Journal de la Société Française de Statistique
Vol. 159 No. 2 (2018)

Reconstruction automatique de formulaires d'enquête médicale sur la culture de sécurité des patients par une méthode de factorisation matricielle bayésienne

Title: Bayesian matrix factorization for reconstruction of removed items in a medical survey

B. Don Bosco Diatta[1], Papa Ngom[1], Bastien Boussat[2] et Olivier François[2]

Résumé : Certaines enquêtes de santé publique souffrent d'un problème d'acceptabilité auprès des personnes interrogées, en particulier à cause de la longueur des questionnaires. Pour aborder ce problème, nous proposons de réduire délibérément les questionnaires en les individualisant de manière randomisée. Afin de compléter automatiquement les questionnaires incomplets générés par cette procédure, nous considérons un modèle de factorisation matricielle bayésienne. Pour estimer les paramètres de ce modèle, nous proposons un algorithme combinant un échantillonneur de Gibbs et une approche variationnelle. En utilisant les résultats d'une enquête portant sur la culture de sécurité des patients réalisée au centre hospitalier universitaire de Grenoble auprès de 3888 travailleurs médicaux, nous comparons les performances de notre méthode à plusieurs approches classiques en santé publique, à la méthode des forêts aléatoires, ainsi qu'à trois autres méthodes de factorisation matricielle. L'erreur de reconstruction de notre algorithme est inférieure à celle des autres algorithmes lorsque la proportion d'items supprimés est supérieure à 40%. Lorsque la proportion d'items supprimés est moins élevée, les histogrammes des lois marginales sont reconstruits de manière satisfaisante. Pour ce second critère, la méthode des forêts aléatoires est la plus performante. En général, nos résultats suggèrent que des enquêtes médicales similaires à celle réalisée pour cet article pourraient réduire substantiellement le nombre de questions posées à chaque travailleur avec une perte d'information limitée pour l'interprétation des résultats.

Abstract: Some public health surveys suffer from an issue of acceptability among respondents leading to a low response rate. To address this problem, we propose to reduce the length of individual questionnaires by randomly removing items. In order to impute the missing data, we introduce a Bayesian model for data imputation based on non-negative matrix factorization. We propose an inference algorithm combining a Gibbs sampler algorithm and a variational approach. Using the results of a survey on patient safety culture conducted at Grenoble University Hospital, we compare the performance of our new method with several classical approaches, with a random forest method, and with three additional matrix factorization methods. The average reconstruction error is lower than for other methods when the proportion of removed items is high (greater than 40%). With lower proportions of removed items (lower than 40%), the histograms of the marginal distributions are reconstructed satisfactorily. In this respect, the best performances were obtained with the random forest approach. Overall, our results suggest that similar surveys could be carried out by substantially reducing the number of questions asked to each worker with limited loss of information and interpretation.

[1] Laboratoire de Mathématiques Appliquées, Faculté des Sciences et Techniques, Université Cheikh Anta Diop BP 5005, Dakar-Fann (Sénégal).
E-mail : badiassiatta.diatta@ucad.edu.sn et E-mail : papa.ngom@ucad.edu.sn

[2] Université Grenoble Alpes, Centre National de la Recherche Scientifique, Centre Hospitalier Universitaire, Unité d'Evaluation Médicale, Laboratoire TIMC-IMAG UM5525, 38000 Grenoble (France).
E-mail : bboussat@chu-grenoble.fr et E-mail : olivier.francois@univ-grenoble-alpes.fr

 ISSN: 2102-6238

Mots-clés : enquête médicale, données manquantes, reconstruction matricielle, modèle Poisson bayésien, échantillonnage de Gibbs
Keywords: medical survey, missing data, matrix completion, Bayesian Poisson model, Gibbs sampler
Classification AMS 2000 : 62-07, 62P10, 92C60

1. Introduction

Plusieurs milliers d'hôpitaux dans le monde évaluent la culture de la sécurité des patients chez leur personnel. Les instruments d'auto-évaluation en termes de qualité et de sécurité des systèmes de santé s'appuient sur des questionnaires dont les données peuvent être incomplètes par le fait de perte, de refus de réponse, ou d'impossibilité d'acquisition (Etchegaray et Thomas, 2012; Kemp *et al.*, 2016). L'enquête type repose sur un instrument appelé HSOPSC (*Hospital Survey on Patient Safety Culture*) développé en 2004 par l'agence américaine AHRQ (*Agency for Healthcare Research and Quality*) (Sorra et Nieva, 2004). Cette enquête a été élaborée sur la base d'une revue de la littérature, raffinée selon la théorie psychométrique et soutenue par des analyses psychométriques effectuées en 2004 sur un personnel de 1437 agents travaillant dans 21 hôpitaux américains. Certaines méthodes d'analyse du questionnaire ne tiennent pas compte des données manquantes, selon la méthode de calcul des scores proposée par l'agence AHRQ (Sorra et Nieva, 2004). De telles approches tendent à réduire le nombre de données et la puissance des enquêtes, tout en favorisant l'introduction de biais (Rotnitzky et Wypij, 1994; Demissie *et al.*, 2003; Joseph *et al.*, 2004).

