Imputation and low-rank estimation with Missing Not At Random data

DAGStat Conference, München

Aude Sportisse [1,3]   Claire Boyer [1,2]   Julie Josse [2,4]

[1]Laboratoire de Probabilités Statistique et Modélisation, Sorbonne Université, France

[2]Département de Mathématiques et applications, Ecole Normale Supérieure, Paris, France

[3]Centre de Mathématiques Appliquées, Ecole Polytechnique, France

[4]XPOP, INRIA, France

21st March 2019

Classical definitions

- ⋆ $Y \in \mathbb{R}^{n \times p}$ the data matrix,
- ⋆ Y_{obs} the observed variables, Y_{mis} the missing variables,
- ⋆ $M \in \mathbb{R}^{n \times p}$ the missing-data pattern:

$$M_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

Toy example:

Realisations of Y and M: $y = \begin{pmatrix} 1 & 6 \end{pmatrix}$ and $\Omega = \begin{pmatrix} 0 & 1 \end{pmatrix}$.
We observe: $y = \begin{pmatrix} NA & 6 \end{pmatrix}$
Observed and missing variables: $Y_{\mathrm{obs}} = Y_2$ and $Y_{\mathrm{mis}} = Y_1$

- ⋆ **Issue: Cause of the missingness ?**
  [Rubin, 1976], [Little and Rubin, 2014]

Toy Example:

$$\begin{array}{cc} Y_1 & Y_2 \end{array}$$
$$\begin{pmatrix} 1 & 2 \\ 3 & 20 \\ 22 & 4 \end{pmatrix}$$

MCAR mechanism

$$p(M|Y;\phi) = p(M;\phi), \quad \forall Y, \phi.$$

ϕ: the unknown parameters of the missingness.

MAR mechanism

$$p(M|Y;\phi) = p(M|Y_{\mathrm{obs}};\phi), \quad \forall Y_{\mathrm{mis}}, \phi.$$

MNAR mechanism

Other case, i.e.

$$p(M|Y;\phi) = p(M|Y_{\mathrm{obs}}, Y_{\mathrm{mis}};\phi), \quad \forall \phi.$$

MCAR

$$\begin{array}{cc} Y_1 & Y_2 \end{array}$$
$$\begin{pmatrix} \mathrm{NA} & 2 \\ 3 & 20 \\ 22 & 4 \end{pmatrix}$$

MAR

$$\begin{array}{cc} Y_1 & Y_2 \end{array}$$
$$\begin{pmatrix} 1 & 2 \\ \mathrm{NA} & 20 \\ 22 & 4 \end{pmatrix}$$

MNAR

$$\begin{array}{cc} Y_1 & Y_2 \end{array}$$
$$\begin{pmatrix} 1 & 2 \\ 3 & 20 \\ \mathrm{NA} & 4 \end{pmatrix}$$

Motivating data in health

Traumabase: 15 000 patients/ 250 var/ 15 hospitals

Center	Age	Sex	Weight	Height	BMI	Lactates	Glasgow
Beaujon	54	m	85	NA	NA	NA	12
Lille	33	m	80	1.8	24.69	4.8	15
Pitie	26	m	NA	NA	NA	3.9	3
Beaujon	63	m	80	1.8	24.69	1.66	15
Pitie	30	w	NA	NA	NA	NA	15

- Aim: predict the Glascow score.
- MNAR case extremely frequent: *missingness of the patient's blood pressure due to his or her health condition.*

Model

$Y \in \mathbb{R}^{n \times p}$ noisy realisation of a low-rank matrix $\Theta \in \mathbb{R}^{n \times p}$:

$$Y = \Theta + \epsilon, \text{where} \begin{cases} \Theta \text{ with rank } r < \min\{n, p\}, \\ \epsilon_i \overset{\perp\!\!\!\perp}{\sim} \mathcal{N}(0_n, \sigma^2 I_{n \times n}), \forall i \in [1, n]\,. \end{cases}$$

⇢ Access only to the missing-data matrix

$$Y \odot \Omega,$$

with $\odot$ the Hadamard product.

Imputation and low-rank estimation issues :

- How to estimate Θ ?
- How to impute the unknown entries of Y ?

Model

Data distribution

"$Y \sim \mathcal{N}(\Theta, \sigma^2)$" entry by entry, i.e. $\forall i \in [1, n]$, $\forall j \in [1, p]$:

$$p(y_{ij}; \Theta_{ij}) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2}\left(\frac{y_{ij} - \Theta_{ij}}{\sigma}\right)^2\right).$$

$\rightsquigarrow$ σ^2 is supposed to be known.

MNAR missing-data mechanism via a Logistic Model

$\forall i \in [1, n]$, $\phi_j = (\phi_{1j}, \phi_{2j})$ denoting a parameter vector:

$$p(\Omega_{ij}|y_{ij}; \phi) = [(1 + e^{-\phi_{1j}(y_{ij} - \phi_{2j})})^{-1}]^{(1-\Omega_{ij})}[1 - (1 + e^{-\phi_{1j}(y_{ij} - \phi_{2j})})^{-1}]^{\Omega_{ij}}$$

$\rightsquigarrow$ self-masked MNAR missing-data mechanism: the lack of a data only depends on the value itself.

