Low-rank methods for multi-source, heterogeneous and incomplete data

Geneviève Robin

Centre de Mathématiques Appliquées, École Polytechnique (UMR 7641), XPOP project-team, INRIA Saclay

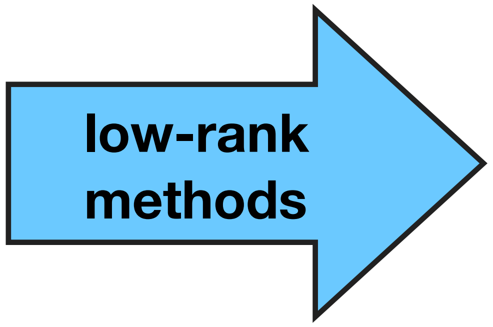
Thèse de doctorat encadrée par Julie Josse et Éric Moulines

11 Juin 2019

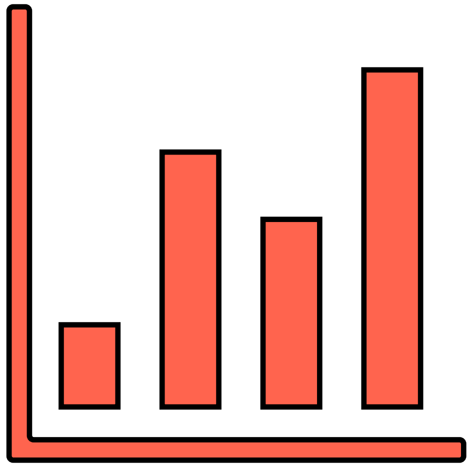
Statistical data table analysis

Patient ID	Weight	Pelvic X-ray	Accident	Time in ICU (h)
1	NA	Normal	Falling (from a height)	NA
2	85	NA	Falling (from a height)	2
3	80	NA	Car-pedestrian accident	NA
4	50	Normal	Falling (from a height)	2
5	73	NA	Falling (from own height)	NA
6	NA	NA	Falling (from own height)	NA

Data table
(multivariate data)



Analysis methods
(estimation)

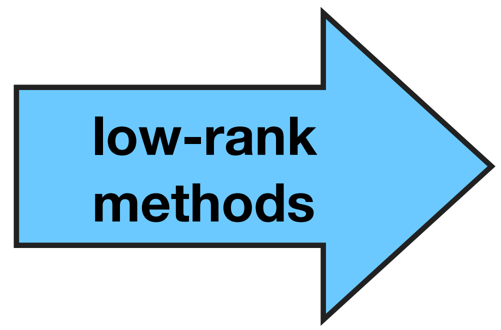


Interpretable
data summaries,
impute missing values

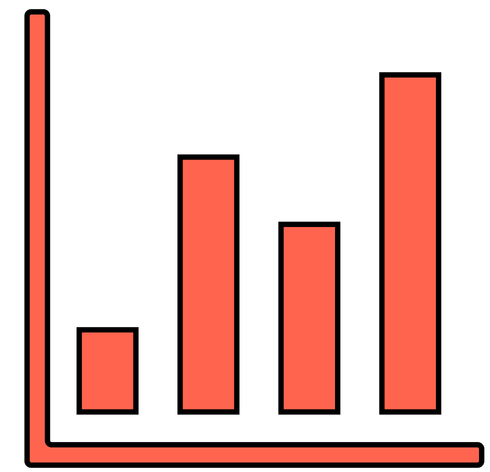
Statistical data table analysis

Patient ID	Weight	Pelvic X-ray	Accident	Time in ICU (h)
1	NA	Normal	Falling (from a height)	NA
2	85	NA	Falling (from a height)	2
3	80	NA	Car-pedestrian accident	NA
4	50	Normal	Falling (from a height)	2
5	73	NA	Falling (from own height)	NA
6	NA	NA	Falling (from own height)	NA

Data table
(multivariate data)



Analysis methods
(estimation)



Interpretable
data summaries,
impute missing values

« Old » problem: multivariate data analysis methods data back to the early 20th century (Pearson, 1901 and Hotelling, 1933)

Modern data tables

High-dimensional	Multi-source	Heterogeneous	Incomplete
• medical registry (20,000x250) • genomics data set (1,000x100,000) • Netflix data (800,000x20,000)	• patients across hospitals • aggregation of experiments • combining data sources (survey data, experimental results, web scraping)	• qualitative attributes (prof. activity) • quantitative features (age, income) • discrete features (species counts)	• nonresponse phenomenon • machine failures • unaccessible data

Modern data tables

High-dimensional

- medical registry (20,000x250)
- genomics data set (1,000x100,000)
- Netflix data (800,000x20,000)

Multi-source

- patients across hospitals
- aggregation of experiments
- combining data sources (survey data, experimental results, web scraping)

Heterogeneous

- qualitative attributes (prof. activity)
- quantitative features (age, income)
- discrete features (species counts)

Incomplete

- nonresponse phenomenon
- machine failures
- unaccessible data

Need for new models, theory, software

Example: Traumabase data set

(20,000 individuals, 250 attributes)

Patient ID	Centre	Weight	Pelvic X-ray	Accident	Time in ICU	Age	On call	DC
1	Beaujon	NA	Normal	Falling (from own height)	NA	84	Non	NA
2	Bicêtre	85	NA	Falling (from a height)	2	64	Non	NA
3	Beaujon	80	NA	Car accident	NA	35	Non	Non
4	Beaujon	50	Normal	Falling (from a height)	2	NA	Non	NA
5	Henri Mondor	73	NA	Car accident	NA	22	Non	NA
6	Pitié-Salpêtrière	NA	NA	Falling (from a height)	NA	14	Non	NA

Multi-source

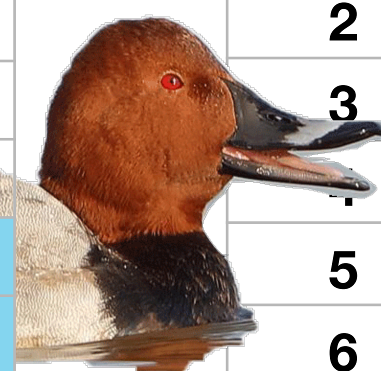
- Finding predictors of mortality = predictive models
- Describe the patients population = exploratory data analysis

Example: Waterbirds data set

(23 species, 785 sites, 28 years, 17 covariates)

Common pochard (canard milouin)

Site	2008	2009	2010
1	NA	0	0
2	4	50	25
3	NA	0	0
4	NA	NA	NA
5	NA	NA	NA
6	0	0	0
7	5	75	870
8	9	34	0
9	10	8	30
10	NA	182	27



Site	Year	Rain	Eco	Country	Agri
1	**2008**	163.7	0.8	Algeria	16.2
2	**2008**	60.7	0.8	Algeria	16.2
3	**2008**	227.9	0.8	Algeria	16.2
4	**2008**	174.8	0.8	Algeria	16.2
5	**2008**	163.7	0.8	Algeria	16.2
6	**2008**	230.7	0.8	Algeria	16.2
7	**2008**	243.5	0.8	Algeria	16.2
8	**2008**	262.6	0.8	Algeria	16.2
9	**2008**	197.3	0.8	Algeria	16.2
10	**2008**	227.9	0.8	Algeria	16.2
1	**2009**	255.1	-1.2	Algeria	16.1
2	**2009**	179.8	-1.2	Algeria	16.1

- Two sources of data: bird censuses and web-scraping
- Estimate population trends and select important covariates

Low-rank matrices

$$\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,m_2} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,m_2} \\ \vdots & & & \\ A_{m_1-1,1} & A_{m_1-1,2} & \cdots & A_{m_1-1,m_2} \\ A_{m_1,1} & A_{m_1-1,2} & \cdots & A_{m_1,m_2} \end{bmatrix} = (A_{i,j}) \in \mathcal{X}^{m_1 \times m_2}$$

$$A_{2,.} \in \mathcal{X}^{m_2}$$

$$A_{.,2} \in \mathcal{X}^{m_1}$$

Rank of a matrix:

A matrix is of rank r, noted $\mathrm{rank}(\mathbf{A}) = r$, if its rows lie in a subspace of dimension r:

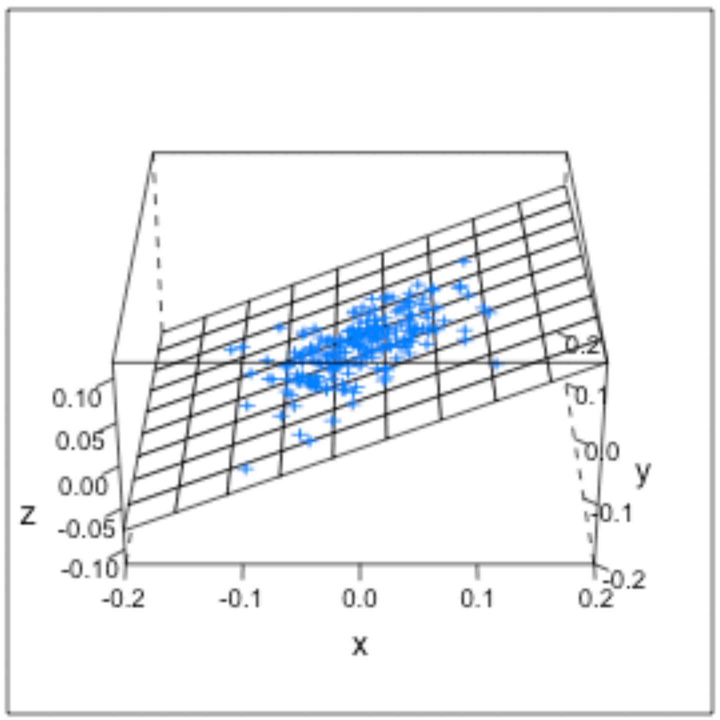
$$\forall i \in \{1, \ldots, m_1\},\ A_{i,.} \in \mathcal{S}_1 \subseteq \mathcal{X}^{m_2},\ \dim(\mathcal{S}_1) = r$$

Low-rank matrix:

A matrix $\mathbf{A}$ of size $m_1 \times m_2$ is of low-rank if

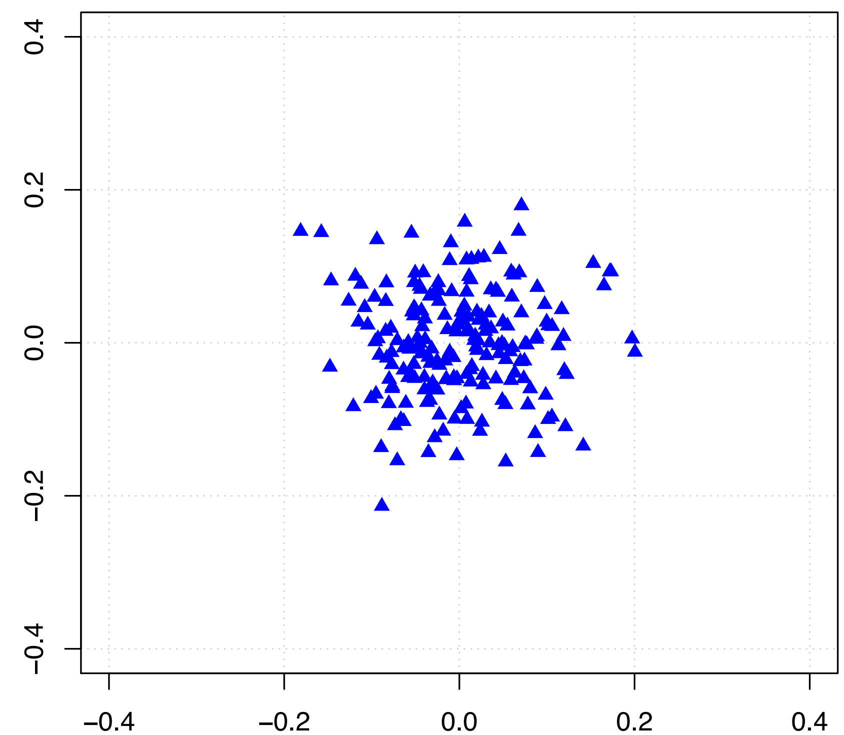
$$\mathrm{rank}(\mathbf{A}) \ll \max(m_1, m_2)$$

Row and column vector spaces



change of

basis



The rows are of dimension 3 but lie in a 2-dimensional subspace

Singular value decomposition

$$\mathbf{A} = \underbrace{\begin{bmatrix} U_1 & \ldots & U_r \end{bmatrix}}_{\text{new coordinates (norm.)}} \begin{bmatrix} \sigma_1(\mathbf{A}) & 0 & \ldots \\ 0 & & \\ \vdots & & \sigma_r(\mathbf{A}) \end{bmatrix} \underbrace{\begin{bmatrix} V_1^\top \\ \vdots \\ V_r^\top \end{bmatrix}}_{\text{new basis}}$$

singular values

$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$: singular value decomposition (SVD)

Number of parameters: $r(m_1 + m_2 - r) \leq m_1 m_2$

The rank controls:
- Computational cost
- Model complexity

Low-rank models and approximations

Main idea: replace a data table by a low-rank matrix

Low-rank models and approximations

Main idea: replace a data table by a low-rank matrix

Example of model:

Data table (observations)

$$\mathbf{Y} = \mathcal{F}(\mathbf{X}^0)$$

Function (noisy)

Low-rank matrix (unknown)

Low-rank models and approximations

Main idea: replace a data table by a low-rank matrix

Example of model:

Data table (observations) $\rightarrow$

$$\mathbf{Y} = \mathcal{F}(\mathbf{X}^0)$$

Function (noisy)

Low-rank matrix (unknown)

Estimate the low-rank matrix:

data fitting term

$$\begin{array}{ll} \text{minimize} & d(\mathbf{Y}, \mathcal{F}(\mathbf{X})) \\ \text{subject to} & \operatorname{rank}(X) \leq r \end{array}$$

Low-rank models and approximations

Main idea: replace a data table by a low-rank matrix

Example of model:

Data table (observations)

$$\mathbf{Y} = \mathcal{F}(\mathbf{X}^0)$$

Function (noisy)

Low-rank matrix (unknown)

Estimate the low-rank matrix:

data fitting term

$$\begin{array}{ll} \text{minimize} & d(\mathbf{Y}, \mathcal{F}(\mathbf{X})) \\ \text{subject to} & \text{rank}(X) \leq r \end{array}$$

Intractable problem in general

Nuclear norm heuristics

$$\begin{array}{ll} \text{minimize} & d(\mathbf{Y}, \mathcal{F}(\mathbf{X})) \\ \text{subject to} & \text{rank}(X) \leq r \end{array}$$

$$\text{minimize} \quad d(\mathbf{Y}, \mathcal{F}(\mathbf{X})) + \lambda \|X\|_\star$$

Intractable

Convex relaxation

Nuclear norm heuristics

$$\begin{array}{ll} \text{minimize} & d(\mathbf{Y}, \mathcal{F}(\mathbf{X})) \\ \text{subject to} & \operatorname{rank}(X) \le r \end{array}$$

Intractable

$$\text{minimize} \quad d(\mathbf{Y}, \mathcal{F}(\mathbf{X})) + \lambda \|X\|_\star$$

Convex relaxation

nuclear norm: $\|X\|_\star = \sum_{k=1}^{\operatorname{rank}(\mathbf{X})} \sigma_k(\mathbf{X})$

Nuclear norm heuristics

$$\begin{array}{ll} \text{minimize} & d(\mathbf{Y}, \mathcal{F}(\mathbf{X})) \\ \text{subject to} & \text{rank}(X) \leq r \end{array}$$

Intractable

$$\text{minimize} \quad d(\mathbf{Y}, \mathcal{F}(\mathbf{X})) + \lambda \|X\|_\star$$

Convex relaxation

nuclear norm: $\|X\|_\star = \sum_{k=1}^{\text{rank}(\mathbf{X})} \sigma_k(\mathbf{X})$

Theory, software, numerous applications:

Candès and Recht (2009), Recht et al. (2010), Candès and Plan (2010), Candès and Tao (2010), Recht (2011), Keshavan et al. (2010), Agarwal et al. (2012), Klopp (2014), Hastie et al. (2015), Udell et al. (2016)

Nuclear norm heuristics

$$\begin{array}{ll} \text{minimize} & d(\mathbf{Y}, \mathcal{F}(\mathbf{X})) \\ \text{subject to} & \operatorname{rank}(X) \leq r \end{array} \quad \rightarrow \quad \text{minimize} \quad d(\mathbf{Y}, \mathcal{F}(\mathbf{X})) + \lambda \|X\|_\star$$

Intractable

Convex relaxation

nuclear norm: $\|X\|_\star = \sum_{k=1}^{\operatorname{rank}(\mathbf{X})} \sigma_k(\mathbf{X})$

Theory, software, numerous applications:

Candès and Recht (2009), Recht et al. (2010), Candès and Plan (2010), Candès and Tao (2010), Recht (2011), Keshavan et al. (2010), Agarwal et al. (2012), Klopp (2014), Hastie et al. (2015), Udell et al. (2016)

Mostly for incomplete numeric data, or heterogeneous data without multi-source aspect

Nuclear norm heuristics

$$\begin{array}{ll} \text{minimize} & d(\mathbf{Y}, \mathcal{F}(\mathbf{X})) \\ \text{subject to} & \text{rank}(X) \leq r \end{array} \quad \longrightarrow \quad \text{minimize} \quad d(\mathbf{Y}, \mathcal{F}(\mathbf{X})) + \lambda \|X\|_\star$$

Intractable

Convex relaxation

nuclear norm: $\|X\|_\star = \sum_{k=1}^{\text{rank}(\mathbf{X})} \sigma_k(\mathbf{X})$

Theory, software, numerous applications:

Candès and Recht (2009), Recht et al. (2010), Candès and Plan (2010), Candès and Tao (2010), Recht (2011), Keshavan et al. (2010), Agarwal et al. (2012), Klopp (2014), Hastie et al. (2015), Udell et al. (2016)

Extend convex low-rank matrix completion to multi-source, and heterogeneous data simultaneously

Objectives of this thesis

1. Provide *theoretically sound* models adapted to multi-source, heterogeneous and incomplete data *simultaneously*
    - Hybrid low-rank structures
    - Heterogeneous data fitting terms
    - Upper and lower bounds on estimation errors

2. For these models, provide estimation methods and empirically robust software solutions
    - Optimization algorithms
    - Implementation of R packages
    - Numerical results

3. Confront the methods to applications in life sciences
    - Analysis of a waterbird abundance data set
    - Imputation of a medical registry

Objectives of this thesis

1. Provide *theoretically sound* models adapted to multi-source, heterogeneous and incomplete data *simultaneously*
    - Hybrid low-rank structures
    - Heterogeneous data fitting terms
    - Upper and lower bounds on estimation errors

2. For these models, provide estimation methods and empirically robust software solutions
    - Optimization algorithms
    - Implementation of R packages
    - Numerical results

3. Confront the methods to applications in life sciences
    - Analysis of a waterbird abundance data set
    - Imputation of a medical registry

= heterogeneous

Main effects and interactions in mixed and incomplete data frames (MIMI)

Site	2008	2009	2010
1	NA	0	0
2	4	50	25
3	NA	0	0
4	NA	NA	NA
5	NA	NA	NA
6	0	0	0
7	5	75	870
8	9	34	0
9	10	8	30
10	NA	182	27

Data frame $\mathbf{Y}\,(m_1 \times m_2)$

Site	Year	Rain	Eco	Country	Agri
1	2008	163.7	0.8	Algeria	16.2
2	2008	60.7	0.8	Algeria	16.2
3	2008	227.9	0.8	Algeria	16.2
4	2008	174.8	0.8	Algeria	16.2
5	2008	163.7	0.8	Algeria	16.2
6	2008	230.7	0.8	Algeria	16.2
7	2008	243.5	0.8	Algeria	16.2
8	2008	262.6	0.8	Algeria	16.2
9	2008	197.3	0.8	Algeria	16.2
10	2008	227.9	0.8	Algeria	16.2
1	2009	255.1	-1.2	Algeria	16.1
2	2009	179.8	-1.2	Algeria	16.1

Side information $\mathbf{U}\,(m_1 m_2 \times N)$

Statistical model

Data frame: *random* (noisy) observations

$$\mathbf{Y} = \begin{bmatrix} Y_{1,1} & \dots & \color{red}{NA} & \dots & Y_{1,m_2} \\ Y_{2,1} & \color{red}{NA} & & & \\ \vdots & & Y_{i,j} & & \\ Y_{m_1,1} & \color{red}{NA} & & & Y_{m_1,m_2} \end{bmatrix}$$

independent entries with parametric model:

$$f_{Y_{ij}}(y) = f_{ij}(y, X_{ij})$$

probability density function

known function

unknown parameter

Independent

Missing data pattern (*random)*

$$\Omega = \begin{bmatrix} 1 & \dots & \color{red}{0} & \dots & 1 \\ 1 & \color{red}{0} & & & \\ \vdots & & 1 & & \\ 1 & \color{red}{0} & & & 1 \end{bmatrix}$$

independent Bernoulli random variables:

$$\mathbb{P}(\Omega_{i,j} = 1) = \pi_{ij} > 0$$

Exponential family model

$$f_{Y_{ij}}(y) = \underbrace{h_j(y)}_{\text{base function: } \mathcal{Y}_j \to \mathbb{R}_+} \exp(yX_{ij} - \underbrace{g_j(X_{ij})}_{\text{link function: } \mathbb{R} \to \mathcal{X}_j}))$$

Exponential family model

$$f_{Y_{ij}}(y) = \underbrace{h_j(y)}_{\text{base function: } \mathcal{Y}_j \to \mathbb{R}_+} \exp(yX_{ij} - \underbrace{g_j(X_{ij})}_{\text{link function: } \mathbb{R} \to \mathcal{X}_j})$$

Example 1:
(numeric variables)

$$\left.\begin{aligned} h_j(y) &= (2\pi\sigma^2)^{-1/2} \exp(-y^2/2\sigma^2) \\ g_j(x) &= x^2\sigma^2/2 \end{aligned}\right\} \begin{array}{c} \mathcal{N}(x,\sigma^2) \\ \text{(Gaussian)} \end{array}$$

Exponential family model

$$f_{Y_{ij}}(y) = \underbrace{h_j(y)}_{\text{base function: } \mathcal{Y}_j \to \mathbb{R}_+} \exp(yX_{ij} - \underbrace{g_j(X_{ij})}_{\text{link function: } \mathbb{R} \to \mathcal{X}_j})$$

Example 2:
(binary variables)

$$\left.\begin{aligned} h_j(y) &= 1 \\ g_j(x) &= \log(1 + \exp(x)) \end{aligned}\right\} \quad \mathcal{B}(1/(1+\exp(-x)))$$

(Bernoulli)

Exponential family model

$$f_{Y_{ij}}(y) = \underbrace{h_j(y)}_{} \exp(yX_{ij} - \underbrace{g_j(X_{ij})}_{})$$

base function: $\mathcal{Y}_j \to \mathbb{R}_+$

link function: $\mathbb{R} \to \mathcal{X}_j$

Example 3:
(discrete variables)

$$\left.\begin{aligned} h_j(y) &= 1/y! \\ g_j(x) &= \exp(ax) \end{aligned}\right\} \mathcal{P}(\exp(ax))$$

(Poisson)

Log-likelihood & side information

$$\mathcal{L}(\boldsymbol{X};\boldsymbol{Y},\Omega) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\Omega_{i,j}(-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + g_j(\boldsymbol{X}_{i,j}))$$