Le questionnaire HSOPSC est conçu pour évaluer la culture de la sécurité des patients du point de vue du personnel hospitalier. Il est auto-administré et comporte quarante-deux items . Bien que les propriétés psychométriques du questionnaire soient bien caractérisées, l'enquête souffre d'un problème d'acceptabilité et peut posséder un taux élevé de non-réponse. Pour aborder ce problème, nous envisageons de réduire considérablement le nombre d'items par participant en individualisant le questionnaire. Nous proposons alors de reconstituer l'intégralité du questionnaire de manière automatique. Dans cette étude, nous évaluons la reconstruction automatique de versions individualisées et aléatoirement réduites de la version française du questionnaire HSOPSC (Occelli *et al.*, 2013). Pour cela, nous utilisons les données issues d'une enquête effectuée à l'hôpital universitaire de Grenoble comportant un grand nombre de réponses avec un très faible taux de données manquantes. Nous simulons la réduction du questionnaire en supprimant des items au hasard, suivant une prodédure de type *MCAR (missing completely at random)* (Rubin, 1976; Little et Rubin, 2002). Ce mécanisme permet de mimer l'individualisation du questionnaire lors d'enquêtes futures. Techniquement, la reconstruction du questionnaire revient à résoudre un problème d'imputation de données manquantes (reconstruction matricielle) lors que le nombre de données manquantes est important.

Plusieurs méthodes ont été proposées pour répondre à la problématique des données manquantes en général, et particulièrement dans les enquêtes médicales (Shrive *et al.*, 2006). Dans notre étude, nous comparons les résultats de plusieurs méthodes de reconstruction. Des stratégies naïves, telles que les méthodes d'imputation par la moyenne, la médiane ou *locf (last observation carried forward)* complètant les données manquantes à partir de la dernière valeur observée, seront considérées comme les méthodes de référence. D'autres méthodes plus élaborées reposent sur des algorithmes statistiques. Nous implémentons un algorithme *EM* pour un modèle de don-

 ISSN: 2102-6238

nées manquantes de type Poisson-Gamma (Dempster *et al.*, 1977; Ghomrawi *et al.*, 2011). Nous utilisons la méthode *MissForest* s'appuyant sur les forêts aléatoires (Breiman, 2001; Stekhoven et Bühlmann, 2011; Waljee *et al.*, 2013). Nous comparons ces méthodes à la méthode d'analyse factorielle *mimca (Multiple Imputation with Multiple Correspondence Analysis)*, s'appuyant sur l'analyse des correspondances multiples (Josse et Husson, 2016; Audigier *et al.*, 2017) et à une méthode de factorisation matricielle *wnmf (Weighted Nonnegative Matrix Factorization)* proposée par Kim et Choi (2009). À notre connaissance, l'application de modèles à la reconstruction automatique d'enquêtes médicales n'a pas été envisagée auparavant, et notre approche se veut novatrice dans ce domaine. De plus, nous proposons une nouvelle méthode fondée sur un modèle poissonnien pour la factorisation de matrices positives. L'algorithme, que nous nommerons *pgnmf (Poisson-Gamma Nonnegative Matrix Factorization)*, correspond à une méthode d'analyse factorielle bayésienne et s'appuie sur l'imputation multiple.

Cet article est structuré de la manière suivante. Dans la section 2, nous décrivons les données sur lesquelles porte notre étude de reconstruction simulée. Nous décrivons ensuite le modèle bayésien utilisé pour la reconstruction matricielle et la méthode d'inférence correspondant à ce modèle. Dans la section 3, nous présentons des méthodes de reconstruction concurrentes pour comparer les performances de notre algorithme. Les résultats de comparaison des performances des différentes méthodes sont présentés dans la section 4.

2. Description des données et du modèle de reconstruction matricielle

2.1. Description des données

Les données de notre étude sont issues d'un questionnaire HSOPSC soumis à l'hôpital universitaire de Grenoble, d'une capacité de 1836 lits et desservant une population de 675 000 habitants. L'enquête a été menée de façon anonyme et bénévole entre avril 2013 et septembre 2014. Les participants admissibles étaient des employés à temps plein ou à temps partiel ayant au moins 6 mois d'emploi dans les services cliniques, de laboratoire, de pathologie, de radiologie ou de pharmacie. Sur 5044 employés admissibles, 3888 (77.08%) ont participé à l'étude. Les données issues de l'enquête présentent environ 1.8% de réponses manquantes.

Chaque item du questionnaire comporte 5 réponses possibles ordonnées selon le niveau de l'accord du répondant par rapport à chaque point. Les réponses possibles sont données par ordre croissant de l'accord : *pas du tout d'accord (strongly disagree)*, *pas d'accord (disagree)*, *ni d'accord ni en désaccord (neither agree nor disagree)*, *en accord (agree)* *tout à fait d'accord (strongly agree)*. Dans le questionnaire, les données qualitatives ordinales sont transformées en données quantitatives discrètes et positives sur une échelle allant de 1 à 5 selon le codage de Likert.

2.2. Modèle bayésien pour l'imputation de données incomplètes

Dans cette section, nous décrivons un modèle bayésien pour des données discrètes représentées par des matrices d'entiers positifs, puis nous décrivons les algorithmes utilisés pour l'estimation des paramètres de ce modèle. Dans la suite, le modèle sera référencé comme le modèle *Poisson-Gamma*. Supposons que l'on dispose d'un échantillon de n individus interrogés sur p items, dont

