sources

Hospital data set

7000 patients/ 250 variables

	Centre	Mecanisme	Age	Sexe	Poids	Taille	BMI	PAS	PAD
1	Beaujon	Chute d'une hauteur	54	m	85	NR	NR	180	110
2	Lille	Autre	33	m	80	1.8	24.69	130	62
3	Pitie Salpetriere	AVP 2 roues motorise	26	m	NR	NR	NR	131	62
4	Beaujon	AVP 2 roues motorise	63	m	80	1.8	24.69	145	89
6	Pitie Salpetriere	AVP bicyclette	33	m	75	NR	NR	104	86
7	Pitie Salpetriere	AVP pieton	30	m	NR	NR	NR	107	66
9	HEGP	Arme blanche	16	m	98	1.92	26.58	118	54
10	Toulon	AVP (voiture. camion. bus)	20	m	NR	NR	NR	124	73
11	Bicetre	Chute d'une hauteur	61	m	84	1.7	29.07	144	105
12	Bicetre	AVP 2 roues motorise	20	m	80	1.85	23.37	124	67
13	Bicetre	AVP pieton	38	m	95	1.9	26.32	88	60
.....									

	SpO2	Temperature	Lactates	Hb	Glasgow
1	97	35.6	<NA>	12.7	12	
2	100	36.5	4.8	11.1	15	
3	100	36	3.9	11.4	3	
4	100	36.7	1.66	13	15	
6	100	36	NF	14.4	15	
7	100	36.6	NF	14.3	15	
9	100	37.5	13	15.9	15	
10	100	36.9	NF	13.7	15	
11	100	36.6	1.2	14.2	14	
12	100	37.2	NF	14.7	15	
13	99	33.6	3.5	12.3	13	
.....						

Recommended methods

⇒ Modify the method, the estimation process to deal with missing values. Maximum likelihood: EM algorithm to obtain point estimates (+ other algorithms for their variability)

One specific algorithm for each statistical method...

⇒ Imputation (multiple) to get a completed data set on which you can perform any statistical method

Missing values in PCA

⇒ PCA: least squares

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

⇒ PCA with missing values: weighted least squares

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

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

- 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) $(\hat{\mu}^\ell)^k = U^\ell D^\ell V^{\ell \top}$ $X^\ell = W * X + (1 - W) * \hat{\mu}^\ell$
- 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) $(\hat{\mu}^\ell)^k = U^\ell D^\ell V^{\ell \top}$ $X^\ell = W * X + (1 - W) * \hat{\mu}^\ell$
- ③ steps of **estimation** and **imputation** are repeated
 - $\Rightarrow \hat{\mu}$ from incomplete data: EM algorithm
 - $X = \mu + \varepsilon$, $\varepsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ 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) $(\hat{\mu}^\ell)^k = U^\ell D^\ell V^{\ell \top}$ $X^\ell = W * X + (1 - W) * \hat{\mu}^\ell$
- ③ steps of **estimation** and **imputation** are repeated
 - $\Rightarrow \hat{\mu}$ from incomplete data: EM algorithm
 - $X = \mu + \varepsilon$, $\varepsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ 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

- ⇒ Overfitting issues of iterative PCA: many parameters ($U_{n \times k}$, $V_{k \times p}$)/observed values (k large - many NA); noisy data
- ⇒ Regularized versions. Init - estimation - imputation steps:

the imputation: $\hat{\mu}_{ij}^{\text{PCA}} = \sum_{l=1}^k d_s u_l v_l^\top$ is replaced by

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

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

SoftImpute for large matrices. T. Hastie, R. Mazumder, 2015

Selecting regularization parameters

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

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

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

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

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

Stein Unbiased Risk Estimate (Candès et al., 2012) (σ^2 known)

$$\text{SURE} = -np\sigma^2 + \text{RSS} + 2\sigma^2 \text{div}(\hat{\mu}_\lambda)$$

Incomplete:

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

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

Selecting regularization parameters

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

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

Stein Unbiased Risk Estimate (Candès et al., 2012) (σ^2 known)

$$\text{SURE} = -np\sigma^2 + \text{RSS} + 2\sigma^2 \text{div}(\hat{\mu}_\lambda)$$

Incomplete:

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

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

\Rightarrow GSURE - Iterative Adaptive (the best for imputing) - Iterative ISA...

Hospital data set

7000 patients/ 250 variables

	Centre	Mecanisme	Age	Sexe	Poids	Taille	BMI	PAS	PAD
1	Beaujon	Chute d'une hauteur	54	m	85	NR	NR	180	110
2	Lille	Autre	33	m	80	1.8	24.69	130	62
3	Pitie Salpetriere	AVP 2 roues motorise	26	m	NR	NR	NR	131	62
4	Beaujon	AVP 2 roues motorise	63	m	80	1.8	24.69	145	89
6	Pitie Salpetriere	AVP bicyclette	33	m	75	NR	NR	104	86
7	Pitie Salpetriere	AVP pieton	30	m	NR	NR	NR	107	66
9	HEGP	Arme blanche	16	m	98	1.92	26.58	118	54
10	Toulon	AVP (voiture. camion. bus)	20	m	NR	NR	NR	124	73
11	Bicetre	Chute d'une hauteur	61	m	84	1.7	29.07	144	105
12	Bicetre	AVP 2 roues motorise	20	m	80	1.85	23.37	124	67
13	Bicetre	AVP pieton	38	m	95	1.9	26.32	88	60
.....									

	SpO2	Temperature	Lactates	Hb	Glasgow
1	97	35.6	<NA>	12.7	12	
2	100	36.5	4.8	11.1	15	
3	100	36	3.9	11.4	3	
4	100	36.7	1.66	13	15	
6	100	36	NF	14.4	15	
7	100	36.6	NF	14.3	15	
9	100	37.5	13	15.9	15	
10	100	36.9	NF	13.7	15	
11	100	36.6	1.2	14.2	14	
12	100	37.2	NF	14.7	15	
13	99	33.6	3.5	12.3	13	
.....						

Filled-in hospital data

7000 patients/ 250 variables

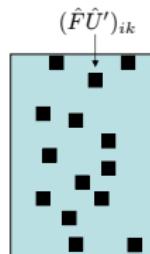
	Centre	Mecanisme	Age	Sexe	Poids	Taille	BMI	PAS	PAD
1	Beaujon	Chute d'une hauteur	54	m	85.00	1.84	27.04	83	13
2	Lille	Autre	33	m	80.00	1.80	24.69	33	98
3	Pitie Salpetriere	AVP 2 roues motorise	26	m	81.78	1.85	24.33	34	98
4	Beaujon	AVP 2 roues motorise	63	m	80.00	1.80	24.69	48	125
6	Pitie Salpetriere	AVP bicyclette	33	m	75.00	1.83	24.53	6	122
7	Pitie Salpetriere	AVP pieton	30	m	81.89	1.82	25.24	9	102
9	HEGP	Arme blanche	16	m	98.00	1.92	26.58	21	90
10	Toulon	AVP (voiture, camion, bus)	20	m	81.68	1.82	25.05	27	109
11	Bicetre	Chute d'une hauteur	61	m	84.00	1.70	29.07	47	8
12	Bicetre	AVP 2 roues motorise	20	m	80.00	1.85	23.37	27	103
13	Bicetre	AVP pieton	38	m	95.00	1.90	26.32	179	96
	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				
12	2	79	276.17	53	6				
13	48	39	388.00	29	14				

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

Multiple imputation (Rubin, 1987)

Single imputation: underestimation of standard errors

⇒ a single value can't reflect the uncertainty of prediction

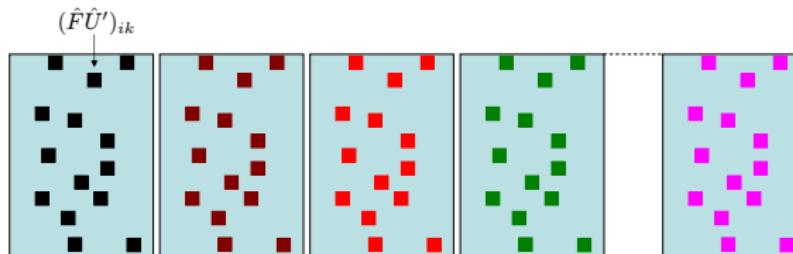


Multiple imputation (Rubin, 1987)

Single imputation: underestimation of standard errors

⇒ a single value can't reflect the uncertainty of prediction

① Generating B imputed data sets



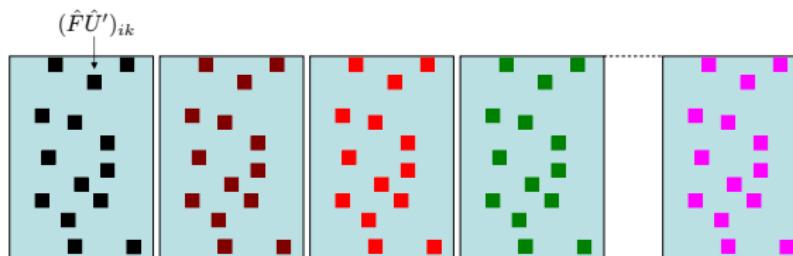
(different plausible values for each missing entry. Ex: one cell:
predict 4.88, empirical interval [3.98 ; 5.89])

Multiple imputation (Rubin, 1987)

Single imputation: underestimation of standard errors

⇒ a single value can't reflect the uncertainty of prediction

① Generating B imputed data sets



(different plausible values for each missing entry. Ex: one cell:
predict 4.88, empirical interval [3.98 ; 5.89])

② Performing the analysis on each imputed data $\hat{\beta}_b, \widehat{Var}(\hat{\beta}_b)$

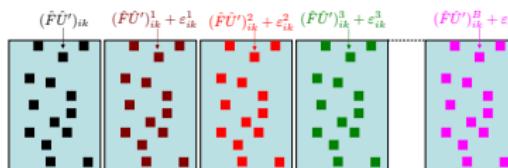
③ Combining: $\hat{\beta} = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_b$

$$T = \frac{1}{B} \sum_{b=1}^B \widehat{Var}(\hat{\beta}_b) + \frac{1}{B-1} \sum_{b=1}^B (\hat{\beta}_b - \hat{\beta})^2$$

PCA multiple imputation (proper)

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

- ① Variability of parameters : $\hat{\mu}^1 = U^1 D^1 V'^1, \dots, \hat{\mu}^B = U^B D^B V^{B'}$
 \Rightarrow bootstrap/bayesian (data augmentation)
- ② Noise: $b = 1, \dots, B$ missing X_{ij}^b drawn from $\mathcal{N}((\hat{\mu}_{ij})^b, (\hat{\sigma}^2)^b)$



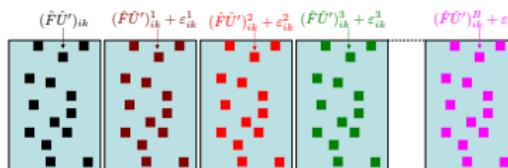
Husson, Josse, Wager (2014). Complete case: variability of PCA estimates: asymptotic, residuals bootstrap, a cell-wise jackknife

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

PCA multiple imputation (proper)

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

- ① **Variability of parameters:** $\hat{\mu}^1 = U^1 D^1 V'^1, \dots, \hat{\mu}^B = U^B D^B V^{B'}$
 \Rightarrow bootstrap/bayesian (data augmentation)
- ② Noise: $b = 1, \dots, B$ missing X_{ij}^b drawn from $\mathcal{N}((\hat{\mu}_{ij})^b, (\hat{\sigma}^2)^b)$



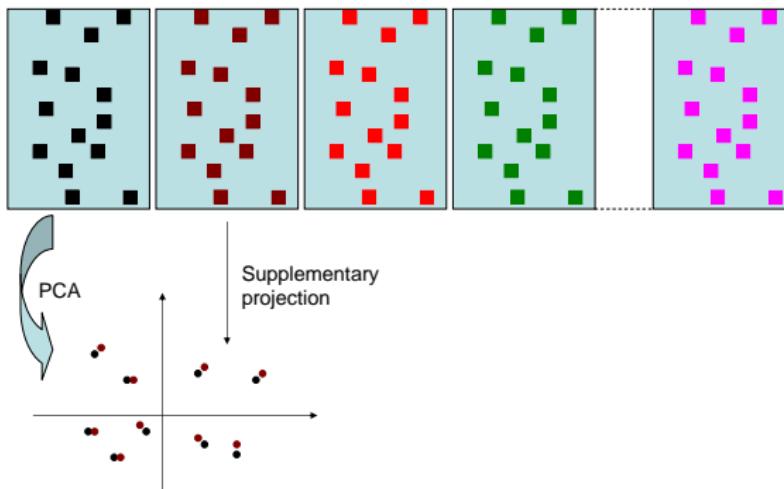
Husson, Josse, Wager (2014). Complete case: variability of PCA estimates: asymptotic, residuals bootstrap, a cell-wise jackknife

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

Supplementary projection

Imputed by Regularized PCA/ B imputed data sets from MIPCA

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



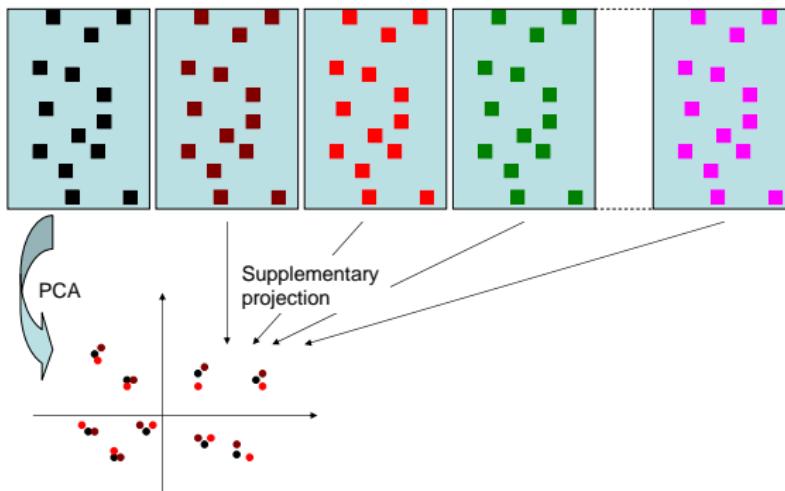
Regularized iterative PCA
⇒ reference configuration

⇒ 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



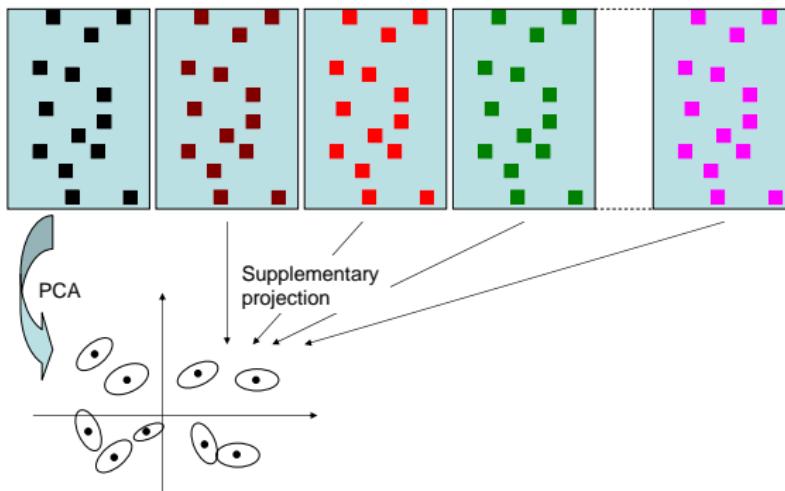
Regularized iterative PCA
⇒ reference configuration

⇒ 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

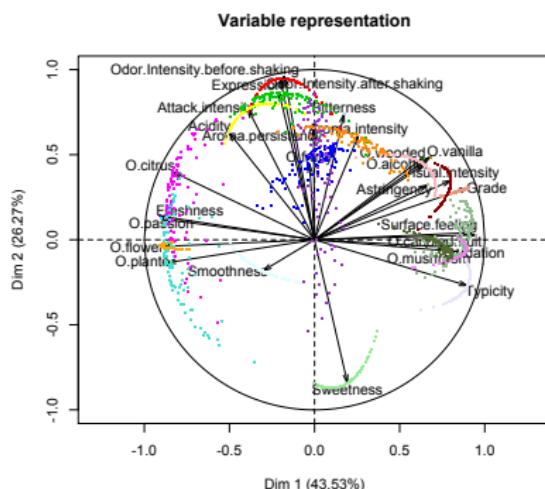
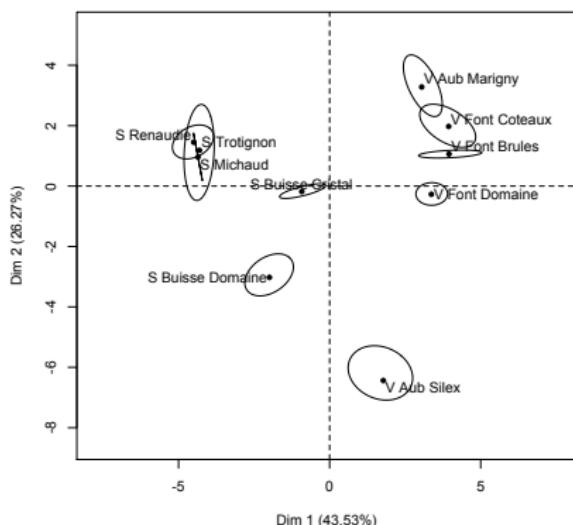


Regularized iterative PCA
⇒ reference configuration

⇒ Individuals' position (and variables) with other predictions

Multiple imputation: visualization of variance of prediction

Supplementary projection

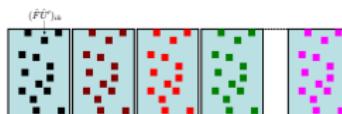


⇒ Does single imputation make sense?