Parameter of parametric model:
side information included in parameter space

Log-likelihood & side information

$$\mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \Omega_{i,j}(-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + g_j(\boldsymbol{X}_{i,j}))$$

Log-likelihood & side information

$$\mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \Omega_{i,j}(-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + g_j(\boldsymbol{X}_{i,j}))$$

Sparse main effects and low-rank interactions:

$$\boldsymbol{X}_{i,j} = \langle u_{ij}, \alpha \rangle + \boldsymbol{\Theta}_{i,j} \qquad \boldsymbol{X} = \sum_{k=1}^{N} \alpha_k \boldsymbol{U}^k + \boldsymbol{\Theta}$$

Log-likelihood & side information

$$\mathcal{L}(\boldsymbol{X};\boldsymbol{Y},\Omega) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\Omega_{i,j}(-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + g_j(\boldsymbol{X}_{i,j}))$$

Sparse main effects and low-rank interactions:

covariates

$$\boldsymbol{X}_{i,j} = \langle \boxed{u_{ij}}, \alpha\rangle + \boldsymbol{\Theta}_{i,j}$$

main effects
of covariates

$$\boldsymbol{X} = \sum_{k=1}^{N}\alpha_k \boldsymbol{U}^k + \boldsymbol{\Theta}$$

Log-likelihood & side information

$$\mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \Omega_{i,j}(-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + g_j(\boldsymbol{X}_{i,j}))$$

Sparse main effects and low-rank interactions:

$$\boldsymbol{X}_{i,j} = \langle u_{ij}, \alpha \rangle + \boldsymbol{\Theta}_{i,j}$$

interactions
(residuals)

$$\boldsymbol{X} = \sum_{k=1}^{N} \alpha_k \boldsymbol{U}^k + \boldsymbol{\Theta}$$

Log-likelihood & side information

$$\mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \Omega_{i,j}(-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + g_j(\boldsymbol{X}_{i,j}))$$

Sparse main effects and low-rank interactions:

$$\boldsymbol{X}_{i,j} = \langle u_{ij}, \alpha \rangle + \boldsymbol{\Theta}_{i,j}$$

fixed dictionary

$$\boldsymbol{X} = \sum_{k=1}^{N} \alpha_k \boxed{\boldsymbol{U}^k} + \boldsymbol{\Theta}$$

sparse

Log-likelihood & side information

$$\mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \Omega_{i,j}(-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + g_j(\boldsymbol{X}_{i,j}))$$

Sparse main effects and low-rank interactions:

$$\boldsymbol{X}_{i,j} = \langle u_{ij}, \alpha \rangle + \boldsymbol{\Theta}_{i,j} \qquad \boldsymbol{X} = \sum_{k=1}^{N} \alpha_k \boldsymbol{U}^k + \underbrace{\boldsymbol{\Theta}}_{\text{low-rank}}$$

Log-likelihood & side information

$$\mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \Omega_{i,j}(-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + g_j(\boldsymbol{X}_{i,j}))$$

Sparse main effects and low-rank interactions:

$$\boldsymbol{X}_{i,j} = \langle u_{ij}, \alpha \rangle + \cancel{\quad}_{i,j} \qquad\qquad \boldsymbol{X} = \sum_{k=1}^{N} \alpha_k \boldsymbol{U}^k + \cancel{\quad}$$

1/ Only main effects: (Sparse) Generalized Linear Model (GLM)

[Friedman et al. (2010), Pannekoek and van Strien (2001)]

Log-likelihood & side information

$$\mathcal{L}(\boldsymbol{X};\boldsymbol{Y},\Omega) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\Omega_{i,j}(-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + g_j(\boldsymbol{X}_{i,j}))$$

Sparse main effects and low-rank interactions:

$$\boldsymbol{X}_{i,j} = \cancel{\langle u, \alpha\rangle} + \boldsymbol{\Theta}_{i,j} \qquad \boldsymbol{X} = \cancel{\sum_{k=1}^{N}\alpha_k\boldsymbol{U}^k} + \boldsymbol{\Theta}$$

2/ Only interactions: Convex low-rank matrix completion

[Candès and Recht (2008), Agarwal et al. (2011), Klopp (2014), Lafond (2015),Udell et al. (2016), Kumar and Schneider (2017)]

Log-likelihood & side information

$$\mathcal{L}(\boldsymbol{X};\boldsymbol{Y},\Omega) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \Omega_{i,j}(-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + g_j(\boldsymbol{X}_{i,j}))$$

Sparse main effects and low-rank interactions:

$$\boldsymbol{X}_{i,j} = \langle u_{ij}, \alpha\rangle + \boldsymbol{\Theta}_{i,j} \qquad \boldsymbol{X} = \sum_{k=1}^{N} \alpha_k \boldsymbol{U}^k + \boldsymbol{\Theta}$$

Low-rank plus sparse decomposition:

$$\begin{array}{ll} (\hat{\alpha}, \hat{\boldsymbol{\Theta}}) \in & \operatorname{argmin} \mathcal{L}(\alpha, \boldsymbol{\Theta}; \boldsymbol{Y}, \Omega) + \lambda_1 \|\boldsymbol{\Theta}\|_\star + \lambda_2 \|\alpha\|_1 \\ \text{subject to} & \|\alpha\|_\infty \leq a, \|\boldsymbol{\Theta}\|_\infty \leq a, \end{array}$$

Low-rank plus sparse matrix decomposition

$$\boldsymbol{Y} = \boldsymbol{L} + \boldsymbol{S}$$

low-rank

Sparse

Low-rank plus sparse matrix decomposition

$$\boldsymbol{Y} = \boldsymbol{L} + \boldsymbol{S}$$

low-rank

Sparse

1/ No noise

Both components can be recovered exactly via convex optimisation

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{L}\|_\star + \lambda \|\boldsymbol{S}\|_1 \\ \text{subject to} & \boldsymbol{L}_{i,j} + \boldsymbol{S}_{i,j} = \boldsymbol{Y}_{i,j} \text{ if } \Omega_{i,j} = 1 \end{array}$$

Chandrasekaran et al. (2011), Hsu et al. (2011), Candès et al. (2011), Xu et al. (2010), Mardani et al. (2013)

Low-rank plus sparse matrix decomposition

$$\boldsymbol{Y} = \boldsymbol{L} + \boldsymbol{S} + \mathcal{E}$$

Additive noise

2/ Noisy observations

Both components can be estimated with minimax optimal error

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \Omega_{i,j} (\boldsymbol{Y}_{i,j} - \boldsymbol{L}_{i,j} - \boldsymbol{S}_{i,j})^2 + \lambda_1 \|\boldsymbol{L}\|_\star + \lambda_2 \|\boldsymbol{S}\|_1 \\ \text{subject to} & \|\boldsymbol{L}\|_\infty \leq a, \ \|\boldsymbol{S}\|_\infty \leq a \end{array}$$

[Agarwal et al. (2012), Klopp et al. (2017)]

Low-rank plus sparse matrix decomposition

$$\boldsymbol{Y} = \boldsymbol{L} + \boldsymbol{S} + \mathcal{E}$$

Additive noise

2/ Noisy observations

Both components can be estimated with minimax optimal error

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\Omega_{i,j}(\boldsymbol{Y}_{i,j} - \boldsymbol{L}_{i,j} - \boldsymbol{S}_{i,j})^2 + \lambda_1\|\boldsymbol{L}\|_\star + \lambda_2\|\boldsymbol{S}\|_1 \\ \text{subject to} & \|\boldsymbol{L}\|_\infty \leq a, \ \|\boldsymbol{S}\|_\infty \leq a \end{array}$$

[Agarwal et al. (2012), Klopp et al. (2017)]

Two-fold generalisation:
- **heterogeneous exponential family noise**
- **general sparsity pattern**

Target parameters

Definition:

$$\forall (i,j) \in [\![m_1]\!] \times [\![m_2]\!],\ \boldsymbol{X}^0_{i,j} = \operatorname{argmin}_{x\in\mathbb{R}}\{-\mathbb{E}[\boldsymbol{Y}_{i,j}]x + g_j(x)\}$$

Target parameters

Definition:

$$\forall (i,j) \in [\![m_1]\!] \times [\![m_2]\!],\ \boldsymbol{X}^0_{i,j} = \operatorname{argmin}_{x\in\mathbb{R}}\{-\mathbb{E}[\boldsymbol{Y}_{i,j}]x + g_j(x)\}$$

Decomposition: $$\boldsymbol{X}^0 = \sum_{k=1}^{N} \alpha_k^0 \boldsymbol{U}^k + \boldsymbol{\Theta}^0$$

Target parameters

Definition:

$$\forall (i,j) \in [\![m_1]\!] \times [\![m_2]\!],\ \boldsymbol{X}^0_{i,j} = \operatorname{argmin}_{x\in\mathbb{R}} \{-\mathbb{E}[\boldsymbol{Y}_{i,j}]x + g_j(x)\}$$

Decomposition: $$\boldsymbol{X}^0 = \sum_{k=1}^{N} \alpha_k^0 \boldsymbol{U}^k + \boldsymbol{\Theta}^0$$

Specification: $$s = \min_{\boldsymbol{X}^0 = \sum_{k=1}^N \alpha_k \boldsymbol{U}^k + \boldsymbol{\Theta}} \{\|\alpha\|_0 + \operatorname{rank}(\boldsymbol{\Theta})\}$$

$$(\alpha^0, \boldsymbol{\Theta}^0) \in \operatorname{argmin}_{\substack{\boldsymbol{X}^0 = \sum_{k=1}^N \alpha_k \boldsymbol{U}^k + \boldsymbol{\Theta} \\ \|\alpha\|_0 + \operatorname{rank}(\boldsymbol{\Theta}) = s}} \|\alpha\|_0$$

Main assumptions

Model: $\forall k \in [\![N]\!],\ \alpha_k \neq 0,\ \langle \boldsymbol{U}^k, \boldsymbol{\Theta}^0 \rangle = 0$

$$\|\alpha^0\|_\infty \leq a,\ \|\boldsymbol{\Theta}^0\|_\infty \leq a$$

Main assumptions

Model: $\forall k \in [\![N]\!],\ \alpha_k \neq 0,\ \langle \boldsymbol{U}^k, \boldsymbol{\Theta}^0 \rangle = 0$

$$\|\alpha^0\|_\infty \leq a,\ \|\boldsymbol{\Theta}^0\|_\infty \leq a$$

Missing values: $\forall (i,j) \in [\![m_1]\!] \times [\![m_2]\!],\ c_1 p \leq \pi_{i,j} \leq c_2 p,\ p > 0$

Main assumptions

Model: $\forall k \in [\![N]\!],\ \alpha_k \neq 0,\ \langle \boldsymbol{U}^k, \boldsymbol{\Theta}^0 \rangle = 0$

$$\|\alpha^0\|_\infty \leq a,\ \|\boldsymbol{\Theta}^0\|_\infty \leq a$$

Missing values: $\forall (i,j) \in [\![m_1]\!] \times [\![m_2]\!],\ c_1 p \leq \pi_{i,j} \leq c_2 p,\ p > 0$

Noise: $\forall j \in [\![m_2]\!],\ g_j$ is $\mathcal{C}^2$

$$\forall x \in \mathbb{R},\ |x| < (1+æ)a, \forall j \in [\![m_2]\!],\ \sigma_-^2 \leq g_j''(x) \leq \sigma_+^2$$

$$\forall z \in \mathbb{R},\ |z| < \gamma,\ \mathbb{E}[\mathrm{e}^{z(\mathbf{Y}_{i,j} - \mathbb{E}[\mathbf{Y}_{i,j}])}] \leq \mathrm{e}^{\sigma^2 z^2/2}$$

Main assumptions

Model: $\forall k \in [\![N]\!],\ \alpha_k \neq 0,\ \langle \boldsymbol{U}^k, \boldsymbol{\Theta}^0 \rangle = 0$

$$\|\alpha^0\|_\infty \leq a,\ \|\boldsymbol{\Theta}^0\|_\infty \leq a$$

Missing values: $\forall (i,j) \in [\![m_1]\!] \times [\![m_2]\!],\ c_1 p \leq \pi_{i,j} \leq c_2 p,\ p > 0$

Noise: $\forall j \in [\![m_2]\!],\ g_j \text{ is } \mathcal{C}^2$

$$\forall x \in \mathbb{R},\ |x| < (1+æ)a, \forall j \in [\![m_2]\!],\ \sigma_-^2 \leq g_j''(x) \leq \sigma_+^2$$

$$\forall z \in \mathbb{R},\ |z| < \gamma,\ \mathbb{E}[\mathrm{e}^{z(\mathbf{Y}_{i,j} - \mathbb{E}[\mathbf{Y}_{i,j}])}] \leq \mathrm{e}^{\sigma^2 z^2/2}$$

Dictionary: $\forall k \in [\![N]\!],\ \|\boldsymbol{U}^k\|_\infty \leq 1$

$$\forall\ (i,j) \in [\![m_1]\!] \times [\![m_2]\!], \sum_{k=1}^{N} |\boldsymbol{U}_{i,j}^k| \leq æ$$

$$\forall \alpha \in \mathbb{R}^N, \alpha^\top G \alpha \geq \kappa^2 \|\alpha\|_2^2,\ \text{where } G \text{ is the Gram matrix of } (U^1, \ldots, U^N)$$

Statistical guarantees

Theorem (Robin et al. 2019)

Set: $\lambda_1 = 2c^* \sigma_+ \sqrt{p m_1 \vee m_2 \log(m_1 + m_2)}, \quad \lambda_2 \geq 24 \max_k \|\boldsymbol{U}^k\|_1 \log(m_1 + m_2)/\gamma$

Assume: $m_1 \vee m_2 \geq \max\{4\sigma_+^2/\gamma^6 \log^2(\sqrt{m_1 \wedge m_2}), 2\exp(\sigma_+^2/\gamma^2 \wedge \sigma_+^2 \gamma(1 + \text{æ}a))\}$

Statistical guarantees

Theorem (Robin et al. 2019)

Set: $\lambda_1 = 2c^*\sigma_+\sqrt{pm_1 \vee m_2 \log(m_1+m_2)}, \quad \lambda_2 \geq 24 \max_k \|\boldsymbol{U}^k\|_1 \log(m_1+m_2)/\gamma$

Assume: $m_1 \vee m_2 \geq \max\{4\sigma_+^2/\gamma^6 \log^2(\sqrt{m_1 \wedge m_2}), 2\exp(\sigma_+^2/\gamma^2 \wedge \sigma_+^2\gamma(1+æa))\}$

Then, with probability at least $1 - 10(m_1+m_2)^{-1}$:

$$\|\alpha^0 - \hat{\alpha}\|_2^2 \lesssim \frac{\|\alpha^0\|_0}{p} \times \frac{\max_k \|U^k\|_1}{\kappa^2}$$

$$\|\boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}}\|_F^2 \lesssim \frac{\mathrm{rank}(\boldsymbol{\Theta}^0)(m_1 \vee m_2)}{p} + \frac{\|\alpha^0\|_0 \max_k \|\boldsymbol{U}^k\|_1}{p}$$

Statistical guarantees

$$\|\alpha^0 - \hat{\alpha}\|_2^2 \lesssim \frac{\|\alpha^0\|_0}{p} \times \frac{\max_k \|U^k\|_1}{\kappa^2}$$

$$\|\mathbf{\Theta}^0 - \hat{\mathbf{\Theta}}\|_F^2 \lesssim \frac{\operatorname{rank}(\mathbf{\Theta}^0)(m_1 \vee m_2)}{p} + \frac{\|\alpha^0\|_0 \max_k \|\boldsymbol{U}^k\|_1}{p}$$

Statistical guarantees

$$\|\alpha^0 - \hat{\alpha}\|_2^2 \lesssim \frac{\|\alpha^0\|_0}{p} \times \frac{\max_k \|U^k\|_1}{\kappa^2}$$

$$\|\mathbf{\Theta}^0 - \hat{\mathbf{\Theta}}\|_F^2 \lesssim \frac{\mathrm{rank}(\mathbf{\Theta}^0)(m_1 \vee m_2)}{p} + \frac{\|\alpha^0\|_0 \max_k \|\boldsymbol{U}^k\|_1}{p}$$

Usual low-rank
matrix completion
rate

Statistical guarantees

$$\|\alpha^0 - \hat{\alpha}\|_2^2 \lesssim \frac{\|\alpha^0\|_0}{p} \times \frac{\max_k \|U^k\|_1}{\kappa^2}$$

$$\|\boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}}\|_F^2 \lesssim \frac{\mathrm{rank}(\boldsymbol{\Theta}^0)(m_1 \vee m_2)}{p} + \frac{\|\alpha^0\|_0 \max_k \|\boldsymbol{U}^k\|_1}{p}$$

Usual low-rank matrix completion rate

Interplay with main effects

Statistical guarantees

$$\|\alpha^0 - \hat{\alpha}\|_2^2 \lesssim \frac{\|\alpha^0\|_0}{p} \times \frac{\max_k \|U^k\|_1}{\kappa^2}$$

$$\|\mathbf{\Theta}^0 - \hat{\mathbf{\Theta}}\|_F^2 \lesssim \frac{\mathrm{rank}(\mathbf{\Theta}^0)(m_1 \vee m_2)}{p} + \frac{\|\alpha^0\|_0 \max_k \|\boldsymbol{U}^k\|_1}{p}$$

Statistical guarantees

$$\|\alpha^0 - \hat{\alpha}\|_2^2 \lesssim \frac{\|\alpha^0\|_0}{p} \times \frac{\max_k \|U^k\|_1}{\kappa^2}$$

Usual sparse rate
in low-rank + sparse

$$\|\boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}}\|_F^2 \lesssim \frac{\operatorname{rank}(\boldsymbol{\Theta}^0)(m_1 \vee m_2)}{p} + \frac{\|\alpha^0\|_0 \max_k \|\boldsymbol{U}^k\|_1}{p}$$

Statistical guarantees

Effect of dictionary

$$\|\alpha^0 - \hat{\alpha}\|_2^2 \lesssim \frac{\|\alpha^0\|_0}{p} \times \frac{\max_k \|U^k\|_1}{\kappa^2}$$

Usual sparse rate
in low-rank + sparse

$$\|\mathbf{\Theta}^0 - \hat{\mathbf{\Theta}}\|_F^2 \lesssim \frac{\mathrm{rank}(\mathbf{\Theta}^0)(m_1 \vee m_2)}{p} + \frac{\|\alpha^0\|_0 \max_k \|\boldsymbol{U}^k\|_1}{p}$$

Objectives of this thesis

1. Provide *theoretically sound* models adapted to multi-source, heterogeneous and incomplete data *simultaneously*
    - Hybrid low-rank structures
    - Heterogeneous data fitting terms
    - Upper and lower bounds on estimation errors

2. For these models, provide estimation methods and empirically robust software solutions
    - Optimization algorithms
    - Implementation of R packages
    - Numerical results

3. Confront the methods to applications in life sciences
    - Analysis of a waterbird abundance data set
    - Imputation of a medical registry

Optimization problem

$$\begin{array}{ll} (\hat{\alpha}, \hat{\boldsymbol{\Theta}}) \in & \operatorname{argmin}_{(\alpha, \boldsymbol{\Theta})} \mathcal{L}(\mathsf{f}_U(\alpha) + \boldsymbol{\Theta}; \boldsymbol{Y}, \Omega) + \lambda_1 \|\boldsymbol{\Theta}\|_* + \lambda_2 \|\alpha\|_1 \\ \text{subject to} & \|\alpha\|_\infty \leq a, \|\boldsymbol{\Theta}\|_\infty \leq a, \end{array}$$

Optimization problem

$$
\begin{aligned}
(\hat{\alpha}, \hat{\mathbf{\Theta}}) \in \quad & \operatorname{argmin}_{(\alpha, \mathbf{\Theta})} \mathcal{L}(\mathsf{f}_U(\alpha) + \mathbf{\Theta}; \boldsymbol{Y}, \Omega) + \lambda_1 \|\mathbf{\Theta}\|_* + \lambda_2 \|\alpha\|_1 \\
\text{subject to} \quad & \|\alpha\|_\infty \leq a, \|\mathbf{\Theta}\|_\infty \leq a,
\end{aligned}
$$

Optimization problem

$$(\hat{\alpha}, \hat{\mathbf{\Theta}}) \in \quad \operatorname{argmin}_{(\alpha, \mathbf{\Theta})} \mathcal{L}(\mathsf{f}_U(\alpha) + \mathbf{\Theta}; \boldsymbol{Y}, \Omega) + \lambda_1 \|\mathbf{\Theta}\|_* + \lambda_2 \|\alpha\|_1$$

$$\text{subject to} \quad \|\alpha\|_\infty \leq a, \|\mathbf{\Theta}\|_\infty \leq a,$$

Drop the constraint

$$(\hat{\alpha}, \hat{\mathbf{\Theta}}) \in \operatorname{argmin} \mathcal{L}(\mathsf{f}_U(\alpha) + \mathbf{\Theta}; \boldsymbol{Y}, \Omega) + \lambda_1 \|\mathbf{\Theta}\|_* + \lambda_2 \|\alpha\|_1$$

Optimization problem

$$(\hat{\alpha}, \hat{\boldsymbol{\Theta}}) \in \quad \operatorname{argmin}_{(\alpha,\boldsymbol{\Theta})} \mathcal{L}(\mathsf{f}_U(\alpha) + \boldsymbol{\Theta}; \boldsymbol{Y}, \Omega) + \lambda_1 \|\boldsymbol{\Theta}\|_* + \lambda_2 \|\alpha\|_1$$
$$\text{subject to} \quad \|\alpha\|_\infty \le a, \|\boldsymbol{\Theta}\|_\infty \le a,$$

Drop the constraint

$$(\hat{\alpha}, \hat{\boldsymbol{\Theta}}) \in \operatorname{argmin} \overbrace{\mathcal{L}(\mathsf{f}_U(\alpha) + \boldsymbol{\Theta}; \boldsymbol{Y}, \Omega)}^{\text{smooth}} + \overbrace{\lambda_1 \|\boldsymbol{\Theta}\|_* + \lambda_2 \|\alpha\|_1}^{\text{separable}}$$

Optimization problem

$$(\hat{\alpha}, \hat{\mathbf{\Theta}}) \in \quad \mathrm{argmin}_{(\alpha, \mathbf{\Theta})} \, \mathcal{L}(\mathsf{f}_U(\alpha) + \mathbf{\Theta}; \boldsymbol{Y}, \Omega) + \lambda_1 \|\mathbf{\Theta}\|_* + \lambda_2 \|\alpha\|_1$$

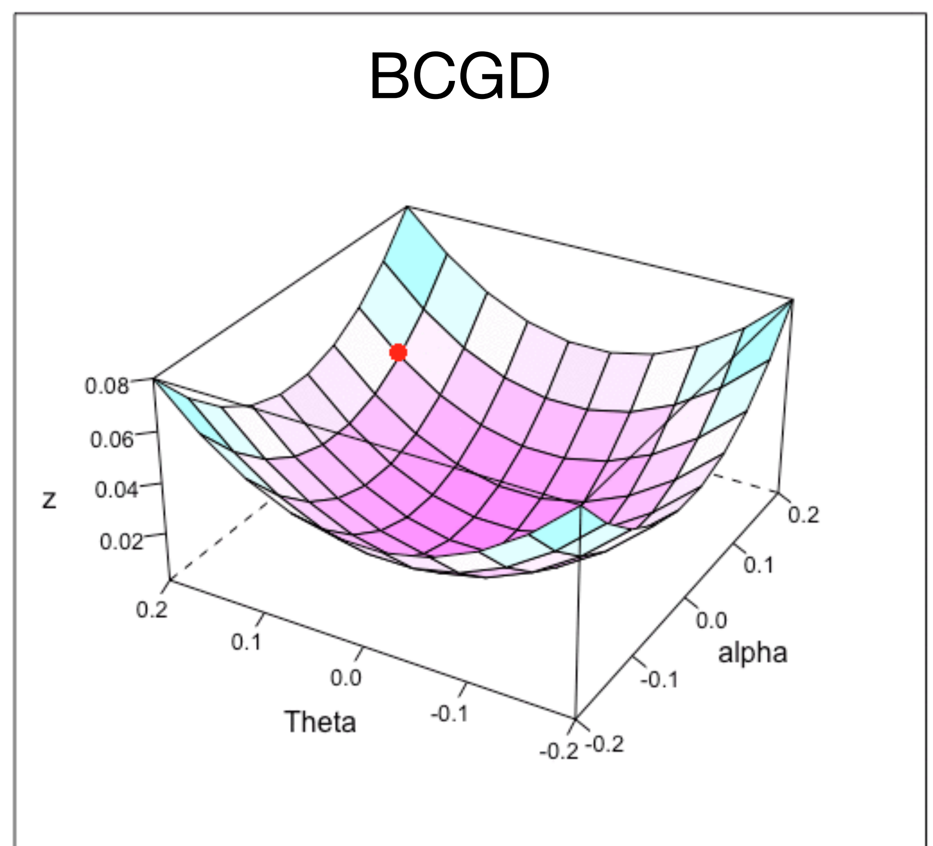
$$\text{subject to} \quad \|\alpha\|_\infty \leq a, \|\mathbf{\Theta}\|_\infty \leq a,$$

Drop the constraint

$$(\hat{\alpha}, \hat{\mathbf{\Theta}}) \in \mathrm{argmin} \, \overbrace{\mathcal{L}(\mathsf{f}_U(\alpha) + \mathbf{\Theta}; \boldsymbol{Y}, \Omega)}^{\text{smooth}} + \overbrace{\lambda_1 \|\mathbf{\Theta}\|_* + \lambda_2 \|\alpha\|_1}^{\text{separable}}$$

Algorithm:
Block coordinate gradient descent (BCGD)

Idea:
Update the parameters α and $\mathbf{\Theta}$ alternatively along descent directions



Sketch of the algorithm

Algorithm 1 BCGD algorithm.