 ISSN: 2102-6238

les réponses sont codées par des entiers positifs. Les réponses au questionnaire sont disposées sous forme d'une matrice, $\mathbf{X}$, de taille $p \times n$. Soit K un entier supérieur à 1 (et inférieur au minimum de n et p). Le modèle probabiliste que nous considérons fait l'hypothèse que les coefficients de la matrice $\mathbf{X}$ sont des réalisations de la loi de Poisson, dont la moyenne matricielle est décrite par le produit matriciel $\mathbf{UV}$, où $\mathbf{U}$ est une matrice de dimension $p \times K$ et $\mathbf{V}$ est une matrice de dimension $K \times n$. Pour l'observation x_{ij}, le modèle génératif est décrit par

$$x_{ij}|\mathbf{U},\mathbf{V} \sim \mathscr{P}(\sum_{k=1}^{K} u_{ik}v_{kj}), \quad i = 1,\ldots,p, \quad j = 1,\ldots,n,$$

où $\mathscr{P}(\lambda)$ désigne la loi de Poisson de paramètre λ. Conditionnellement aux matrices $\mathbf{U}$ et $\mathbf{V}$, les variables x_{ij} sont indépendantes. Les paramètres du modèle sont les matrices $\mathbf{U}$, $\mathbf{V}$. Elles sont considérées comme des variables latentes dont les coefficients ont pour lois a priori des lois Gamma

$$u_{ik} \sim \mathscr{G}(a^u, b^u/a^u), \quad i = 1,\ldots,p, \quad k = 1,\ldots,K,$$

et

$$v_{kj} \sim \mathscr{G}(a^v, b^v/a^v), \quad j = 1,\ldots,n, \quad k = 1,\ldots,K.$$

Les variables u_{ik}, v_{kj} sont indépendantes. Les hyperparamètres correspondant aux paramètres de forme et d'échelle de la loi Gamma sont positifs. Les paramètres de forme déterminent la parcimonie du modèle. Ils seront généralement choisis inférieurs à 1, et sélectionnés par une heuristique de choix de modèle. Les paramètres d'échelle seront déterminés par une procédure empirique (approximation variationnelle) que nous décrirons plus loin. Dans le modèle probabiliste, l'opération de reconstruction des réponses manquantes correspond à l'estimation des matrices $\mathbf{U}$, $\mathbf{V}$, suivie du calcul du produit $\mathbf{UV}$. Pour prendre en compte les données manquantes, nous utilisons une matrice binaire de coefficients δ_{ij} telle que $\delta_{ij} = 0$ si la donnée x_{ij} est manquante et $\delta_{ij} = 1$ sinon. La fonction de vraisemblance des paramètres $\mathbf{U},\mathbf{V}$ est alors définie de la manière suivante

$$p(\mathbf{X}\,|\,\mathbf{U},\mathbf{V}) = \prod_{i=1}^{p}\prod_{j=1}^{n} p(x_{ij}\,|\,\mathbf{U},\mathbf{V})^{\delta_{ij}}.$$

2.3. Échantillonnage de Gibbs

Dans cette section, nous présentons l'algorithme *pgnmf* pour l'échantillonnage de la loi a posteriori des matrices latentes $\mathbf{U}$, $\mathbf{V}$. L'algorithme effectue des tirages aléatoires grâce à une méthode de Gibbs pour le modèle de factorisation matricielle décrit dans la section précédente.

L'algorithme d'échantillonnage de Gibbs est construit à partir d'un modèle augmenté où sont introduites des matrices auxilliaires $(\mathbf{S}^k)$, $k = 1,\ldots,K$. Chaque matrice $\mathbf{S}^k$ possède la même dimension que la matrice des données. Nous supposons que

$$s_{ij}^k|\mathbf{U},\mathbf{V} \sim \mathscr{P}(u_{ik}v_{kj}), i = 1,\ldots,p, \quad j = 1,\ldots,n.$$

Chaque observation x_{ij} est donc considérée comme étant la somme de K réalisations indépendantes $s_{ij}^1,\ldots,s_{ij}^K$ de loi de Poisson. L'introduction des variables auxilliaires ne change pas le

 ISSN: 2102-6238

modèle de factorisation matricielle proposé précédement et facilite le calcul des lois conditionnelles.

L'algorithme nécessite la détermination des lois conditionnelles a posteriori de chacune des variables latentes $\mathbf{U}$, $\mathbf{V}$ et variables auxiliaires $\mathbf{S}$ sachant les autres variables. Pour tout i, j et pour toute donnée x_{ij} (non-manquante), la loi conditionnelle a posteriori du vecteur latent $\mathbf{s}_{ij}$ est une loi multinomiale

$$\mathbf{s}_{ij} \equiv (s_{ij}^1, \ldots, s_{ij}^K) \,|\, \mathbf{U}, \mathbf{V}, \mathbf{X} \sim \mathscr{M}\left(\delta_{ij} x_{ij}\,,\, p_{ij}^1\,,\, \ldots\,,\, p_{ij}^K\right)$$

où les probabilités p_{ij}^ℓ sont données par

$$p_{ij}^\ell = \frac{u_{i\ell} v_{\ell j}}{\sum_{k=1}^K u_{ik} v_{kj}}\,, \quad \ell = 1, \ldots, K.$$