Properties

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

- ① Variability of parameters. Bootstrap rows: X^1, \dots, X^B
EM algorithm: $(\hat{\mu}^1, \hat{\Sigma}^1), \dots, (\hat{\mu}^B, \hat{\Sigma}^B)$
 - ② Noise. imputation: X_{ij}^b drawn from $\mathcal{N}(\hat{\mu}^b, \hat{\Sigma}^b)$
- ⇒ Conditional modeling: one model/variable
- Imputation model: B multiple imputed sets with PCA, JM, CM



- Analysis model: $\hat{\theta}_b, \widehat{\text{Var}}(\hat{\theta}_b)$ - combine Rubin's rules: $\hat{\theta}, T$
- ⇒ Good estimates of θ (bias) and coverage ≈ 0.95 (CI width OK): variability due to missing values is taken into account
- ⇒ PCA: small - large n/p ; strong - weak relation; low-high % NA

Multiple Correspondence Analysis (MCA)

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

$$A = \begin{array}{|c|c|c|} \hline 1 & 0 & \dots & 1 & 0 & 0 \\ \hline 1 & 0 & \dots & 1 & 0 & 0 \\ \hline 1 & 0 & \dots & 1 & 0 & 0 \\ \hline 0 & 1 & \dots & 0 & 1 & 0 \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hline 0 & 1 & \dots & 0 & 0 & 1 \\ \hline 0 & 1 & \dots & 0 & 1 & 0 \\ \hline \end{array} \quad D_p =$$

$$\begin{array}{ccccc} & & p_1 & & 0 \\ & & \ddots & & \ddots \\ 0 & & & & p_J \end{array}$$

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

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

Benzecri, 1973 : "All in all, doing a data analysis, in good mathematics, is simply finding eigenvectors; all the science (or the art) of it is just to find 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
 - 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
 - estimation: MCA on the completed data $\rightarrow U, D, V$
 - 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
 - estimation: MCA on the completed data $\rightarrow U, D, V$
 - imputation with the fitted matrix $\hat{\mu} = U_k D_k V'_k$
 - 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	...

⇒ the imputed values can be seen as degree of membership

Regularized iterative MCA

- Initialization: imputation of the indicator matrix (proportions)
- Iterate until convergence
 - estimation: MCA on the completed data $\rightarrow U, D, V$
 - imputation with the fitted matrix $\hat{\mu} = U_k D_k V'_k$
 - 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
 - estimation: MCA on the completed data $\rightarrow U, D, V$
 - imputation with the fitted matrix $\hat{\mu} = U_k D_k V'_k$
 - 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
 Inference with missing values (ex logistic regression)
 ⇒ Multiple imputation with MCA (Audigier, Husson & Josse, 2016)

Take home message

Principal component methods are powerful for single and multiple imputation with continuous, categorical variables: require a small number of parameters and capture the similarities between rows and relationship between variables.

From a practical point of view:

- very quick (1 min for MIMCA and 1h30 for forests on a data with 6000 rows/14 variables)
- can be applied on data of various size (many categories)
- provide correct inferences for analysis model based on relationships between pairs of variables
- require to select tuning parameters with missing values (k, λ)

Outline

1 Low-rank matrix estimation

- Singular values shrinkage
- Bootstrap framework

2 Inference with missing values using PC methods

- Single and multiple imputation with PCA
- Imputation with MCA

3 A multilogit bilinear model for MCA

Multilogit-bilinear model

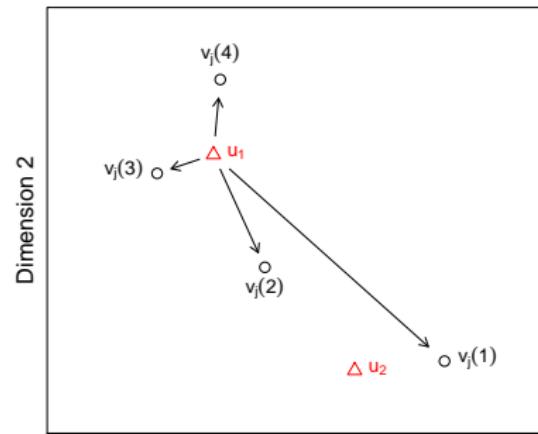
$$\mathbb{P}(x_{ij} = c) = \pi_{ijc} = \frac{e^{\theta_{ij}(c)}}{\sum_{c'=1}^{C_j} e^{\theta_{ij}(c')}},$$

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

Latent Space

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

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



Relationship with MCA

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

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: $\zeta = (\begin{smallmatrix} \beta \\ \text{vec}(\Gamma) \end{smallmatrix})$; $\zeta_0 = (\begin{smallmatrix} \log(p) \\ 0 \end{smallmatrix})$

Relationship with MCA

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

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: $\zeta = (\begin{smallmatrix} \beta \\ \text{vec}(\Gamma) \end{smallmatrix})$; $\zeta_0 = (\begin{smallmatrix} \log(p) \\ 0 \end{smallmatrix})$

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

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: $\zeta = (\begin{smallmatrix} \beta \\ \text{vec}(\Gamma) \end{smallmatrix})$; $\zeta_0 = (\begin{smallmatrix} \log(p) \\ 0 \end{smallmatrix})$

\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:

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

Solution: rank k SVD of $(A - \mathbf{1} p^T) D_p^{-1/2}$ \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 $(\beta_0, \widehat{\Gamma}_{MCA})$.*

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 $\mathbf{H}_1, \mathbf{H}_2 \succ 0$.

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

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

Maximizing the likelihood

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

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 $\theta_{ij(c)} \rightarrow \infty$

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

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 $H = \nabla^2 f_{ij}(\theta_i) = (\text{Diag}(\pi_{ij}) - \pi_{ij}\pi_{ij}')$

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 $1/2$)

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

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

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

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

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

⇒ 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

⇒ 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

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

Perspectives

- Multiple imputation for mixed data:
 - uncertainty of the parameters: PCA residuals bootstrap, MCA nonparametric bootstrap
 - extend the multilogit model
- Multiple imputation with "big data":
 - mixture (PCA or MCA) or biclustering to impute
 - theory of MI: asymptotic $n \ll p$
- Distributed computation with missing values (medical databases)

Outline

1 Low-rank matrix estimation

- Singular values shrinkage
- Bootstrap framework

2 Inference with missing values using PC methods

- Single and multiple imputation with PCA
- Imputation with MCA

3 A multilogit bilinear model for MCA

Estimation of σ^2

⇒ Number of independent parameters

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

⇒ Trace of the PCA projection matrix

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

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

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

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

Pazman & Denis, 2002; Candes & Tao, 2009

Cross-validation to select S

?			
?	?	?	?
?	?	?	
?	?	?	?
?	?	?	
?	?	?	
			?

⇒ EM-CV (Bro *et al.* 2008)

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

⇒ Computational costly

Cross-validation to select S

?	■		
?	?	?	?
?	?	?	?
?	?	?	?
?	?	?	?
?	?	?	?
?	?	?	?
?	?	?	?

⇒ EM-CV (Bro *et al.* 2008)

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

⇒ Computational costly

Cross-validation to select S

?	■	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?

⇒ EM-CV (Bro *et al.* 2008)

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

⇒ Computational costly

Cross-validation to select S

?	■	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?

⇒ EM-CV (Bro *et al.* 2008)

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

⇒ Computational costly

⇒ 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

?	■	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?

⇒ EM-CV (Bro *et al.* 2008)

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

⇒ Computational costly

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

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

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

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

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

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

Pazman & Denis, 2002; Candes & Tao, 2009

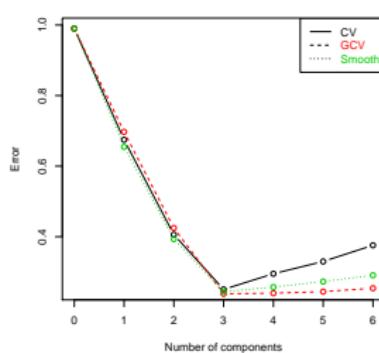
\Rightarrow Number of independent parameters:

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

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



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

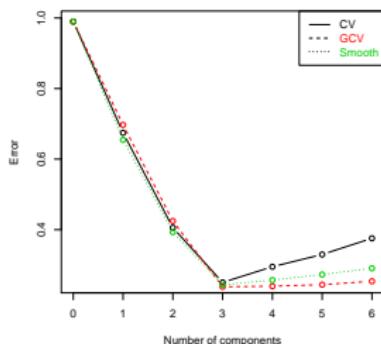
Josse, J. & Husson, F. Selecting the number of components in PCA using cross-validation approximations.
Computational Statistics and Data Analysis.

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	



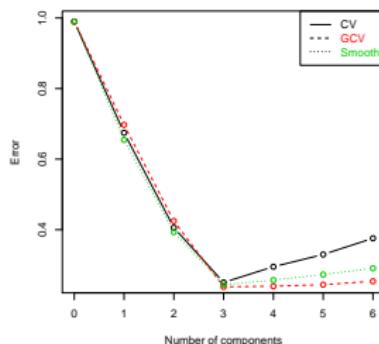
$$\begin{aligned} \text{CV}_S &= \frac{1}{np} \sum_{i,j} \left(X_{ij} - (\hat{\mu}_{ij}^S)^{-ij} \right)^2 \\ \text{ACV}_S &= \frac{1}{np} \sum_{i,j} \left(\frac{X_{ij} - (\hat{\mu}_{ij}^S)}{1 - P_{ij,ij}} \right)^2 \\ \text{GCV}_S &= \frac{1}{np} \times \frac{\sum_{i,j} (X_{ij} - \hat{\mu}_{ij}^S)^2}{(1 - \text{tr}(P^{(S)})/np)^2} \end{aligned}$$

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



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

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

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

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

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

Outline

1 Low-rank matrix estimation

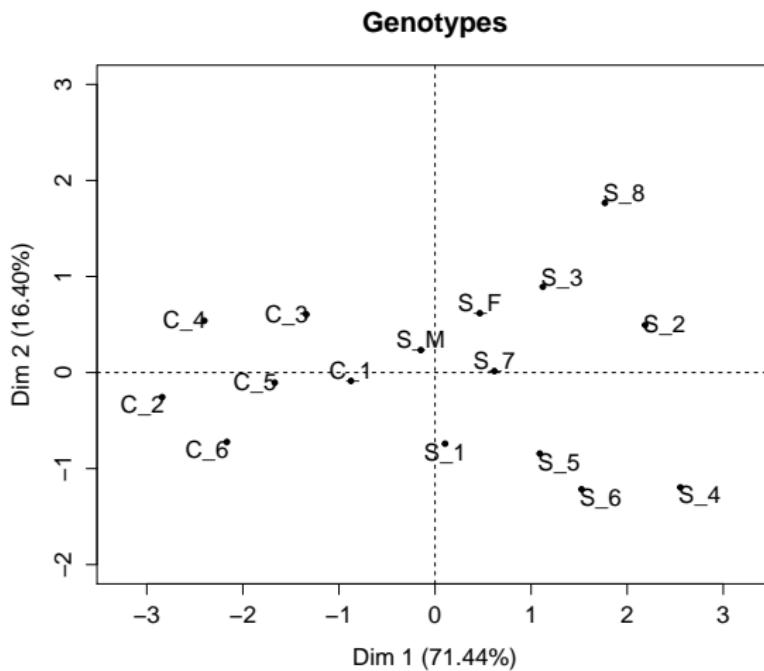
- Singular values shrinkage
- Bootstrap framework

2 Inference with missing values using PC methods

- Single and multiple imputation with PCA
- Imputation with MCA

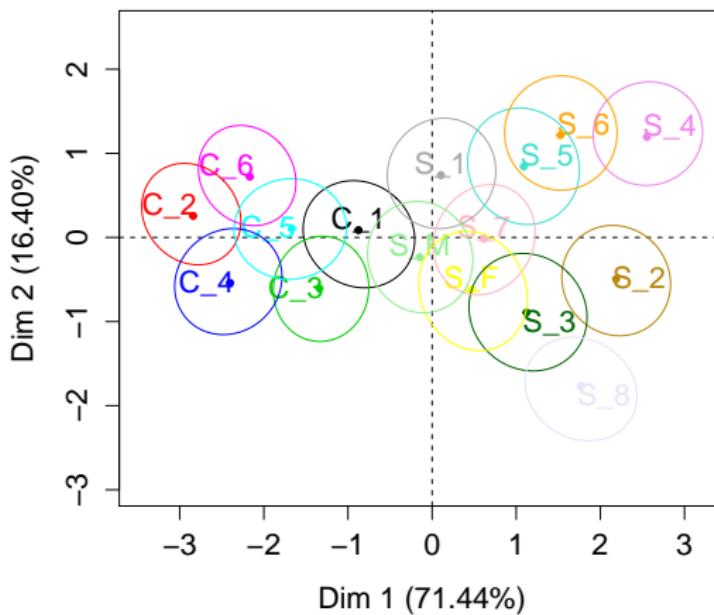
3 A multilogit bilinear model for MCA

Inference in PCA



Inference in PCA

Bootstrap



Confidence ellipses

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

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

Rq: Population data. (PPCA: random effects)

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

Confidence ellipses

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

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

Rq: Population data. (PPCA: random effects)

\Rightarrow **Asymptotic** $\sigma^2 \rightarrow 0$ (Denis, Gower, 1994; Papadopoulou & Louraki 2000)

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

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

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

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

Confidence ellipses

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

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

Rq: Population data. (PPCA: random effects)

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

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

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

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

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

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

Confidence ellipses

⇒ **Bootstrap**

⇒ Rows bootstrap (Holmes, 1985, Timmerman *et al*, 2007)

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

⇒ **Residuals bootstrap**

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

Confidence ellipses

⇒ **Bootstrap**

⇒ Rows bootstrap (Holmes, 1985, Timmerman *et al*, 2007)

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

⇒ **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 $X^b \rightarrow (\hat{\mu}^1 = U^1 D^1 V^{1'}, \dots, \hat{\mu}^B = U^B D^B V^{B'})$

Confidence ellipses

⇒

Jackknife

Confidence ellipses

⇒ Cell-wise Jackknife

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

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

Represent pseudo-values: $(\hat{\mu}^1, \dots, \hat{\mu}^{np})$

Requires a method to deal with missing values - costly

Confidence ellipses

⇒ Cell-wise Jackknife

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

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

Represent pseudo-values: $(\hat{\mu}^1, \dots, \hat{\mu}^{np})$

Requires a method to deal with missing values - costly

⇒ Approximate Jackknife (Craven & Wahba, 1979 lemma)

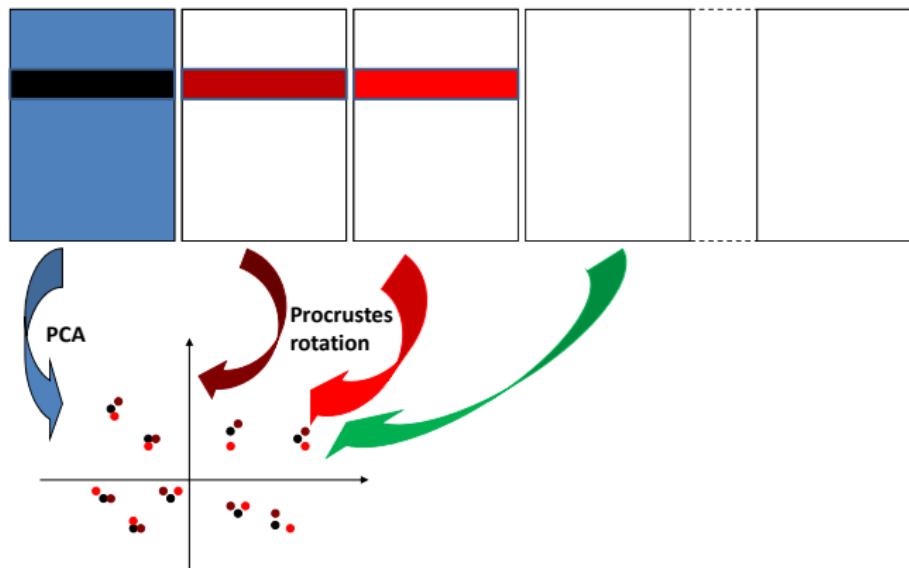
In regression $\hat{y} = Py$: $\hat{y}_i - y_i = \frac{\hat{y}_i - y_i}{1 - P_{i,i}}$

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

⇒ Spread of $(\hat{\mu}^1, \dots, \hat{\mu}^B)$ gives the variability of PCA $\hat{\mu}_{n \times p}^k$

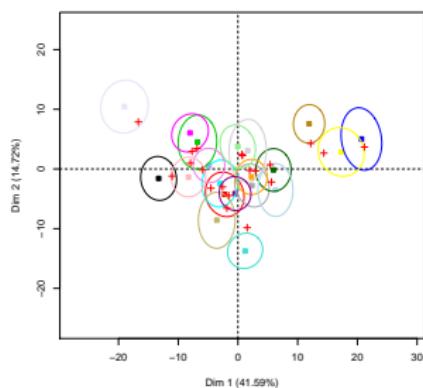


Simulations

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

- 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

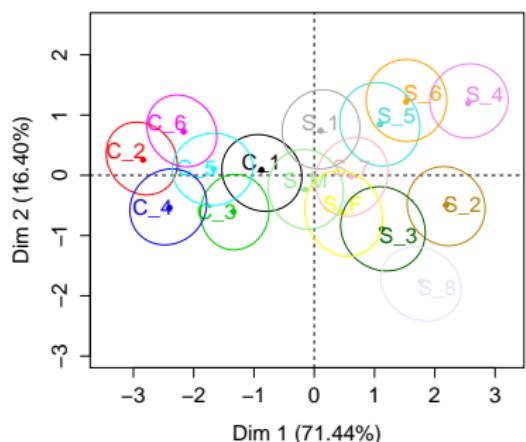


⇒ Different regimes:

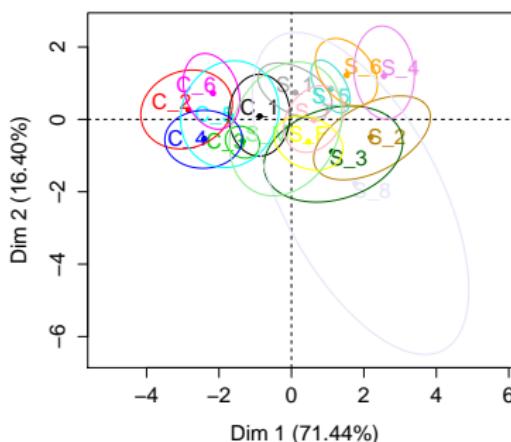
- High SNR - n large:
Asymptotic \approx Bootstrap
- Small SNR: Jackknives

Inference in PCA graphs

Bootstrap



Jackknife



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

Bayesian treatment of SVD

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

⇒ Overparametrization - priors and posteriors meeting constraints?

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

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

Bayesian treatment of SVD

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

⇒ Overparametrization - priors and posteriors meeting constraints?