Initialize: — $\alpha^{(0)}, \mathbf{\Theta}^{(0)}$. E.g., $(\alpha^{(0)}, \mathbf{\Theta}^{(0)}) = (0, \mathbf{0})$.
for $t = 1, 2, \ldots, T$ **do**
    // Compute quadratic approximation //
    Taylor expansion with additional strongly convex quadratic term
    // Update for α *//*
    Compute descent direction (weighted LASSO problem)
    Perform Armijo line search to compute the step size
    // Update for $\mathbf{\Theta}$ *//*
    Compute descent direction (weighted softImpute problem)
    Perform Armijo line search to compute the step size
end for
Return: $\alpha^{[T]}, \mathbf{\Theta}^{[T]}$

Quadratic approximation

of $f(\alpha, \boldsymbol{\Theta}) = \mathcal{L}(\mathsf{f}_U(\alpha) + \boldsymbol{\Theta}; \boldsymbol{Y}, \Omega)$ around $(\alpha, \boldsymbol{\Theta})$

$$f(\alpha + d_\alpha, \boldsymbol{\Theta} + d_{\boldsymbol{\Theta}}) = f(\alpha, \boldsymbol{\Theta}) + \overbrace{\mathcal{A}(\mathsf{f}_U(\alpha) + \boldsymbol{\Theta}, d_\alpha, d_{\boldsymbol{\Theta}})}^{\text{Taylor expansion + quadratic term}}$$

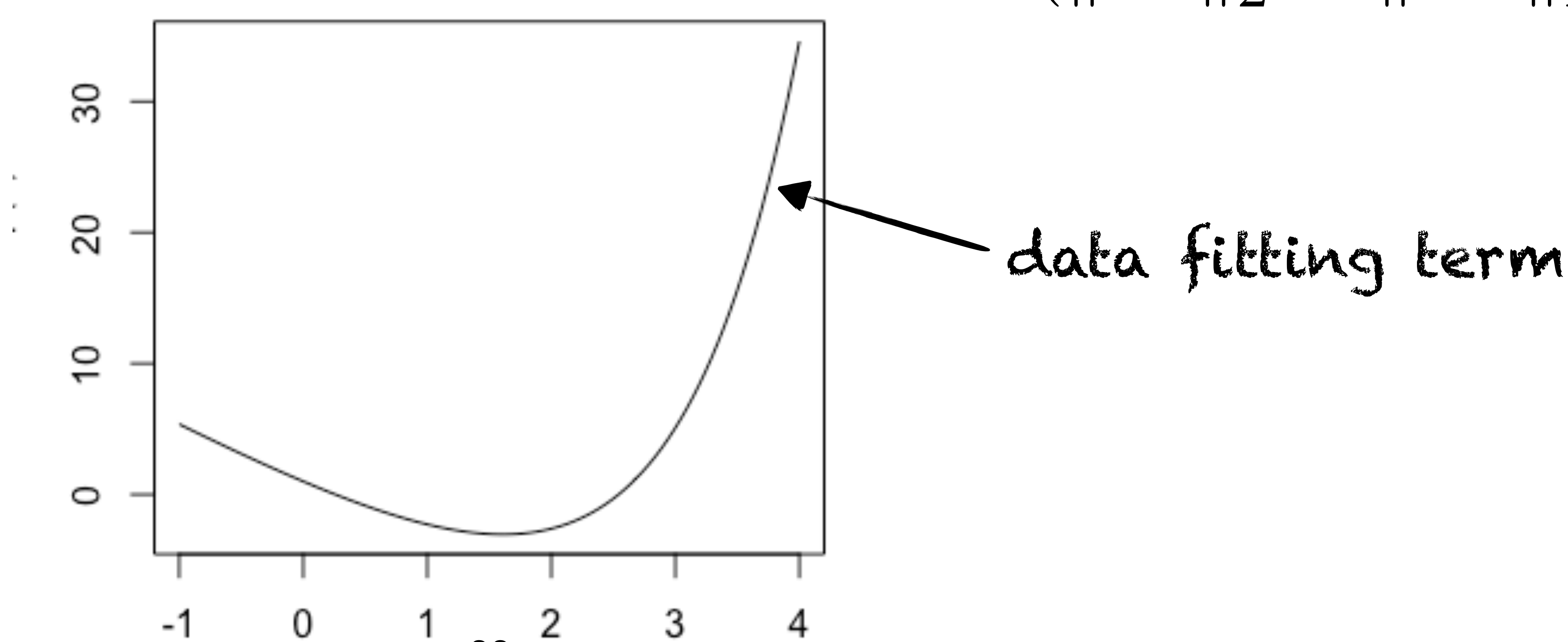
$$+ \underbrace{o(\|d_\alpha\|_2^2 + \|d_{\boldsymbol{\Theta}}\|_F^2)}_{\text{residual}}$$

Quadratic approximation

of $f(\alpha, \boldsymbol{\Theta}) = \mathcal{L}(\mathsf{f}_U(\alpha) + \boldsymbol{\Theta}; \boldsymbol{Y}, \Omega)$ **around** $(\alpha, \boldsymbol{\Theta})$

$$f(\alpha + d_\alpha, \boldsymbol{\Theta} + d_{\boldsymbol{\Theta}}) = f(\alpha, \boldsymbol{\Theta}) + \mathcal{A}(\mathsf{f}_U(\alpha) + \boldsymbol{\Theta}, d_\alpha, d_{\boldsymbol{\Theta}})$$

$$+ o(\|d_\alpha\|_2^2 + \|d_{\boldsymbol{\Theta}}\|_F^2)$$

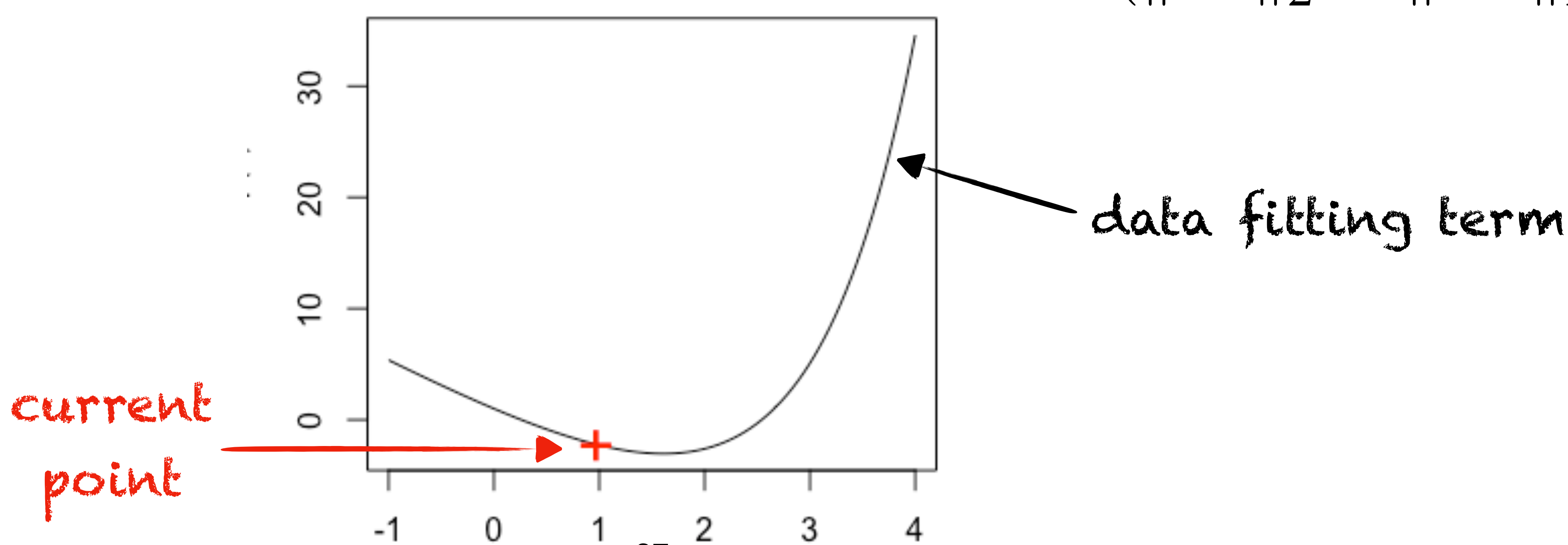


Quadratic approximation

of $f(\alpha, \boldsymbol{\Theta}) = \mathcal{L}(\mathsf{f}_U(\alpha) + \boldsymbol{\Theta}; \boldsymbol{Y}, \Omega)$ **around** $(\alpha, \boldsymbol{\Theta})$

$$f(\alpha + d_\alpha, \boldsymbol{\Theta} + d_{\boldsymbol{\Theta}}) = f(\alpha, \boldsymbol{\Theta}) + \mathcal{A}(\mathsf{f}_U(\alpha) + \boldsymbol{\Theta}, d_\alpha, d_{\boldsymbol{\Theta}})$$

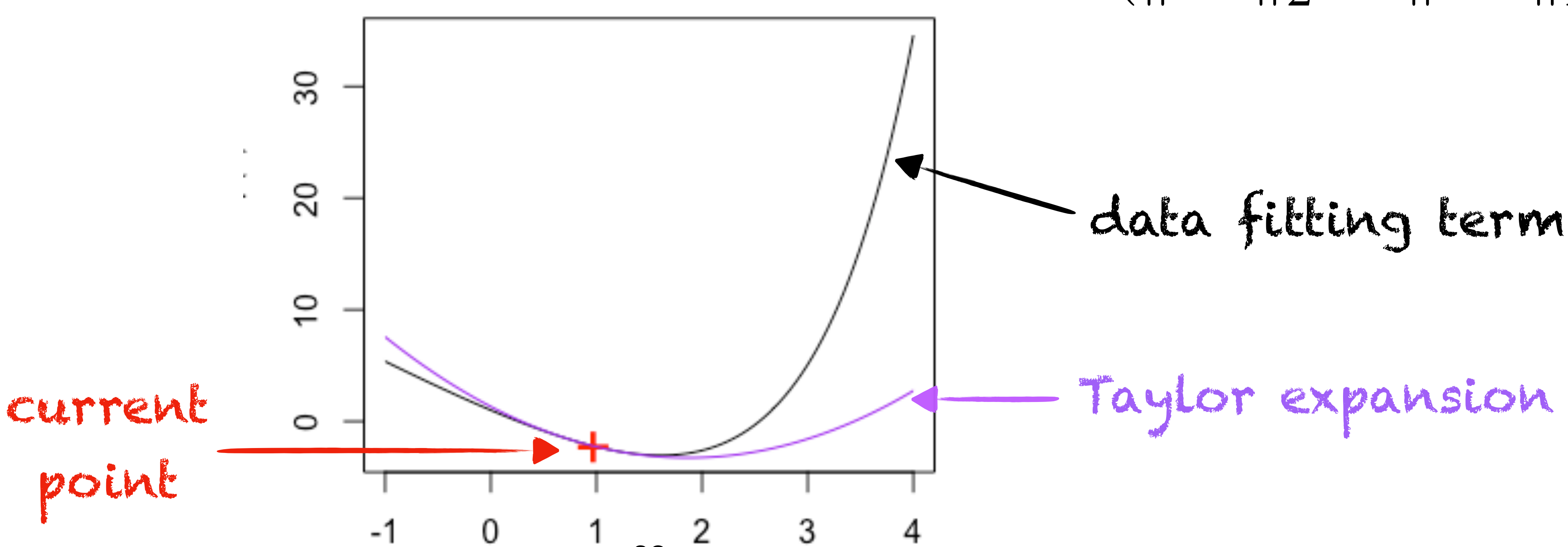
$$+o(\|d_\alpha\|_2^2 + \|d_{\boldsymbol{\Theta}}\|_F^2)$$



Quadratic approximation

of $f(\alpha, \mathbf{\Theta}) = \mathcal{L}(\mathsf{f}_U(\alpha) + \mathbf{\Theta}; \boldsymbol{Y}, \Omega)$ **around** $(\alpha, \mathbf{\Theta})$

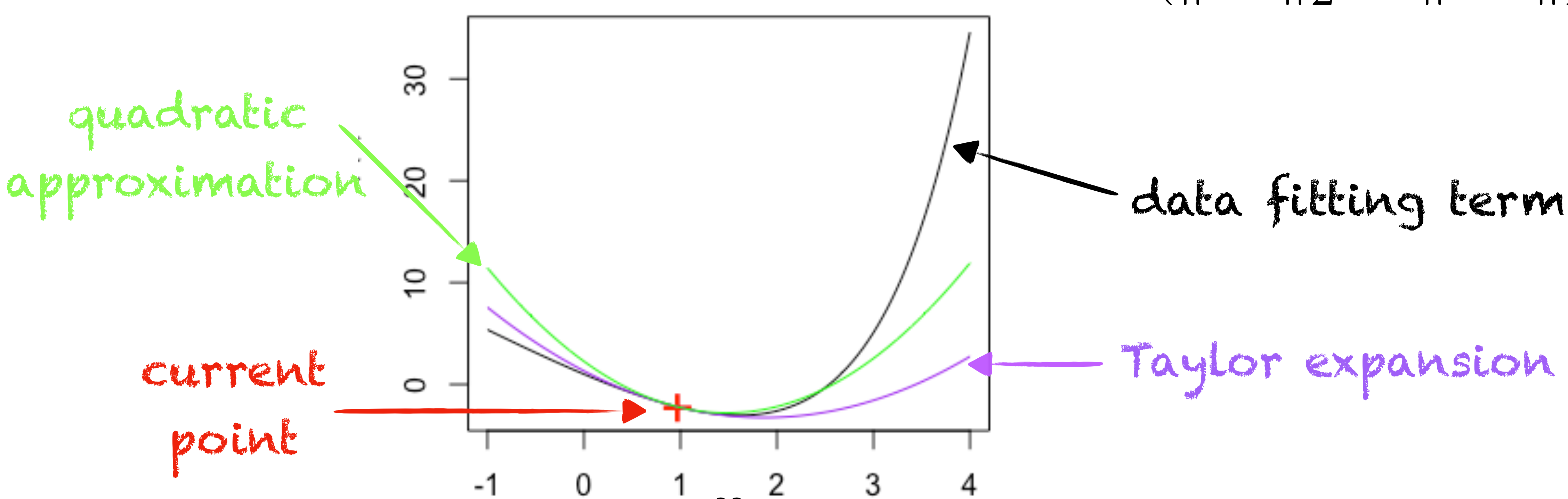
$$f(\alpha + d_\alpha, \mathbf{\Theta} + d_{\mathbf{\Theta}}) = f(\alpha, \mathbf{\Theta}) + \mathcal{A}(\mathsf{f}_U(\alpha) + \mathbf{\Theta}, d_\alpha, d_{\mathbf{\Theta}})$$
$$+o(\|d_\alpha\|_2^2 + \|d_{\mathbf{\Theta}}\|_F^2)$$



Quadratic approximation

of $f(\alpha, \mathbf{\Theta}) = \mathcal{L}(\mathsf{f}_U(\alpha) + \mathbf{\Theta}; \boldsymbol{Y}, \Omega)$ **around** $(\alpha, \mathbf{\Theta})$

$$f(\alpha + d_\alpha, \mathbf{\Theta} + d_{\mathbf{\Theta}}) = f(\alpha, \mathbf{\Theta}) + \mathcal{A}(\mathsf{f}_U(\alpha) + \mathbf{\Theta}, d_\alpha, d_{\mathbf{\Theta}})$$
$$+o(\|d_\alpha\|_2^2 + \|d_{\mathbf{\Theta}}\|_F^2)$$

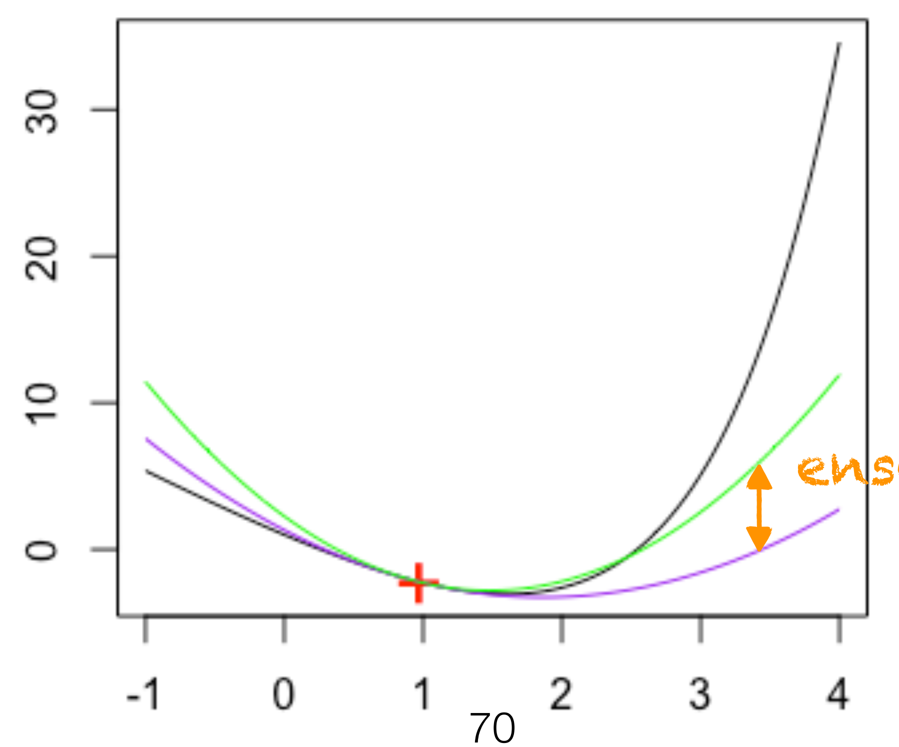


Quadratic approximation

of $f(\alpha, \boldsymbol{\Theta}) = \mathcal{L}(\mathsf{f}_U(\alpha) + \boldsymbol{\Theta}; \boldsymbol{Y}, \Omega)$ **around** $(\alpha, \boldsymbol{\Theta})$

$$f(\alpha + d_\alpha, \boldsymbol{\Theta} + d_{\boldsymbol{\Theta}}) = f(\alpha, \boldsymbol{\Theta}) + \mathcal{A}(\mathsf{f}_U(\alpha) + \boldsymbol{\Theta}, d_\alpha, d_{\boldsymbol{\Theta}})$$

$$+ o(\|d_\alpha\|_2^2 + \|d_{\boldsymbol{\Theta}}\|_F^2)$$



ensures strong convexity
in each iteration

Quadratic approximation

$$\mathcal{A}(X, d_\alpha, d_{\mathbf{\Theta}}) = -2\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}[\boldsymbol{X}_{i,j}]Z_{ij}[\boldsymbol{X}_{i,j}](\mathsf{f}_U(d_\alpha)_{i,j} + d_{\mathbf{\Theta}\,i,j})$$
$$+\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}[\boldsymbol{X}_{i,j}](\mathsf{f}_U(d_\alpha)_{i,j} + d_{\mathbf{\Theta}\,i,j})^2 + \nu\|d_\alpha\|_2^2 + \nu\|d_{\mathbf{\Theta}}\|_F^2.$$

$$w_{ij}[x] = \Omega_{i,j} g_j''(x)/2 \;, \quad Z_{ij}[x] = (\boldsymbol{Y}_{i,j} - g_j'(x))/g_j''(x) \;.$$

Quadratic approximation

$$\mathcal{A}(X, d_\alpha, d_{\boldsymbol{\Theta}}) = -2\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}[\boldsymbol{X}_{i,j}] Z_{ij}[\boldsymbol{X}_{i,j}](\mathsf{f}_U(d_\alpha)_{i,j} + d_{\boldsymbol{\Theta}\,i,j})$$
$$+ \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}[\boldsymbol{X}_{i,j}](\mathsf{f}_U(d_\alpha)_{i,j} + d_{\boldsymbol{\Theta}\,i,j})^2 + \nu\|d_\alpha\|_2^2 + \nu\|d_{\boldsymbol{\Theta}}\|_F^2.$$

$$w_{ij}[x] = \Omega_{i,j} g_j''(x)/2\;, \quad Z_{ij}[x] = (\boldsymbol{Y}_{i,j} - g_j'(x))/g_j''(x)\;.$$

Important point: it is quadratic

Update for α

1/ Search direction: $$d_\alpha^{[t]} \in \operatorname{argmin}_{d \in \mathbb{R}^N} \Big\{ \underbrace{\mathcal{A}(\boldsymbol{X}^{[t]}, d, 0)}_{\text{quadratic term}} + \underbrace{\lambda_2 \|\alpha^{[t]} + d\|_1}_{\ell_1 \text{ penalty}} \Big\} \ .$$

Update for α

1/ Search direction: $d_\alpha^{[t]} \in \operatorname{argmin}_{d \in \mathbb{R}^N} \Big\{ \underbrace{\mathcal{A}(\boldsymbol{X}^{[t]}, d, 0)}_{\text{quadratic term}} + \underbrace{\lambda_2 \|\alpha^{[t]} + d\|_1}_{\ell_1 \text{ penalty}} \Big\}$.

