Multiple Imputation with Bayesian PCA

Vincent Audigier & Julie Josse & François Husson

Agrocampus Ouest, Rennes

46e journées de Statistique, Rennes, 02 Juin 2014

Missing values

			X				Y
NA					*NA*	*NA*	
			NA				
	NA						*NA*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		NA	*NA*	*NA*			

Missing values

			X				Y
NA					*NA*	*NA*	
			NA				
	NA						*NA*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		NA	*NA*	*NA*			

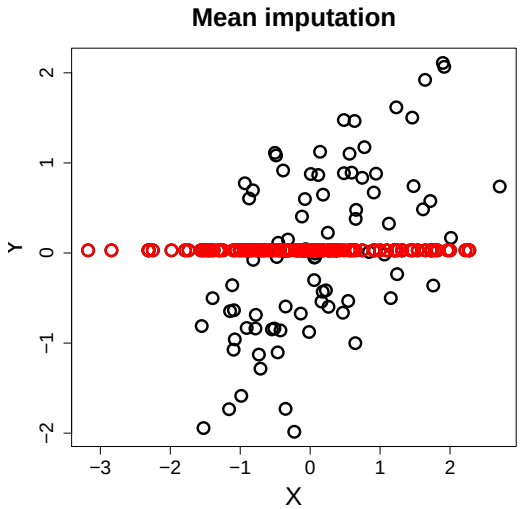
- Deletion of individuals: complete case

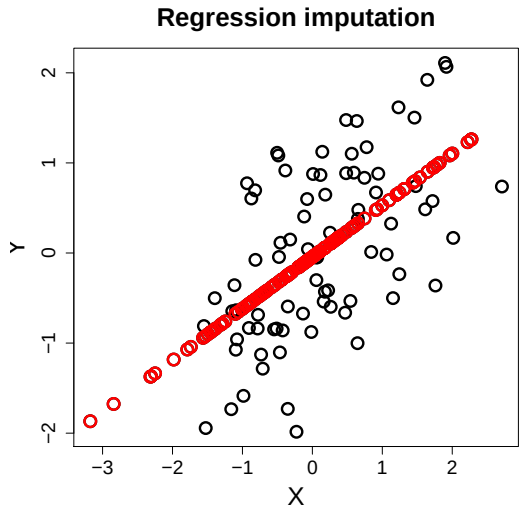
Missing values

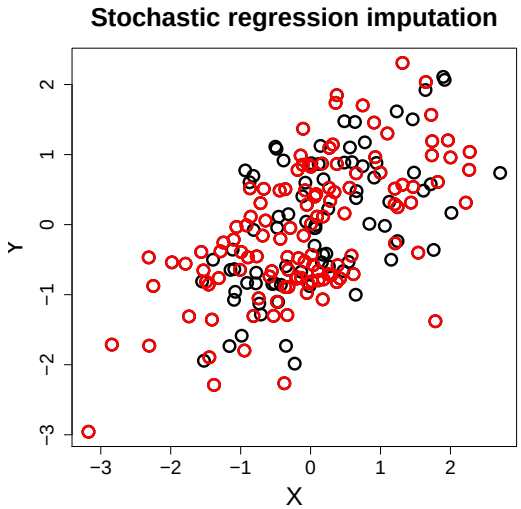
			X				Y
NA					*NA*	*NA*	
			NA				
	NA						*NA*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		NA	*NA*	*NA*			

- Deletion of individuals: complete case
- Imputation

Single imputation methods





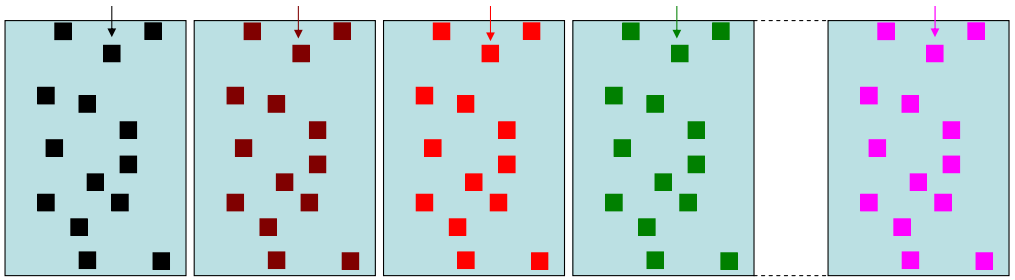


$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59
$CI\,\mu_y 95\%$	39.4	61.6	70.8

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

Multiple imputation (Rubin, 1987)