- **Hoff 2009** uniform prior for U and V special cases of von Mises-Fisher distributions (Stiefel manifold); $(d_l)_{l=1\dots k} \sim \mathcal{N}(0, s_\lambda^2)$
- 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 $\mu^1, \dots, \mu^B \Rightarrow$ postprocessing step
SVD on each matrix $\mu^1 = U^1 D^1 V^1, \dots, \mu^B = U^B D^B V^B$

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

Random effect models: Probabilistic PCA

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

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

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

⇒ Bayesian SVD with *a priori* on U : Regularized PCA

The log-bilinear model & CA

⇒ The saturated log-linear models (Christensen, 1990; Agresti, 2013).

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

⇒ 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

⇒ The saturated log-linear models (Christensen, 1990; Agresti, 2013).

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

⇒ 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.

⇒ 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 $Z = \tilde{U}_K D_K \tilde{V}_K^T$. Standard row and col coord

$U_K = D_r^{-1/2} \tilde{U}_K$, $V_K = D_c^{-1/2} \tilde{V}_K$. If the low-rank approx is good:

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

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

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

CA approximates the log-bilinear model

\Rightarrow SVD $Z = \tilde{U}_K D_K \tilde{V}_K^T$. Standard row and col coord

$U_K = D_r^{-1/2} \tilde{U}_K$, $V_K = D_c^{-1/2} \tilde{V}_K$. If the low-rank approx is good:

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

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

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

\Rightarrow Connection (Escofier, 1982) : when $\sum_{k=1}^K d_k u_{ik} v_{jk} \ll 1$, 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

1 Low-rank matrix estimation

- Singular values shrinkage
- Bootstrap framework

2 Inference with missing values using PC methods

- Single and multiple imputation with PCA
- Imputation with MCA

3 A multilogit bilinear model for MCA

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
Rhone Alpes	:5421	M:23165	26_34: 9401	2010:25034	E2:23521	1-2 : 4952	0 : 6133
Provence Alpes	:4116		35_44:10899		E3:6563	10-19: 839	1-2/m: 7583
Nord Pas de Calais	:3819		45_54: 9505		E4:10100	20-29: 212	1-2/w: 9526
Pays de Loire	:3152		55_64: 9503		NA:73	3-5 : 1908	3-4/w: 6815
Bretagne	:3038		65_+ : 6713			30+ : 404	5-6/w: 3402
(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

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
 - 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
 - estimation: MCA on the completed data $\rightarrow U, D, V$
 - 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
 - estimation: MCA on the completed data $\rightarrow U, D, V$
 - imputation with the fitted matrix $\hat{\mu} = U_k D_k V'_k$
 - 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
 - estimation: MCA on the completed data $\rightarrow U, D, V$
 - imputation with the fitted matrix $\hat{\mu} = U_k D_k V'_k$
 - 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	...

⇒ the imputed values can be seen as degree of membership

Regularized iterative MCA

- Initialization: imputation of the indicator matrix (proportions)
- Iterate until convergence
 - estimation: MCA on the completed data $\rightarrow U, D, V$
 - imputation with the fitted matrix $\hat{\mu} = U_k D_k V'_k$
 - 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 \hat{X}_2 \hat{X}_B

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

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

1	0	...	1	0
1	0	...	1	0
1	0	...	0.11	0.77
0.20	0.80		0	0
0	1		0	0

- ② Categories drawn from multinomial distribution using the values in $(\hat{X}_b)_{1 \leq b \leq B}$

A	...	A
A	...	A
A	...	B
B	...	C
B	...	B

A	...	A
A	...	A
A	...	A
B	...	C
B	...	B

A	...	A
A	...	A
A	...	B
B	...	C
B	...	B

Multiple imputation for categorical data

⇒ Joint modeling:

- Log-linear model (Schafer, 1997): pb many levels
- Latent class models (Vermunt J.K, 2014) - nonparametric Bayesian (Y. Si & J. P. Reiter, 2014, Dunson, 2015)

⇒ Conditional model: logistic, multinomial logit, random forests

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

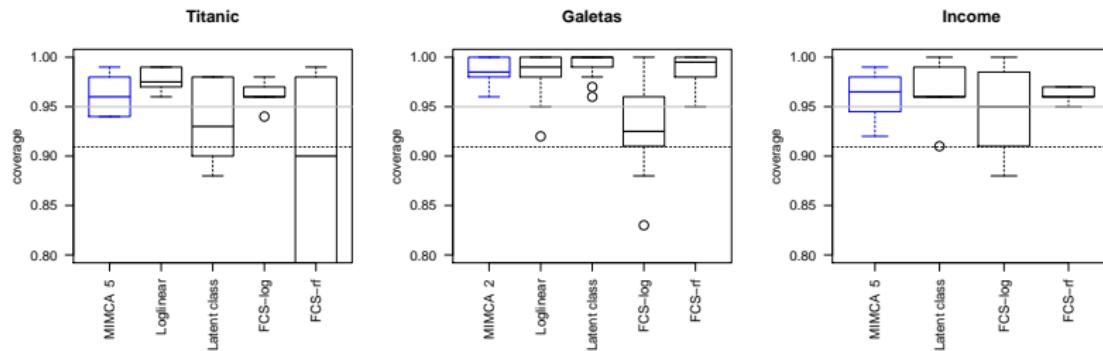
- reduce the dimensionality
- applied to data of various size (many levels, rare levels)

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

Simulations

- Quantities of interest: θ = 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: $\mathbb{P}(X = (x_1, \dots, x_K); \theta) = \sum_{\ell=1}^L \left(\theta_\ell \prod_{k=1}^K \theta_{x_k}^{(\ell)} \right)$
- ① Variability of the parameters:
 - a hierachic prior on θ :

$$\alpha \sim \mathcal{G}(.25, .25) \quad \zeta_\ell \sim \mathcal{B}(1, \alpha) \quad \theta_\ell = \zeta_\ell \prod_{g < \ell} (1 - \zeta_g) \text{ for } \ell \text{ in } 1, \dots, \infty$$
 - 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.*)

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

⇒ Conditional modeling based on RF

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

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

⇒ with Random Forests

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

⇒ with PCA

⇒ Non linear relationship well handled by forests

Bayesian PCA complete case

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

⇒ Priors on Q, D, R : overparametrization issue

- Hoff (2009): Uniform prior for Q and R von Mises-Fisher distributions (Stiefel manifold); $(d_l)_{l=1\dots k} \sim \mathcal{N}(0, s_\lambda^2)$
- Josse & Denis (2013): disregarding the constraints;
 $Q_{il} \sim \mathcal{N}(0, 1)$, $R_{jl} \sim \mathcal{N}(0, 1)$, $(d_l)_{l=1\dots k} \sim \mathcal{N}(0, s_\lambda^2)$

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

Empirical Bayes: $(X_{ij}^{(s)}) \sim \mathcal{N}(0, \tau_l^2 + \sigma^2)$ 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 $(\hat{\mu}_{ij})^1, \dots, (\hat{\mu}_{ij})^B$

⇒ Posterior distribution: Bayesian PCA

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

⇒ 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

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

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

⇒ 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 $\mathcal{N}(X_{-j}\beta^{-j}, \sigma^{-j})$
- ② Cycling through variables - Repeat B times

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

Properties

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

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

⇒ 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 $\mathcal{N}(X_{-j}\beta^{-j}, \sigma^{-j})$
- ② Cycling through variables - Repeat B times

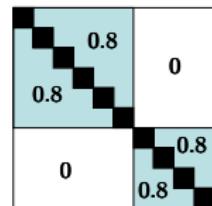
With continuous variables and a regression/variable: $\mathcal{N}(\mu, \Sigma)$

More flexible? Tedious?

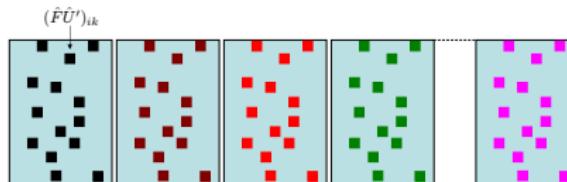
Simulations

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)



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



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

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

⇒ Bias, CI width, coverage - 1000 simulations

Results for the expectation

	parameters				confidence interval width			coverage		
	n	p	ρ	%	Joint	Cond	MPCA	Joint	Cond	MPCA
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

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

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

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

1 Low-rank matrix estimation

- Singular values shrinkage
- Bootstrap framework

2 Inference with missing values using PC methods

- Single and multiple imputation with PCA
- Imputation with MCA

3 A multilogit bilinear model for MCA

Feature Noising = Ridge = Shrinkage (Proof)

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

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

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

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

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

Simulations $X_{200 \times 500} = \mu + \varepsilon$

k	SNR	Bootstrap		TSVD (k)	Adaptive (λ, γ) (σ) - SURE	Asympt (σ)	Soft (λ) (σ) - SURE
		SA (σ, k)	ISA (σ)				
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 - Environment data

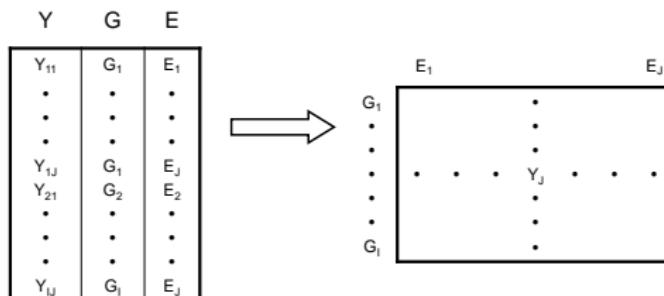
16 triticales (6 complete - 8 substituted - 2 checks)

10 locations in Spain (heterogeneous pH)

⇒ Adaptation of the triticales to the Spanish soil

⇒ AMMI (Additive Main effects and Multiplicative Interaction)

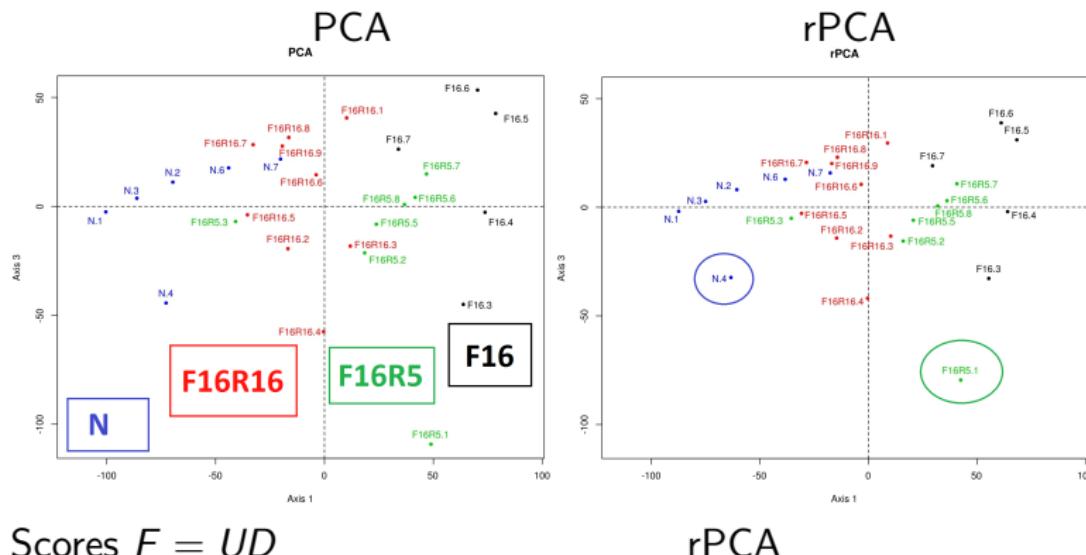
$$y_{ij} = \mu + \alpha_i + \beta_j + \sum_{l=1}^k d_l u_{il} v_{jl} + \varepsilon_{ij}$$



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

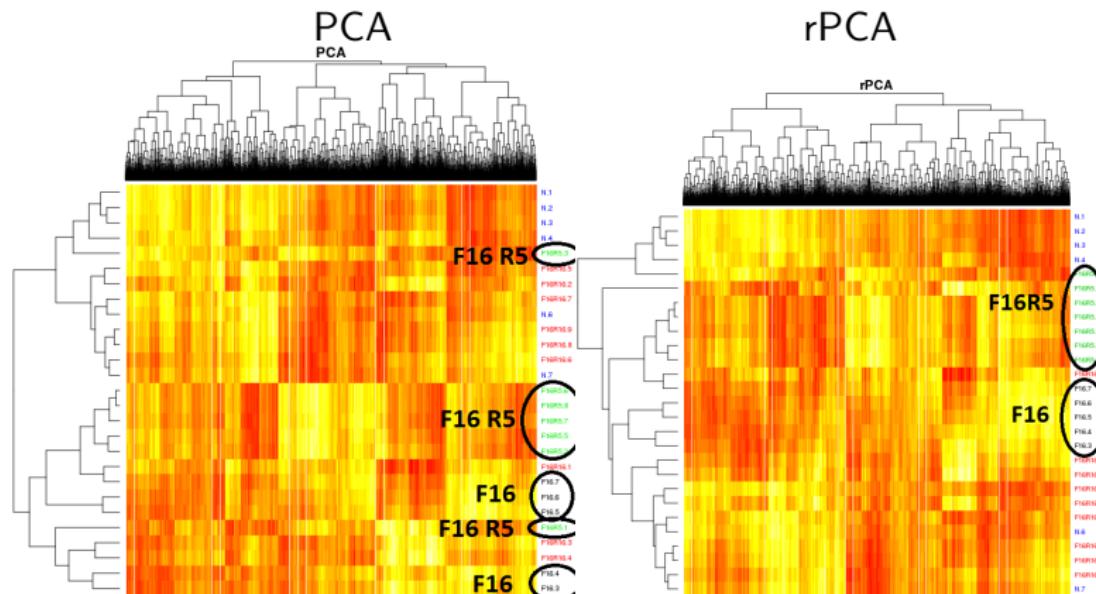


A 2x5 grid of 10 small circles, arranged in two rows of five.

A horizontal row of ten small circles, with three additional small circles positioned directly below the first three.

Expression data

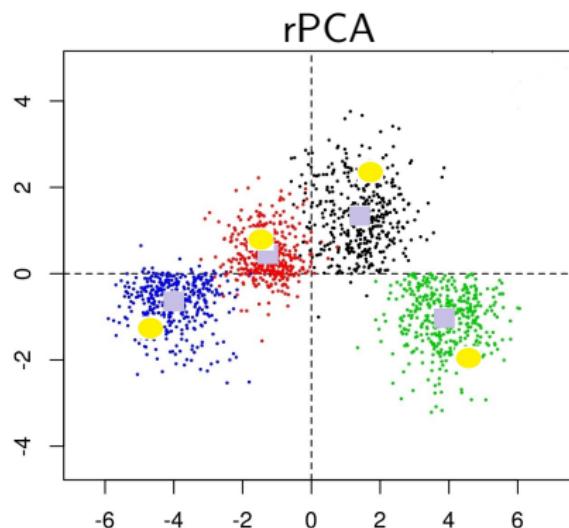
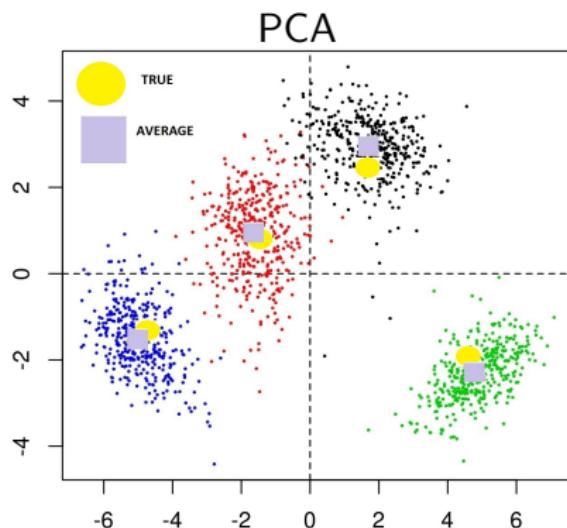
Clustering on the denoised matrix $\hat{\mu}_k = \sum_{l=1}^k u_l d_l v_l^\top$



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

Illustration of the regularization

$$X_{4 \times 10}^{sim} = \mu_{4 \times 10} + \varepsilon_{4 \times 10}^{sim} \quad sim = 1, \dots, 1000$$

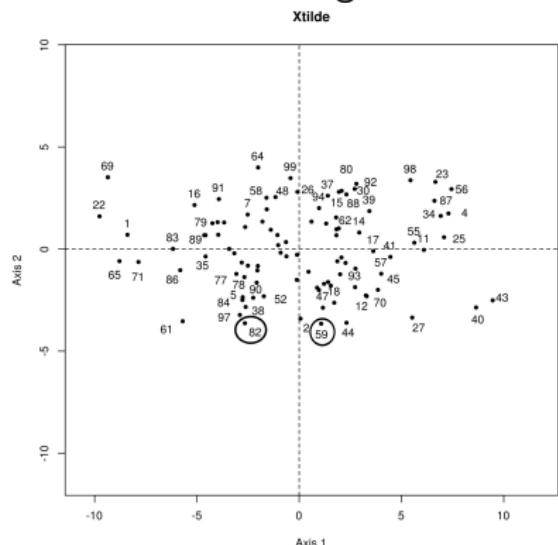


⇒ rPCA more biased less variable

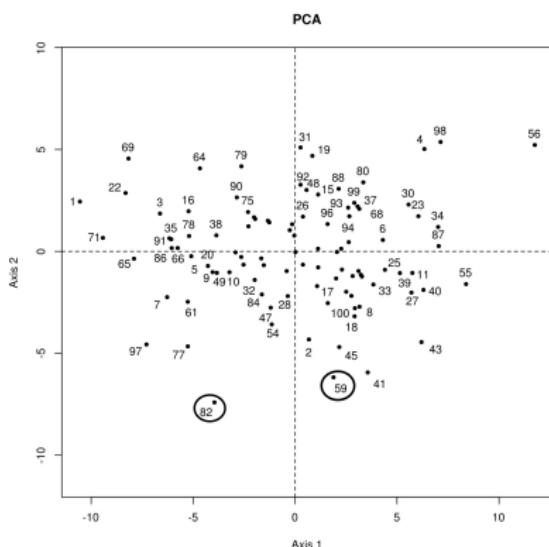
Rows representation

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

True configuration

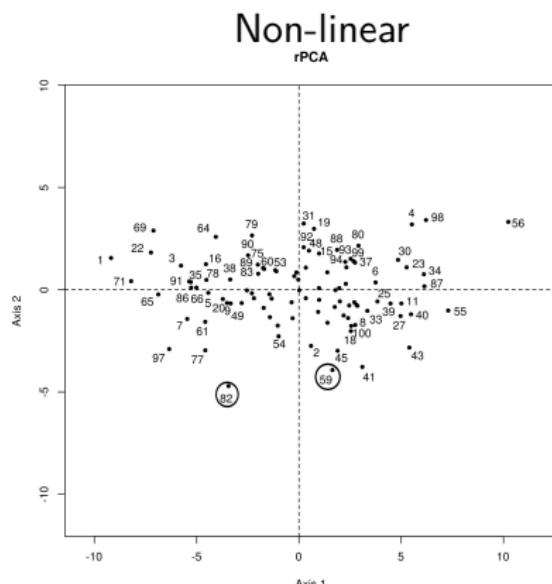
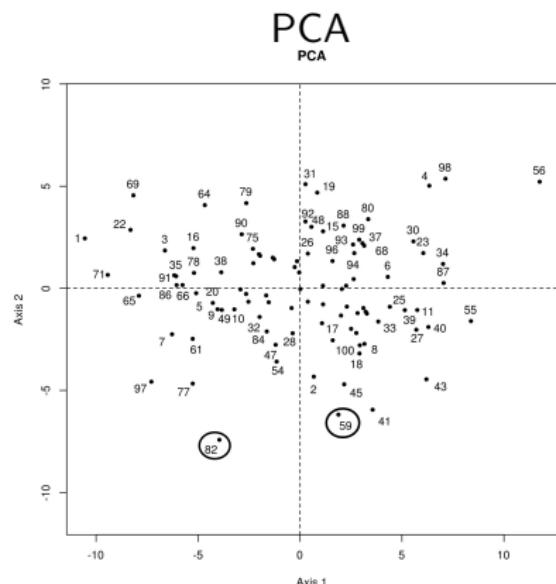