Les lois conditionnelles a posteriori des coefficients des matrices $\mathbf{U}$ et $\mathbf{V}$ sont des lois Gamma

$$u_{ik} \,|\, \mathbf{S}, \mathbf{V}, \mathbf{X} \sim \mathscr{G}(\alpha_{ik}^u, \beta_{ik}^u)\,, \quad i = 1, \ldots, p\,,\, k = 1, \ldots, K,$$

et

$$v_{kj} \,|\, \mathbf{S}, \mathbf{U}, \mathbf{X} \sim \mathscr{G}(\alpha_{kj}^v, \beta_{kj}^v)\,, \quad j = 1, \ldots, n\,,\, k = 1, \ldots, K,$$

dont les paramètres de forme et d'échelle sont donnés par

$$\alpha_{ik}^u = a^u + \sum_{j=1}^n \delta_{ij} s_{ij}^k\,, \quad \beta_{ik}^u = (a^u/b^u + \sum_{j=1}^n \delta_{ij} v_{kj})^{-1}$$

et

$$\alpha_{kj}^v = a^v + \sum_{i=1}^p \delta_{ij} s_{ij}^k\,, \quad \beta_{kj}^v = (a^v/b^v + \sum_{i=1}^p \delta_{ij} u_{ik})^{-1}.$$

La mise en œuvre pratique de l'algorithme *pgnmf* demande d'observer une période de chauffe (*burn-in*) comprenant entre 100 et 300 cycles de mise à jour des paramètres avant d'entrer dans une phase stationnaire (Algorithme 1). La période de chauffe est suivie de 1500 à 2000 cycles de mise à jour pour calculer les estimateurs. L'algorithme *pgnmf* est utilisé comme une méthode d'imputation multiple car les trois règles décrites par Little et Rubin (1987) peuvent être appliquées : 1) plusieurs estimations $\mathbf{U}^{(t)}$ et $\mathbf{V}^{(t)}$ sont enregistrées lors des itérations de l'algorithme, 2) les estimations ponctuelles sont moyennées afin d'obtenir les estimations finales, $\bar{\mathbf{U}}$ et $\bar{\mathbf{V}}$, 3) la reconstruction automatique de la matrice des données est obtenue en effectuant le produit matriciel $\tilde{\mathbf{X}} = \bar{\mathbf{U}}\bar{\mathbf{V}}$.

2.4. Estimation des hyperparamètres du modèle

Les hyperparamètres du modèle Poisson-Gamma englobent les paramètres de forme et les paramètres d'échelle des lois a priori des matrices $\mathbf{U}$ et $\mathbf{V}$. Nous notons $\theta = (a^u, b^u, a^v, b^v)$. Dans une

 ISSN: 2102-6238

Données : La matrice de données $\mathbf{X}$ et les valeurs des hyperparamètres a^u, b^u, a^v et b^v, où a^u, a^v représentent les paramètres de forme et b^u, b^v les moyennes des lois a priori.
Résultat : Les estimations des matrices $\mathbf{U}$ et $\mathbf{V}$.
Initialisation des matrices latentes : $u_{ik}^{(0)} \sim \mathscr{G}(a^u, b^u/a^u)$ et $v_{kj}^{(0)} \sim \mathscr{G}(a^v, b^v/a^v)$;
pour *t de* 1 *à* $t_{\max}$ **faire**
  Simulation des variables latentes $\mathbf{S}$;
  pour *i de 1 à p et j de 1 à n* **faire**
    pour *k de 1 à K* **faire**
      $p_{ij}^{k,(t)} = \frac{u_{ik}^{(t-1)} v_{kj}^{(t-1)}}{\sum_{\ell=1}^{K} u_{i\ell}^{(t-1)} v_{\ell j}^{(t-1)}}$;
    fin
    $p_{ij}^{(t)} = \left(p_{ij}^{1,(t)}, \ldots, p_{ij}^{K,(t)}\right)$;
    $s_{ij}^{(t)} \sim \mathscr{M}\left(\delta_{ij} x_{ij}\,,\, p_{ij}^{(t)}\right)$;
  fin
  pour *i de 1 à p et j de 1 à n* **faire**
    pour *k de 1 à K* **faire**
      Simulation des coefficients de $\mathbf{U}$;
        $\alpha_{ik}^{u,(t)} = a^u + \sum_j \delta_{ij} s_{ij}^{k,(t)}$;
        $\beta_{ik}^{u,(t)} = \left(a^u/b^u + \sum_j \delta_{ij} v_{kj}^{(t-1)}\right)^{-1}$;
        $u_{ik}^{(t)} \sim \mathscr{G}\left(\alpha_{ik}^{u,(t)}, \beta_{ik}^{u,(t)}\right)$;
      Simulation des coefficients de $\mathbf{V}$;
        $\alpha_{kj}^{v,(t)} = a^v + \sum_i \delta_{ij} s_{ij}^{k,(t)}$;
        $\beta_{kj}^{v,(t)} = \left(a^v/b^v + \sum_i \delta_{ij} u_{ik}^{(t)}\right)^{-1}$;
        $v_{kj}^{(t)} \sim \mathscr{G}\left(\alpha_{kj}^{v,(t)}, \beta_{kj}^{v,(t)}\right)$;
    fin
  fin
fin

Algorithme 1 : Échantillonneur de Gibbs (*pgnmf*).