- Provide a set of M parameters to generate M plausible imputed data sets



- Perform the analysis on each imputed data set: $\hat{\theta}_m, \widehat{Var}\left(\hat{\theta}_m\right)$
- Combine the results: $\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$

$$T = \frac{1}{M}\sum_{m=1}^{M}\widehat{Var}\left(\hat{\theta}_m\right) + \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \hat{\theta}\right)^2$$

$\Rightarrow$ Aim: provide estimation of the parameters and of their variability

A multiple imputation procedure requires a single imputation method

PCA

$\Rightarrow$ Geometrical point of view

$$\|\mathbf{X}_{n\times p} - \hat{\mathbf{X}}^S_{n\times p}\|^2 \qquad \hat{\mathbf{X}}^S = \widehat{\mathbf{U}}_{n\times S}\widehat{\mathbf{\Lambda}}^{\frac{1}{2}}_{S\times S}\widehat{\mathbf{V}}^\top_{p\times S}$$

- $\widehat{\mathbf{F}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}^{\frac{1}{2}}$ principal components - scores
- $\widehat{\mathbf{V}}$ principal axes - loadings

$\Rightarrow$ Model point of view: $\mathbf{X}_{n\times p} = \tilde{\mathbf{X}}_{n\times p} + \varepsilon_{n\times p}$

$$\begin{aligned} x_{ij} &= \tilde{x}_{ij} + \varepsilon_{ij} \ , \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ &= \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js} + \varepsilon_{ij} \end{aligned}$$

Max Likelihood = Min least squares

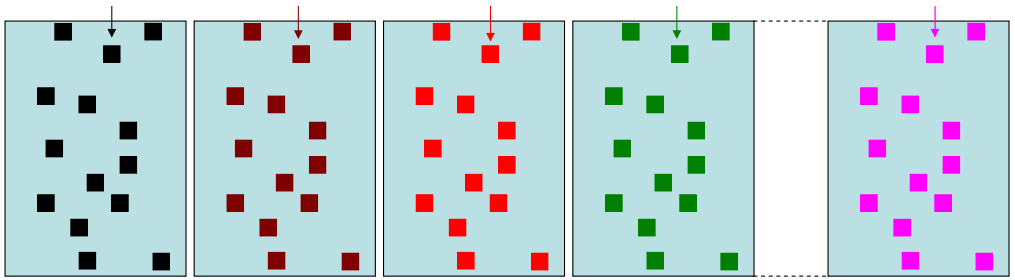
Imputation with PCA

$$
\begin{aligned}
x_{ij} &= \tilde{x}_{ij} + \varepsilon_{ij} \ , \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\
&= \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js} + \varepsilon_{ij}
\end{aligned}
$$

- An EM algorithm to estimate the parameters: iterative PCA (Kiers 1997)
- Fitted values as imputed values
- Good properties as single imputation method

Multiple imputation (Rubin, 1987)

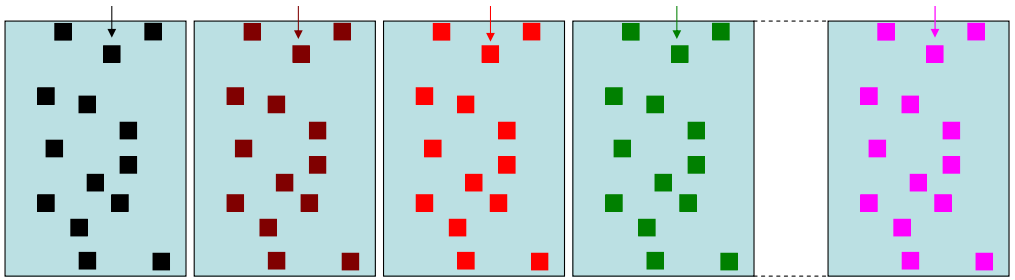
- Provide a set of M parameters to generate M plausible imputed data sets



- Perform the analysis on each imputed data set
- Combine the results and provide an estimate of the associated variability

Multiple imputation (Rubin, 1987)

- Provide a set of M parameters to generate M plausible imputed data sets



Bootstrap or Bayesian approach

- Perform the analysis on each imputed data set

- Combine the results and provide an estimate of the associated variability

Bayesian PCA complete case

$$\begin{aligned} \text{Model: } \mathbf{X}_{n\times p} &= \tilde{\mathbf{X}}_{n\times p} + \varepsilon_{n\times p} \\ x_{ij} &= \tilde{x}_{ij} + \varepsilon_{ij} \ , \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ &= \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js} + \varepsilon_{ij} \\ &= \sum_{s=1}^{S} \tilde{x}_{ij}^{(s)} + \varepsilon_{ij} \end{aligned}$$