PCA



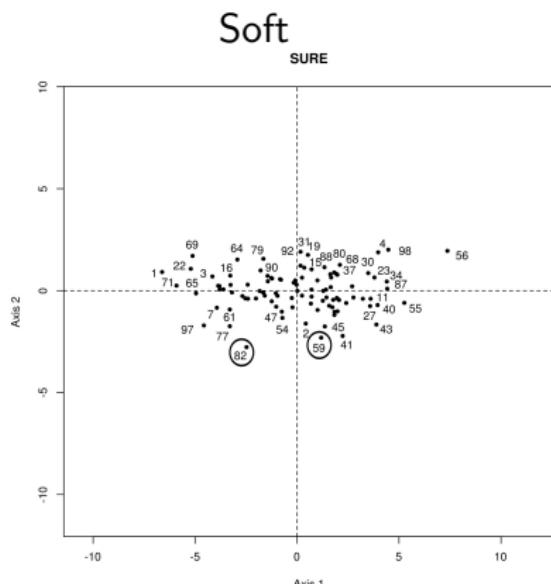
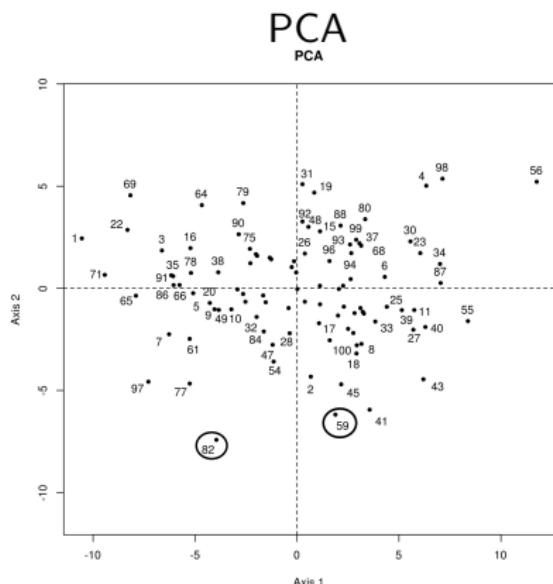
Rows representation

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



Rows representation

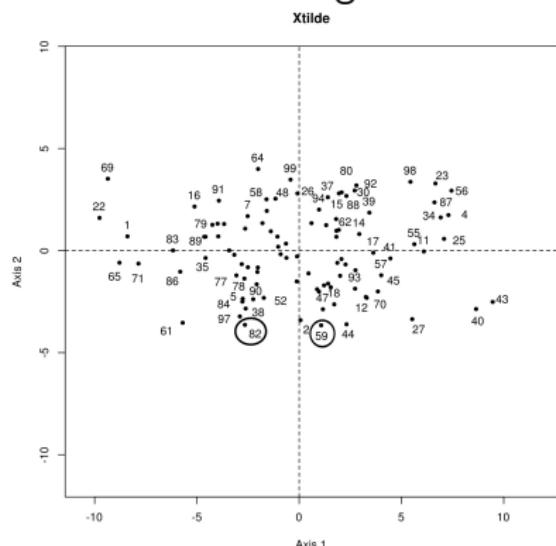
100 rows, 20 variables, 2 dimensions, SNR=0.8, $d_1/d_2 = 4$



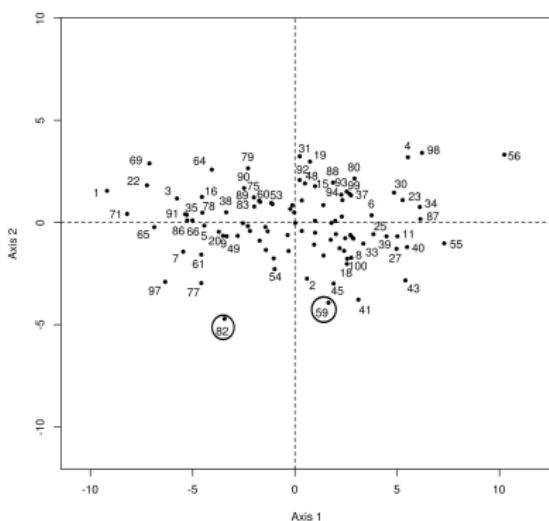
Rows representation

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

True configuration



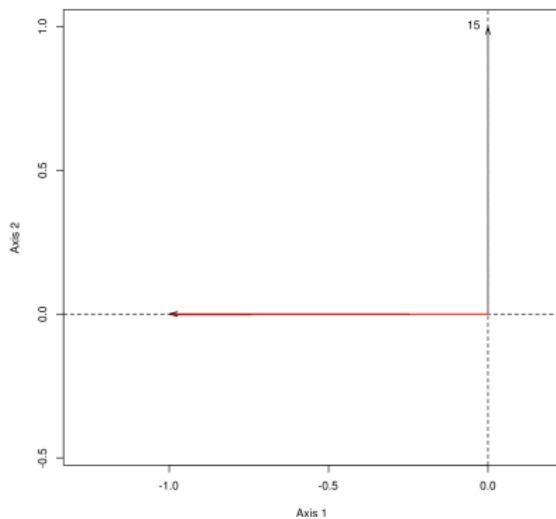
Non-linear
rPCA



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

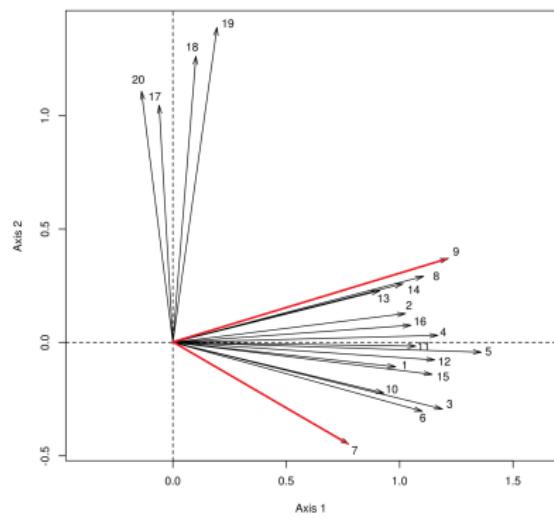
Variables representation

True configuration



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

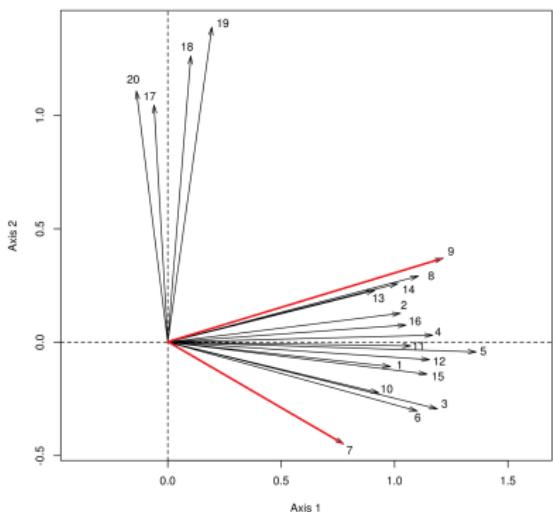
PCA



$$0.68$$

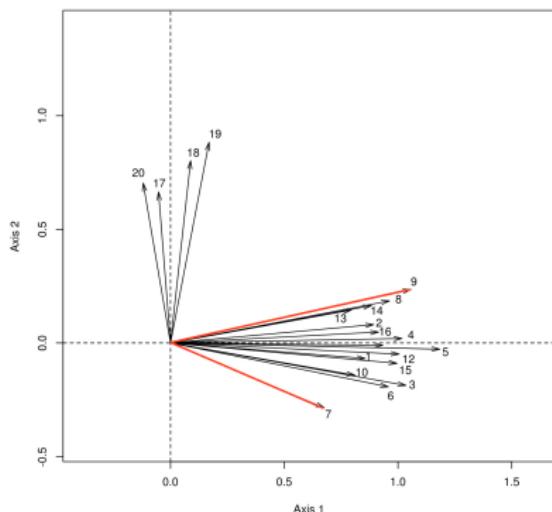
Variables representation

PCA
PCA



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

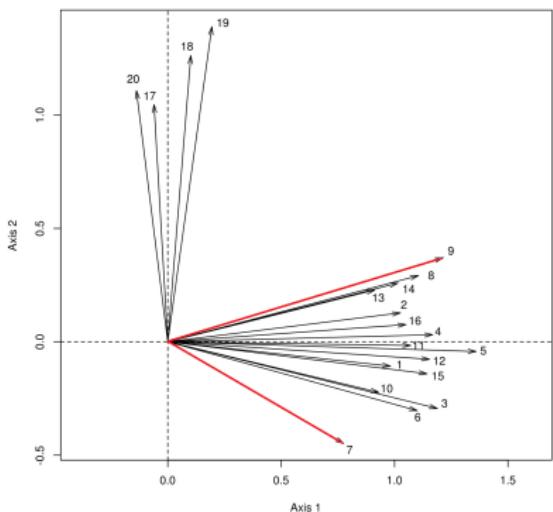
Non-linear
rPCA



$$0.81$$

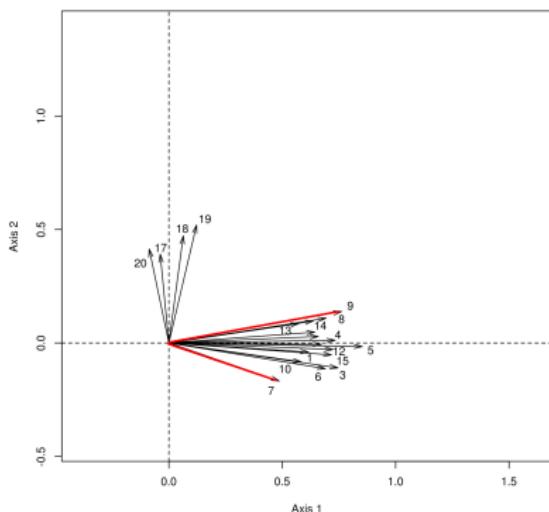
Variables representation

PCA
PCA



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

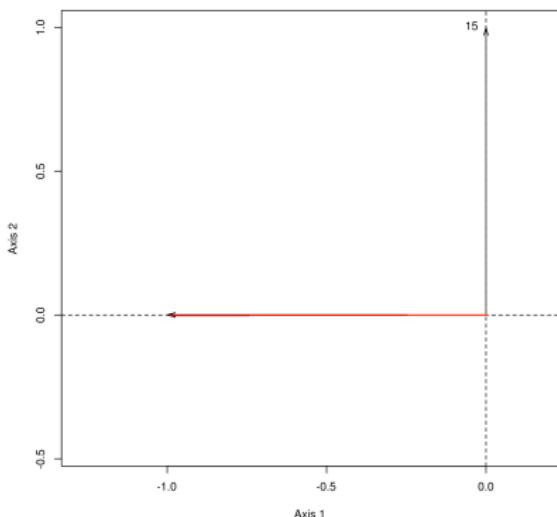
Soft
SURE



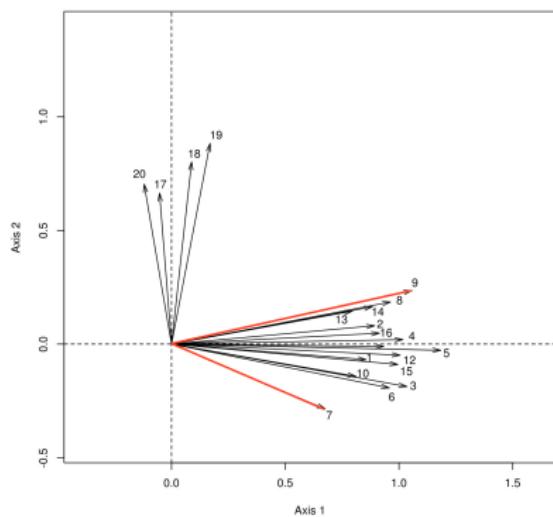
$$0.82$$

Variables representation

True configuration



Non-linear
rPCA



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

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

⇒ 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

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

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

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

⇒ Bootstrap estimator = noising: $\tilde{X}_{ij} \sim \text{Poisson}(X_{ij})$

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

⇒ Estimator robust to subsampling the obs used to build X

Bootstrap estimators

⇒ Feature noising = regularization

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

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

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

⇒ 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

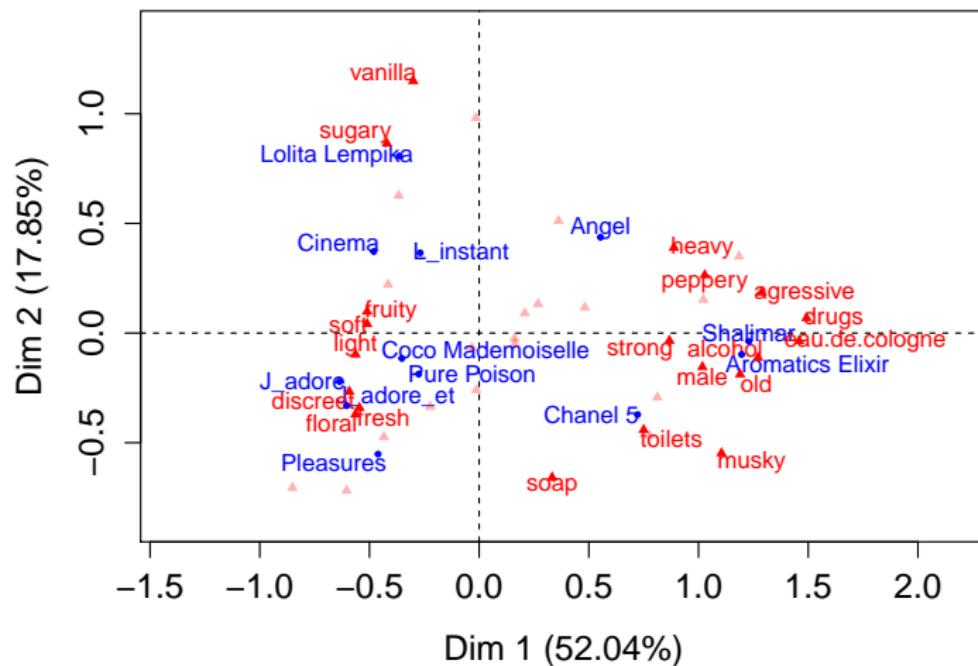
$$M = R^{-\frac{1}{2}} \left(X - \frac{1}{N} rc^\top \right) C^{-\frac{1}{2}}, \quad \text{where } R = \text{diag}(r), C = \text{diag}(c)$$

If the low-rank approx is good:

$$\hat{\mu}_k^{CA} = R^{\frac{1}{2}} \widehat{M}_k C^{\frac{1}{2}} + \frac{1}{N} rc^\top. \quad (3)$$

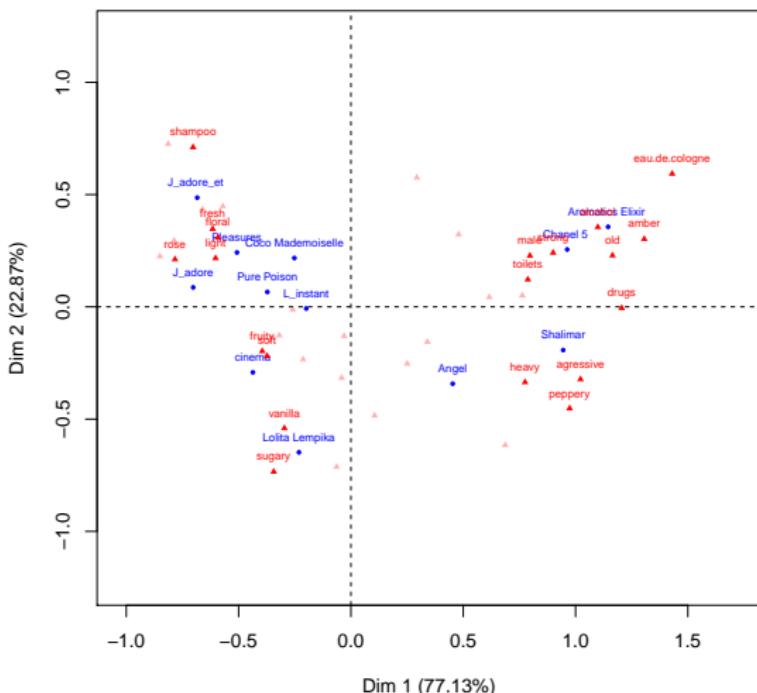
Perfume data set

CA



Perfume data set

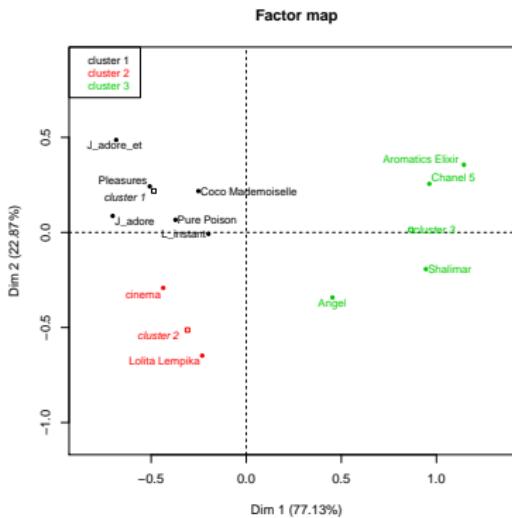
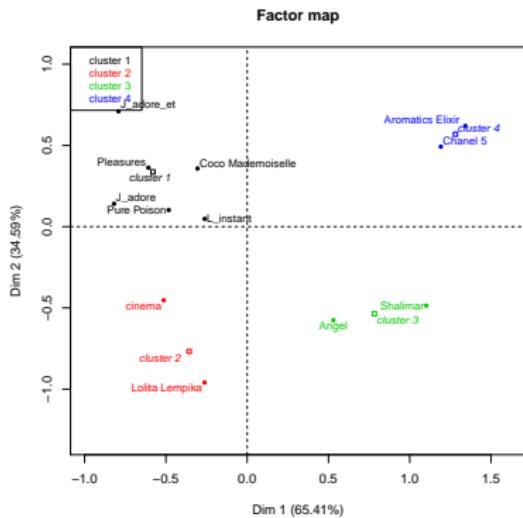
FREE



A 2x5 grid of 10 small circles, arranged in two rows of five.