Likelihood approach

- Θ: unknown parameter.
- Likelihood-approach: maximizing the joint log-likelihood

$$\ell(\Theta, \phi; y; \Omega) = p(y; \Theta) p(\Omega | y; \phi)$$

- Missing data: basing the statistical inference on the observed joint log-likelihood:

$$\ell(\Theta, \phi; y_{\mathrm{obs}}, \Omega) = \int \ell(\Theta, \phi; y, \Omega) dy_{\mathrm{mis}}$$

✓ In the MAR setting, one can ignore the mechanism since

$$p(\Omega | y; \phi) = p(\Omega | y_{\mathrm{obs}}; \phi).$$

$$\Rightarrow \ell(\Theta, \phi; y_{\mathrm{obs}}, \Omega) \propto \ell(\Theta; y_{\mathrm{obs}}) = \int \ell(\Theta; y) dy_{\mathrm{mis}}.$$

✗ In the MNAR setting, one should consider the mechanism.

Methods

- (Naive method) Ignoring the MNAR mechanism and applying classical methods.
- (New method) Modelling the MNAR mechanism.
- (New method) Implicitly modelling the MNAR mechanism by adding the mask to the data matrix and applying classical methods,

Classical approach with MAR

MAR: maximize the observed penalized log-likelihood

$\rightsquigarrow$ the missing-data mechanism is ignorable.

$$\hat{\Theta} \in \operatorname{argmin}_{\Theta} \|(Y - \Theta) \odot \Omega\|_F^2 + \lambda \|\Theta\|_\star,$$

Classical algorithm: iterative soft-thresholding algorithm (ISTA) of the singular value decomposition:

- `softImpute` [Hastie et al., 2015],
- **its accelerated version: FISTA** [Beck and Teboulle, 2009].

Equivalence with an Expectation-Maximization algorithm:

$\rightsquigarrow$ maximizing $l(\Theta; y_{\mathrm{obs}})$, integral which has no closed form.

- **E-step:** $Q(\Theta|\hat{\Theta}^{(t)}) = -\mathbb{E}_{Y_{\mathrm{mis}}}[\ell(\Theta; y)|Y_{\mathrm{obs}}; \hat{\Theta} = \hat{\Theta}^{(t)}]$.
- **M-step:** $\hat{\Theta}^{(t+1)} \in \operatorname{argmin}_{\Theta} Q(\Theta|\hat{\Theta}^{(t)}) + \lambda \|\Theta\|_\star$.

Modelling the missing-data mechanism

$\rightsquigarrow$ maximizing $\ell(\Theta, \phi; y_{\mathrm{obs}}, \Omega) = \int p(y;\Theta) p(\Omega|y;\phi) dy_{\mathrm{mis}}$.

EM algorithm [S., Boyer, Josse 2018]

- **E-step:**

$$Q(\Theta, \phi|\hat{\Theta}^{(t)}, \hat{\phi}^{(t)}) = -\mathbb{E}_{Y_{\mathrm{mis}}}\left[\ell(\Theta, \phi; y, \Omega)|Y_{\mathrm{obs}}, M; \Theta = \hat{\Theta}^{(t)}, \phi = \hat{\phi}^{(t)}\right]$$

- **M-step:**

$$\hat{\Theta}^{(t+1)}, \hat{\phi}^{(t+1)} \in \mathrm{argmin}_{\Theta,\phi}\, Q(\Theta, \phi|\hat{\Theta}^{(t)}, \hat{\phi}^{(t)}) + \lambda\|\Theta\|_\star$$

- ▶ E-step: Monte-Carlo approximation and SIR algorithm.
- ▶ M-step: Separability of Q:
  - Θ: `softImpute`, FISTA.
  - ϕ: Newton-Raphson algorithm.

✓ Handling MNAR data (under a self-masked logistic model).

✗ Computationally costly.

Implicitly modeling

Adding the mask to the data matrix

$$\hat{\Theta} \in \mathrm{argmin}_{\Theta} \frac{1}{2} \| [\Omega \odot Y | \Omega] - [\Omega | \mathbf{1}] \odot [\Theta | \Omega] \|_F^2 + \lambda \|\Theta\|_{\star},$$

where $\mathbf{1} \in \mathbb{R}^{n \times p}$ denotes the matrix with all values equal to 1, and $[X_1 | X_2]$ denotes the column-concatenation of matrices X_1 and X_2.

- `softImpute`, FISTA.
- taking into account the mask binary type, with a Penalized Iteratively Reweighted Least Squares algorithm [Robin et al., 2018].

✗ No theoretical modeling.

✓ Computationally efficient.

Numerical experiments

- $\star$ $N = 50$ simulations.

