

ESTIMATION AVEC DES DONNÉES INCOMPLÈTES INFORMATIVES DANS UN CAS DE FAIBLE RANG

Aude Sportisse ^{1,2}, Claire Boyer ^{1,3} & Julie Josse ^{2,4}

¹ *Laboratoire de Probabilités Statistique et Modélisation, Sorbonne Université, 4 place
Jussieu, 75252 Paris,*

² *Centre de Mathématiques Appliquées, Ecole Polytechnique,*

³ *Département de Mathématiques et Applications, Ecole Normale Supérieure, Paris*

⁴ *XPOP, Inria Saclay, Bâtiment Turing, 1 Rue Honoré d'Estienne d'Orves, 91120
Palaiseau.*

Résumé. Pour traiter les données manquantes, la complétion de matrices basée sur des modèles de faible rang est récemment devenue très populaire puisqu'elle s'appuie sur des garanties algorithmiques, méthodologiques et théoriques. Cependant, les méthodes existantes ne s'appliquent pas au cas de valeurs manquantes informatives de type MNAR (Missing Not At Random). Le cas MNAR nécessite de spécifier la distribution du mécanisme de données manquantes et donc d'avoir un fort a priori sur les causes du manque. Par conséquent, cette approche est difficile à mettre en oeuvre en pratique.

Nous reformulerons le problème de données manquantes dans une matrice de faible rang à l'aide de modèles graphiques à variables latentes. Ce nouveau cadre établi nous permet d'obtenir des garanties théoriques d'estimation d'espérance d'une variable avec données manquantes informatives de type MNAR auto-masqué, sans avoir à modéliser le mécanisme de données manquantes. Nous présentons aussi une méthode pour l'estimation de la variance et des covariances liées d'une telle variable. Nous proposons également une méthode d'imputation et d'estimation de la matrice paramètre de faible rang. Nous comparerons notre proposition avec les méthodes classiques, comme les algorithmes itératifs de décomposition en valeurs singulières qui ignore le mécanisme de données manquantes, ou la méthode qui le modélise, le plus souvent grâce à un modèle paramétrique de régression logistique.

Mots-clés. modèles graphiques, modèle de faible rang, données manquantes, données manquantes informatives, complétion de matrice, variables latentes.

Abstract. For handling missing values, matrix completion based on low-rank models has recently become very popular since it ensures algorithmic, methodological and theoretical guarantees to deal with missing values. However, existing methods do not consider the case of informative missing values of type MNAR (Missing Not At Random) which

{aude.sportisse, claire.boyer}@sorbonne-universite.fr,
julie.josse@polytechnique.edu

are widely encountered in practice. The MNAR case requires specifying the distribution of the missing data mechanism and therefore having a strong a priori on the lack causes. This approach is thus difficult to implement in practice.

The missing values problem of a low-rank matrix will be positioned within the graphical models with latent variables framework. This new established framework allows us to obtain theoretical guarantees for estimating the expectation of a variable with missing data of the self-masked MNAR type, without having to model the missing data mechanism. We also present a method for the estimation of the variance and the related covariances for such a variable. We propose then a method for imputation and estimation of the low rank parameter matrix. We will compare our proposal with the classical methods, such as the iterative singular values decomposition algorithms which ignores the mechanism, or the method that models it, most often using a parametric logistic regression model.

Keywords. graphical models, low-rank model, missing values, informative missing values, matrix completion, latent variables.

1 Introduction

Les données manquantes représentent un véritable enjeu en science des données. Pour traiter des jeux de données incomplets, une solution classique est d'imputer les données manquantes. Les garanties théoriques et ainsi le choix des méthodes d'imputation reposent généralement sur des hypothèses concernant les causes de l'absence de données. Rubin (1976) a décomposé ces causes en trois types de données manquantes. Les données sont dites (i) manquantes complètement aléatoirement (MCAR pour *Missing Completely At Random*), si la probabilité qu'une observation soit manquante est la même pour toutes les observations ; (ii) manquante aléatoirement (MAR pour *Missing At Random*), si l'absence d'une observation dépend seulement des variables observées ; et (iii) manquante non aléatoirement (MNAR pour *Missing Not At Random*), si l'absence d'une observation peut dépendre de toutes les variables (observées et manquantes). Dans cet exposé, nous nous focaliserons sur les données manquantes de type MNAR, et en particulier de type auto-masqué, c'est-à-dire dont l'absence ne dépend que de sa valeur elle-même. Par exemple, dans un questionnaire, l'absence d'observations sur une variable représentant le salaire peut s'expliquer par la réticence des personnes riches à répondre.

Lorsque les données manquantes sont de type MCAR ou MAR, l'inférence statistique peut être réalisée en ignorant la cause de l'absence de données (Rubin et Little (2014)). Dans le cas MNAR, les données observées ne sont souvent pas représentatives de la population. Pour éviter un biais dans l'estimation, il est généralement nécessaire de prendre en compte la distribution particulière des données manquantes. Cela conduit à faire des hypothèses paramétriques comme la modélisation par une régression logistique et à appliquer dans la plupart des cas des algorithmes qui peuvent être coûteux en temps de calcul.

Récemment, Mohan, Thoemmes et Pearl (2018) ont présenté une nouvelle approche par modèles graphiques dans le cas spécifique du modèle linéaire avec une variable manquante de type MNAR auto-masqué. Dans ce contexte, ils ont montré que l’espérance et la variance peuvent être estimées de manière consistante en utilisant des régressions linéaires basées sur les données observées seulement.