A horizontal row of ten small, light-gray circles. Below this row, there is a vertical column of three small, light-gray circles.

Regularized Correspondence Analysis



Outline

1 Low-rank matrix estimation

- Singular values shrinkage
- Bootstrap framework

2 Inference with missing values using PC methods

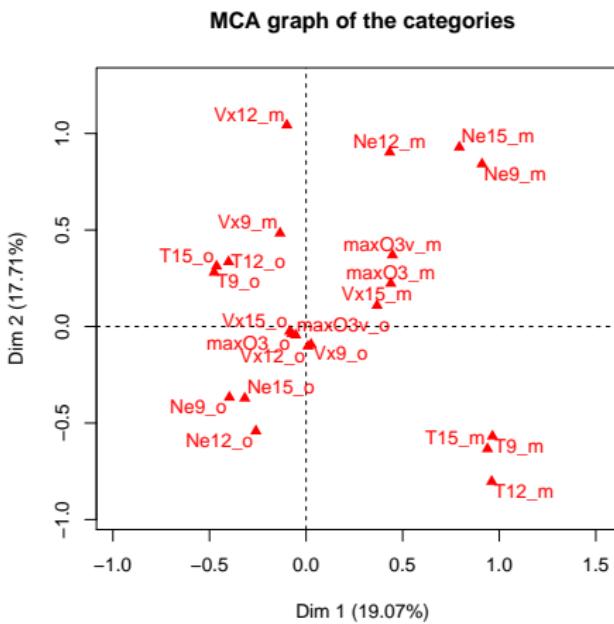
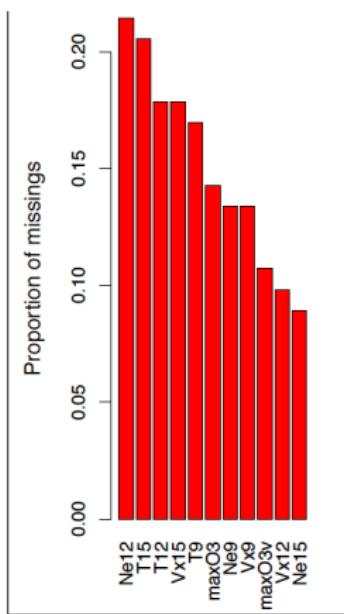
- Single and multiple imputation with PCA
- Imputation with MCA

3 A multilogit bilinear model for MCA

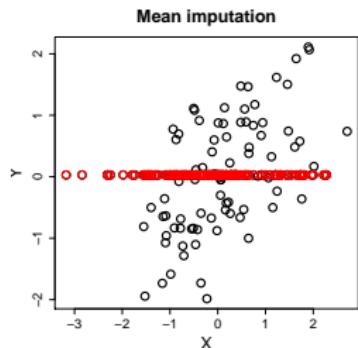
Ozone data set

	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

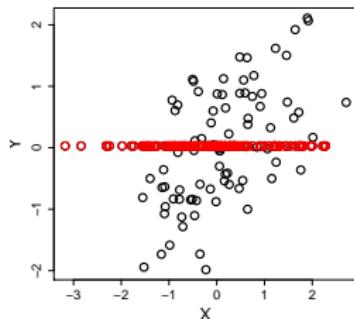


$\mu_y = 0$	0.01
$\sigma_y = 1$	0.5
$\rho = 0.6$	0.30
$CI\mu_y^{95\%}$	39.4

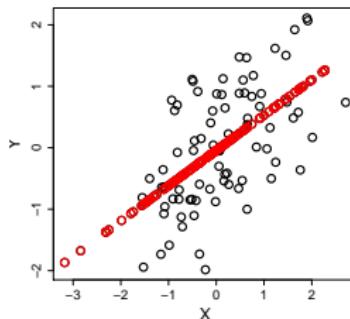
The idea of imputation is both seductive and dangerous (Dempster and Rubin, 1983)

Single imputation methods

Mean imputation



Regression imputation



$$\begin{aligned}\mu_y &= 0 \\ \sigma_y &= 1 \\ \rho &= 0.6 \\ CI\mu_y 95\% &= 39.4\end{aligned}$$

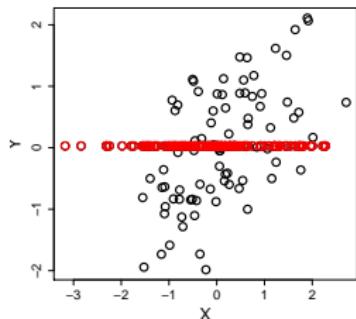
0.01
0.5
0.30
39.4

0.01
0.72
0.78
61.6

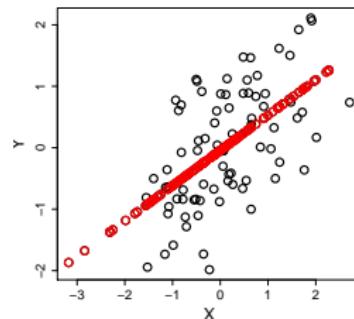
The idea of imputation is both seductive and dangerous (Dempster and Rubin, 1983)

Single imputation methods

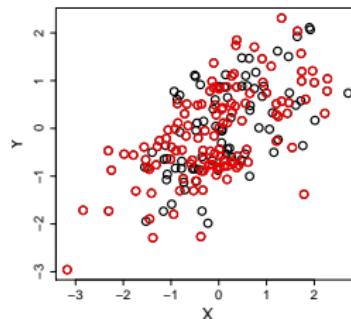
Mean imputation



Regression imputation



Stochastic regression imputation



$$\mu_y = 0$$

$$\sigma_y = 1$$

$$\rho = 0.6$$

$$CI\mu_y 95\%$$

0.01
0.5
0.30
39.4

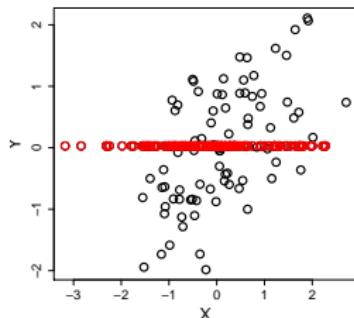
0.01
0.72
0.78
61.6

0.01
0.99
0.59
70.8

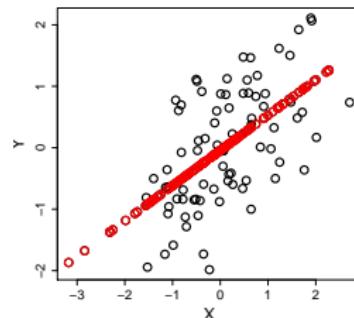
The idea of imputation is both seductive and dangerous (Dempster and Rubin, 1983)

Single imputation methods

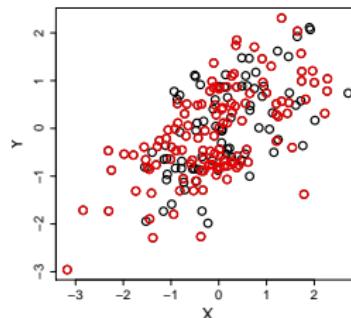
Mean imputation



Regression imputation



Stochastic regression imputation



$$\begin{aligned}\mu_y &= 0 \\ \sigma_y &= 1 \\ \rho &= 0.6 \\ CI\mu_y 95\% &= 39.4\end{aligned}$$

0.01
0.5
0.30
39.4

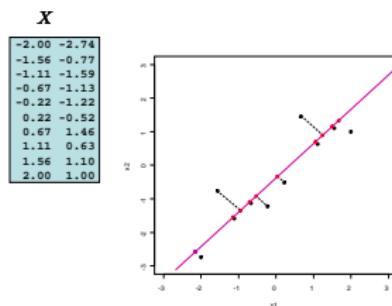
0.01
0.72
0.78
61.6

0.01
0.99
0.59
70.8

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

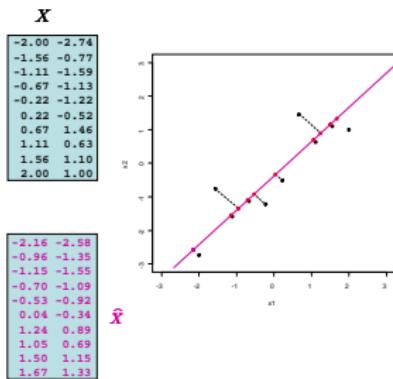
⇒ Standard errors of the parameters ($\hat{\sigma}_{\hat{\mu}_y}$) calculated from the imputed data set are underestimated

PCA reconstruction



⇒ Minimizes distance between observations and their projection

PCA reconstruction



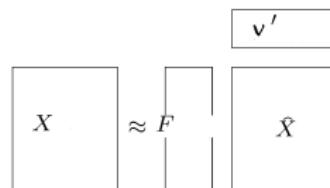
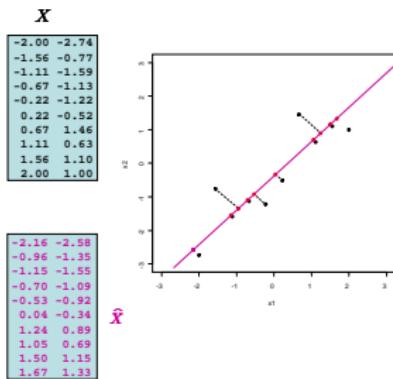
- ⇒ Minimizes distance between observations and their projection
- ⇒ Approx $X_{n \times p}$ with a low rank matrix $S < p$:

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

```
oooo
oooooooooo
```

```
oooooooooooo
ooo
```

PCA reconstruction



- ⇒ Minimizes distance between observations and their projection
- ⇒ Approx $X_{n \times p}$ with a low rank matrix $S < p$:

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

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

$$= F_{n \times S} V_{p \times S}'$$

V principal axes - loadings

Missing values in PCA

⇒ PCA: least squares

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

⇒ PCA with missing values: weighted least squares

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

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

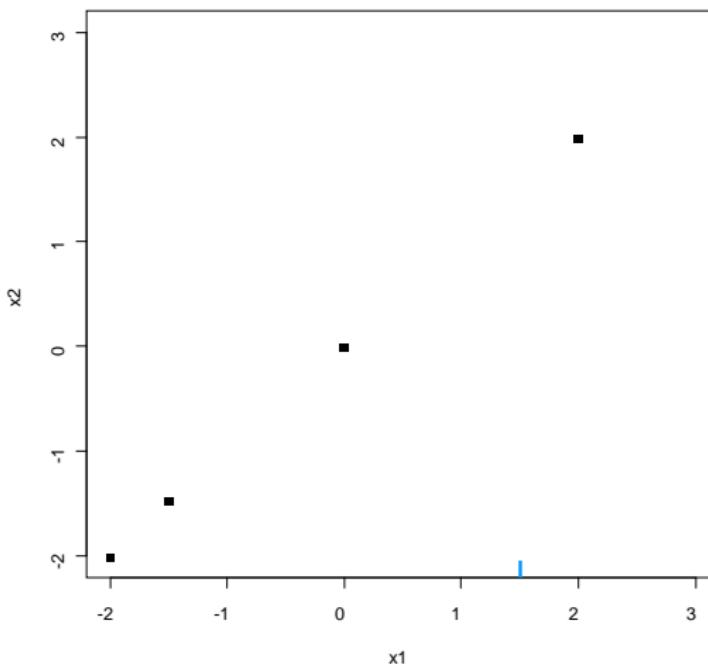
Many algorithms: weighted alternating least squares (Gabriel & Zamir, 1979); iterative PCA (Kiers, 1997)

oooo
oooooooooo

oooooooooooo
ooo

Iterative PCA

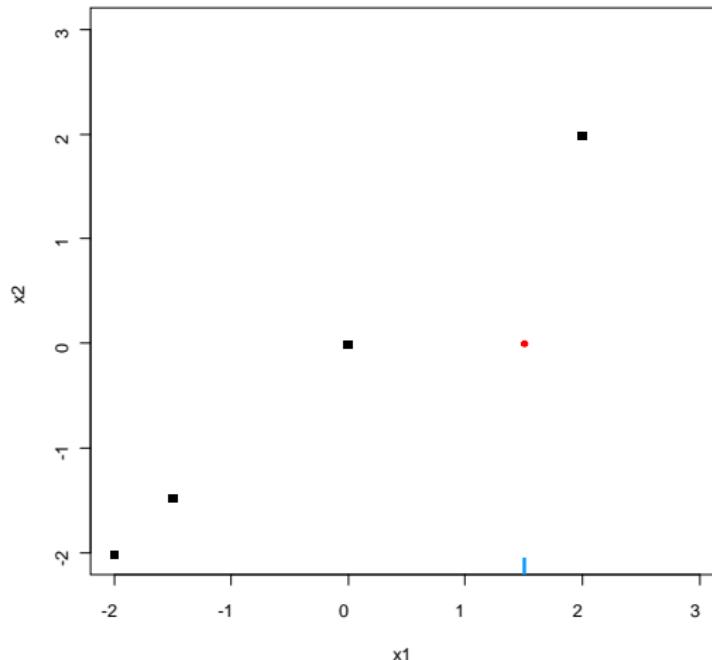
	x1	x2
-	2.0	-2.01
-	1.5	-1.48
0.0	-0.01	
1.5		NA
2.0	1.98	



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)

```
oooo
oooooooooo
```

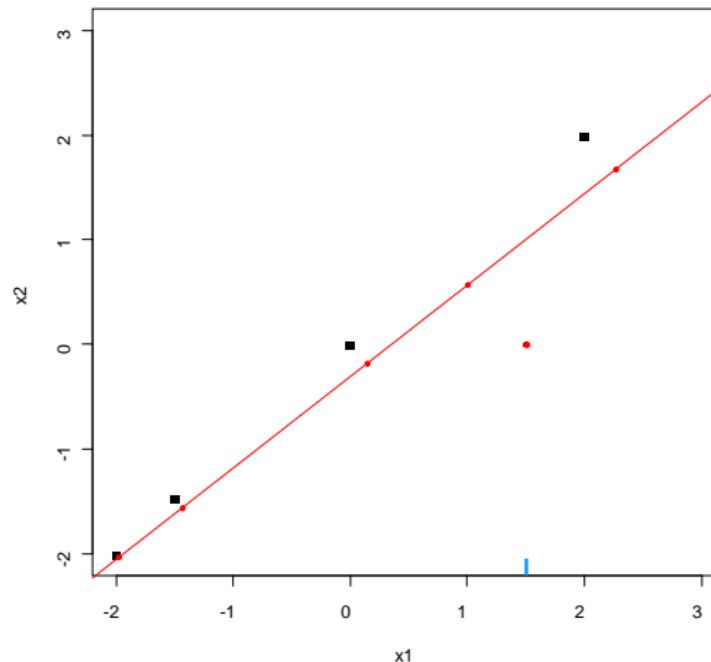
```
oooooooooooo
ooo
```

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{x}_1	\hat{x}_2
-	-1.98	-2.04
-	-1.44	-1.56
0.15	0.18	
1.00	0.57	
2.27	1.67	



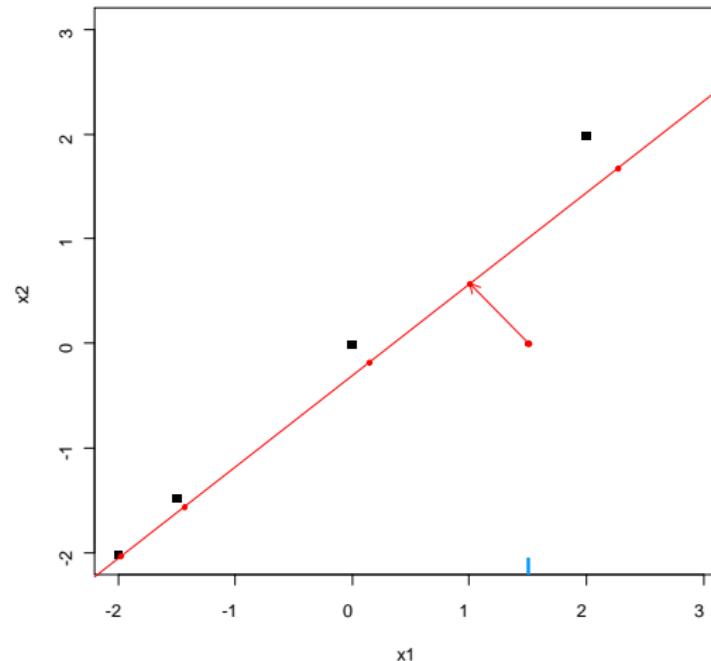
PCA on the completed data set $\rightarrow (U^\ell, \Lambda^\ell, V^\ell)$;

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{x}_1	\hat{x}_2
-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 $\hat{U}^\ell \Lambda^{1/2} V^\ell$