optique bayésienne empirique, les hyperparamètres sont estimés en effectuant une approximation variationnelle de la log-vraisemblance marginale $p(\mathbf{X}|\theta)$, puis en calculant le maximum de la fonction approchée (Ghahramani et Beal, 2000; Bishop, 2006; Cemgil, 2009; Gopalan *et al.*, 2014, 2015). La méthode que nous utilisons est inspirée d'algorithmes développés dans plusieurs travaux antérieurs, notamment par Cemgil (2009). Les détails des calculs peuvent être trouvés dans cette référence bibliographique. Nous résumons les étapes principales de la méthode d'estimation des hyperparamètres ci-dessous. Tout d'abord, la log-vraisemblance marginale peut être minorée de la manière suivante

$$\log p(\mathbf{X}\,|\,\theta) \geq \mathbb{E}_q\left[\log \frac{p(\mathbf{X},\mathbf{S},\mathbf{U},\mathbf{V}\,|\,\theta)}{q(\mathbf{S},\mathbf{U},\mathbf{V}\,|\,\theta)}\right] \tag{1}$$

où l'espérance $\mathbb{E}_q(.)$ est calculée par rapport à une loi instrumentale, $q(\mathbf{S},\mathbf{U},\mathbf{V}\,|\,\theta)$, choisie de sorte à approcher la loi a posteriori au sens de la divergence de Kullback-Leibler. Suivant Bishop (2006), la loi instrumentale repose sur un produit de lois marginales

$$q(\mathbf{S},\mathbf{U},\mathbf{V}\,|\,\theta) = q(\mathbf{S}\,|\,\theta)q(\mathbf{U}\,|\,\theta)q(\mathbf{V}\,|\,\theta)\,. \tag{2}$$

 ISSN: 2102-6238

Les lois marginales sont déterminées par une procédure itérative aboutissant à une approximation de la loi a posteriori $p(\mathbf{S},\mathbf{U},\mathbf{V}\,|\,\mathbf{X},\theta)$ (Cemgil, 2009). Comme dans l'algorithme EM, nous considèrons alors le calcul de l'espérance suivante

$$Q(\theta^*;\theta) = \mathbb{E}_q\left[\log p(\mathbf{X},\mathbf{S},\mathbf{U},\mathbf{V}\,|\,\theta^*)\right],$$

suivi de la maximisation de cette espérance en la variable θ^*. En fixant le vecteur des hyperparamètres, $\theta^{(t)}$, l'espérance s'exprime de la manière suivante

$$\begin{aligned}
Q(\theta^*;\theta^{(t)}) \;=\; & \sum_i\sum_j \delta_{ij} \sum_k \Big(-\mathbb{E}_q^{(t)}[u_{ik}]\mathbb{E}_q^{(t)}[v_{kj}] + \mathbb{E}_q^{(t)}[s_{ij}^k]\mathbb{E}_q^{(t)}[\log u_{ik}] \\
& + \mathbb{E}_q^{(t)}[s_{ij}^k]\mathbb{E}_q^{(t)}[\log v_{kj}] - \mathbb{E}_q^{(t)}[\log\Gamma(s_{ij}^k+1)]\Big) \\
& + \sum_i\sum_j \delta_{ij}\mathbb{E}_q^{(t)}[\log\delta(x_{ij}-\sum_k s_{ij}^k)] \\
& + \sum_i\sum_k \Big((a^u-1)\mathbb{E}_q^{(t)}[\log u_{ik}] - \frac{a^u}{b^u}\mathbb{E}_q^{(t)}[u_{ik}] - \log\Gamma(a^u) - a^u\log\frac{b^u}{a^u}\Big) \\
& + \sum_k\sum_j \Big((a^v-1)\mathbb{E}_q^{(t)}[\log v_{kj}] - \frac{a^v}{b^v}\mathbb{E}_q^{(t)}[v_{kj}] - \log\Gamma(a^v) - a^v\log\frac{b^v}{a^v}\Big).
\end{aligned}$$

où $\delta(x)$ est la fonction de Kronecker. L'estimation des hyperparamètres consiste alors à résoudre itérativement le problème d'optimisation suivant

$$\theta^{(t+1)} = \arg\max_\theta Q(\theta;\theta^{(t)}). \tag{3}$$

Lors d'expériences préliminaires, nous avons obtenu de meilleurs résultats lorsque les paramètres de forme étaient fixés à des valeurs garantissant la parcimonie (valeurs inférieures à 1, fixées à $a^v = 0.8$ et $a^u = 0.5$). Dans ce cas, l'algorithme se simplifie et l'équation (3) revient à mettre à jour les moyennes des loi Gamma de la manière suivante

$$b^{u,(t+1)} = \frac{1}{pK}\sum_{i=1}^{p}\sum_{k=1}^{K}\mathbb{E}_q^{(t)}[u_{ik}],$$

et

$$b^{v,(t+1)} = \frac{1}{nK}\sum_{k=1}^{K}\sum_{j=1}^{n}\mathbb{E}_q^{(t)}[v_{kj}].$$

L'algorithme est exécuté avec plusieurs valeurs initiales $b^{u,(0)}$ et $b^{v,(0)}$. Les estimations donnant la valeur la plus élevée de la borne variationnelle sont retenues pour lancer ensuite l'algorithme de Gibbs. Notons que l'approximation variationnelle fournit aussi des estimations des matrices latentes $\mathbf{U}$ et $\mathbf{V}$, mais nous avons observé que ces estimations étaient moins bonnes que celles données par l'algorithme de Gibbs.

3. Comparaison de méthodes

Nous comparons l'algorithme *pgnmf* à sept autres algorithmes d'imputation de données manquantes. Trois de ces méthodes reposent sur des principes de factorisation matricielle similaires à l'approche proposée dans la section précédente.

 ISSN: 2102-6238

Méthodes naïves. Les méthodes naïves regroupent l'ensemble des stratégies d'imputation ne nécessitant pas d'analyse structurée des données. Nous considérons les méthodes d'imputation par la *moyenne (moy)*, la *médiane (med)*, le *mode (mod)* (Shrive *et al.*, 2006; Ben Othman Amroussi, 2011), et la méthode *last observation carried forward (locf)* (Streiner, 2008). La méthode *locf* complète chaque valeur manquante par la dernière valeur observée en parcourant la matrice des observations. Pour *locf*, nous utilisons la version implémentée dans le programme *zoo* du logiciel R. Il est connu que les méthodes d'imputation naïves induisent des biais dans les analyses (Streiner, 2008; Little *et al.*, 2012).