2 Modèle

Dans cet exposé, nous considérons le modèle de faible rang, pour lequel les méthodes d’imputation ont été démontrées performantes pour des données de type MCAR ou MAR (Josse, Sardy et Wager (2016)).

Plus particulièrement, nous considérons un modèle de faible rang, via le formalisme de l’ACP probabiliste avec variables latentes. Supposons que nous observons la matrice centrée $Y \in \mathbb{R}^{n \times p}$ qui peut se décomposer à l’aide d’une matrice de coefficients $B \in \mathbb{R}^{p \times r}$ et r variables latentes regroupées dans la matrice $W \in \mathbb{R}^{n \times r}$, donnant ainsi le modèle à effets aléatoires suivant :

$$Y = \alpha + WB^T + \epsilon, \text{ avec } \begin{cases} W = (W_1 | \dots | W_n)^T, \text{ avec } W_i \sim \mathcal{N}(0_r, I_{r \times r}) \text{ de dimension } r, \\ B \text{ de rang } r, \\ \alpha \in \mathbb{R}^{n \times p} \text{ dont les lignes sont égales : } \alpha_1. = \dots = \alpha_n., \\ \epsilon = (\epsilon_1 | \dots | \epsilon_n)^T, \text{ avec } \epsilon_i \sim \mathcal{N}(0_p, \sigma^2 I_{p \times p}) \text{ de dimension } p. \end{cases} \quad (1)$$

pour σ^2 et r connus.

On peut alors représenter le modèle (1) à l’aide de modèles graphiques, et obtenir ainsi une équivalence graphique entre les modèles de faible rang et les modèles linéaires, cf Figure 1 pour un exemple jouet.

3 Contributions

En utilisant les équivalences de modèles graphiques, nous étendons les résultats récents de Mohan et al. (2018) d’estimation consistante avec données manquantes informatives de type MNAR auto-masqué au cas de modèle de faible rang. À notre connaissance, ce résultat est le premier à garantir une estimation consistante sans faire des hypothèses paramétriques supplémentaires dans un modèle de faible rang avec données manquantes informatives.

Nous proposons également une méthode d’estimation de la variance des variables manquantes de type MNAR et des covariances qui leur sont liées. En découlent alors une méthode d’estimation des paramètres B , ainsi qu’une méthode d’imputation de Y .

Nous comparons numériquement cette approche avec (i) le cas naïf de suppression de toutes les observations qui contiennent des variables manquantes, (ii) la méthode

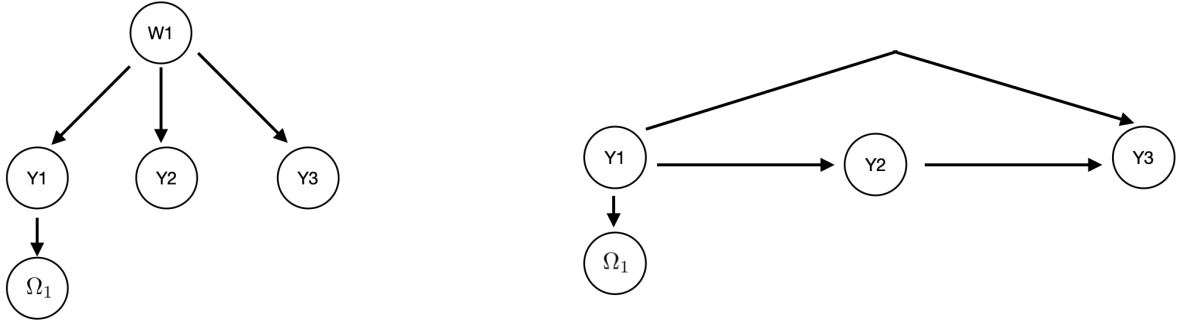


FIGURE 1 – Modèles graphiques causaux pour un exemple jouet avec $p = 3$, $r = 1$ et une variable manquante (Y_1) de type MNAR auto-masqué : ces deux modèles graphiques sont équivalents (Pearl (2003)), celui de gauche correspond à notre approche de modèle de faible rang (1) et celui de droite à une approche de modèle linéaire.

de complétion classique utilisant l’algorithme itératif de décomposition en valeurs singulières, et (iii) une méthode paramétrique modélisant le mécanisme de données manquantes de type MNAR auto-masqué proposée dans Sportisse et al. (2018). Les simulations numériques laissent à penser que la méthode proposée ici donne d’aussi bons résultats pour l’estimation de la moyenne et de la matrice de covariance que la méthode paramétrique. Notons que cette dernière peut être coûteuse numériquement, ce qui rend la méthode proposée ici d’autant plus compétitrice.

En éprouvant numériquement la stabilité au modèle de notre méthode, la séparation entre les hypothèses de données manquantes de type MAR ou MNAR peut être à nuancer dès lors que des matrices sont de rang faible et que le bruit est négligeable.

Bibliographie

- Julie Josse, Sylvain Sardy, and Stefan Wager. denoiser : A package for low rank matrix estimation. *Journal of Statistical Software*, 2016.
- Karthika Mohan, Felix Thoemmes, and Judea Pearl. Estimation with incomplete data : The linear case. In *IJCAI*, pages 5082–5088, 2018.
- Judea Pearl. Causality : models, reasoning, and inference. *Econometric Theory*, 19(675-685) :46, 2003.
- Donald B Rubin. Inference and missing data. *Biometrika*, 63(3) :581–592, 1976.
- Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 333. John Wiley & Sons, 2014.
- Aude Sportisse, Claire Boyer, and Julie Josse. Imputation and low-rank estimation with missing non at random data. *arXiv preprint arXiv :1812.11409*, 2018.