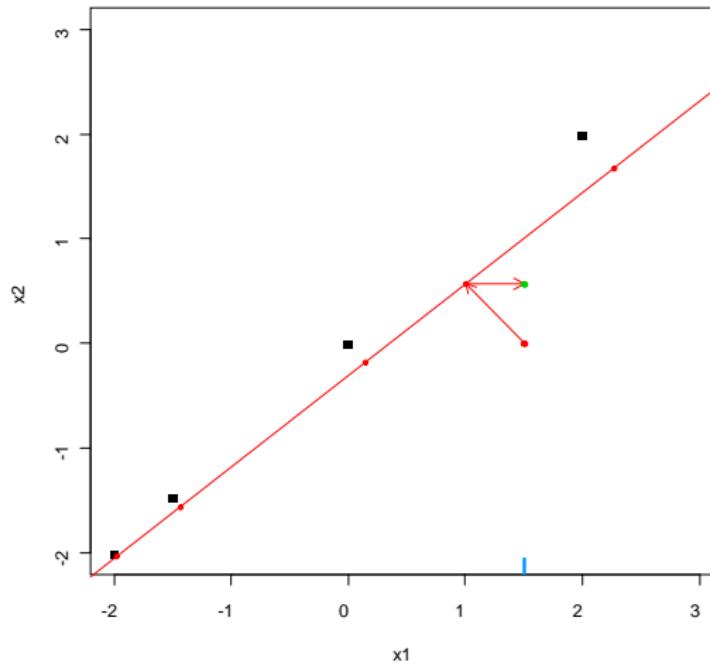
Iterative PCA

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{x}_1	\hat{x}_2
-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 $\hat{X}^\ell = W * X + (1 - W) * \hat{\mu}^\ell$

oooo
oooooooooooo

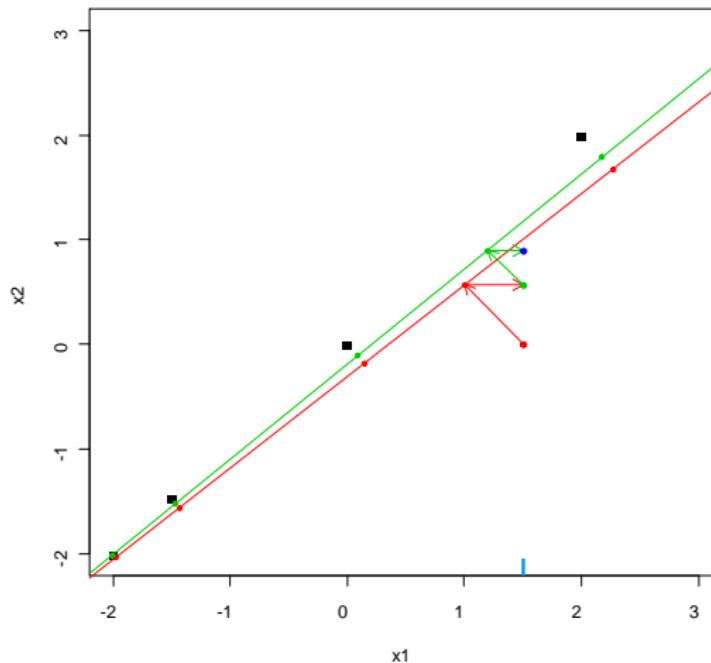
oooooooooooo
ooo

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



oooo
oooooooooo

oooooooooooo
ooo

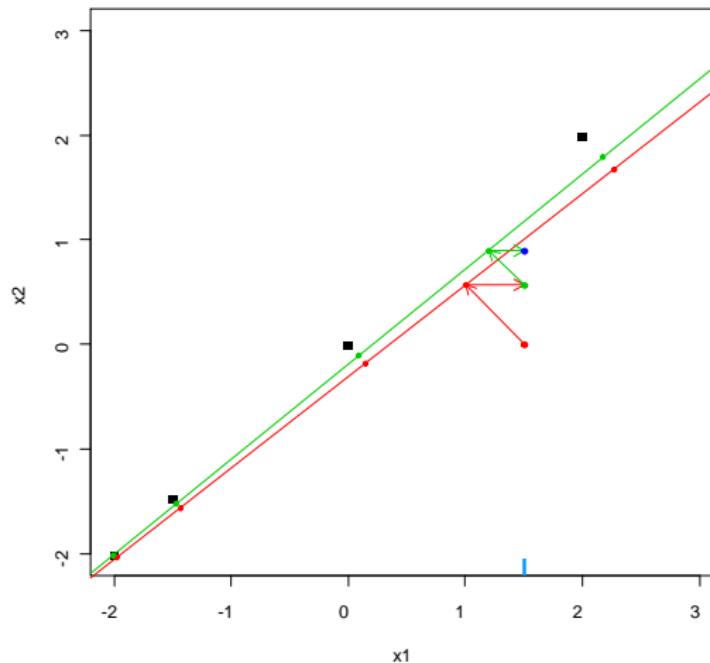
Iterative PCA

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

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

\hat{x}_1	\hat{x}_2
-2.00	-2.01
-1.47	-1.52
0.09	-0.11
1.20	0.90
2.18	1.78

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



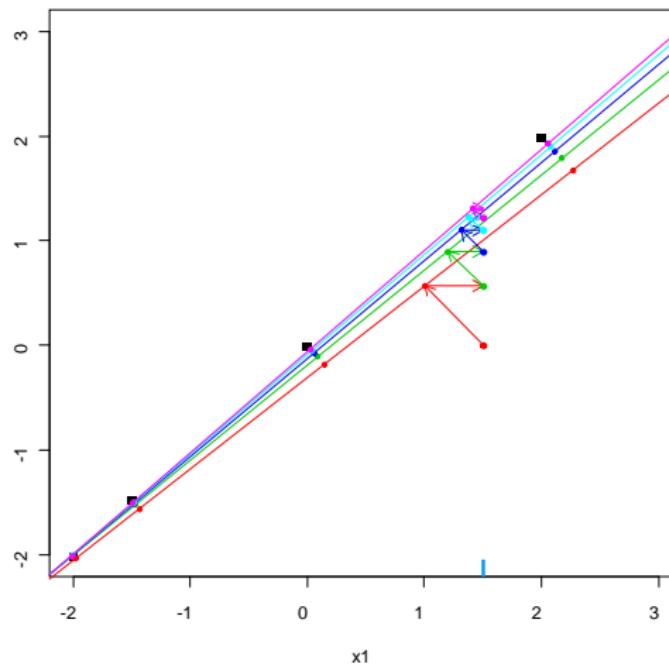
Iterative PCA

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

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

	\hat{x}_1	\hat{x}_2
-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	



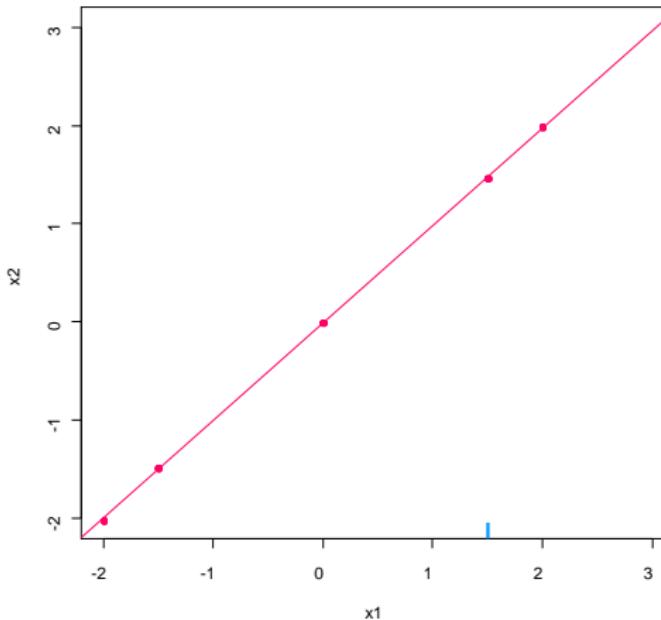
Steps are repeated until convergence

```
oooo
oooooooooo
```

```
oooooooooooo
ooo
```

Iterative PCA

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



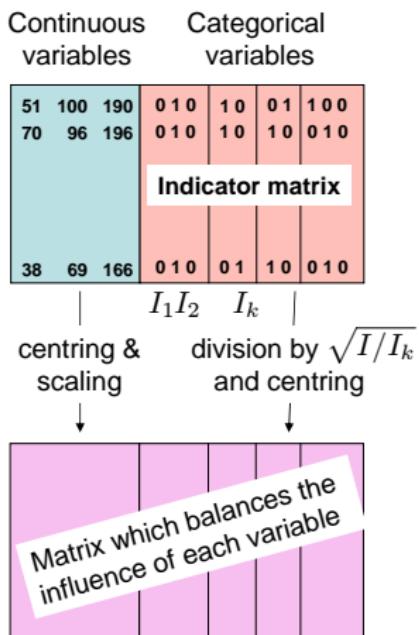
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	1.46
2.0	1.98

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

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

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:

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

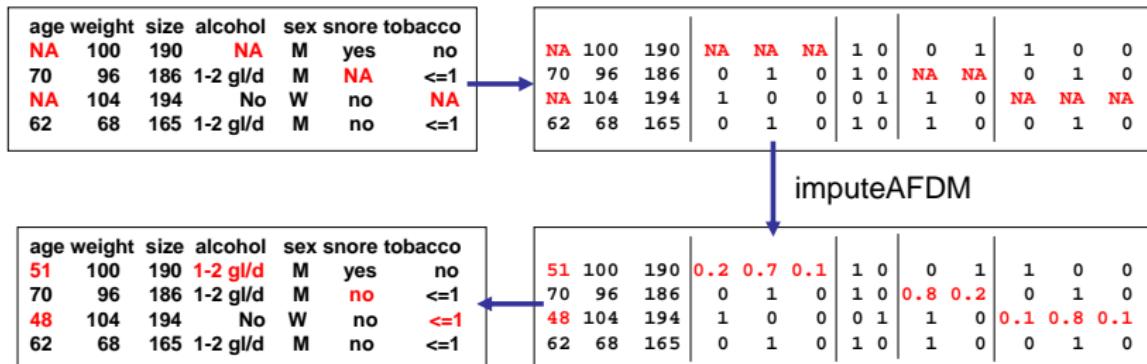
- The principal component F_s maximises:

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

Iterative FAMD algorithm

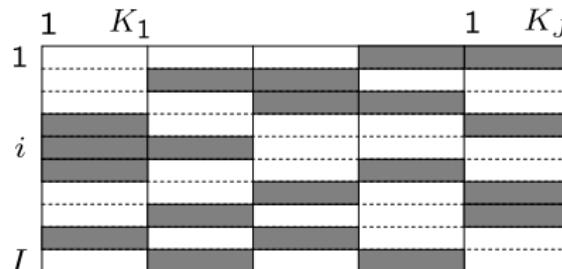
- ① Initialization: imputation mean (continuous) and proportion (dummy)
- ② Iterate until convergence
 - (a) estimation: FAMD on the completed data $\Rightarrow U, \Lambda, V$
 - (b) imputation of the missing values with the fitted matrix

$$\hat{X} = U_S \Lambda_S^{1/2} V_S'$$
 - (c) means, standard deviations and column margins are updated



⇒ Imputed values can be seen as degrees of membership

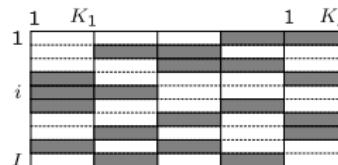
Multi-blocks data set



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

⇒ Missing rows per subtable

⇒ Regularized iterative Multiple Factor Analysis (Husson & Josse, 2013)

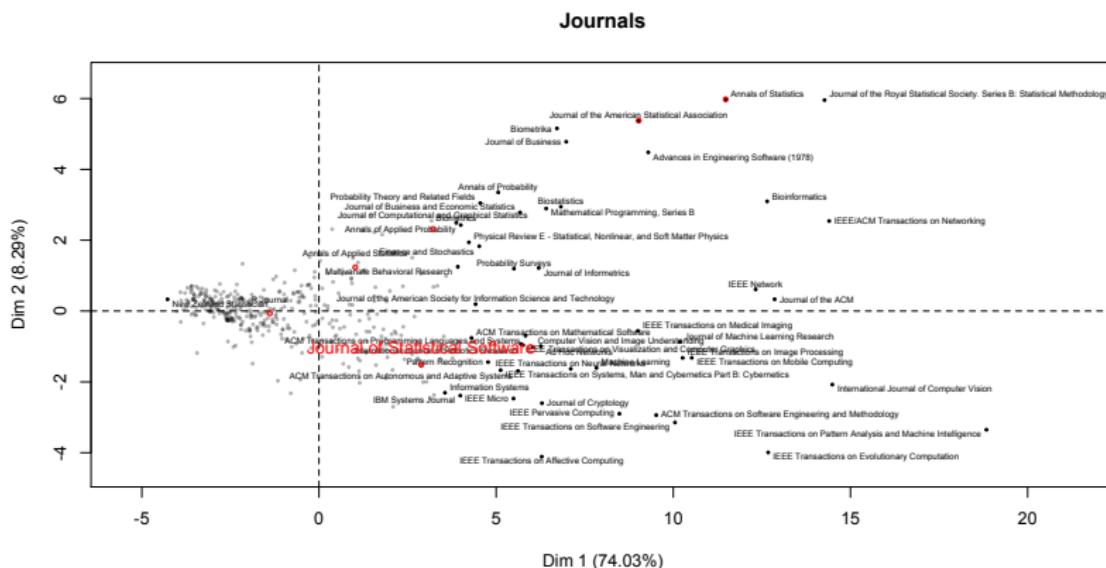
Journal impact factors

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

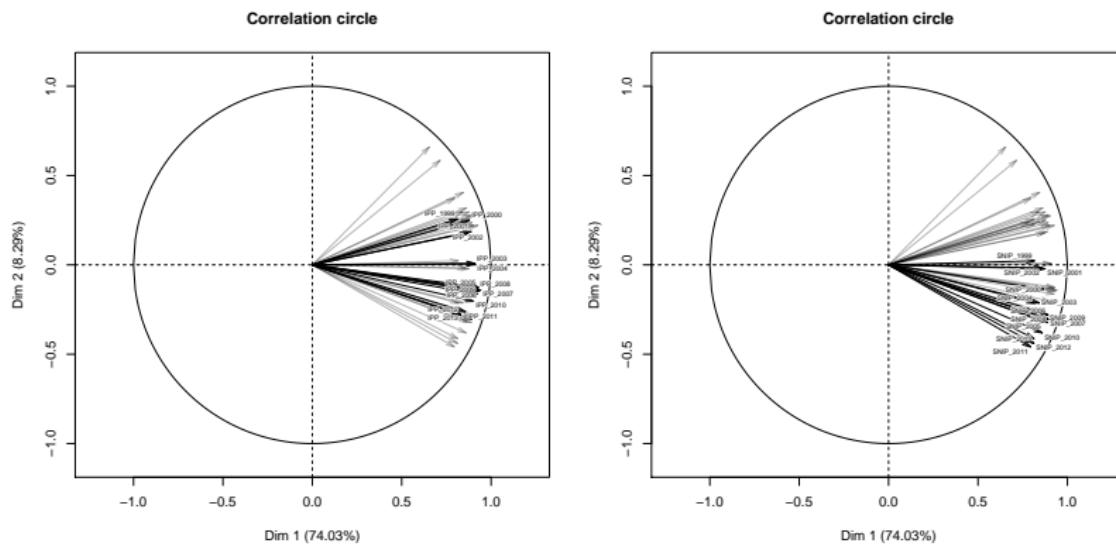
443 journals (Computer Science, Statistics, Probability and Mathematics). 45 metrics, some may be NA, 15 years by 3 types of measures:

- 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

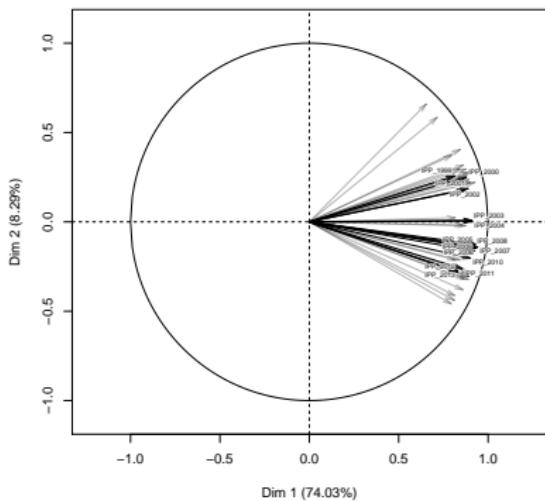


MFA with missing values

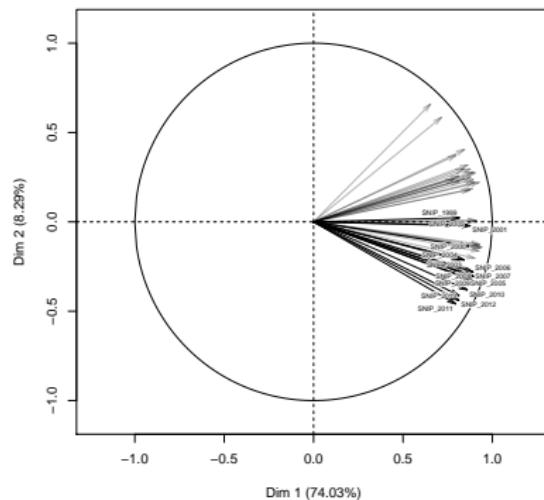


MFA with missing values

Correlation circle



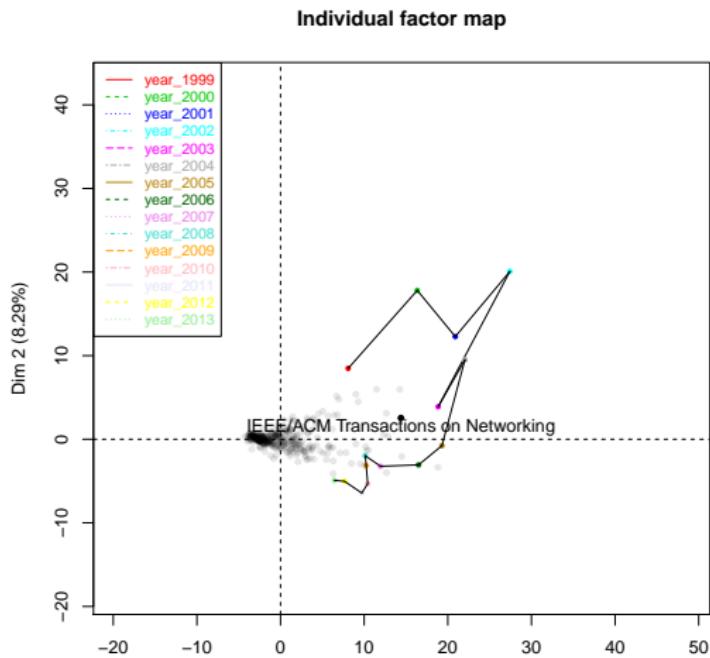
Correlation circle



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



Outline

1 Low-rank matrix estimation

- Singular values shrinkage
- Bootstrap framework

2 Inference with missing values using PC methods

- Single and multiple imputation with PCA
- Imputation with MCA

3 A multilogit bilinear model for MCA

Models for categorical variables

- Log-linear models - gold standard (Christensen, 1990; Agresti, 2013)
⇒ 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)
- ⇒ 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 $\mathbf{H}_1, \mathbf{H}_2 \succ 0$.

$$\begin{aligned} & \operatorname{argmax}_{\Gamma: \operatorname{rank}(\Gamma) \leq K} \langle \Gamma, \mathbf{G} \rangle - \frac{1}{2} \|\mathbf{H}_1 \Gamma \mathbf{H}_2\|_F^2 \\ & \Gamma^* = \mathbf{H}_1^{-1} \left[\text{SVD}_K(\mathbf{H}_1^{-1} \mathbf{G} \mathbf{H}_2^{-1}) \right] \mathbf{H}_2^{-1} \end{aligned}$$

Thus, using Lemma ??, the solution $\langle \Gamma, \mathbf{A} - \mathbf{1}\mathbf{p}^T \rangle - \frac{1}{2} \|\Gamma \mathbf{D}_p^{1/2}\|_F^2$ is given by the rank K SVD of $(\mathbf{A} - \mathbf{1}\mathbf{p}^T) \mathbf{D}_p^{-1/2}$ 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 ($\beta_0 = \log \mathbf{p}$, $\Gamma_0 = 0$), is $(\beta_0, \widehat{\Gamma}_{MCA})$.