Measuring the performance: normalized Mean Square Errors (MSE)

Prediction error:

$$\mathbb{E}\left[\left\|(\hat{\Theta} - Y) \odot (1 - \Omega)\right\|_F^2\right] \Big/ \ \mathbb{E}\left[\|Y \odot (1 - \Omega)\|_F^2\right]$$

Total error:

$$\mathbb{E}\left[\left\|\hat{\Theta} - \Theta\right\|_F^2\right] \Big/ \ \mathbb{E}\left[\|\Theta\|_F^2\right]$$

Bivariate MNAR missing data

- ⋆ $n = 100$, $p = 50$, rank $r = 4$, $\sigma^2 = 0.8$
- ⋆ 2 missing variables, 1.5% missing values in the whole matrix.

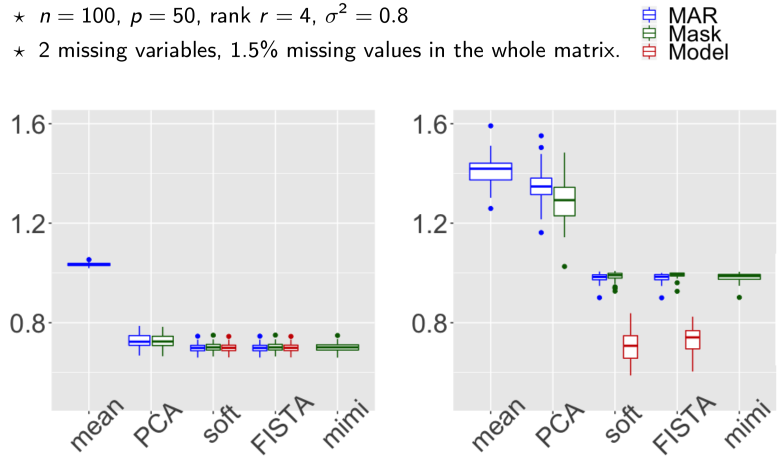

Figure: Total error (left), prediction error (right).

Multivariate MNAR missing data

- ⋆ $n = 100$, $p = 20$, rank $r = 4$, $\sigma^2 = 0.8$
- ⋆ 10 missing variables, 25% missing values in the whole matrix.

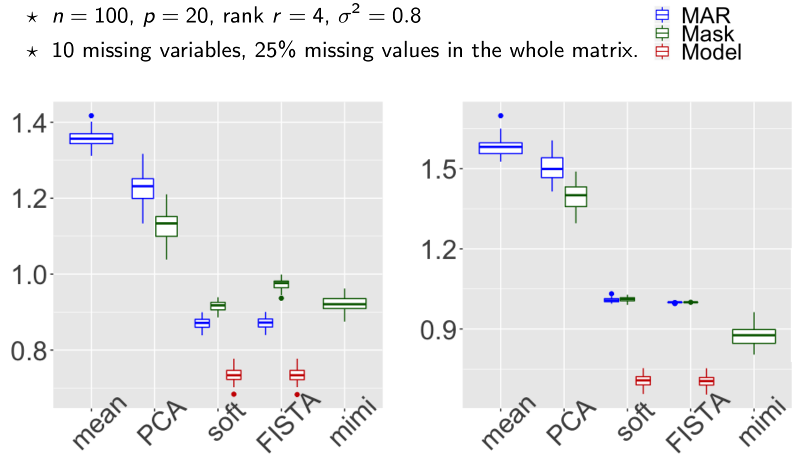

Figure: Total error (left), prediction error (right).

Conclusion

[Sportisse et al., 2018]

Take-home messages:

- Dealing with MNAR data in a low-rank context.
- Two methods: explicit modeling of the mechanism, implicit consideration by adding the mask.
- few missing variables $\Rightarrow$ model-based.
- many variables are missing $\Rightarrow$ add the mask and model it with a binomial distribution.

On-going work:

☼ Confidence interval.

☼ Variational EM.

References I

Beck, A. and Teboulle, M. (2009).
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
SIAM journal on imaging sciences, 2(1):183–202.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015).
Matrix completion and low-rank svd via fast alternating least squares.
The Journal of Machine Learning Research, 16(1):3367–3402.

Little, R. J. and Rubin, D. B. (2014).
Statistical analysis with missing data, volume 333.
John Wiley & Sons.

References II

Robin, G., Klopp, O., Josse, J., Moulines, É., and Tibshirani, R. (2018).
Main effects and interactions in mixed and incomplete data frames.
arXiv preprint arXiv:1806.09734.

Rubin, D. B. (1976).
Inference and missing data.
Biometrika, 63(3):581–592.

Sportisse, A., Boyer, C., and Josse, J. (2018).
Imputation and low-rank estimation with missing non at random data.
arXiv preprint arXiv:1812.11409.

Process time

with a processor Intel Core i5 of 2,3 GHz

For estimating one matrix Θ when 50% of the variables are missing:

- 0.0549 seconds for the MAR method with `softImpute`,
- 3.215 seconds for the implicit method with `mimi`,
- 13.069 minutes for the model-based method with `softImpute`.