2/ Line search: $\tau_\alpha^{[t]}$ largest element of $\{\tau_{\text{init}} \beta^j\}_{j=0}^{\infty}$ satisfying

$$f(\alpha^{[t]} + \tau_\alpha^{[t]} d^{[t]}, \boldsymbol{\Theta}^{[t]}) + \lambda_2 \|\alpha^{[t]} + \tau_\alpha^{[t]} d^{[t]}\|_1 \leq f(\alpha^{[t]}, \boldsymbol{\Theta}^{[t]}) + \lambda_2 \|\alpha^{[t]}\|_1 + \overbrace{\tau_\alpha^{[t]} \zeta \Gamma_\alpha^{[t]}}^{\text{strict descent}}$$

Update for α

1/ Search direction: $d_\alpha^{[t]} \in \operatorname{argmin}_{d \in \mathbb{R}^N} \left\{ \underbrace{\mathcal{A}(\boldsymbol{X}^{[t]}, d, 0)}_{\text{quadratic term}} + \underbrace{\lambda_2 \|\alpha^{[t]} + d\|_1}_{\ell_1 \text{ penalty}} \right\}.$

2/ Line search: $\tau_\alpha^{[t]}$ largest element of $\{\tau_{\text{init}} \beta^j\}_{j=0}^{\infty}$ satisfying

$$f(\alpha^{[t]} + \tau_\alpha^{[t]} d^{[t]}, \boldsymbol{\Theta}^{[t]}) + \lambda_2 \|\alpha^{[t]} + \tau_\alpha^{[t]} d^{[t]}\|_1 \leq f(\alpha^{[t]}, \boldsymbol{\Theta}^{[t]}) + \lambda_2 \|\alpha^{[t]}\|_1 + \overbrace{\tau_\alpha^{[t]} \zeta \Gamma_\alpha^{[t]}}^{\text{strict descent}}$$

3/ Update: $$\alpha^{[t+1]} = \alpha^{[t]} + \tau_\alpha^{[t]} d_\alpha^{[t]}$$

Update for $\boldsymbol{\Theta}$

1/ Search direction: $$d_{\boldsymbol{\Theta}}^{[t]} := \operatorname{argmin}\left\{\mathcal{A}(\boldsymbol{X}^{[t+1/2]}, 0, d) + \lambda_1 \|\boldsymbol{\Theta}^{[t]} + d\|_*\right\}$$

$$\Leftrightarrow \quad \operatorname{argmin}_{\boldsymbol{\Theta} \in \mathbb{R}^{m_1 \times m_2}} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (\nu + \underbrace{w_{ij}[\boldsymbol{X}_{i,j}^{[t+1/2]}])(Z_{ij}^{[t+1/2]} - \boldsymbol{\Theta}_{i,j}}_{\text{weighted norm (positive weights)}})^2 + \underbrace{\lambda_1 \|\boldsymbol{\Theta}\|_*}_{\text{nuclear norm penalty}}$$

weighted version of softImpute (Hastie et al. 2015) $\Longrightarrow$ iterative SVD

Update for $\boldsymbol{\Theta}$

1/ Search direction: $d_{\boldsymbol{\Theta}}^{[t]} := \operatorname{argmin}\left\{\mathcal{A}(\boldsymbol{X}^{[t+1/2]}, 0, d) + \lambda_1 \|\boldsymbol{\Theta}^{[t]} + d\|_*\right\}$

$$\Leftrightarrow \quad \operatorname{argmin}_{\boldsymbol{\Theta}\in\mathbb{R}^{m_1\times m_2}} \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \underbrace{(\nu + w_{ij}[\boldsymbol{X}_{i,j}^{[t+1/2]}])(Z_{ij}^{[t+1/2]} - \boldsymbol{\Theta}_{i,j})^2}_{\text{weighted norm (positive weights)}} + \underbrace{\lambda_1 \|\boldsymbol{\Theta}\|_*}_{\text{nuclear norm penalty}}$$

weighted version of softImpute (Hastie et al. 2015) $\Longrightarrow$ iterative SVD

2/ Line search: $\tau_{\boldsymbol{\Theta}}^{[t]}$ largest element of $\{\tau_{\text{init}}\beta^j\}_{j=0}^{\infty}$ satisfying

$$\begin{aligned} f(\alpha^{[t+1]}, \boldsymbol{\Theta}^{[t]} + \tau_L^{[t]} d_{\boldsymbol{\Theta}}^{[t]}) + \lambda_1 \|\boldsymbol{\Theta}^{[t]} + \tau_{\boldsymbol{\Theta}}^{[t]} d_{\boldsymbol{\Theta}}^{[t]}\|_* \\ \leq f(\alpha^{[t+1]}, \boldsymbol{\Theta}^{[t]}) + \lambda_1 \|\boldsymbol{\Theta}^{[t]}\|_* + \tau_{\boldsymbol{\Theta}}^{[t]} \zeta \Gamma_{\boldsymbol{\Theta}}^{[t]} \end{aligned}$$

Update for $\boldsymbol{\Theta}$

1/ Search direction: $d_{\boldsymbol{\Theta}}^{[t]} := \operatorname{argmin} \left\{ \mathcal{A}(\boldsymbol{X}^{[t+1/2]}, 0, d) + \lambda_1 \|\boldsymbol{\Theta}^{[t]} + d\|_* \right\}$

$$\Leftrightarrow \quad \operatorname{argmin}_{\boldsymbol{\Theta} \in \mathbb{R}^{m_1 \times m_2}} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \underbrace{(\nu + w_{ij}[\boldsymbol{X}_{i,j}^{[t+1/2]}])(Z_{ij}^{[t+1/2]} - \boldsymbol{\Theta}_{i,j})^2}_{\text{weighted norm (positive weights)}} + \underbrace{\lambda_1 \|\boldsymbol{\Theta}\|_*}_{\text{nuclear norm penalty}}$$

weighted version of softImpute (Hastie et al. 2015) $\Rightarrow$ iterative SVD

2/ Line search: $\tau_{\boldsymbol{\Theta}}^{[t]}$ largest element of $\{\tau_{\text{init}}\beta^j\}_{j=0}^{\infty}$ satisfying

$$f(\alpha^{[t+1]}, \boldsymbol{\Theta}^{[t]} + \tau_L^{[t]} d_{\boldsymbol{\Theta}}^{[t]}) + \lambda_1 \|\boldsymbol{\Theta}^{[t]} + \tau_{\boldsymbol{\Theta}}^{[t]} d_{\boldsymbol{\Theta}}^{[t]}\|_*$$
$$\leq f(\alpha^{[t+1]}, \boldsymbol{\Theta}^{[t]}) + \lambda_1 \|\boldsymbol{\Theta}^{[t]}\|_* + \tau_{\boldsymbol{\Theta}}^{[t]} \zeta \Gamma_{\boldsymbol{\Theta}}^{[t]}$$

3/ Update: $\boldsymbol{\Theta}^{[t+1]} = \boldsymbol{\Theta}^{[t]} + \tau_{\boldsymbol{\Theta}}^{[t]} d_{\boldsymbol{\Theta}}^{[t]}$

Convergence of BCGD algorithm

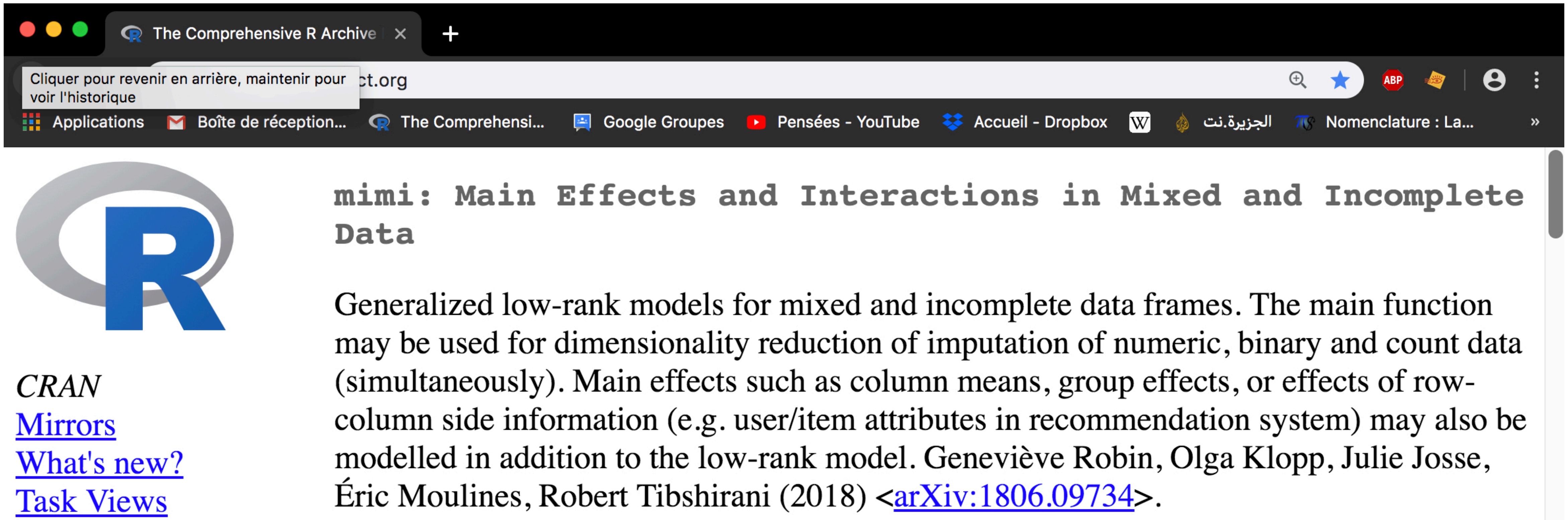
$$\mathcal{F}(\alpha, \boldsymbol{\Theta}) = \mathcal{L}(\mathsf{f}_U(\alpha) + \boldsymbol{\Theta}; \boldsymbol{Y}, \Omega) + \lambda_1 \|\boldsymbol{\Theta}\|_* + \lambda_2 \|\alpha\|_1$$

Theorem (Robin et al. 2019)
Under the assumptions previously stated:

$$\mathcal{F}(\alpha^{[t]}, \boldsymbol{\Theta}^{[t]}) \to \mathcal{F}(\hat{\alpha}, \hat{\boldsymbol{\Theta}})$$

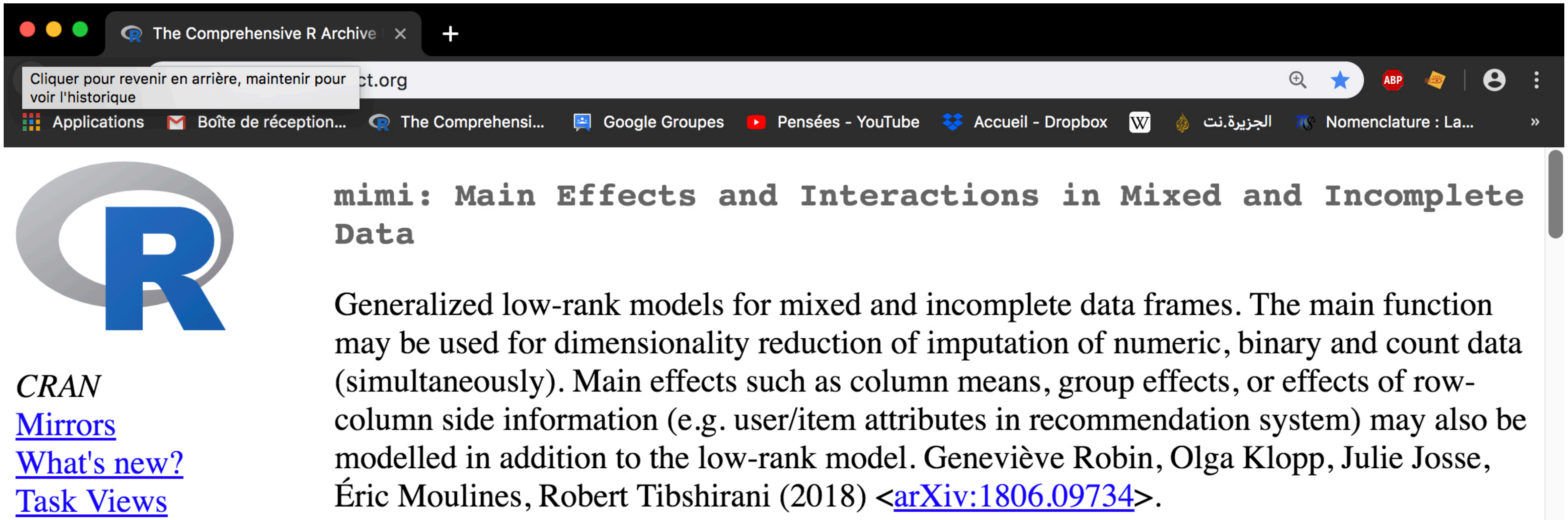
Proof: Tseng and Yun (2009) + compact level sets

R package mimi




```r
install.packages("mimi")
library(mimi)
data <- read.table("mydatafile.txt")
var.type <- c(rep("gaussian", 15), rep("binomial", 10))
model <- "low-rank"
rescv <- cv.mimi(y, model=model, var.type=var.type)
res <- mimi(y, model=model, var.type=var.type, lambda1=rescv$lambda,
            algo="bcgd")
```

R package mimi


The Comprehensive R Archive

Cliquer pour revenir en arrière, maintenir pour voir l'historique

ct.org

Applications | Boîte de réception... | The Comprehensi... | Google Groupes | Pensées - YouTube | Accueil - Dropbox | الجزيرة.نت | Nomenclature : La...

CRAN
Mirrors
What's new?
Task Views

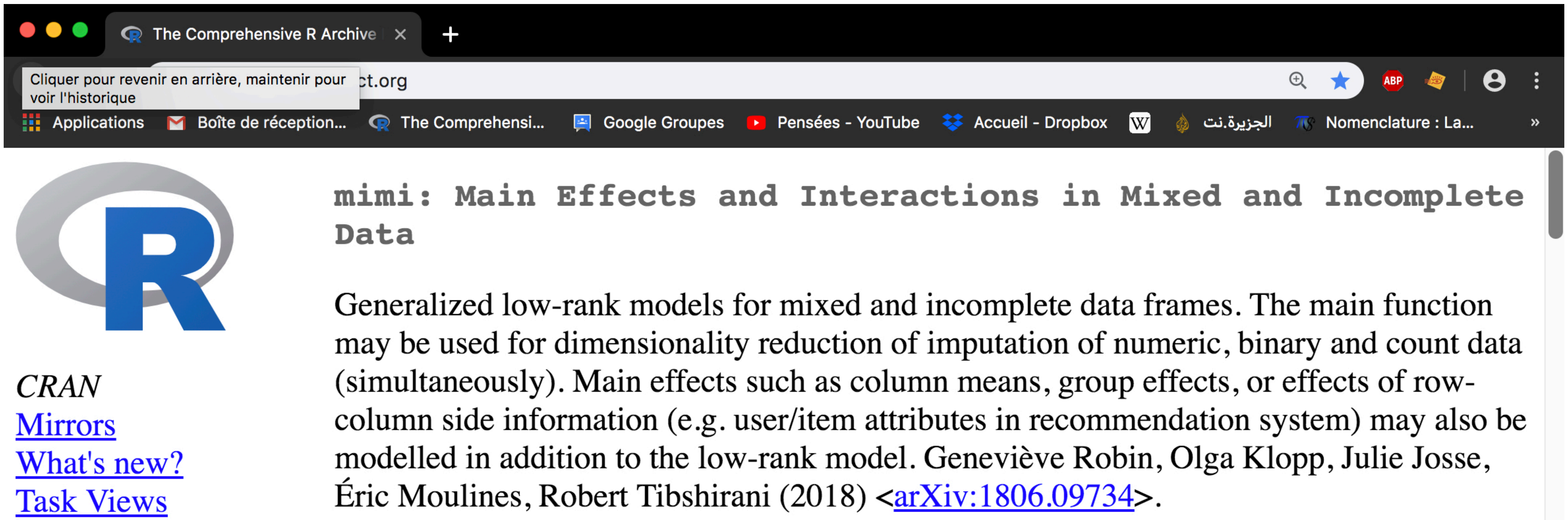
mimi: Main Effects and Interactions in Mixed and Incomplete Data

Generalized low-rank models for mixed and incomplete data frames. The main function may be used for dimensionality reduction of imputation of numeric, binary and count data (simultaneously). Main effects such as column means, group effects, or effects of row-column side information (e.g. user/item attributes in recommendation system) may also be modelled in addition to the low-rank model. Geneviève Robin, Olga Klopp, Julie Josse, Éric Moulines, Robert Tibshirani (2018) <arXiv:1806.09734>.

```r
install.packages("mimi")
library(mimi)
data <- read.table("mydatafile.txt")
var.type <- c(rep("gaussian", 15), rep("binomial", 10))
model <- "low-rank"
rescv <- cv.mimi(y, model=model, var.type=var.type)
res <- mimi(y, model=model, var.type=var.type, lambda1=rescv$lambda,
            algo="bcgd")
```

Cross-validation to select regularization parameters

R package mimi




```r
install.packages("mimi")
library(mimi)
data <- read.table("mydatafile.txt")
var.type <- c(rep("gaussian", 15), rep("binomial", 10))
model <- "low-rank"
rescv <- cv.mimi(y, model=model, var.type=var.type)
res <- mimi(y, model=model, var.type=var.type, lambda1=rescv$lambda,
            algo="bcgd")
```

Another algorithm (Mixed coordinate gradient descent) implemented for large data frames

[Robin et al. 2018]

Simulations: multilevel mixed data

$$\boldsymbol{Y} = \left(\begin{array}{c} \boldsymbol{Y}_1 \\ \hline \boldsymbol{Y}_2 \\ \hline \vdots \\ \hline \boldsymbol{Y}_K \end{array} \right) \begin{array}{l} \updownarrow n_1 \\ \updownarrow n_2 \\ \vdots \\ \updownarrow n_K \end{array}$$

150 individuals in 5 groups
(schools, hospitals, etc.)

$$\boldsymbol{Y} = \left(\begin{array}{cccc} \boldsymbol{Y}_{.,1} & \boldsymbol{Y}_{.,2} & \cdots & \boldsymbol{Y}_{.,m_2} \end{array} \right)$$

Columns of different types
(numeric, binary, etc.)

Simulations: multilevel mixed data

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \hline \boldsymbol{Y}_2 \\ \hline \vdots \\ \hline \boldsymbol{Y}_K \end{pmatrix} \begin{matrix} \updownarrow n_1 \\ \updownarrow n_2 \\ \vdots \\ \updownarrow n_K \end{matrix}$$

150 individuals in 5 groups
(schools, hospitals, etc.)

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_{.,1} & \boldsymbol{Y}_{.,2} & \dots & \boldsymbol{Y}_{.,m_2} \end{pmatrix}$$

Columns of different types
(numeric, binary, etc.)

$$\boldsymbol{Y}_{i,j} \sim \mathcal{N}(\alpha^0_{c(i),j} + \boldsymbol{\Theta}^0_{i,j}, \sigma^2)$$

Columns 1-15

Simulations: multilevel mixed data

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \hline \boldsymbol{Y}_2 \\ \hline \vdots \\ \hline \boldsymbol{Y}_K \end{pmatrix} \begin{array}{l} \updownarrow n_1 \\ \updownarrow n_2 \\ \vdots \\ \updownarrow n_K \end{array}$$

150 individuals in 5 groups
(schools, hospitals, etc.)

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_{.,1} & \boldsymbol{Y}_{.,2} & \ldots & \boldsymbol{Y}_{.,m_2} \end{pmatrix}$$

Columns of different types
(numeric, binary, etc.)

effect of group c(i) on variable j

$$\boldsymbol{Y}_{i,j} \sim \mathcal{N}(\alpha^0_{c(i),j} + \boldsymbol{\Theta}^0_{i,j}, \sigma^2)$$

individual i
in group c(i)

interaction/individual effect

Columns 1-15

Simulations: multilevel mixed data

$$\boldsymbol{Y} = \left(\begin{array}{c} \boldsymbol{Y}_1 \\ \hline \boldsymbol{Y}_2 \\ \hline \vdots \\ \hline \boldsymbol{Y}_K \end{array}\right) \begin{array}{l} \updownarrow n_1 \\ \updownarrow n_2 \\ \vdots \\ \updownarrow n_K \end{array}$$

150 individuals in 5 groups (schools, hospitals, etc.)

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_{.,1} & \boldsymbol{Y}_{.,2} & \dots & \boldsymbol{Y}_{.,m_2} \end{pmatrix}$$

Columns of different types (numeric, binary, etc.)

$$\boldsymbol{Y}_{i,j} \sim \mathcal{N}(\alpha^0_{c(i),j} + \boldsymbol{\Theta}^0_{i,j}, \sigma^2)$$

Columns 1-15

$$\mathbb{P}(\boldsymbol{Y}_{i,j} = 1) = \frac{\mathrm{e}^{\boldsymbol{X}^0_{i,j}}}{1 + \mathrm{e}^{\boldsymbol{X}^0_{i,j}}}, \quad \boldsymbol{X}^0_{i,j} = \alpha^0_{c(i)j} + \boldsymbol{\Theta}^0_{i,j}$$

Columns 16-30

Simulations: multilevel mixed data

$$\boldsymbol{Y} = \left(\begin{array}{c} \boldsymbol{Y}_1 \\ \hline \boldsymbol{Y}_2 \\ \hline \vdots \\ \hline \boldsymbol{Y}_K \end{array} \right) \begin{array}{l} \updownarrow n_1 \\ \updownarrow n_2 \\ \vdots \\ \updownarrow n_K \end{array}$$

150 individuals in 5 groups
(schools, hospitals, etc.)

$$\boldsymbol{Y} = \left(\boldsymbol{Y}_{.,1} \quad \boldsymbol{Y}_{.,2} \quad \ldots \quad \boldsymbol{Y}_{.,m_2} \right)$$

Columns of different types
(numeric, binary, etc.)

$$\boldsymbol{Y}_{i,j} \sim \mathcal{N}(\alpha^0_{c(i),j} + \boldsymbol{\Theta}^0_{i,j}, \sigma^2)$$

Columns 1-15

interaction/
individual
effect

effect of group c(i)
on variable j

$$\mathbb{P}(\boldsymbol{Y}_{i,j} = 1) = \frac{e^{\boldsymbol{X}^0_{i,j}}}{1 + e^{\boldsymbol{X}^0_{i,j}}}, \quad \boldsymbol{X}^0_{i,j} = \alpha^0_{c(i)j} + \boldsymbol{\Theta}^0_{i,j}$$

Columns 16-30

individual i
in group c(i)

Simulations: multilevel mixed data

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \hline \boldsymbol{Y}_2 \\ \hline \vdots \\ \hline \boldsymbol{Y}_K \end{pmatrix} \begin{array}{l} \updownarrow n_1 \\ \updownarrow n_2 \\ \vdots \\ \updownarrow n_K \end{array}$$

150 individuals in 5 groups
(schools, hospitals, etc.)