Relationship with MCA

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

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

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

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

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

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 $\theta^+ = \text{argmin}_\theta g(\theta, \theta_0)$
- Any majorization algorithm is guaranteed to descent.

Selecting λ

⇒ 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

⇒ 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, biased!

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

\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

⇒ 2 steps : QUT to select the rank K - Shrinkage with CV

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

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

- 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

	<i>n</i>	<i>p</i>	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	0.000	0.012

Simulations results

	<i>n</i>	<i>p</i>	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

	<i>n</i>	<i>p</i>	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	0.000	0.012

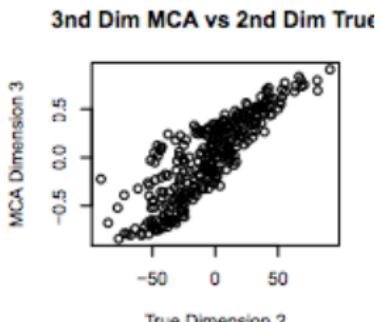
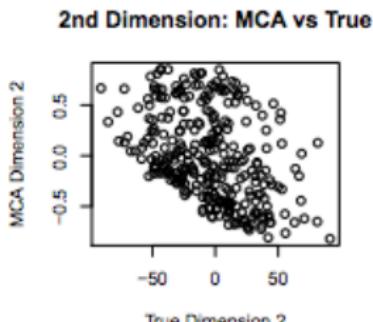
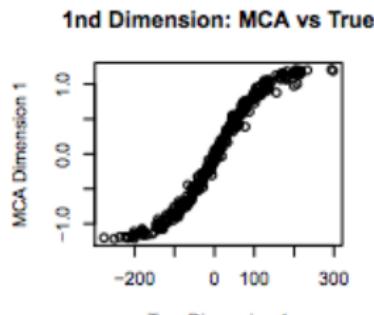
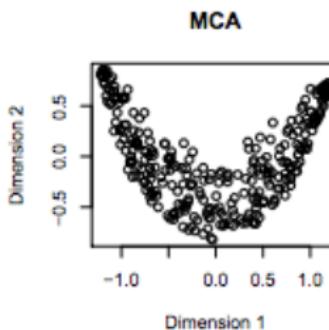
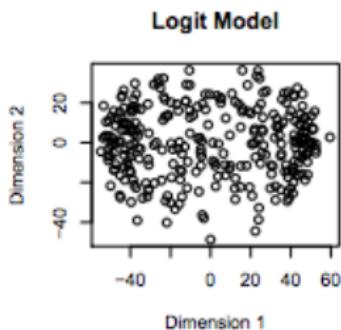
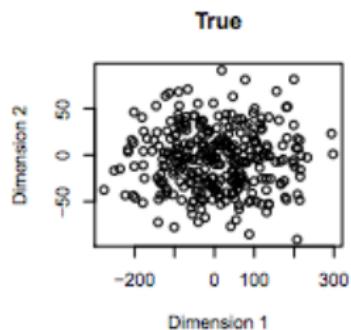
Simulations results

	<i>n</i>	<i>p</i>	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	0.000	0.012

```
oooo
oooooooooo
```

```
oooooooooooo
ooo
```

Simulation

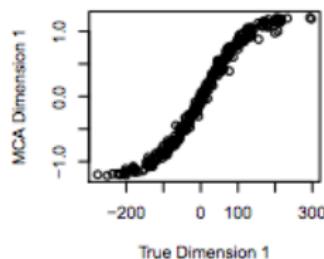


oooo
oooooooooooo

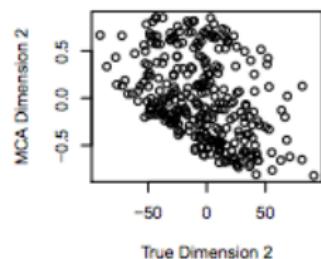
oooooooooooo
ooo

Simulation

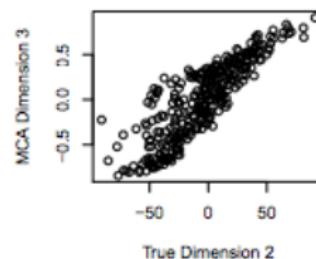
1nd Dimension: MCA vs True



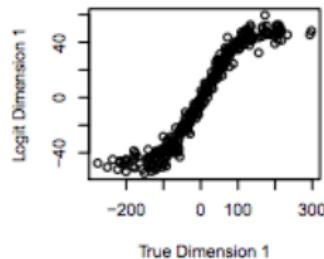
2nd Dimension: MCA vs True



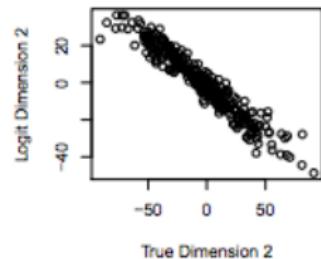
3rd Dim MCA vs 2nd Dim True



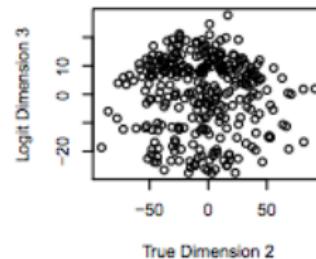
1nd Dimension: Logit vs True



2nd Dimension: Logit vs True

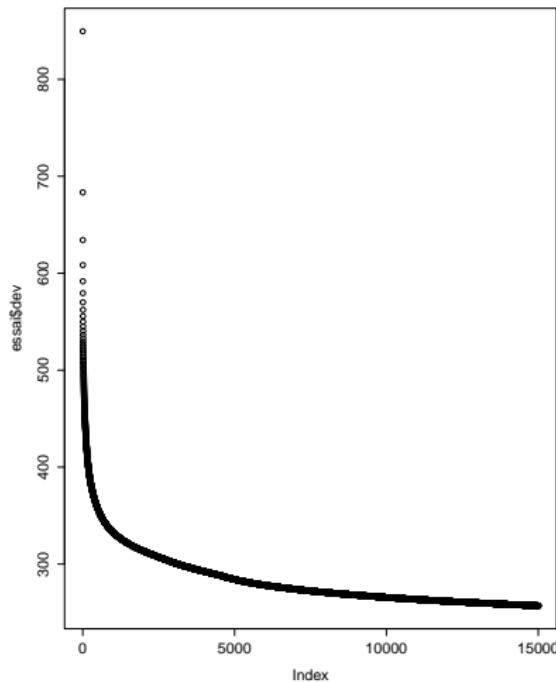


3rd Dim Logit vs 2nd Dim True



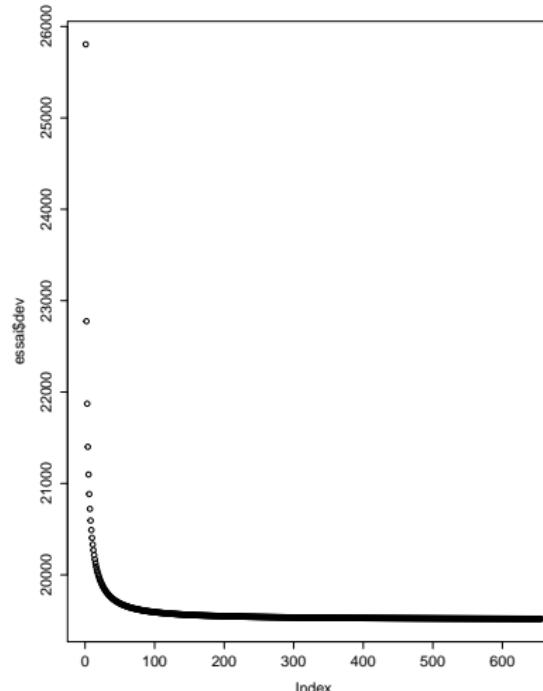
Overfitting

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



Overfitting

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

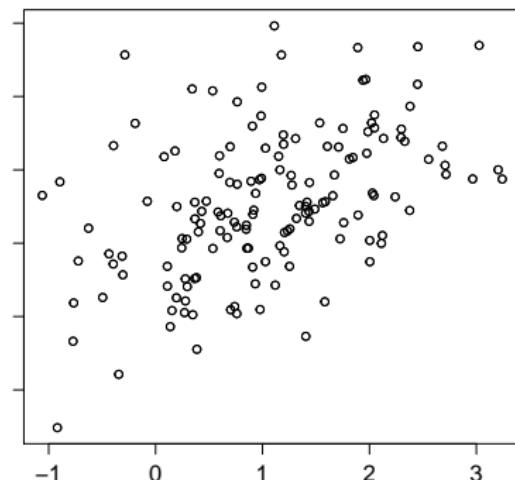
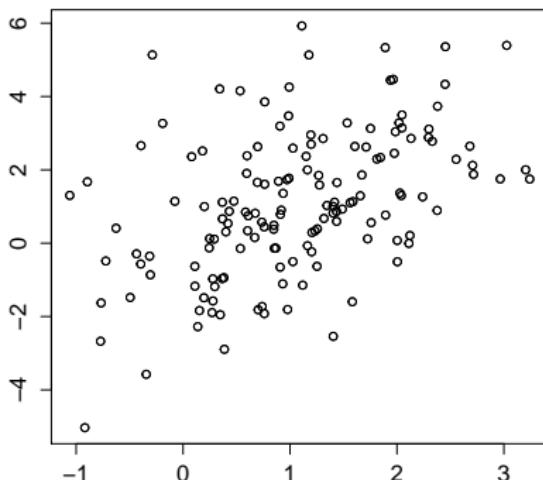


Imputation close to data and extrapolation

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

Missing completely at random

Missing (not) at random



◆ – regression

● – zonoid depth

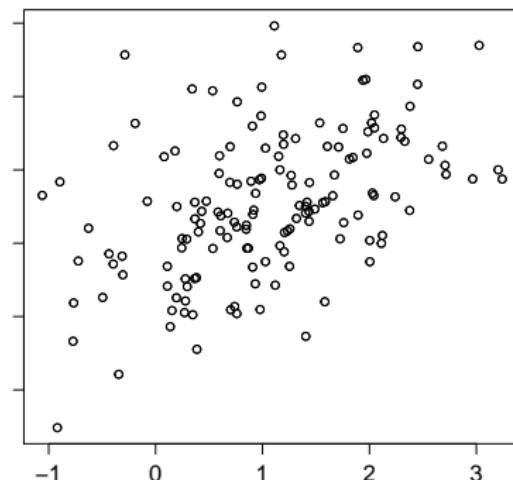
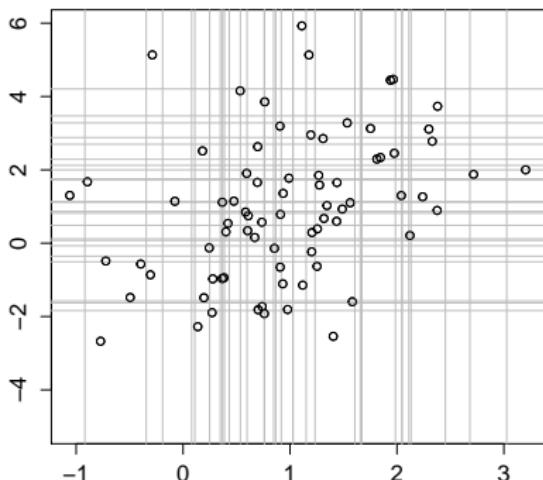
▲ – random forest

Imputation close to data and extrapolation

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

Missing completely at random

Missing (not) at random



◆ – regression

● – zonoid depth

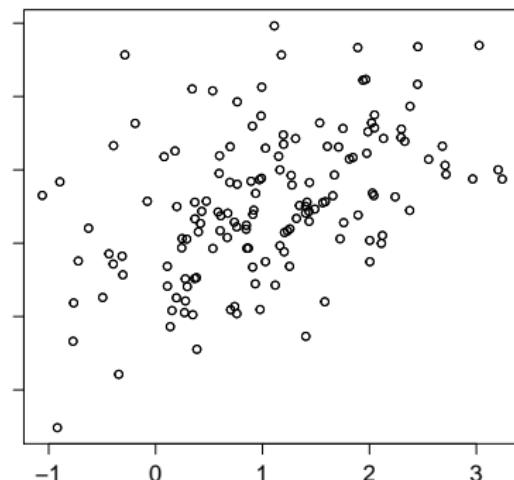
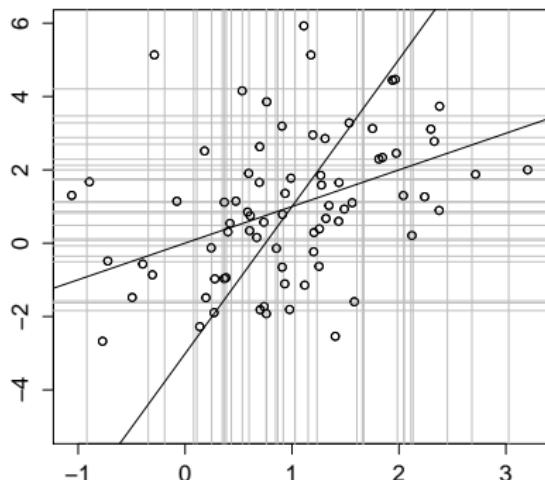
▲ – random forest

Imputation close to data and extrapolation

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

Missing completely at random

Missing (not) at random



◆ – regression

● – zonoid depth

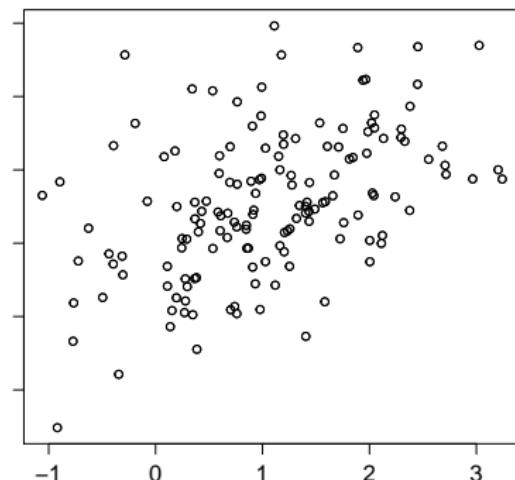
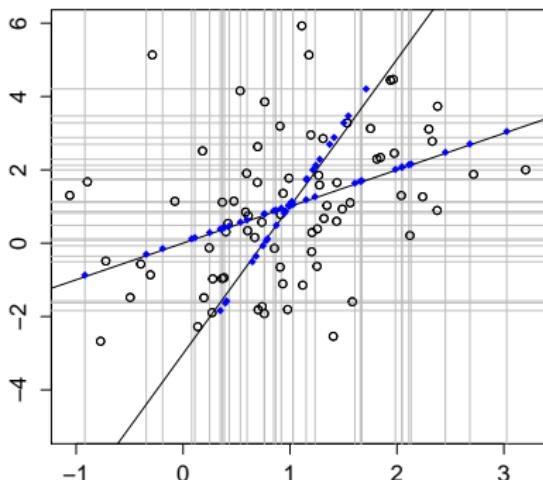
▲ – random forest

Imputation close to data and extrapolation

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

Missing completely at random

Missing (not) at random



◆ – regression

● – zonoid depth

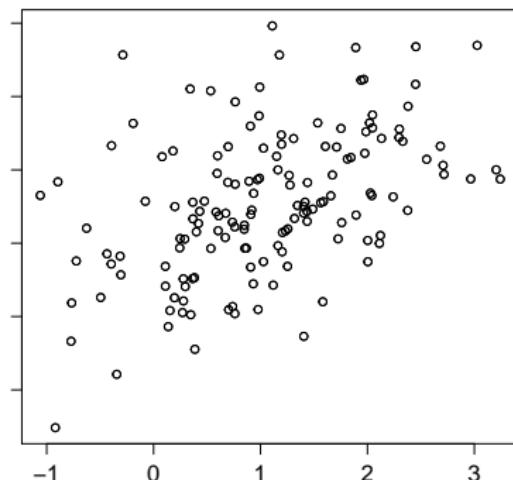
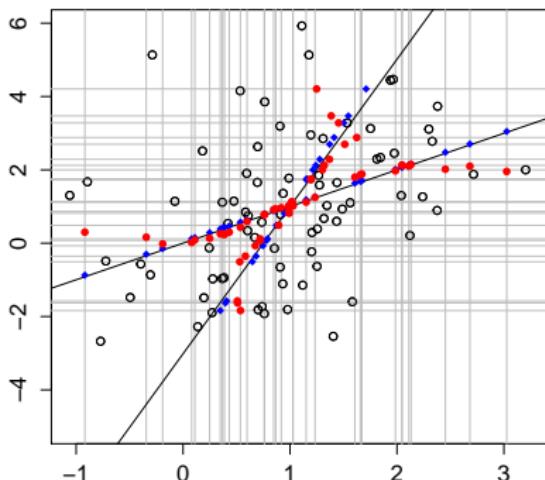
▲ – random forest