Méthode des forêts aléatoires : *missf*. La méthode des forêts aléatoires, *MissForest* (abbréviation *missf*), est une technique d'imputation de données mixtes, continues ou qualitatives. Introduite par Stekhoven et Bühlmann (2011), elle a été proposée comme une alternative à d'autres méthodes traitant des données mixtes telles que la méthode *mice (multivariate imputation by chained equations)* de Van Buuren et Oudshoorn (1999). La méthode *MissForest* s'appuie sur la technique de régression de Breiman (2001) pour prédire les données manquantes. Elle est implémentée dans le programme *missForest* du logiciel R.

Méthode Poisson Espérance-Maximisation : *pem*. La méthode *Espérance-Maximisation (EM)* permet l'estimation des paramètres d'un modèle probabiliste par le maximum de vraisemblance lorsque des variables sont non-observées (Dempster *et al.*, 1977). Dans notre étude, nous avons implémenté l'algorithme EM pour le modèle Poisson-Gamma (algorithme *pem*). L'algorithme opère les tâches d'imputation et d'estimation des matrices latentes $\mathbf{U}$ et $\mathbf{V}$ simultanément. La dérivation des équations de cette méthode est détaillée par Cemgil (2009). La mise à jour itérative des coefficients des matrices latentes est effectuée de la manière suivante

$$u_{ik}^{(t+1)} = u_{ik}^{(t)} \frac{\sum_{j=1}^{n} \delta_{ij} x_{ij} v_{kj}^{(t)} / \sum_{\ell=1}^{K} u_{i\ell}^{(t)} v_{\ell j}^{(t)}}{\sum_{j=1}^{n} \delta_{ij} v_{kj}^{(t)}}, \quad i = 1,\ldots,p, k = 1,\ldots,K,$$

$$v_{kj}^{(t+1)} = v_{kj}^{(t)} \frac{\sum_{i=1}^{p} \delta_{ij} x_{ij} u_{ik}^{(t)} / \sum_{\ell=1}^{K} u_{i\ell}^{(t)} v_{\ell j}^{(t)}}{\sum_{i=1}^{p} \delta_{ij} u_{ik}^{(t)}}, \quad k = 1,\ldots,K, j = 1,\ldots,n.$$

L'algorithme *pem* est proche de l'algorithme de mise à jour multiplicative de Lee et Seung (2001) pour la factorisation de matrice non-négative.

Méthode d'imputation multiple par analyse des correspondances : *mimca*. La méthode *mimca* effectue l'imputation de données catégorielles en utilisant l'analyse des correspondances multiples (ACM) (Josse *et al.*, 2010; Josse et Husson, 2016; Audigier *et al.*, 2017). Cette approche prend en compte la variabilité des paramètres en utilisant une approche de bootstrap non-paramétrique. La méthode *mimca* génère plusieurs ensembles de données imputées à l'aide d'un algorithme d'ACM itératif régularisé. Il s'agit une méthode d'imputation multiple propre au sens de Little et Rubin (2002).

Méthode de factorisation de matrice non-négative pondérée : *wnmf*. La méthode *wnmf* (weighted nonnegative matrix factorization) implémente un algorithme de factorisation de matrice non-négative introduit par Kim et Choi (2009). Initialement développées pour l'analyse

 ISSN: 2102-6238

exploratoire, les méthodes de factorisation de matrice non-négative ont récemment été adaptées pour répondre à la problématique des données manquantes. L'implémentation de Kim et Choi (2009) s'appuie sur une version pondérée de l'algorithme des moindres-carrés alternés pour la factorisation de matrice non-négative (Kim et Park, 2008). L'algorithme *wnmf* effectue une régularisation de type *ridge* sur les facteurs. Nous choisirons le paramètre de régularisation par défaut du programme. La méthode *wnmf* s'apparente à la méthode EM et à l'algorithme d'imputation par ACM (Cemgil, 2009; Josse *et al.*, 2010).

Pour toutes les méthodes produisant une matrice non-entière, nous ré-échantillonnons les coefficients matriciels en leur attribuant l'entier le plus proche proportionnellement à la décimale de la valeur imputée. Pour éviter la création de données non-présentes dans l'échantillon initial, les valeurs en dehors du rang sont ramenées aux valeurs minimales et maximales respectivement.

3.1. Simulations et critères d'évaluation

Simulations. Dans un premier temps, nous avons utilisé des données simulées selon le modèle génératif présenté dans la section 2. Nous avons simulé une matrice comportant $p = 30$ lignes (items) et $n = 1500$ colonnes (individus). Le nombre de facteurs a été fixé à $K = 2$. Les coefficients des matrices latentes $\mathbf{U}$ et $\mathbf{V}$ ont été tirés aléatoirement et indépendamment suivant les lois Gamma $\mathscr{G}(65, 0.8/65)$ et $\mathscr{G}(50, 0.7/50)$ respectivement. La matrice complète a été conditionnée pour que ses coefficients soient strictement positifs. Elle contenait finalement des valeurs entières comprises entre 1 et 7. Huit matrices ont été créées en supprimant 10% à 80% des données simulées par un mécanisme MCAR. Les données manquantes ont été supprimées par des tirages de Bernoulli indépendants en conservant au moins une donnée par colonne.