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_{.,1} & \boldsymbol{Y}_{.,2} & \dots & \boldsymbol{Y}_{.,m_2} \end{pmatrix}$$

Columns of different types
(numeric, binary, etc.)

$$\boldsymbol{Y}_{i,j} \sim \mathcal{N}(\alpha^0_{c(i),j} + \boldsymbol{\Theta}^0_{i,j}, \sigma^2)$$

Columns 1-15

sparse: 5 non zero coefficients

low-rank: rank 3

$$\mathbb{P}(\boldsymbol{Y}_{i,j} = 1) = \frac{\mathrm{e}^{\boldsymbol{X}^0_{i,j}}}{1 + \mathrm{e}^{\boldsymbol{X}^0_{i,j}}}, \quad \boldsymbol{X}^0_{i,j} = \alpha^0_{c(i)j} + \boldsymbol{\Theta}^0_{i,j}$$

Columns 16-30

Compared methods

- **softImpute** (Hastie et al., 2015): method for numeric data based on soft-thresholding of singular values (R package softImpute).
- **Generalized Low-Rank Model** (**GLRM**, Udell et al. 2016): matrix factorization framework for mixed data (h2o package glrm).
- **Factorial Analysis of Mixed Data** (**FAMD**, Pagès 2015): principal component method for mixed data (R package missMDA, Josse and Husson, 2016).
- **Multilevel Factorial Analysis of Mixed Data** (**MLFAMD**, Husson et al. 2018): extension of FAMD to multilevel data (R package missMDA).
- **Multivariate Imputation by Chained Equations** (**mice**, van Buuren and Groothuis- Oudshoorn 2011): multiple imputation using Fully Conditional Specification (R package mice).

Compared methods

- **softImpute** (Hastie et al., 2015): method for numeric data based on soft-thresholding of singular values (R package softImpute).

- **Generalized Low-Rank Model** (**GLRM**, Udell et al. 2016): matrix factorization framework for mixed data (h2o package glrm).

- **Factorial Analysis of Mixed Data** (**FAMD**, Pagès 2015): principal component method for mixed data (R package missMDA, Josse and Husson, 2016).

Also part of this thesis [Husson et al. 2018]

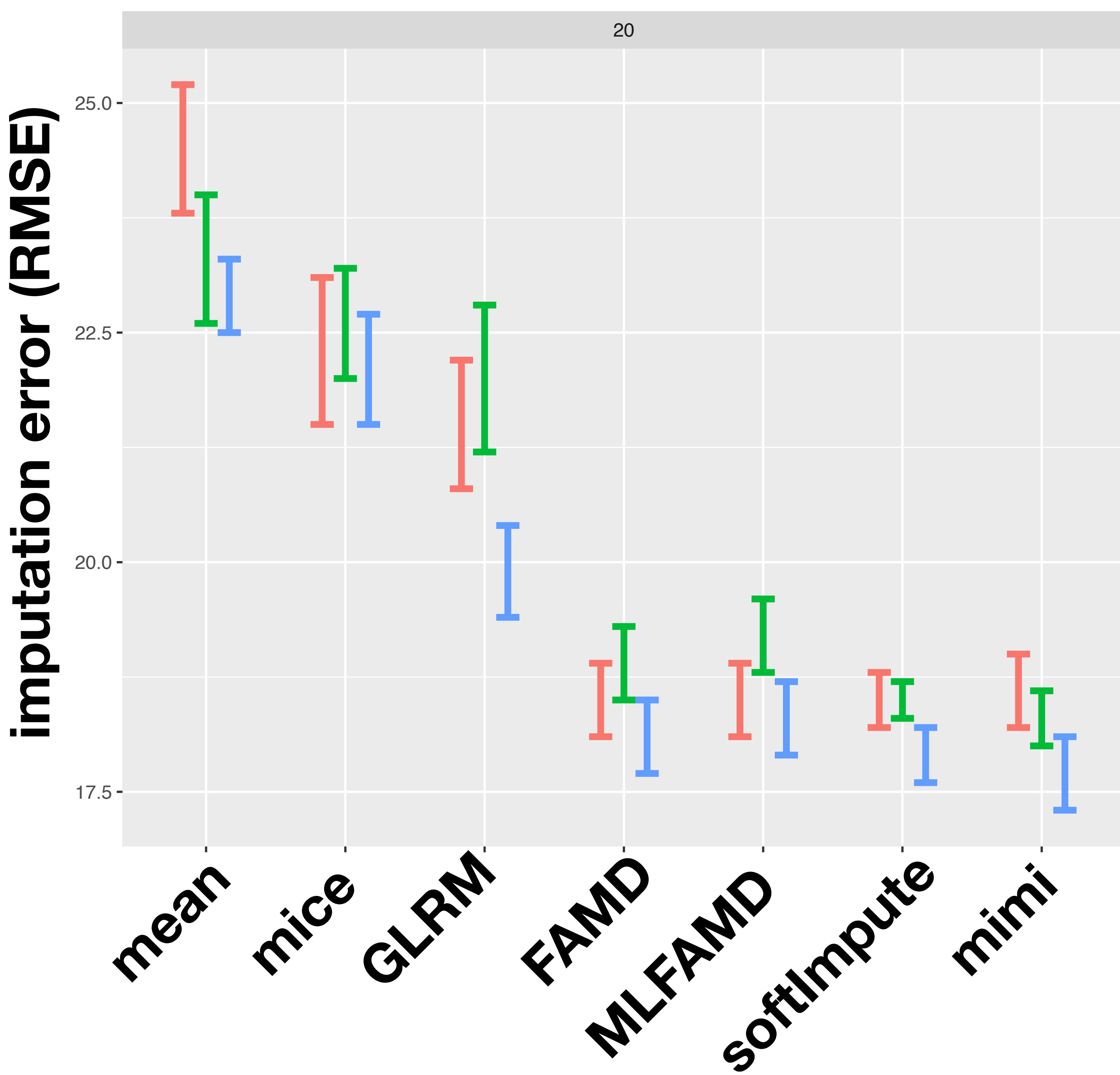
- **Multilevel Factorial Analysis of Mixed Data** (**MLFAMD**, Husson et al. 2018): extension of FAMD to multilevel data (R package missMDA).

- **Multivariate Imputation by Chained Equations** (**mice**, van Buuren and Groothuis- Oudshoorn 2011): multiple imputation using Fully Conditional Specification (R package mice).

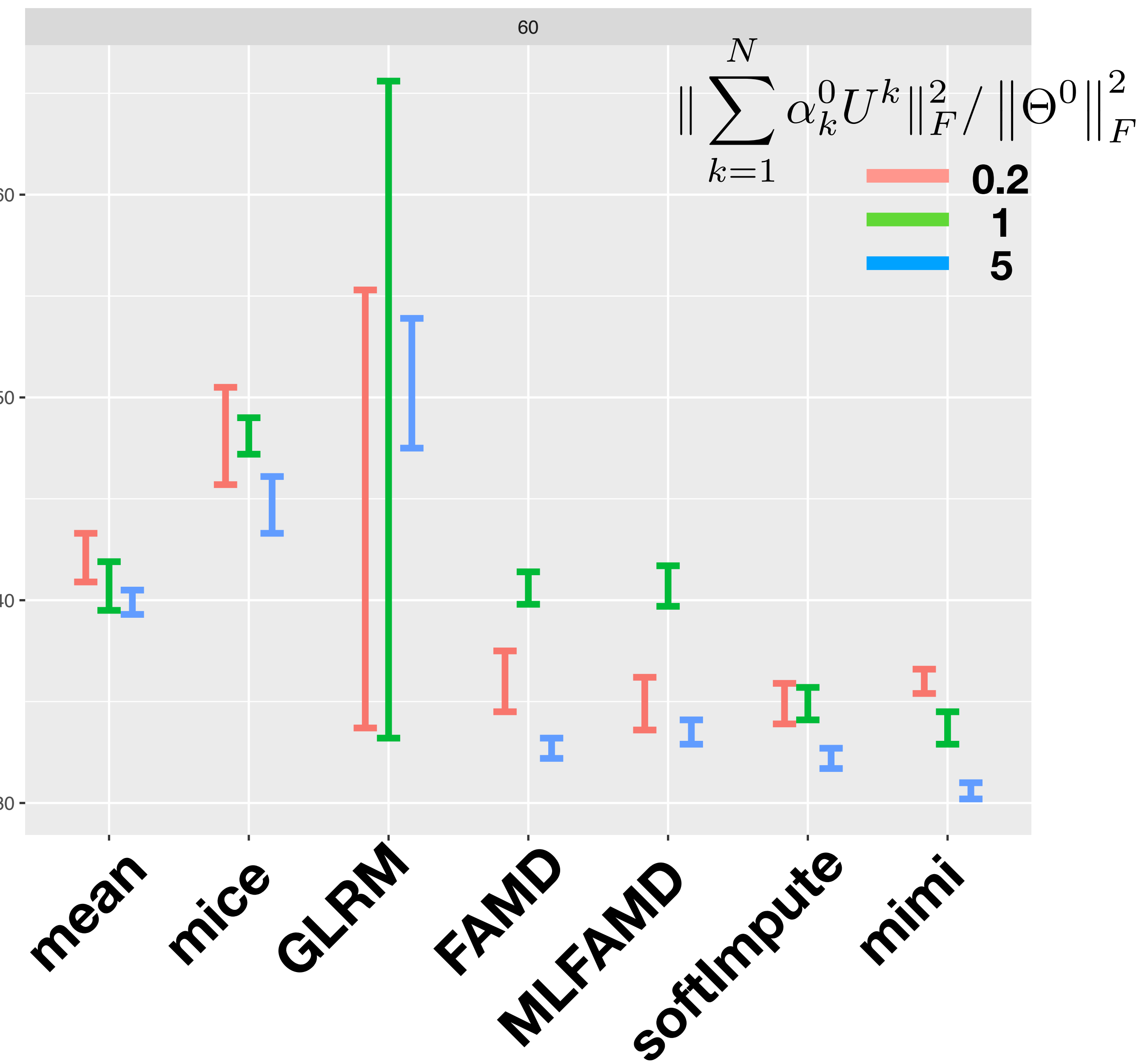
Numerical results

Imputation error (averaged across 100 rep)

20% missing values



60% missing values



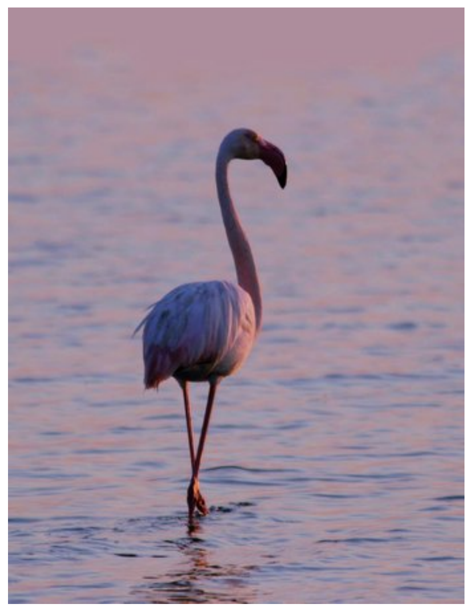
Objectives of this thesis

1. Provide *theoretically sound* models adapted to multi-source, heterogeneous and incomplete data *simultaneously*
    - Hybrid low-rank structures
    - Heterogeneous data fitting terms
    - Upper and lower bounds on estimation errors

2. For these models, provide estimation methods and empirically robust software solutions
    - Optimization algorithms
    - Implementation of R packages
    - Numerical results

3. Confront the methods to applications in life sciences
    - Analysis of a waterbird abundance data set
    - Imputation of a medical registry

Waterbirds monitoring



- Waterbirds depend upon wetland sites for at least part of their life cycle
- Important ecosystem service providers (disperser of seeds, sentinel for epidemics)
- Waterbird monitoring used as surrogate to evaluate global state of biodiversity

Waterbirds monitoring

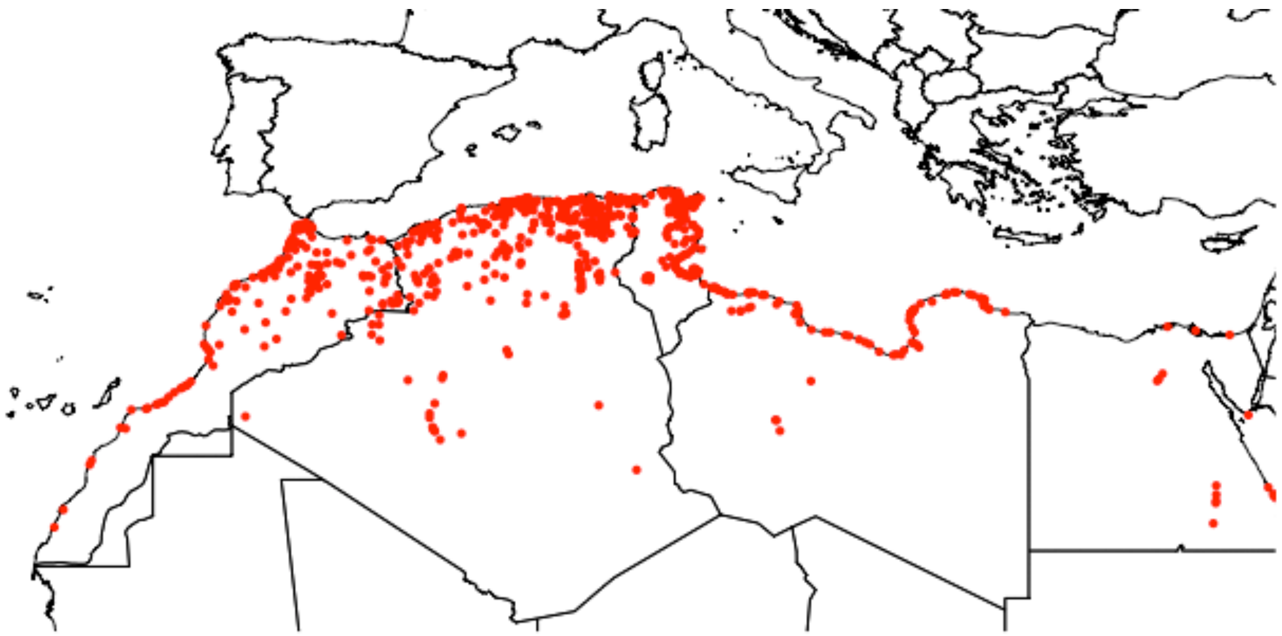



- Yearly censuses supervised by Wetlands International
- First census in 1967
- 25,000 sites counted yearly
- Provide information to international conservation organizations

Waterbirds monitoring in North Africa

- Biodiversity hotspot
- Last stopover before crossing the Sahara or the Mediterranean Sea
- Censuses regular since 1983 in Morocco, 1985 in Algeria, 2002 in Tunisia
- Spatial coverage remains variable for financial and political reasons (Etayeb et al. 2015)

The waterbirds data set



- Collaboration with the Tour du Valat Institute (Camargue)
- 785 sites in Morocco, Algeria, Tunisia, Libya and Egypt
- Counts between 1990 and 2018 (28 years)
- 23 waterbird species, between 40 and 60% missing values
- Side information: covariates about sites and years
- Goal: estimate yearly totals for each species

Objectives and approach

$\boldsymbol{Y}$

Site	2008	2009	2010
1	NA	0	0
2	4	50	25
3	NA	0	0
4	NA	NA	NA
5	NA	NA	NA
6	0	0	0
7	5	75	870

$\boldsymbol{U}$

Site	Year	Rain	Eco	Country	Agri
1	2008	163.7	0.8	Algeria	16.2
2	2008	60.7	0.8	Algeria	16.2
3	2008	227.9	0.8	Algeria	16.2
4	2008	174.8	0.8	Algeria	16.2
5	2008	163.7	0.8	Algeria	16.2
6	16.2	16.2	16.2	16.2	16.2
7	2008	243.5	0.8	Algeria	16.2

- Impute the missing values, then compute yearly sums
- Include side information to improve the predictions
- Estimate covariate effects, select important factors
- Compute empirical intervals of variability

Objectives and approach

Y

Site	2008	2009	2010
1	**15**	0	0
2	4	50	25
3	**7**	0	0
4	**2**	**60**	**160**
5	**5**	**10**	**70**
6	0	0	0
7	5	75	870
	38	**195**	**1125**

U

Site	Year	Rain	Eco	Country	Agri
1	2008	163.7	0.8	Algeria	16.2
2	2008	60.7	0.8	Algeria	16.2
3	2008	227.9	0.8	Algeria	16.2
4	2008	174.8	0.8	Algeria	16.2
5	2008	163.7	0.8	Algeria	16.2
6	2008	177.3	0.8	Algeria	16.2
7	2008	243.5	0.8	Algeria	16.2

- Impute the missing values, then compute yearly sums
- Include side information to improve the predictions
- Estimate covariate effects, select important factors
- Compute empirical intervals of variability

Objectives and approach

Y

Site	2008	2009	2010
1	**15**	0	0
2	4	50	25
3	**7**	0	0
4	**2**	**60**	**160**
5	**5**	**10**	**70**
6	0	0	0
7	5	75	870
	38	**195**	**1125**

U

Site	Year	Rain	Eco	Country	Agri
1	2008	163.7	0.8	Algeria	16.2
2	2008	60.7	0.8	Algeria	16.2
3	2008	227.9	0.8	Algeria	16.2
4	2008	174.8	0.8	Algeria	16.2
5	2008	163.7	0.8	Algeria	16.2
6	2008	177.3	0.8	Algeria	16.2
7	2008	243.5	0.8	Algeria	16.2

- Impute the missing values, then compute yearly sums
- Include side information to improve the predictions
- Estimate covariate effects, select important factors
- Compute empirical intervals of variability

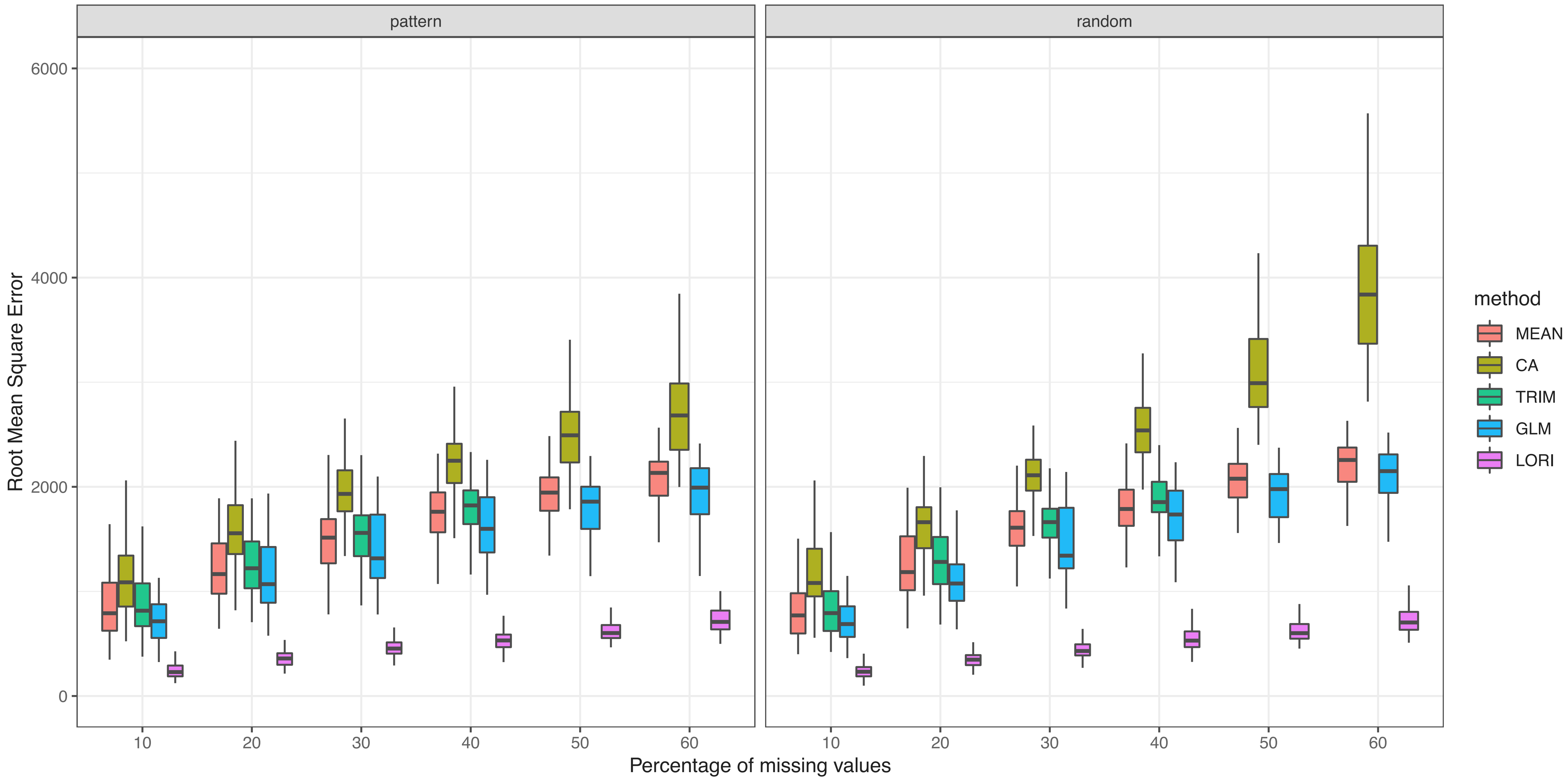
Special case of general model with Poisson entries

R package lori

Empirical performance: Poisson data

Missing values accumulated along rows/columns

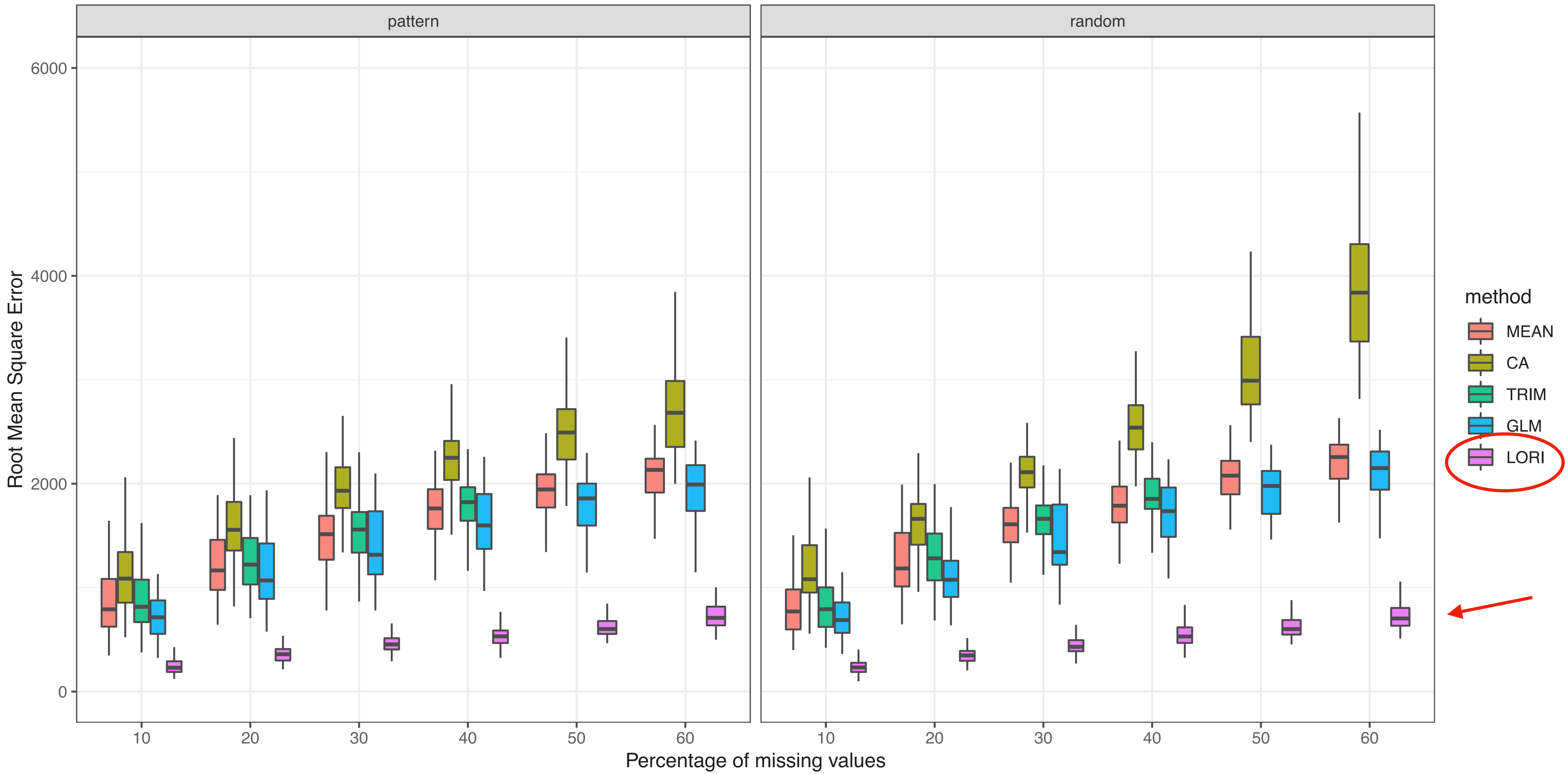
Missing uniformly at random missing values