Imputation close to data and extrapolation

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

Missing completely at random

Missing (not) at random



◆ – regression

● – zonoid depth

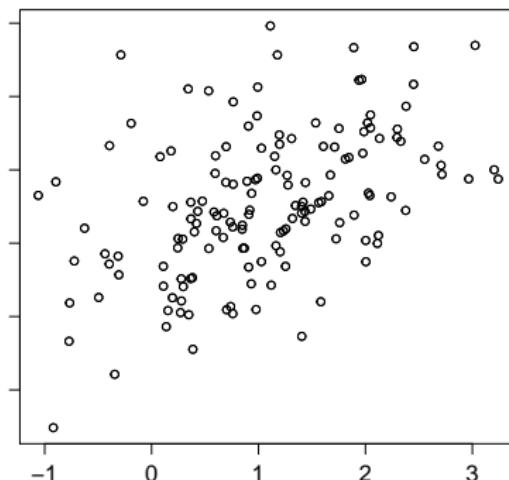
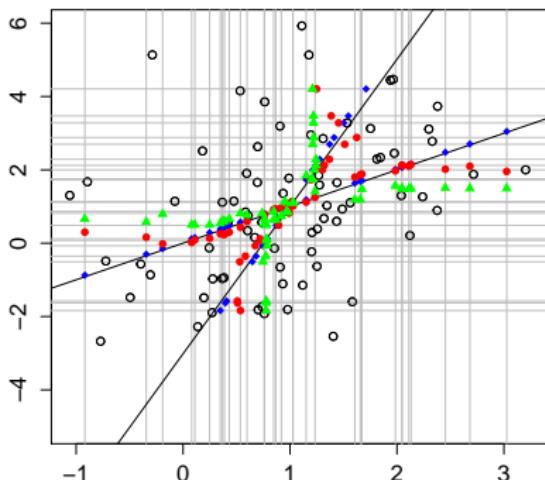
▲ – random forest

Imputation close to data and extrapolation

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

Missing completely at random

Missing (not) at random



◆ – regression

● – zonoid depth

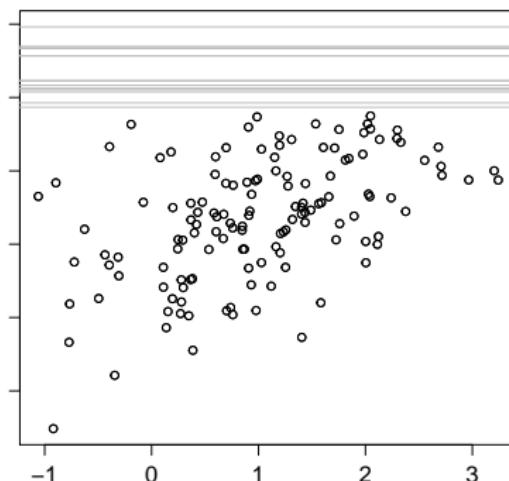
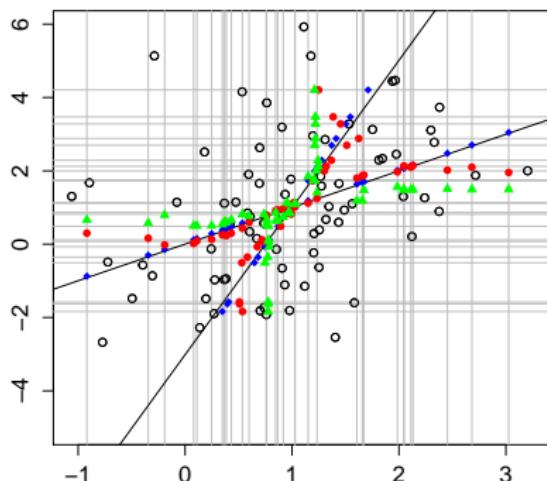
▲ – random forest

Imputation close to data and extrapolation

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

Missing completely at random

Missing (not) at random



◆ – regression

● – zonoid depth

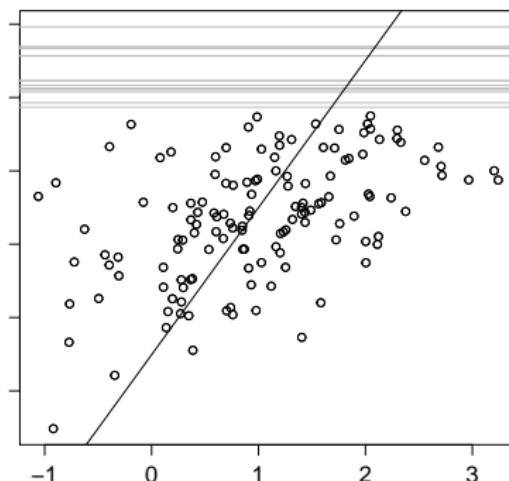
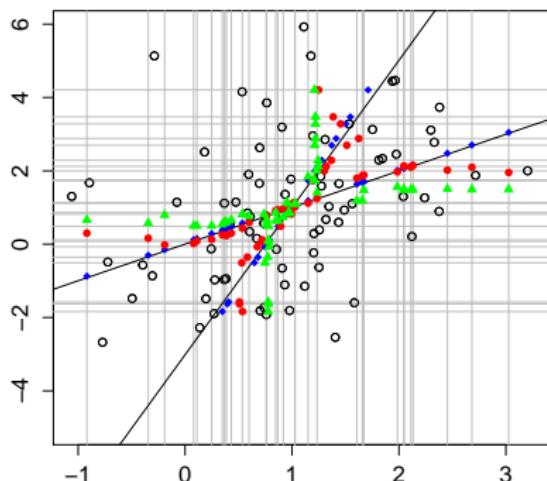
▲ – random forest

Imputation close to data and extrapolation

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

Missing completely at random

Missing (not) at random



◆ – regression

● – zonoid depth

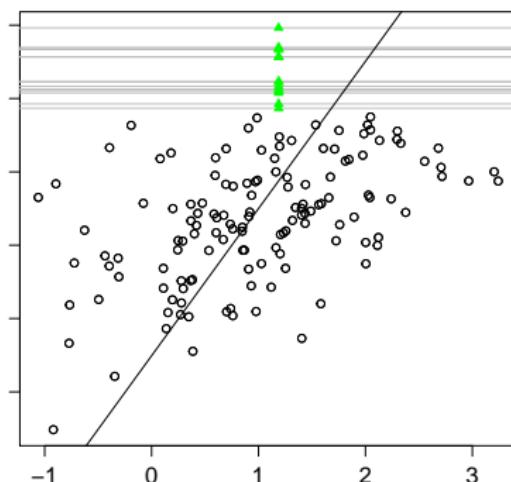
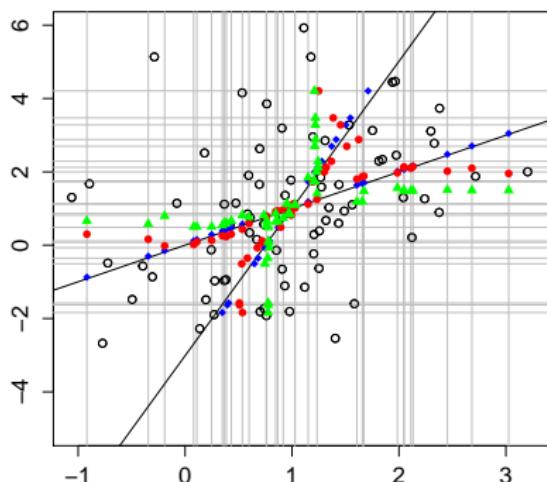
▲ – random forest

Imputation close to data and extrapolation

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

Missing completely at random

Missing (not) at random



◆ – regression

● – zonoid depth

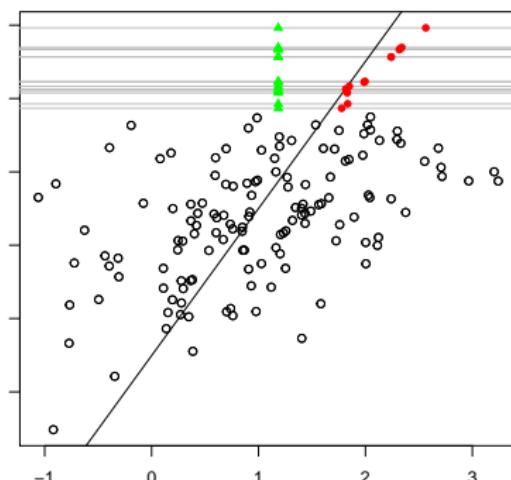
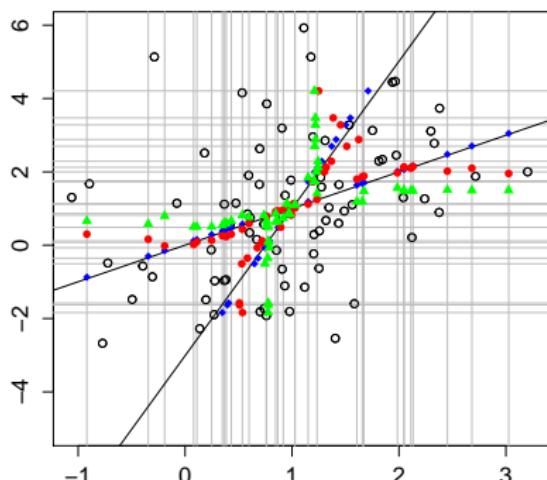
▲ – random forest

Imputation close to data and extrapolation

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

Missing completely at random

Missing (not) at random



◆ – regression

● – zonoid depth

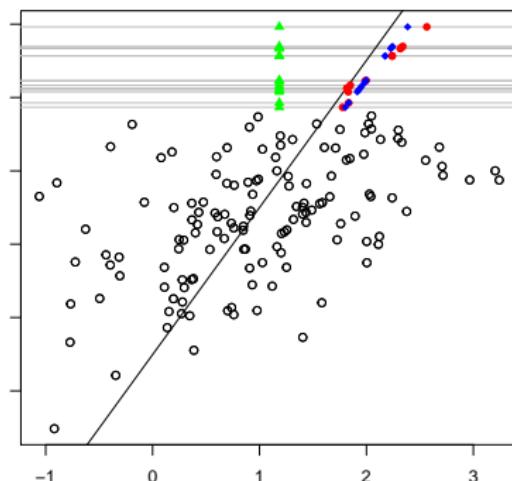
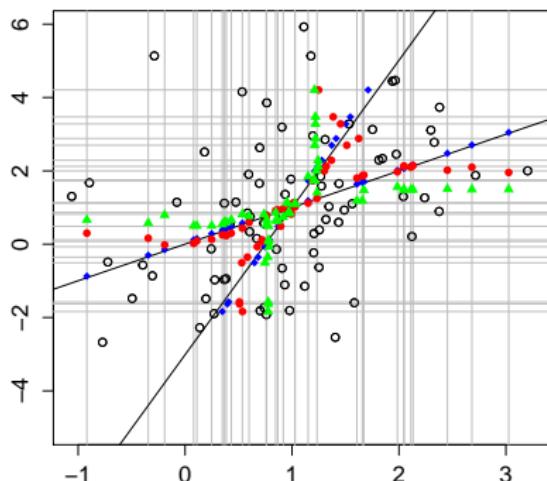
▲ – random forest

Imputation close to data and extrapolation

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

Missing completely at random

Missing (not) at random



◆ – regression

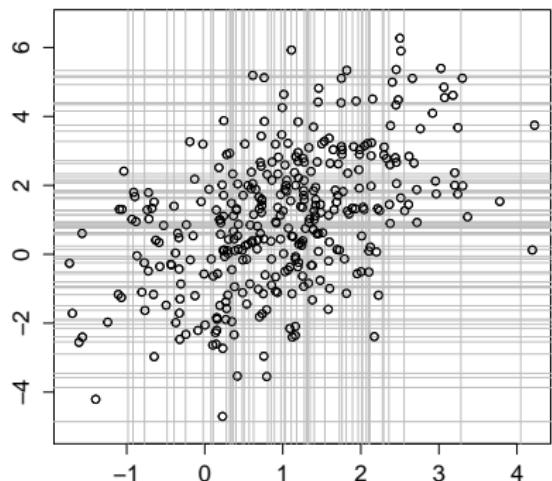
● – zonoid depth

▲ – random forest

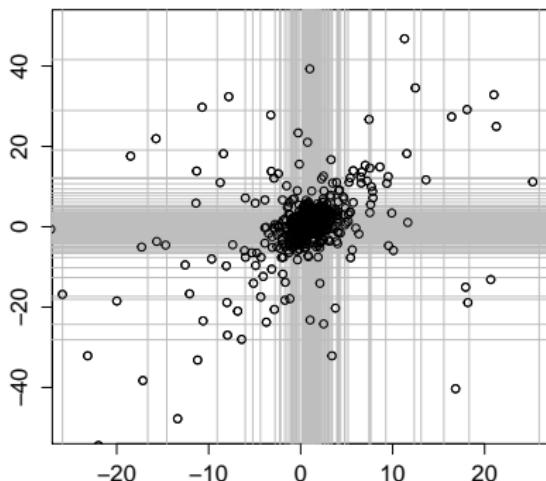
Robust imputation

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 distribution (MCAR)



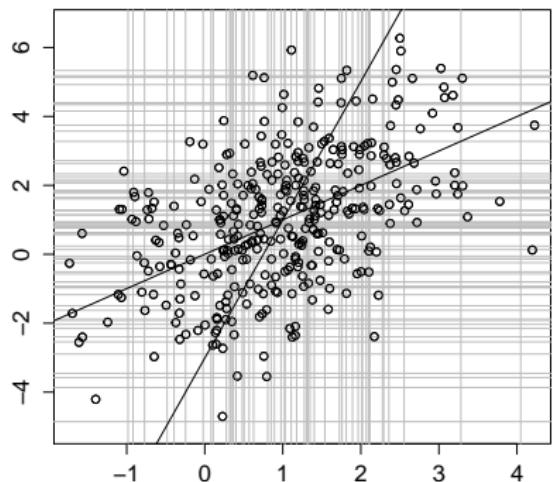
◆ – regression

● – Tukey depth

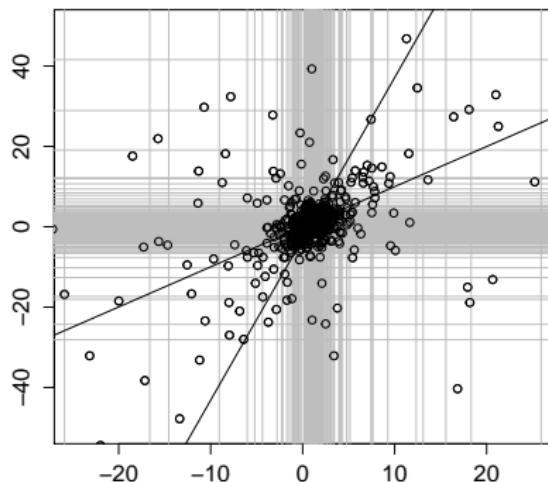
Robust imputation

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 distribution (MCAR)



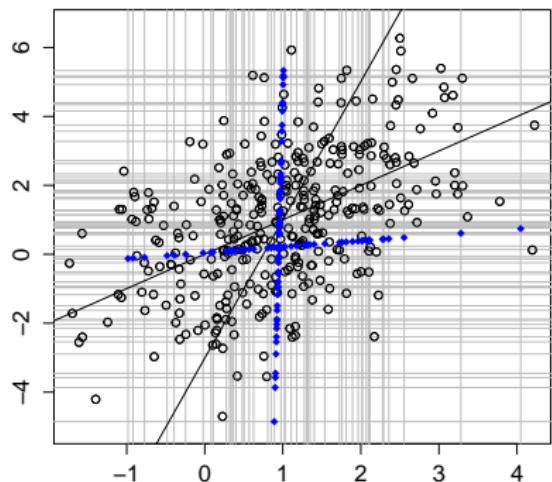
◆ – regression

● – Tukey depth

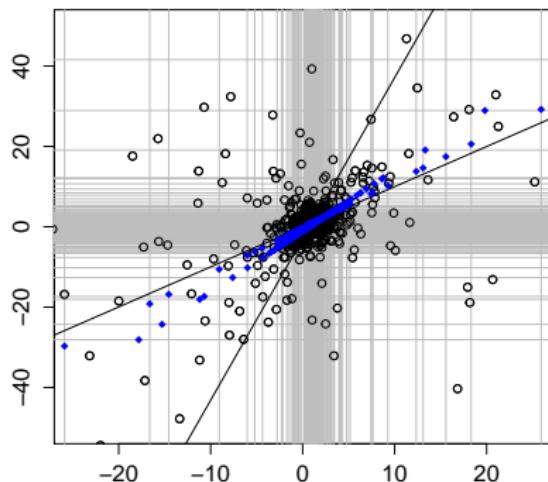
Robust imputation

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 distribution (MCAR)



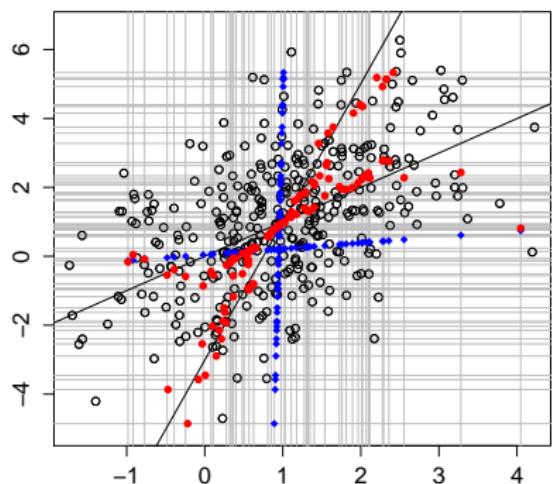
◆ – regression

● – Tukey depth

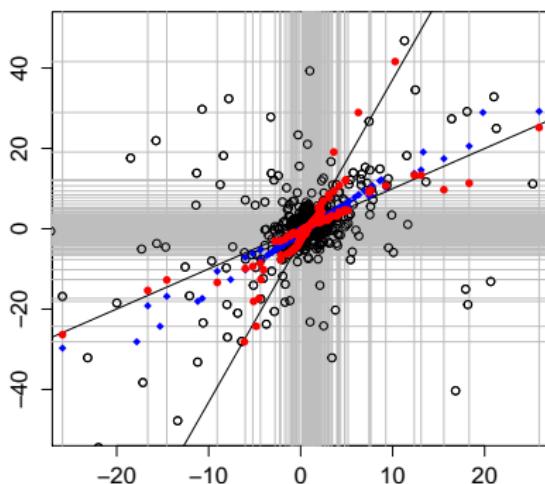
Robust imputation

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 distribution (MCAR)



◆ – regression

● – Tukey depth