Dans un second temps, nous avons utilisé les données issues de l'enquête HSOPSC effectuée à l'hôpital universitaire de Grenoble auprès de travailleurs hospitaliers ($n = 3888$). Une analyse préliminaire des données du questionnaire HSOPSC a été effectuée par ACM utilisant la méthode de Josse et Husson (2016). Afin de quantifier la perte d'information liée à la reconstruction automatique de ce questionnaire, nous avons masqué une partie des données, puis reconstruit la matrice à l'aide d'une méthode s'appuyant sur l'ACM. Nous avons ensuite calculé l'indice de corrélation multiple entre les trois premiers axes principaux de la matrice disjonctive complète et ceux des tableaux de probabilités des modalités pour les matrices reconstruites à partir de données masquées. Les valeurs de corrélation multiple ont été utilisées pour évaluer la perte d'information induite par la reconstruction matricielle. Cette étude préliminaire a été répliquée 100 fois en utilisant la méthode de bootstrap pour estimer l'incertitude liée à la perte d'information. Finalement, huit matrices ont été créées en supprimant 10% à 80% des données par un mécanisme similaire à celui utilisé pour les données simulées par le modèle génératif.

Critères d'évaluation. Dans nos analyses comparatives, les performances des différentes méthodes de reconstruction ont été mesurées à l'aide de deux quantités : la racine de l'erreur quadratique moyenne (RMSE, root mean square error) et la divergence de Kullback-Leibler moyenne. La critère RSME mesure la différence entre les valeurs reconstruites par un algorithme et celles observées dans la matrice complète (ou presque complète dans le cas des données réelles)

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\text{Vec}(\mathbf{X})_i - \text{Vec}(\tilde{\mathbf{X}})_i\right)^2}.$$

 ISSN: 2102-6238

Dans cette formule, $\mathrm{Vec}(\mathbf{X})_i$ et $\mathrm{Vec}(\tilde{\mathbf{X}})_i$ représentent respectivement les N valeurs retirées de la matrice originale et les N valeurs correspondantes dans la matrice reconstruite. Le critère RMSE est classiquement utilisé pour évaluer les méthodes de reconstruction matricielle en apprentissage automatique lorsqu'il s'agit, par exemple, d'évaluer des systèmes de recommandation. L'analyse d'une enquête HSOPSC se présente généralement sous la forme d'un compte-rendu dans lequel les histogrammes des réponses aux $p = 42$ questions sont discutées séquentiellement. Une seconde mesure est introduite dans le but de refléter la pratique mise en œuvre lors de l'analyse des résultats d'un tel questionnaire en santé publique. Nous utilisons la divergence de Kullback-Leibler pour estimer la perte d'information induite par la reconstruction des histogrammes de réponse à chaque item du questionnaire. Pour un item particulier, la divergence de Kullback-Leibler est définie de la manière suivante

$$D_{\mathrm{KL}}(P\|Q) = \sum_{j=1}^{n_R} p_j \log(p_j/q_j),$$

où P représente la loi empirique des réponses codées dans le questionnaire complet et Q représente la loi empirique des réponses codées dans le questionnaire reconstruit (n_R est le nombre de choix possibles). La mesure finale est la moyenne des valeurs de divergence calculées pour chacun des p items du questionnaire. La divergence de Kullback-Leibler mesure la qualité de reconstruction des lois marginales de la matrice originale, tandis que l'erreur RMSE peut tenir compte des structures de corrélation existant dans la matrice originale.

4. Résultats

Dans cette section, nous évaluons les performances de plusieurs approches de reconstruction automatique de questionnaires en supposant que l'on dispose de questionnaires arbitrairement réduits. Pour cela, nous avons considéré une matrice de données complète (ou possédant un fort taux de complétude), et nous avons supprimé des données au hasard suivant une procédure de type MCAR. Dans le but de valider l'implémentation de l'algorithme *pgnmf*, nous avons tout d'abord utilisé une matrice de données simulées selon le modèle Poisson-Gamma. Dans un second temps, nous avons utilisé les données issues d'une l'enquête effectuée à l'hôpital universitaire de Grenoble, considérées représentatives d'une enquête à l'échelle nationale.

4.1. Données simulées

Pour les matrices simulées selon le modèle Poisson-Gamma, les valeurs de RMSE varient entre 0.973 (*mimca*) pour un taux de suppression de 10% et 1.210 (*missf*) pour un taux de suppression de 80% (Table 1). Comme attendu, les méthodes naïves présentent les performances les plus faibles. Les algorithmes reposant sur la factorisation matricielle (*pem*, *pgnmf*, *wnmf* et *mimca*) obtiennent les valeurs de RMSE les plus basses. Une faible variabilité des erreurs de reconstruction est observée pour l'ensemble de ces méthodes. La méthode *mimca* obtient les meilleures performances pour des taux de suppression d'items situés entre 10% et 40% avec des valeurs de RMSE situées entre 0.9726 et 0.9819. L'algorithme *pgnmf* obtient les meilleurs résultats lorsque le taux de suppression est plus important. Pour des taux compris entre 50% et 80%, les valeurs de RMSE pour *pgnmf* se situent entre 0.9778 et 0.9889 (Table 1).