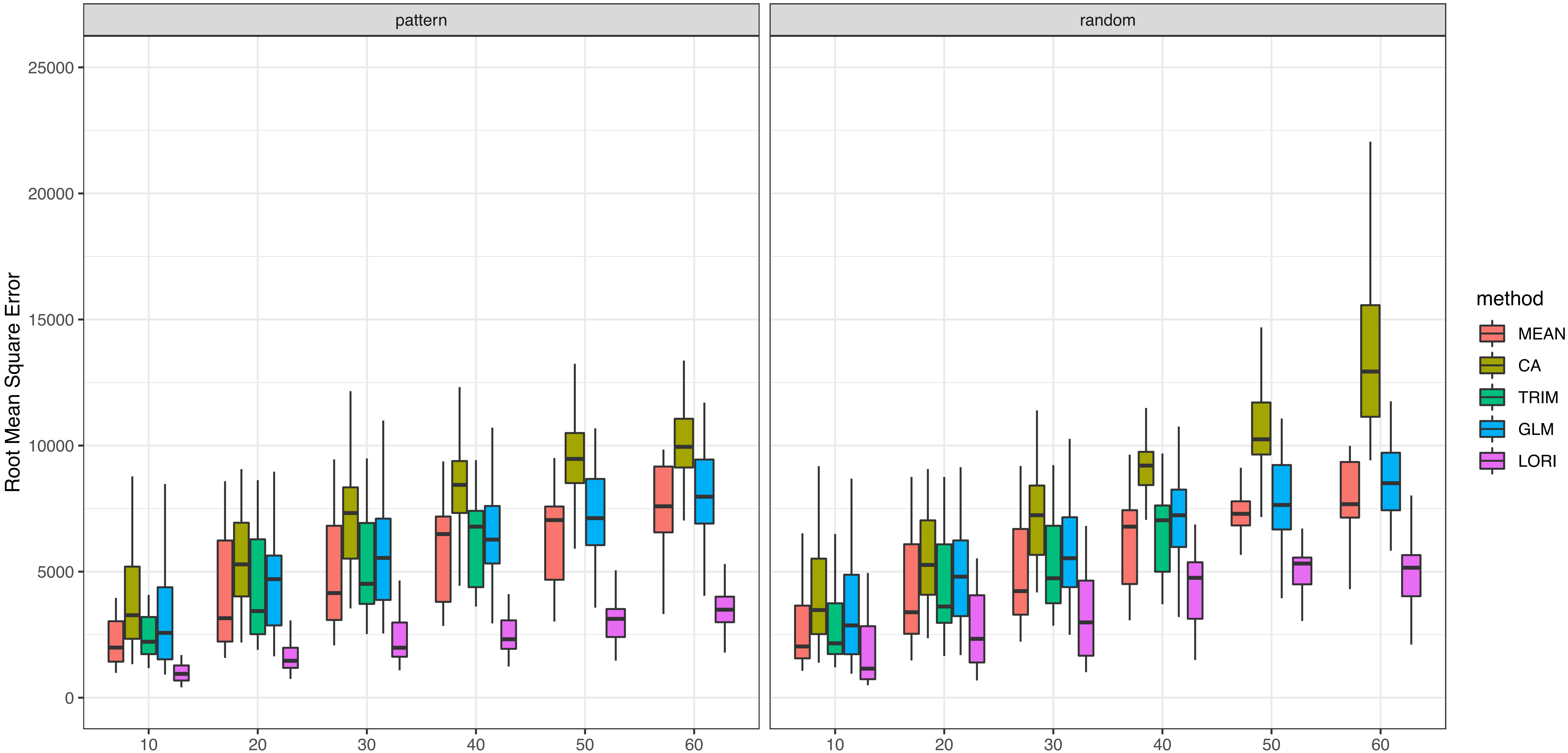
Empirical performance: Poisson data

Missing values accumulated along rows/columns

Missing uniformly at random missing values

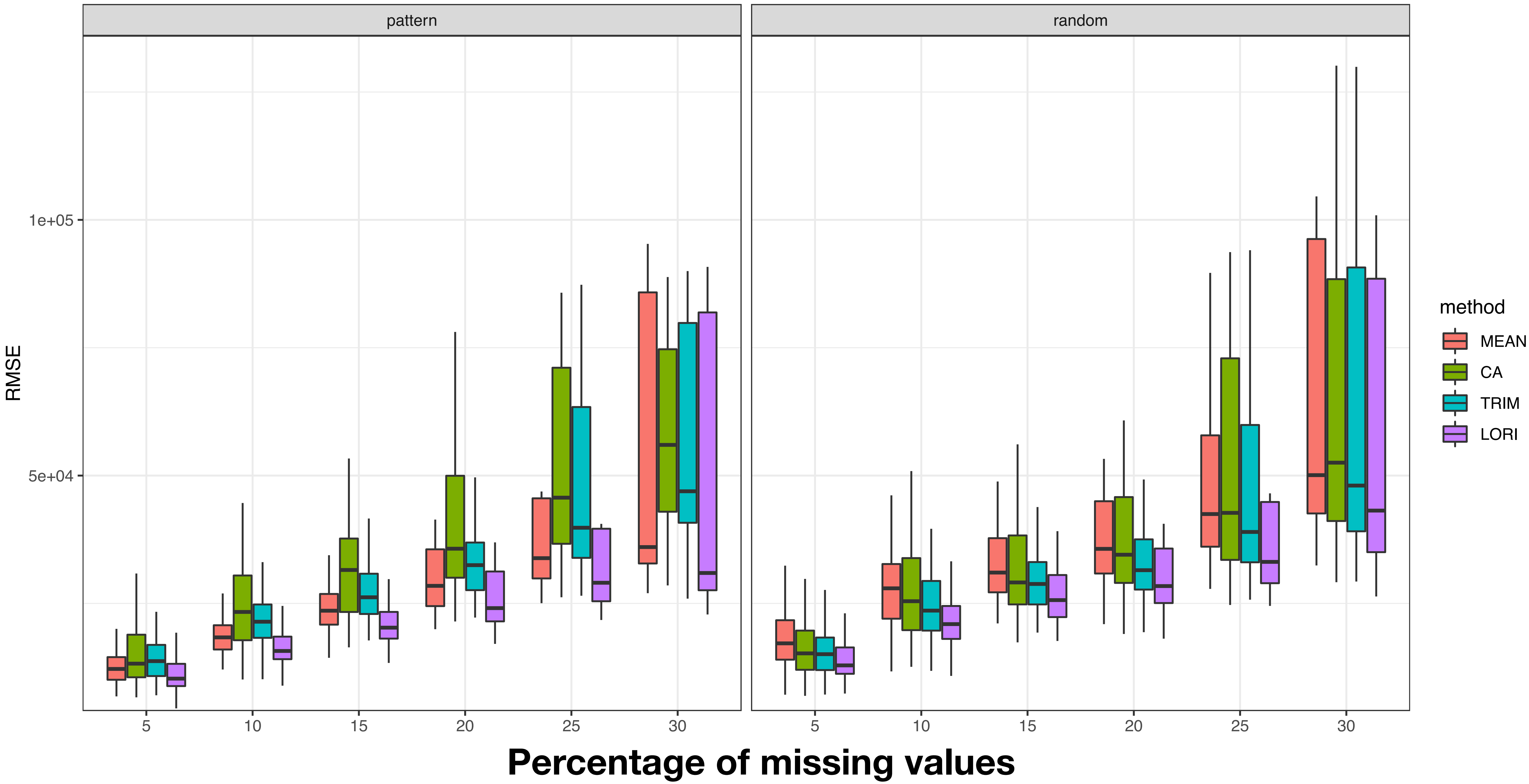


Empirical performance: Zero-inflated negative binomial

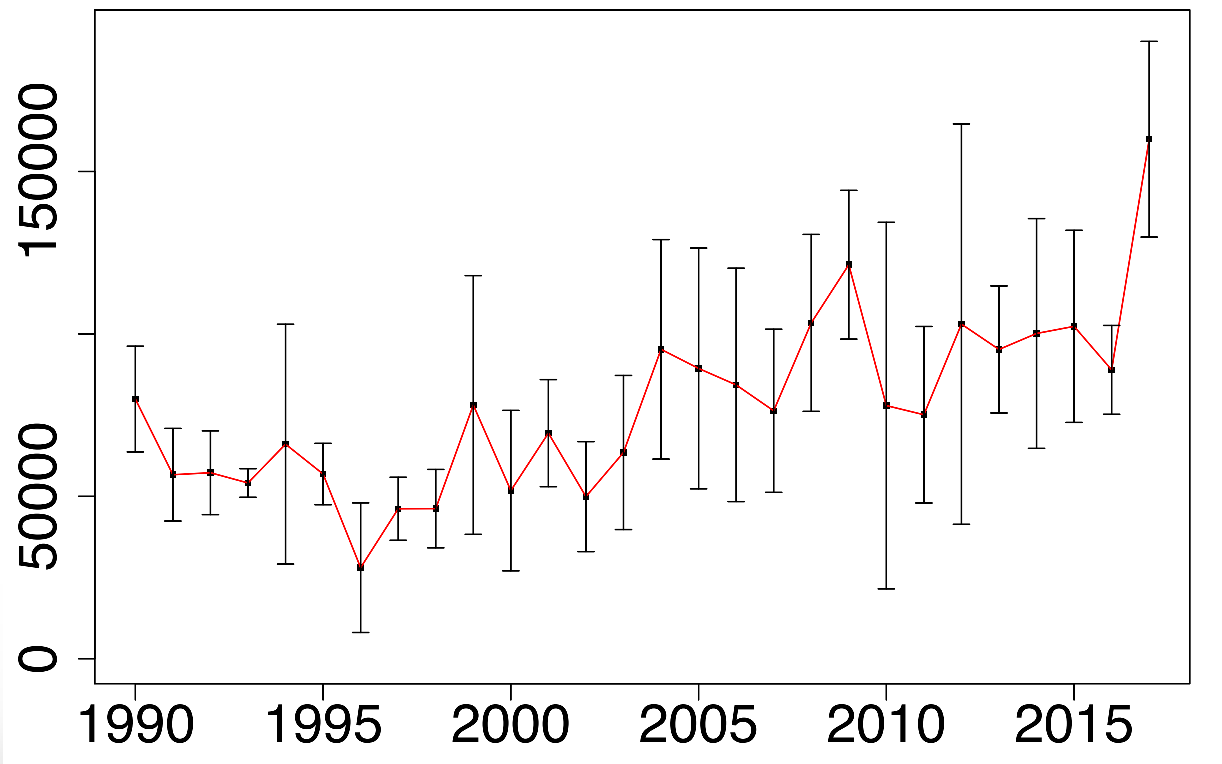


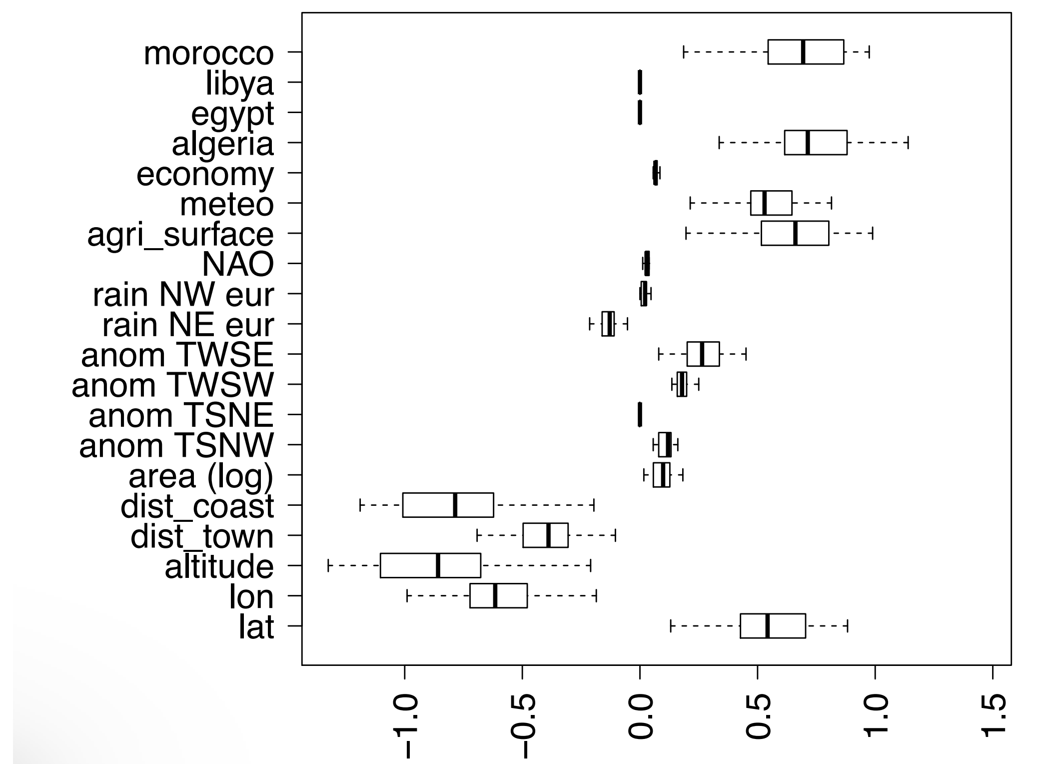
Percentage of missing values

Empirical performance: waterbirds data

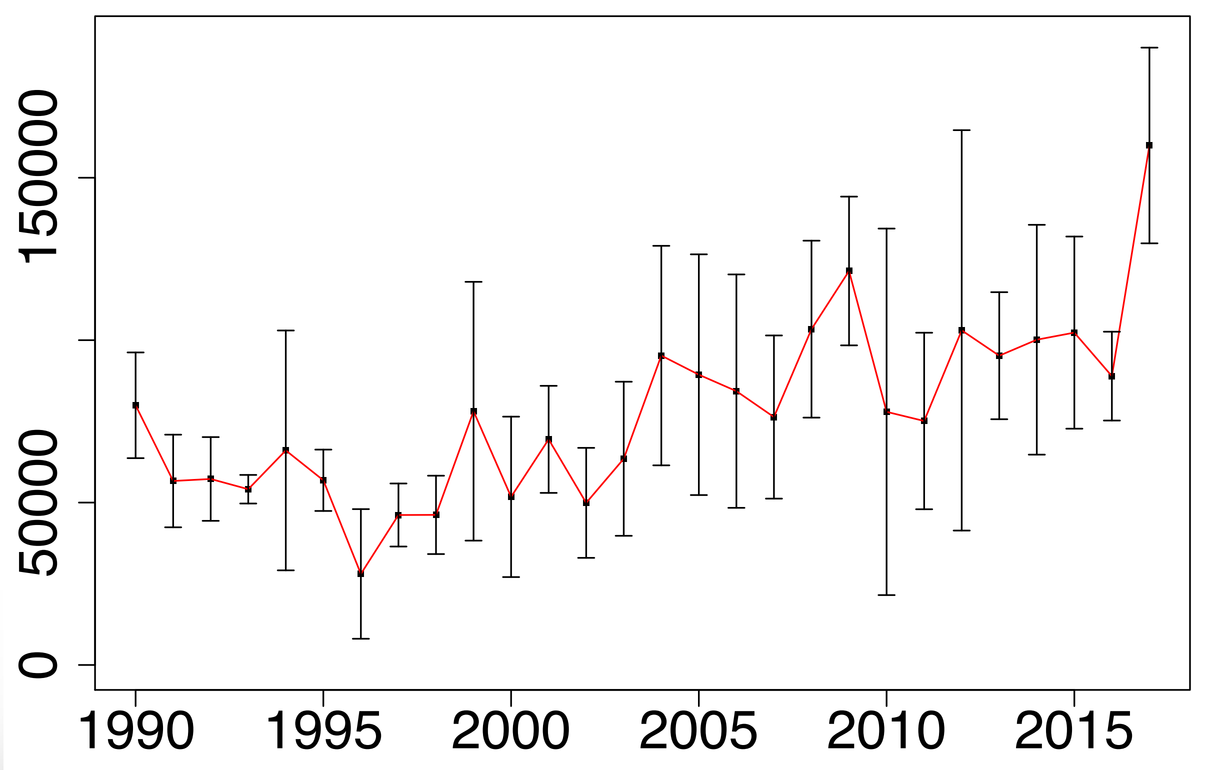


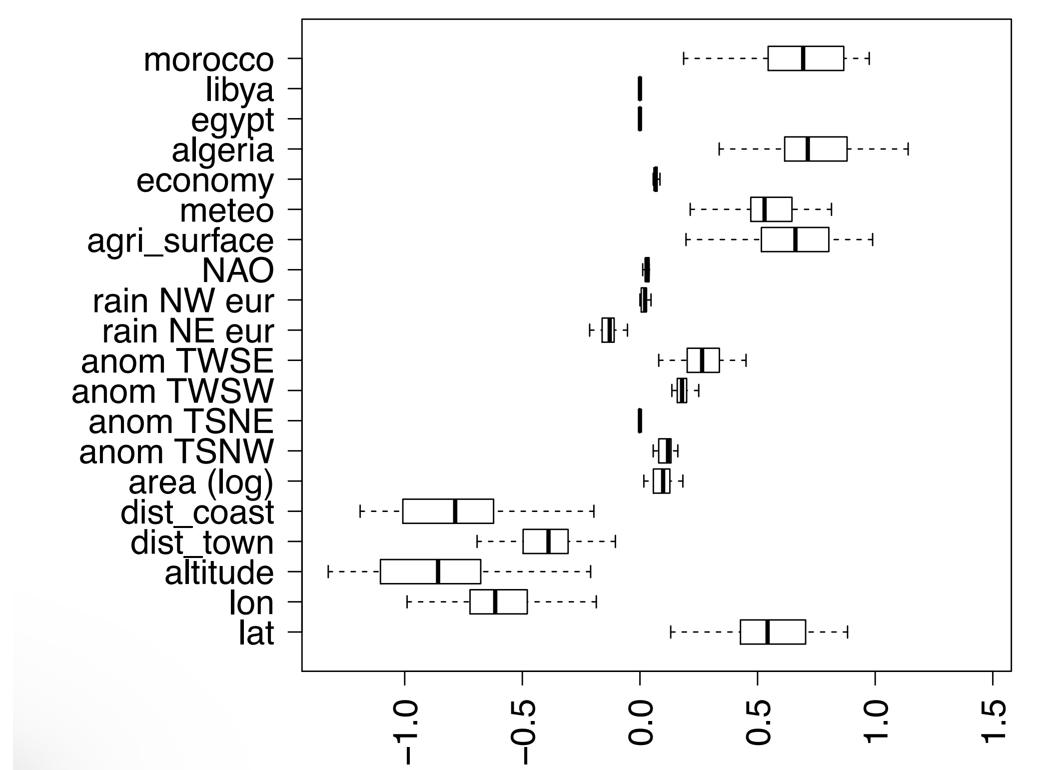
Temporal trends: northern shoveler





Temporal trends: northern shoveler

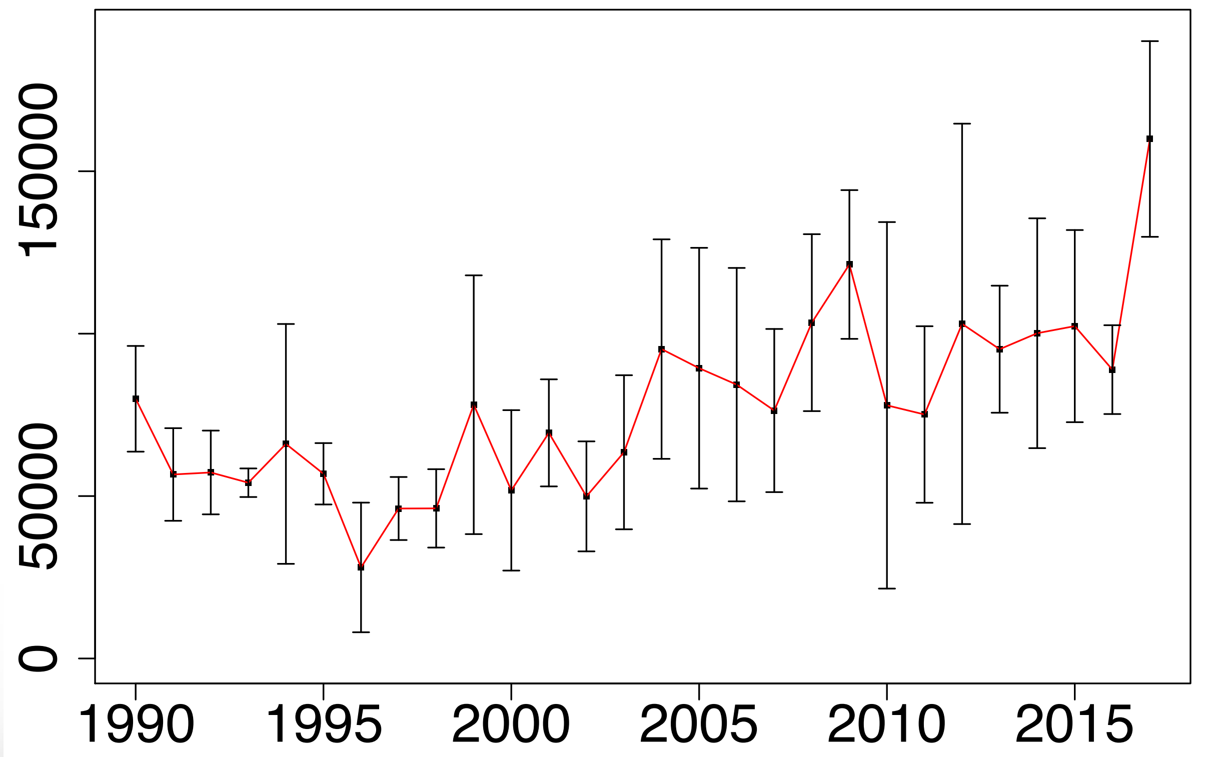


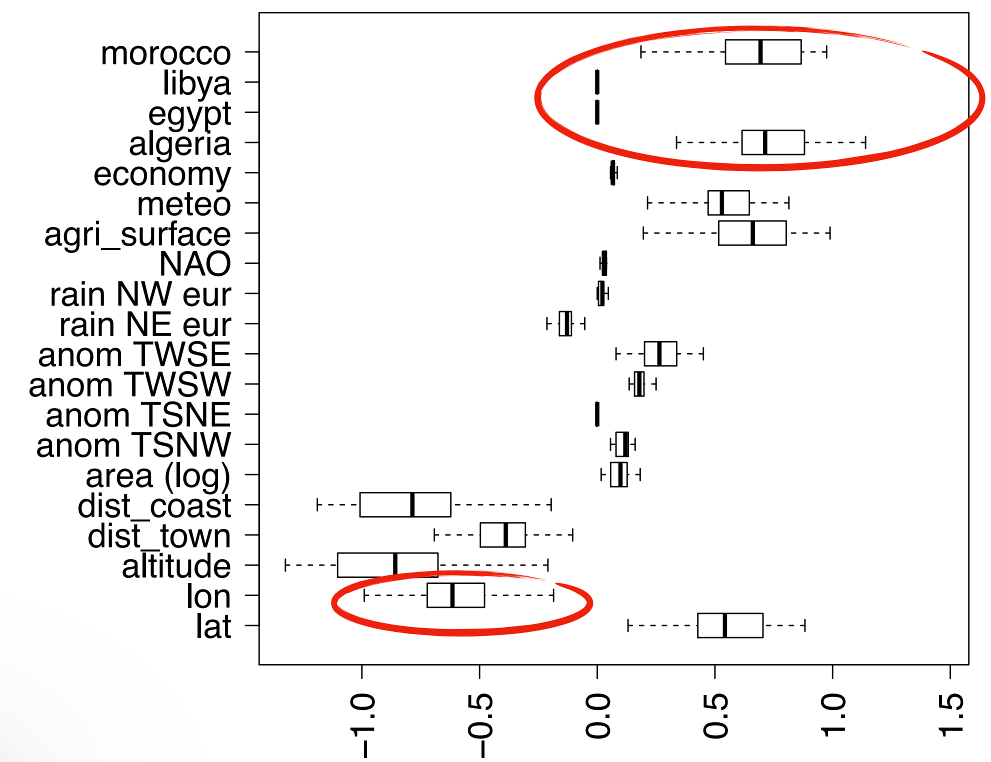


Increasing in North-Africa

Temporal trends: northern shoveler

country effect

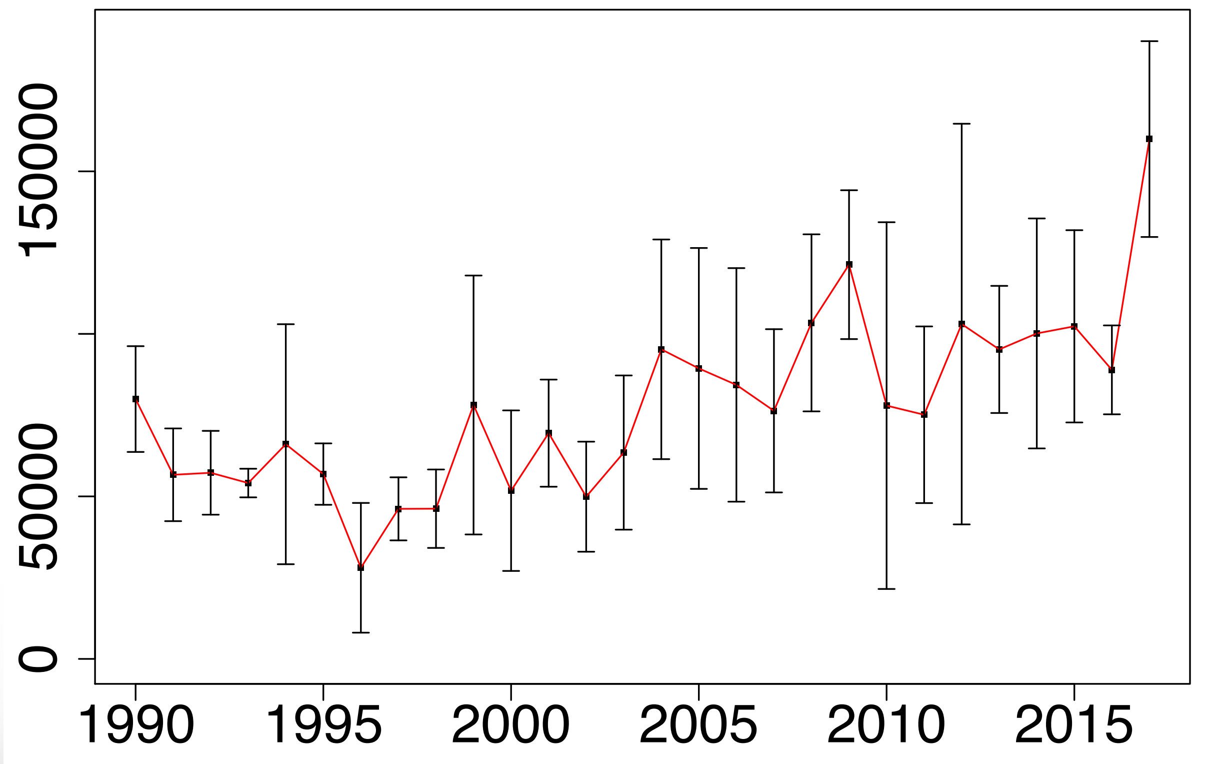


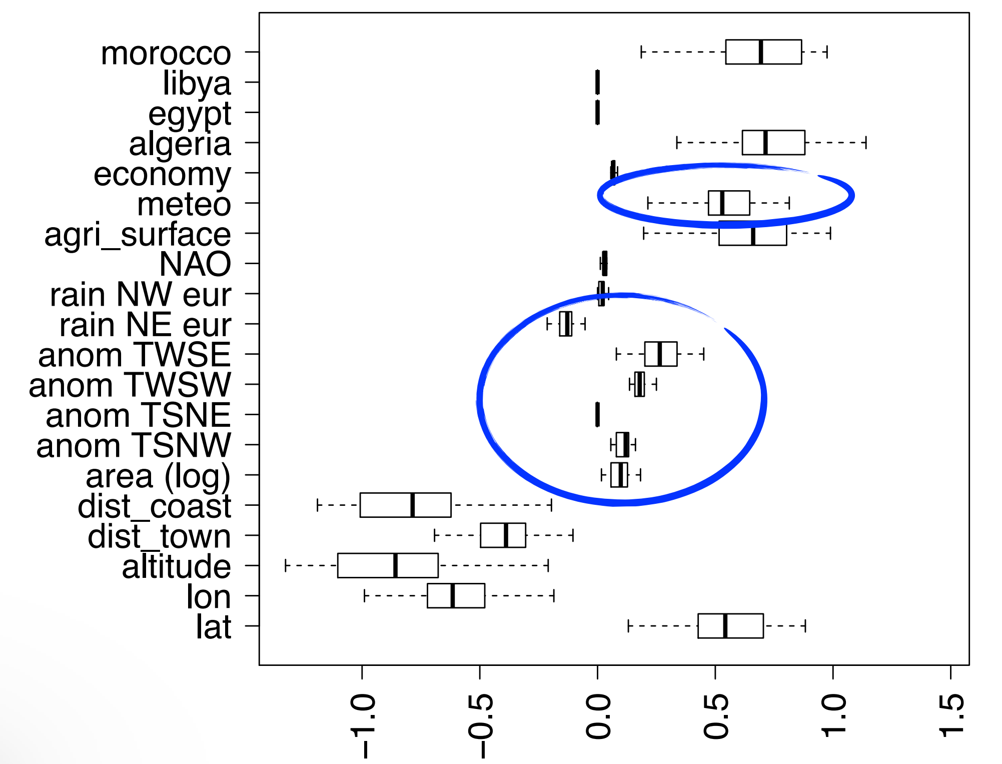


Increasing in North-Africa

Temporal trends: northern shoveler

Effect of meteorological anomalies





Increasing in North-Africa

General conclusion

- New data analysis tools adapted to modern data collection processes
- General framework based on hybrid low-rank models
  and heterogeneous exponential families
- Theoretical guarantees, implementations and ecological application

General conclusion

- New method for count data analysis with covariates and missing values
  - Model, estimation, theoretical results
  - R package lori
- Analysis of a waterbirds abundance data set
  - Results presented at the African-Eurasian Waterbirds Agreement (AEWA) meeting of parties
  - Also at the 21st Conference of the European Bird Census Council
- New method for heterogenous data with missing values and side information
  - Model, estimation, theoretical results
  - R package mimi
- Alternative method to impute missing values in multilevel heterogeneous data (MLFAMD, package missMDA)

Perspectives

- Extension of the framework to exponential families with multiple parameters (incorporate a scale parameter)
- Extension to more complex models (zero-inflation and overdispersion)
- Extension to non-sparse dictionary matrices (multivariate Gaussians)
- Uncertainty measurement (post-selection inference, Bayesian perspective, multiple imputation)
- Analysis of several other bird species (ongoing)

Publications

- Geneviève Robin, Hoi-To Wai, Julie Josse, Olga Klopp, Éric Moulines (2018) *Low-rank interactions and sparse additive effects for large data frames*. Advances in Neural Information Processing Systems 31, pp. 5496–5506. Curran Associates, Inc.

- François Husson, Julie Josse, Balasubramanian Narasimhan, Geneviève Robin (2019). *Imputation of multilevel mixed data using multilevel singular value decomposition*. Journal of Computational and Graphical Statistics.

- Geneviève Robin, Julie Josse, Éric Moulines, Sylvain Sardy (2019). *Low-rank models with covariates for count data with missing values*. Journal of Multivariate Analysis 173*,* 416-434

- Geneviève Robin, Olga Klopp, Julie Josse, Éric Moulines, Robert Tibshirani (2019). *Main effects and interactions in mixed and incomplete data frames*. Journal of the American Statistical Association (accepted)

Thank you for your attention !

Acknowledgements

Éric Moulines

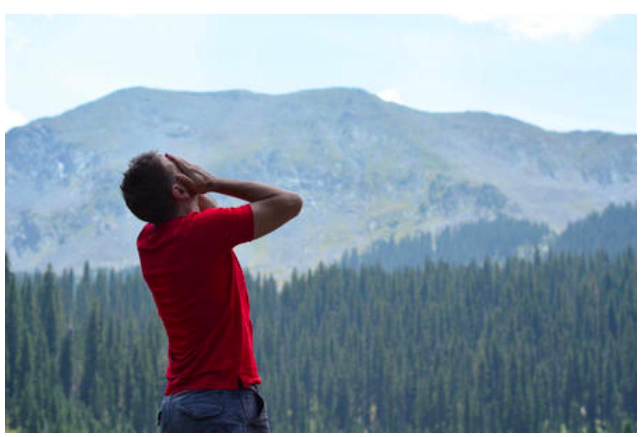
Julie Josse

Acknowledgements

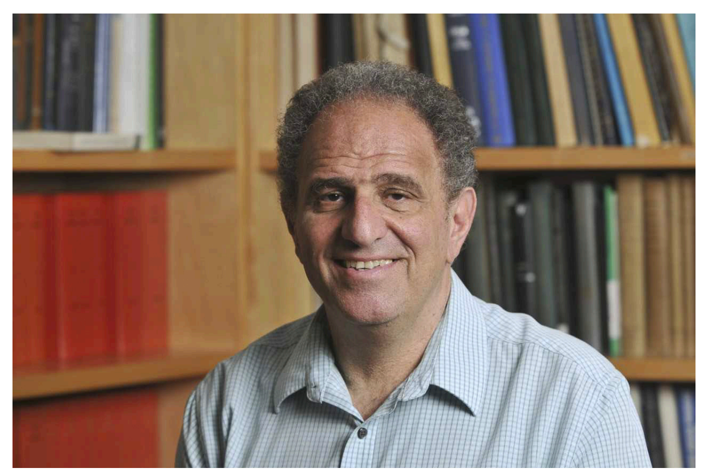
François Husson

Olga Klopp

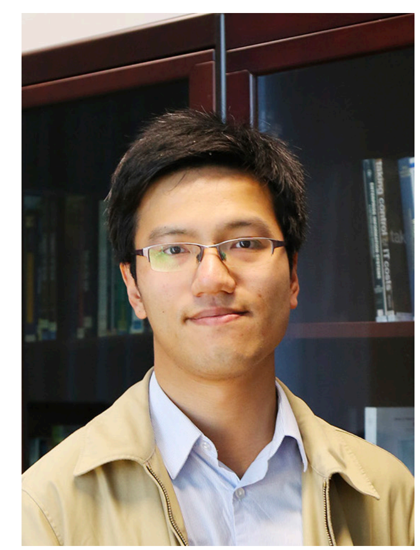
Balasubramanian Narasimhan



Sylvain Sardy



Rob Tibshirani



Hoi-To Wai

Acknowledgements