Prior: $\tilde{x}_{ij}^{(s)} \sim \mathcal{N}(0, \tau_s^2)$

Posterior: $\left(\tilde{x}_{ij}^{(s)} | x_{ij}^{(s)}\right) \sim \mathcal{N}(\Phi_s x_{ij}^{(s)}, \Phi_s \sigma^2)$ with $\Phi_s = \frac{\tau_s^2}{\tau_s^2 + \sigma^2}$

Empirical Bayes for τ_s^2: $\hat{\tau}_s^2 = \left(\hat{\lambda}_s - \hat{\sigma}^2\right)$ (max likelihood)

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

Bayesian PCA complete case

$$
\begin{aligned}
\mathbb{E}\left[\tilde{x}_{ij}|x_{ij}\right] &= \sum_{s=1}^{S}\left(\frac{\hat{\lambda}_s - \hat{\sigma}^2}{\hat{\lambda}_s}\right)\sqrt{\hat{\lambda}_s}\,\hat{u}_{is}\,\hat{v}_{js} \\
&= \sum_{s=1}^{S}\left(\sqrt{\hat{\lambda}_s} - \frac{\hat{\sigma}^2}{\sqrt{\hat{\lambda}_s}}\right)\hat{u}_{is}\,\hat{v}_{js} \\
\hat{x}_{ij}^{PCA} &= \sum_{s=1}^{S}\left(\sqrt{\hat{\lambda}_s}\right)\hat{u}_{is}\,\hat{v}_{js}
\end{aligned}
$$

Multiple imputation Bayes-PCA

1. Variability of the parameters, M plausible $(\hat{x}_{ij})^1, \ldots, (\hat{x}_{ij})^M$
   - Posterior distribution: Bayesian PCA

$$\left(\tilde{x}_{ij}^{(s)} | x_{ij}^{(s)}\right) = \mathcal{N}(\Phi_s x_{ij}^{(s)}, \Phi_s \sigma^2)$$

   - Data Augmentation (Tanner and Wong, 1987)
     - (0) Max likelihood estimate
     - (I) impute using the PCA model
     - (P) draw new parameters from the posterior

2. Imputation according to the PCA model using the set of M parameters

Expectation Maximization Bootstrap Algorithm

- Hypothesis $X_{n\times p}$: $x_{i.} \sim \mathcal{N}(\mu, \Sigma)$
- Algorithm:
  1. Bootstrap rows: $X^1, \ldots, X^M$
     EM algorithm: $(\mu^1, \Sigma^1), \ldots, (\mu^M, \Sigma^M)$
  2. Imputation: $x_{i.}^m$ drawn from $\mathcal{N}(\mu^m, \Sigma^m)$
- Implemented: `R` package `Amelia` (J. Honaker, G. King, M. Blackwell)

Conditional modeling using Bayesian regressions

- Hypothesis: one model/variable
  all variables continuous and a regression/variable: $\mathcal{N}(\mu, \Sigma)$

- Algorithm:

  One variable with missing values:

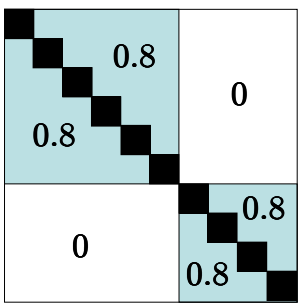
  1. Bayesian approach: (β^m, σ^m)
  2. Imputation: stochastic regression x_{ij}^m drawn from $\mathcal{N}(X_{-j}\beta^m, \sigma^m)$

  Many variables with missing values: cycles through variables

- Implemented: `R` package `MICE` (Stef van Buuren)

Simulations

- Quantities of interest: $\theta_1 = \mathbb{E}[Y], \theta_2 = \beta_1, \theta_3 = \rho$
- 1000 simulations
  - data set drawn from $\mathcal{N}_p(\mu, \Sigma)$ with a two-block structure, varying n (30 or 200), p (6 or 60) and ρ (0.3 or 0.9)



  - 10% or 30% of missing values using a MCAR mechanism
  - multiple imputation using $M = 20$ imputed arrays
- Criteria
  - bias, rmse
  - CI width, coverage

Results for the expectation

	parameters				confidence interval width			coverage		
	n	p	ρ	%	*Amelia*	*MICE*	*BayesMIPCA*	*Amelia*	*MICE*	*BayesMIPCA*
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

Conclusion - Perspectives

A new multiple imputation method based on PCA

- Good results and works when $n < p$
- Linear relationships
- Requires to define the number of dimensions

Conclusion - Perspectives

A new multiple imputation method based on PCA

- Good results and works when $n < p$
- Linear relationships
- Requires to define the number of dimensions

- Categorical data?