Laura Dami, Jean-Yves Mondain-Monval, Marie Suet, Pierre Defos du Rau, Clémence Deschamps

References

[1] Alekh Agarwal, Sahand Negahban and martin J. Wainwright. *Noisy Matrix decomposition via convex relaxation: Optimal rates in high dimensions*. Ann. Statist, April 2012.

[2] Emmanuel J. Candès, Xiaodong Li, Yi Ma, John Wright. *Robust principal component analysis?* J. ACM, 58(3):11:1-11:37, June 2011.

[3] Emmanuel J. Candes, Yaniv Plan (2010, June). *Matrix completion with noise*. Proceedings of the IEEE 98(6), 925–936

[4] Emmanuel J. Candès, Benjamin Recht (2009). *Exact matrix completion via convex optimization*. Foundations of Computational mathematics 9(6), 717–772.

[5] Emmanuel J. Candès, Terence Tao (2010). *The power of convex relaxation: Near-optimal matrix completion*. IEEE Transactions on Information Theory 56(5), 2053–2080.

[6] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parillo and Alan S. Willsky. *Rank-Sparsity incoherence for matrix decomposition*. SIAM Journal on Optimization, 2011.

References

[7] Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). *Regularization paths for generalized linear models via coordinate descent*. Journal of Statistical Software 33(1), 1.

[8] Trevor Hastie, Rahul Mazumder, Jason D. Lee, Reda Zadeh. *Matrix Completion and Low-rank SVD via fast alternating least squares*.The Journal of Machine Learning research, jan 2015.

[9] Harold Hotelling (1933). *Analysis of a complex of statistical variables into principal components*. Journal of Educational Psychology 24(6), 417–441.

[10] Hsu, D., S. M. Kakade, and T. Zhang (2011). *Robust matrix decomposition with sparse corruptions*. EEE Transactions on Information Theory 57(11), 7221–7234.

[11] Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh (2010). *Matrix completion from noisy entries*. J. Mach. Learn. Res. 11, 2057–2078.

[12] Olga Klopp, Karim Lounici and Alexandre B. Tsybakov. *Robust matrix completion*. Probability Theory and Related Fields, October 2017.

References

[13] N. Kishore Kumar, Jan Schneider (2017). *Literature survey on low rank approximation of matrices*. Linear and Multilinear Algebra 65(11), 2212–2244.

[14] Jean Lafond (2015). *Low rank matrix completion with exponential family noise*. Journal of Machine Learning Research: Workshop and Conference Proceedings 40, 1–18.

[15] Mardani, M., G. Mateos, and G. B. Giannakis (2013, Aug). *Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies*. IEEE Transactions on Information Theory 59(8), 5186–5205.

[16] Jeroen Pannekoek, Arco van Strien (2001). *Trim 3 manual (trends & indices for monitoring data)*. Statistics Netherlands.

[17] Karl Pearson (1901). *Liii. on lines and planes of closest fit to systems of points in space*. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2(11), 559–572.

[18] Madeleine Udell, Catherine Horn, Reda Zadeh and Stephen Boyd. *Generalized low rank models.* arXiv preprint arXiv:1410.0342

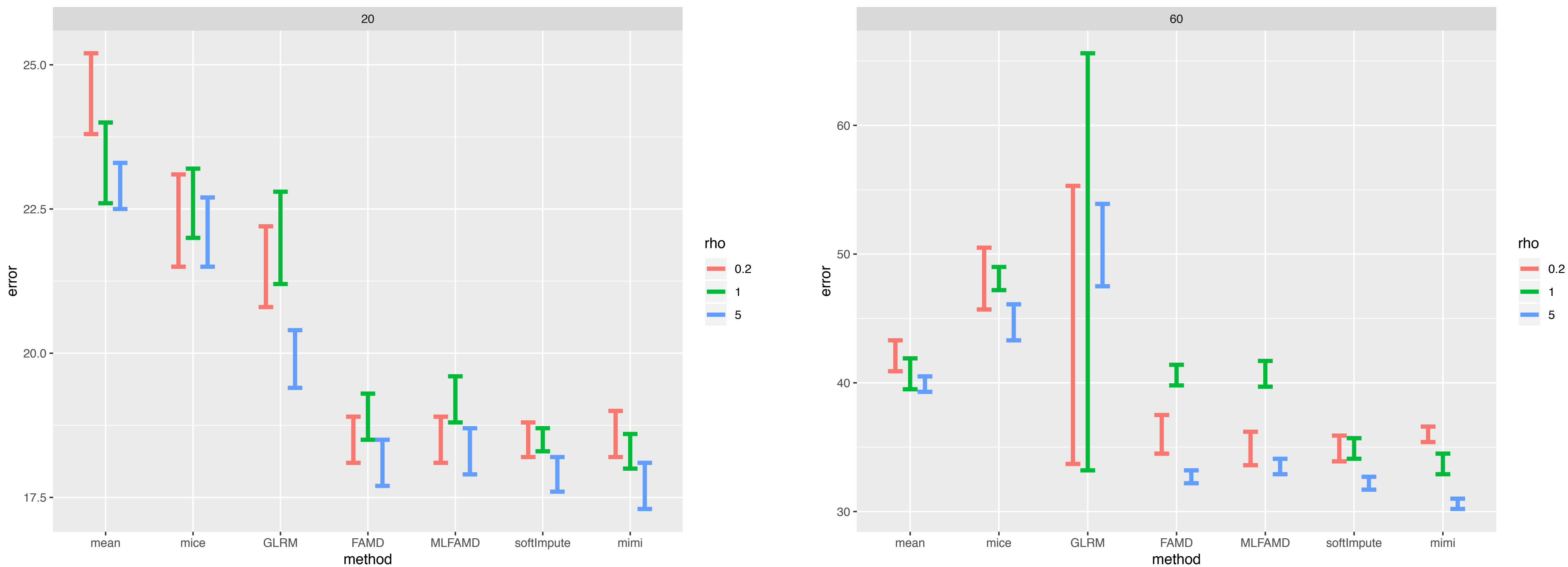
References

[19] Paul Tseng, S. Yun (2009). *A coordinate gradient descent method for nonsmooth separable minimization*. Math. Program. 117(1-2, Ser. B), 387–423

[20] Xu, H., C. Caramanis, and S. Sanghavi (2010). *Robust pca via outlier pursuit*. In Proceedings of the 23rd International Conference on Neural Information Processing Systems, NIPS'10, USA, pp. 2496–2504. Curran Associates Inc.

Numerical results

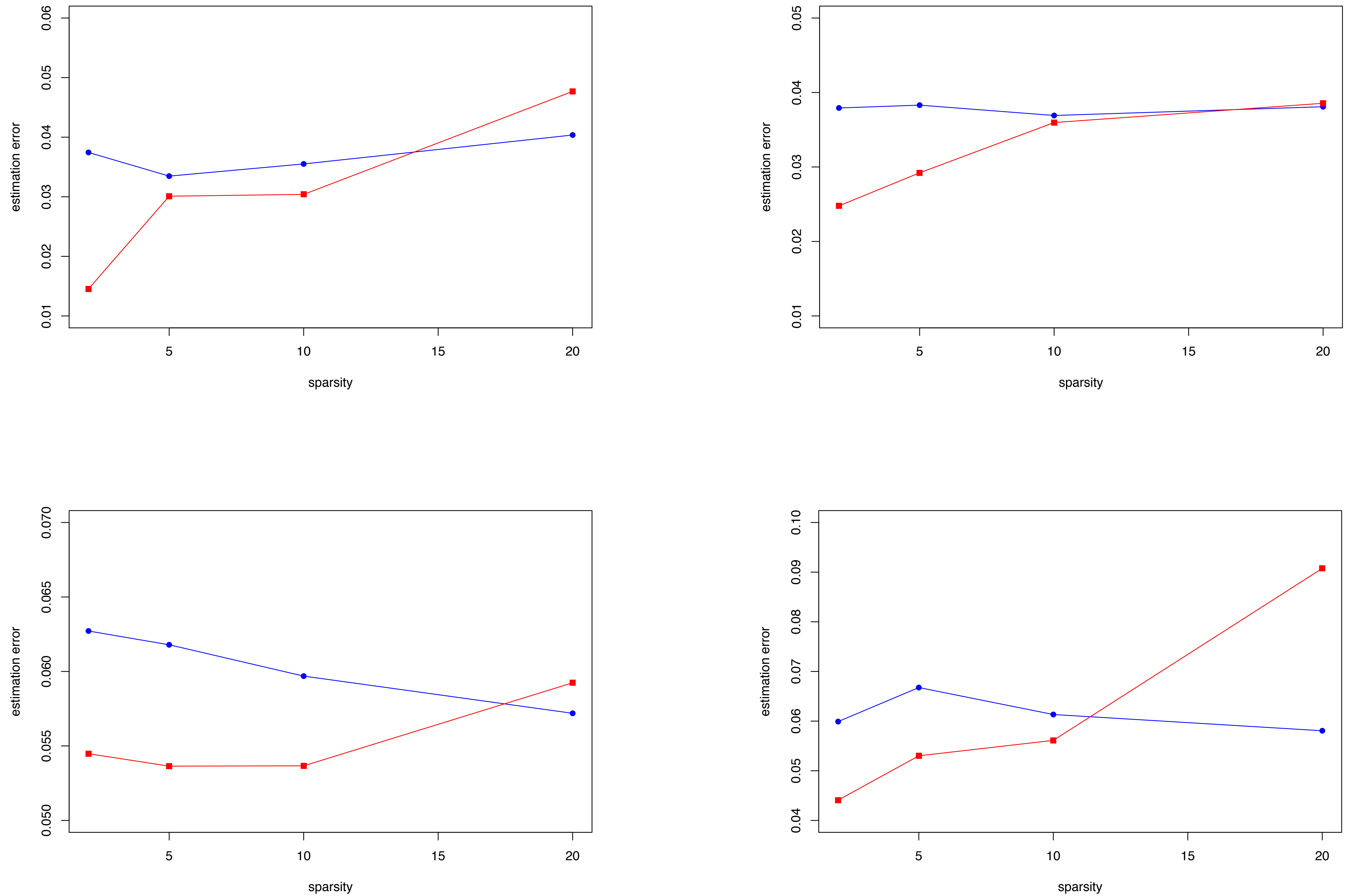
Imputation error (average across 100 rep)



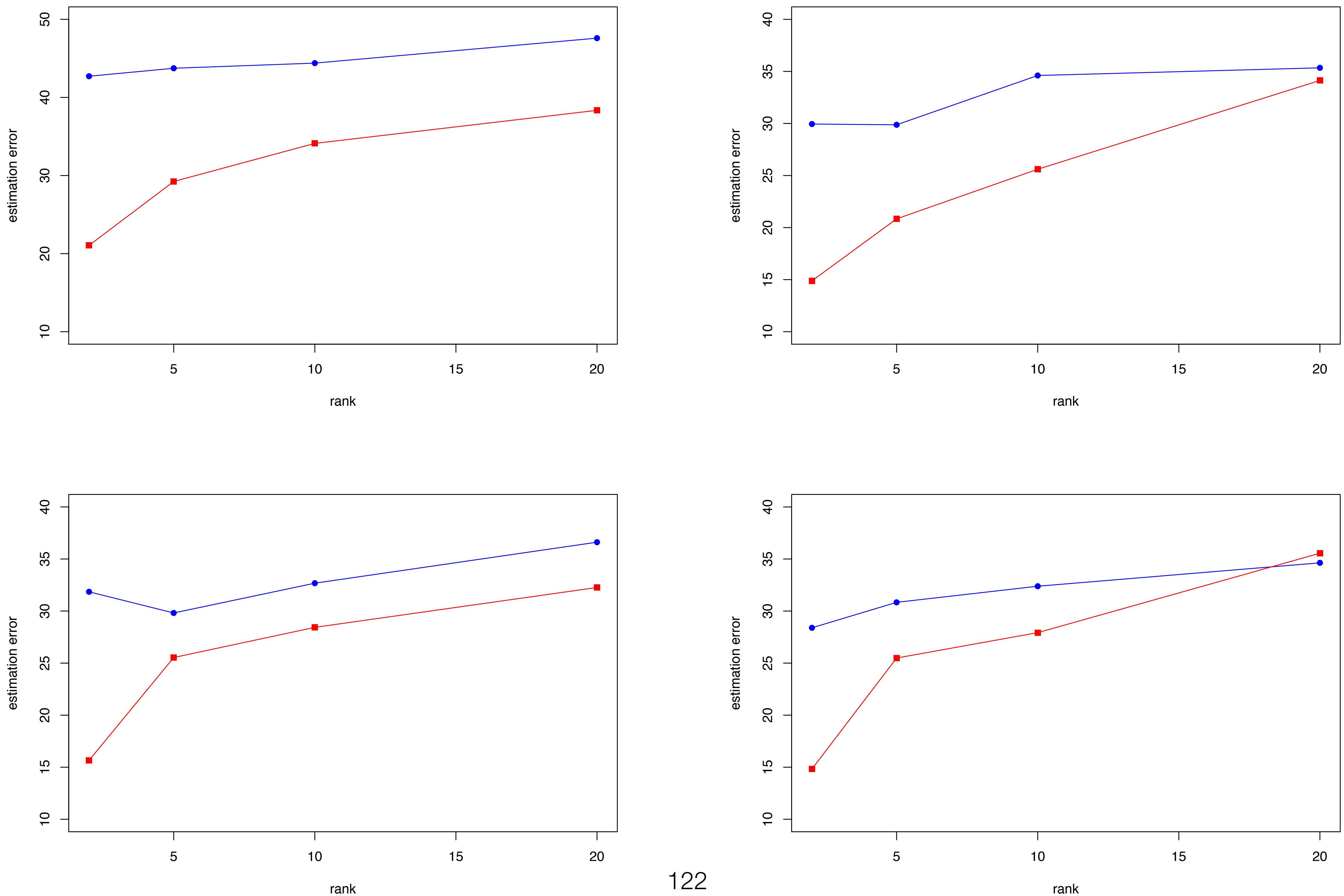
Computational time (average across 100 rep)

method	mean	mice	GLRM	FAMD	MLFAMD	softImpute	mimi
time(s)	1.7e-4	0.2	5.5	2.6	3.5	0.1	6.6

MIMI: estimation results (main effects)



MIMI: estimation results (interactions)



MCGD algorithm

$$(\hat{\alpha}, \hat{\mathbf{\Theta}}) \in \operatorname{argmin} \mathcal{L}(\alpha, \mathbf{\Theta}; \boldsymbol{Y}, \Omega) + \lambda_1 \|\mathbf{\Theta}\|_\star + \lambda_2 \|\alpha\|_1$$

Algorithm 1 MCGD algorithm

1: Initialize: — $\mathbf{\Theta}^{(0)}, \alpha^{(0)}, R^{(0)}$. E.g., $\mathbf{\Theta}^{(0)}, \alpha^{(0)}, R^{(0)} = (\mathbf{0}, \mathbf{0}, 0)$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3:     *// Update for* α *//*
    Compute proximal update using to obtain $\alpha^{(t)}$.
4:     *// Update for* $(\mathbf{\Theta}, R)$ *//*
    Compute the upper bound $R_{\mathsf{UB}}^{(t)} := \lambda_1^{-1} F(\alpha^{(t)}, \mathbf{\Theta}^{(t-1)}, R^{(t-1)})$.
5:     Compute the conditional gradient update direction, $(\hat{\mathbf{\Theta}}^{(t)}, \hat{R}^{(t)})$.
6: **end for**
7: **Return:** $\mathbf{\Theta}^{(T)}, \alpha^{(T)}, R^{(T)}$.

MCGD algorithm

$$(\hat{\alpha}, \hat{\mathbf{\Theta}}) \in \operatorname{argmin} \mathcal{L}(\alpha, \mathbf{\Theta}; \boldsymbol{Y}, \Omega) + \lambda_1 \|\mathbf{\Theta}\|_\star + \lambda_2 \|\alpha\|_1$$

upper bound on $\|\mathbf{\Theta}\|_\star$

Algorithm 1 MCGD algorithm

1: Initialize: — $\mathbf{\Theta}^{(0)}, \alpha^{(0)}, R^{(0)}$. E.g., $\mathbf{\Theta}^{(0)}, \alpha^{(0)}, R^{(0)} = (\mathbf{0}, \mathbf{0}, 0)$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: *// Update for α //*
Compute proximal update using to obtain $\alpha^{(t)}$.
4: *// Update for $(\mathbf{\Theta}, R)$ //*
Compute the upper bound $R_{\mathsf{UB}}^{(t)} := \lambda_1^{-1} F(\alpha^{(t)}, \mathbf{\Theta}^{(t-1)}, R^{(t-1)})$.
5: Compute the conditional gradient update direction, $(\hat{\mathbf{\Theta}}^{(t)}, \hat{R}^{(t)})$.
6: **end for**
7: **Return:** $\mathbf{\Theta}^{(T)}, \alpha^{(T)}, R^{(T)}$.

MCGD algorithm

$$(\hat{\alpha}, \hat{\boldsymbol{\Theta}}) \in \operatorname{argmin} \mathcal{L}(\alpha, \boldsymbol{\Theta}; \boldsymbol{Y}, \Omega) + \lambda_1 \|\boldsymbol{\Theta}\|_\star + \lambda_2 \|\alpha\|_1$$

Algorithm 1 MCGD algorithm

1: Initialize: — $\boldsymbol{\Theta}^{(0)}, \alpha^{(0)}, R^{(0)}$. E.g., $\boldsymbol{\Theta}^{(0)}, \alpha^{(0)}, R^{(0)} = (\mathbf{0}, \mathbf{0}, 0)$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: *// Update for α //*
Compute proximal update using to obtain $\alpha^{(t)}$.
4: *// Update for $(\boldsymbol{\Theta}, R)$ //*
Compute the upper bound $R_{\text{UB}}^{(t)} := \lambda_1^{-1} F(\alpha^{(t)}, \boldsymbol{\Theta}^{(t-1)}, R^{(t-1)})$.
5: Compute the conditional gradient update direction, $(\hat{\boldsymbol{\Theta}}^{(t)}, \hat{R}^{(t)})$.
6: **end for**
7: **Return:** $\boldsymbol{\Theta}^{(T)}, \alpha^{(T)}, R^{(T)}$.

gradient computation
+
soft-thresholding

MCGD algorithm

$$(\hat{\alpha}, \hat{\mathbf{\Theta}}) \in \operatorname{argmin} \mathcal{L}(\alpha, \mathbf{\Theta}; \boldsymbol{Y}, \Omega) + \lambda_1 \|\mathbf{\Theta}\|_\star + \lambda_2 \|\alpha\|_1$$

Algorithm 1 MCGD algorithm

1: Initialize: — $\mathbf{\Theta}^{(0)}, \alpha^{(0)}, R^{(0)}$. E.g., $\mathbf{\Theta}^{(0)}, \alpha^{(0)}, R^{(0)} = (\mathbf{0}, \mathbf{0}, 0)$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: `// Update for` α `//`
Compute proximal update using to obtain $\alpha^{(t)}$.
4: `// Update for` $(\mathbf{\Theta}, R)$ `//`
Compute the upper bound $R_{\mathsf{UB}}^{(t)} := \lambda_1^{-1} F(\alpha^{(t)}, \mathbf{\Theta}^{(t-1)}, R^{(t-1)})$.
5: Compute the conditional gradient update direction, $(\hat{\mathbf{\Theta}}^{(t)}, \hat{R}^{(t)})$.
6: **end for**
7: **Return:** $\mathbf{\Theta}^{(T)}, \alpha^{(T)}, R^{(T)}$.

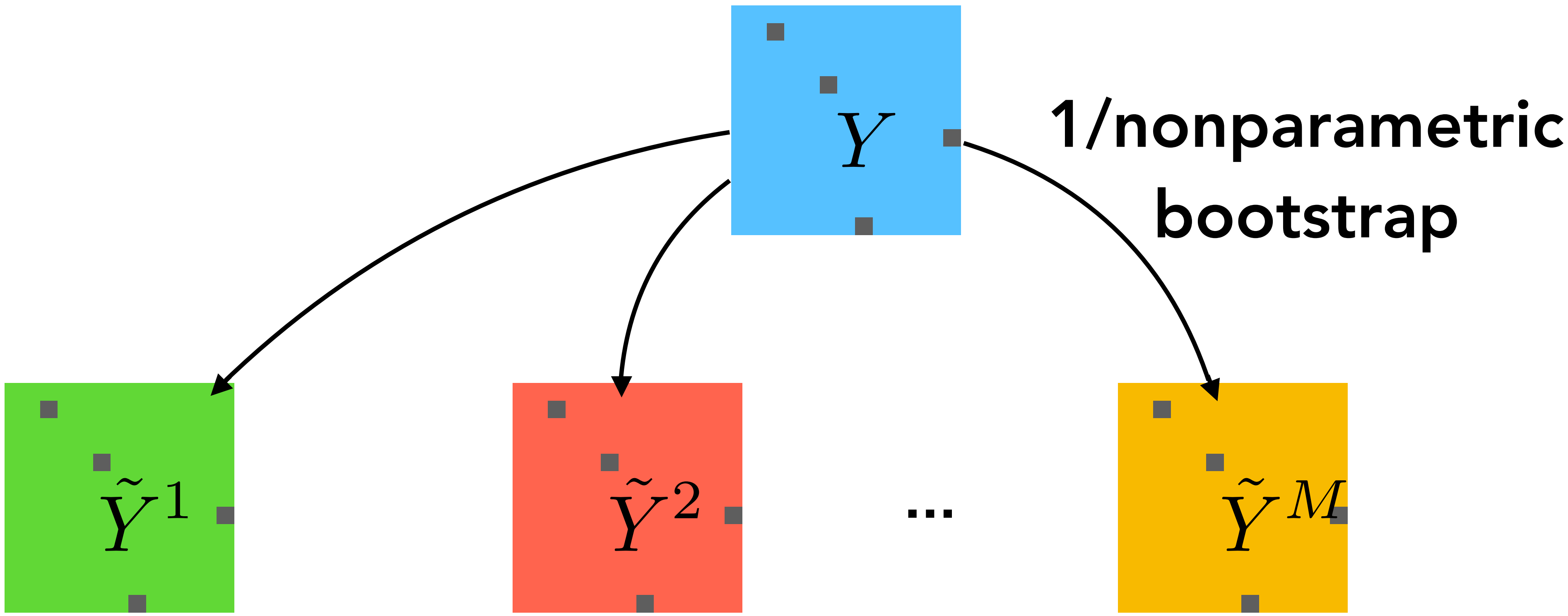
top singular vectors
+
soft-thresholding

Convergence of MCGD

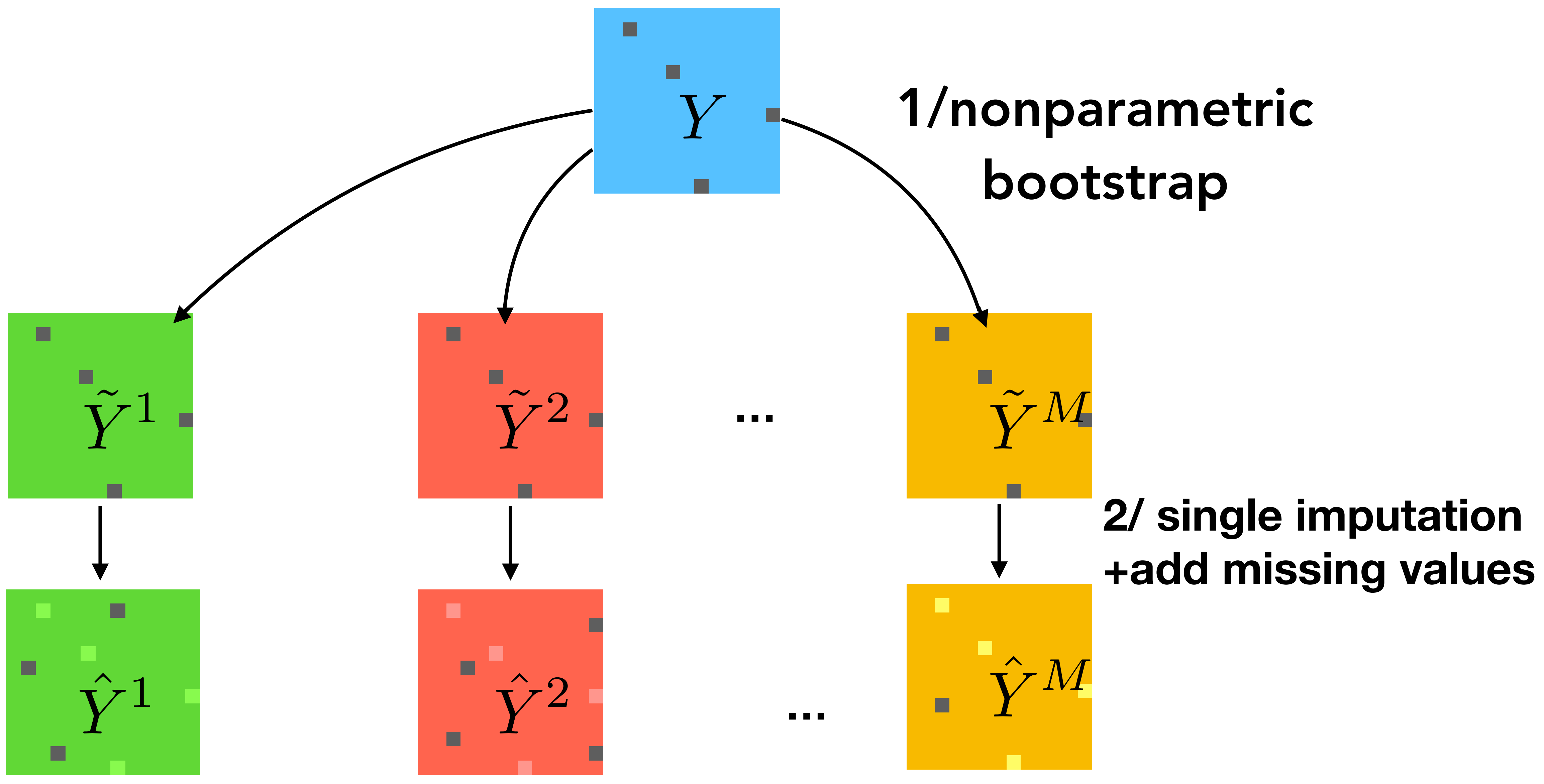
Theorem 1 (Robin et al. 2018)

The MCGD algorithm converges to an ϵ-solution in $\mathcal{O}(1/\epsilon)$ iterations

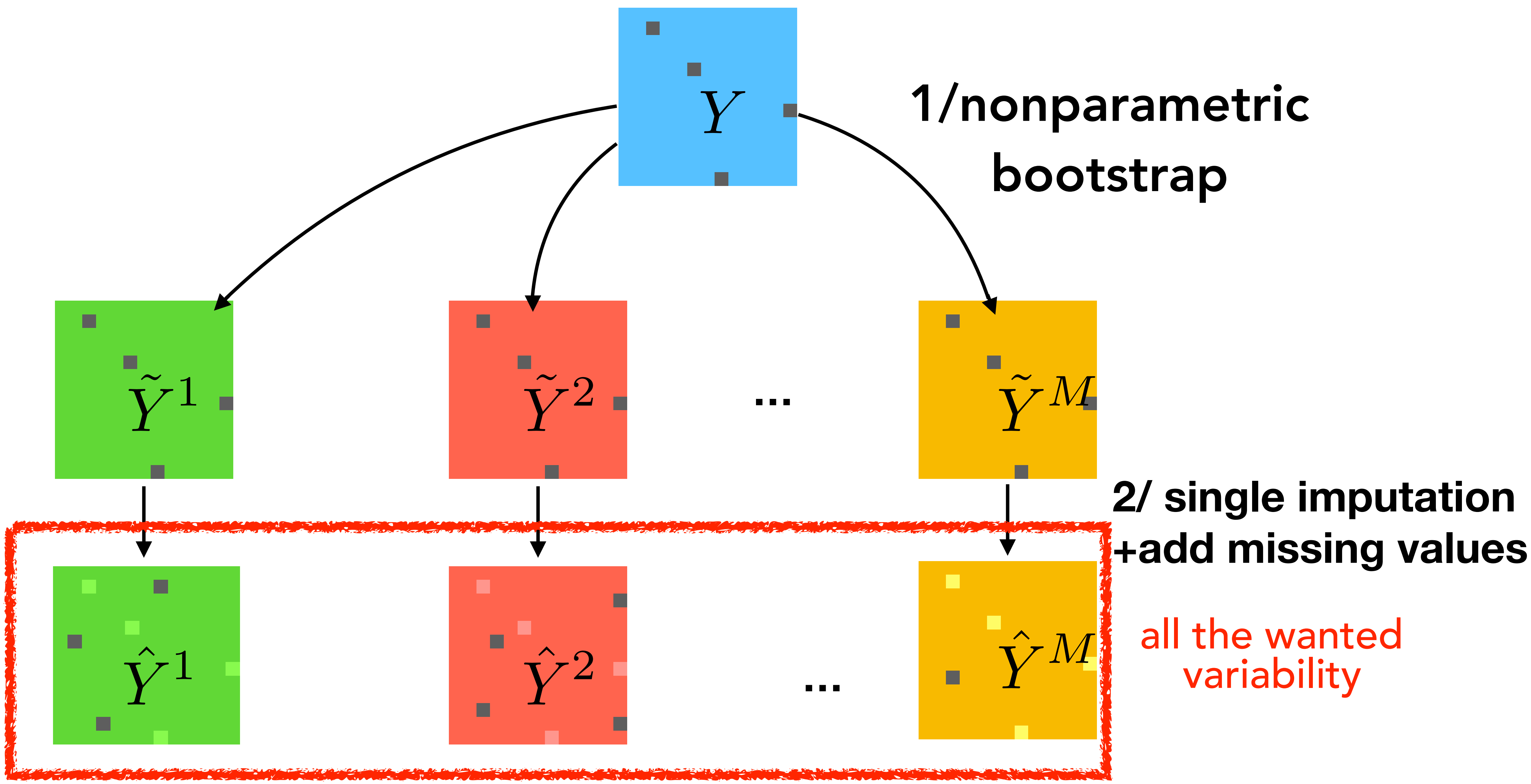
Variability of observations: fixed missing data pattern



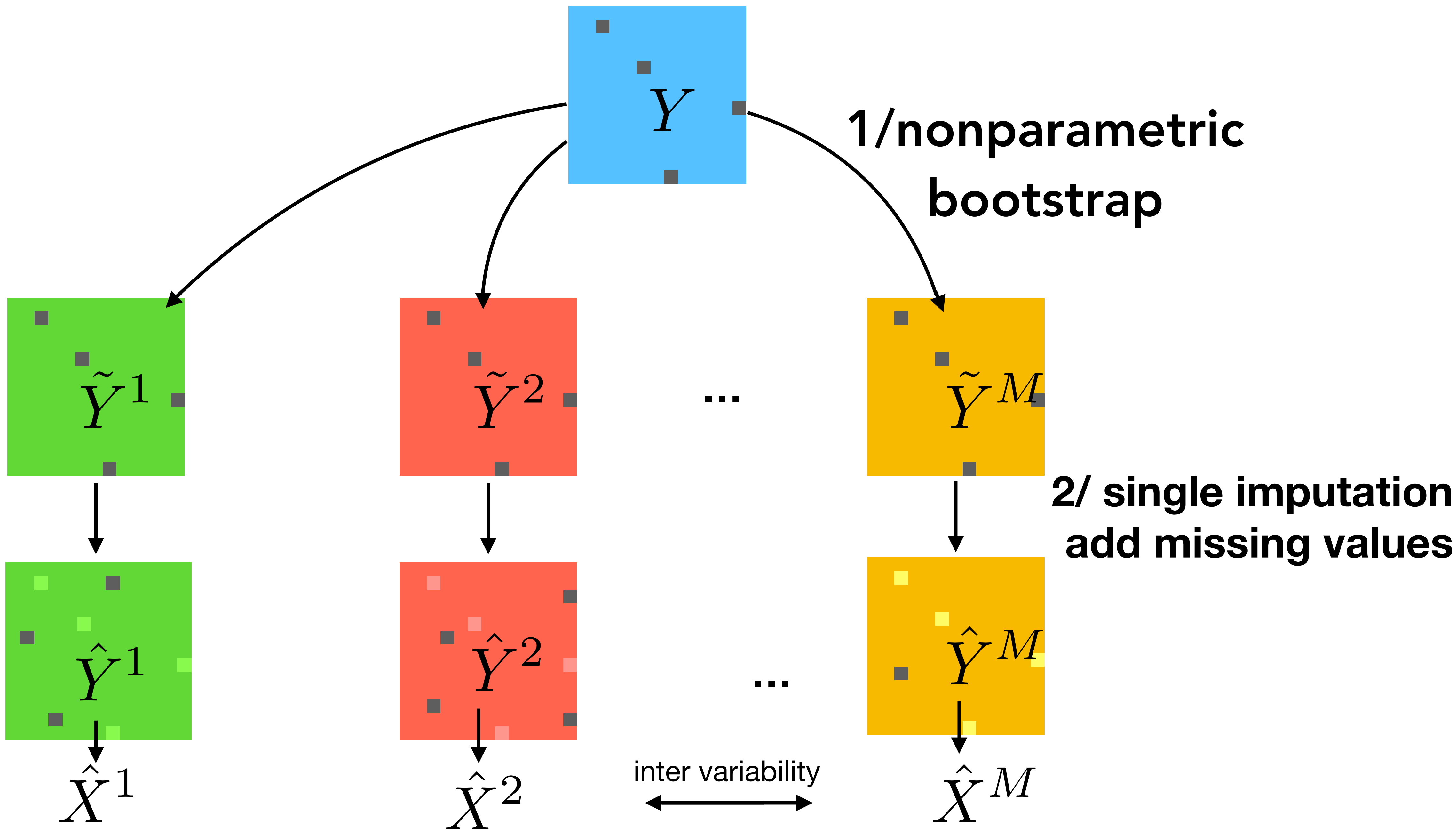
Variability of Observations & missing data pattern



Variability of Observations & missing data pattern



Variability between imputation models (Inter Variability)



Variability between imputations (Intra Variability)