 ISSN: 2102-6238

Concernant la divergence de Kullback-Leibler, les valeurs moyennées sur les 30 items varient entre 0.002 (*pgnmf*) pour un taux de suppression de 10% et 0.489 (*med*) pour un taux de suppression de 80% (Table 2). À nouveau, les méthodes naïves présentent des performances plus faibles que les méthodes matricielles. La méthode *pgnmf* obtient les meilleures performances pour des taux de suppression situés entre 10% et 30% avec des valeurs situées entre 0.002 et 0.012. L'algorithme *missf* obtient les meilleurs résultats lorsque le taux de suppression est plus important. Pour des taux compris entre 40% et 80%, les valeurs de divergence pour *missf* se situent entre 0.018 et 0.057 (Table 2). Relativement aux méthodes de reconstruction précédentes, les résultats des simulations mettent en évidence les bonnes propriétés de reconstruction matricielle de l'algorithme *pgnmf* et valident son implémentation numérique.

TABLE 1. *Simulation du modèle génératif Poisson-Gamma. Erreur de reconstruction (RMSE) en fonction du taux de données supprimées. Abbrévations utilisées. missf : missForest, pem : Poisson-EM, pgnmf : Poisson-Gamma NMF, wnmf : Weighted NMF, mimca : Multiple Imputation using Multiple Correspondence Analysis. *L'algorithme n'a pas convergé pour les taux 70-80%. Les valeurs pour les taux 10-60% ont été obtenues en augmentant la valeur du paramètre de régularisation à 100.*

	10%	20%	30%	40%	50%	60%	70%	80%
locf	1.2130	1.2185	1.2186	1.2037	1.2139	1.2080	1.2016	1.2082
med	1.0559	1.0515	1.0526	1.0576	1.0499	1.0636	1.0553	1.0548
missf	1.0478	1.0483	1.0513	1.0834	1.1086	1.1306	1.1644	1.2101
pem	0.9761	0.9805	0.9781	0.9837	0.9843	0.9945	1.0023	1.0241
pgmnf	0.9780	0.9791	0.9778	0.9821	**0.9778**	**0.9838**	**0.9818**	**0.9889**
wnmf	0.9969	0.9964	0.9983	1.0043	1.0135	1.0247	1.0424	1.0592
mimca*	**0.9726**	**0.9733**	**0.9753**	**0.9819**	0.9835	0.9987	-	-

TABLE 2. *Simulation du modèle génératif Poisson-Gamma. Divergence de Kullback-Leibler en fonction du taux de données supprimées. Abbrévations utilisées : missf : missForest, pem : Poisson-EM, pgnmf : Poisson-Gamma NMF, wnmf : Weighted NMF, mimca : Multiple Imputation using Multiple Correspondence Analysis. *L'algorithme n'a pas convergé pour les taux 70-80%. Les valeurs pour les taux 10-60% ont été obtenues en augmentant la valeur du paramètre de régularisation à 100.*

	10%	20%	30%	40%	50%	60%	70%	80%
moy	0.0039	0.0136	0.0273	0.0525	0.0790	0.1181	0.1707	0.2450
med	0.0051	0.0189	0.0411	0.0789	0.1249	0.1912	0.3038	0.4891
missf	0.0028	0.0091	0.0167	**0.0183**	**0.0179**	**0.0203**	**0.0339**	**0.0570**
pem	0.0038	0.0128	0.0257	0.0484	0.0675	0.0934	0.1186	0.1313
pgmnf	**0.0018**	**0.0060**	**0.0117**	0.0244	0.0348	0.0581	0.0942	0.1589
wnmf	0.0032	0.0106	0.0198	0.0373	0.0471	0.0640	0.0698	0.0702
mimca*	0.0040	0.0137	0.0274	0.0539	0.0866	0.1227	-	-

4.2. Reconstruction du questionnaire HSOPSC

Dans un second temps, nous avons évalué la possibilité de reconstruire automatiquement des questionnaires réduits obtenus à partir de l'enquête HSOPSC effectuée entre avril 2013 et septembre 2014 à l'hôpital universitaire de Grenoble. Une analyse par ACM des données du questionnaire a tout d'abord été effectuée après une imputation des données manquantes utilisant la

 ISSN: 2102-6238

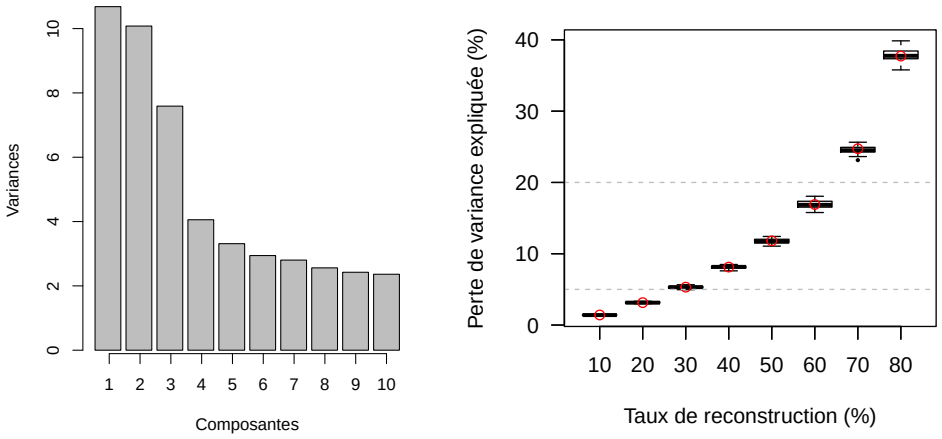

FIGURE 1. *Gauche : Variances correspondant aux axes principaux de l'analyse des correspondances multiples des données du questionnaire HSOPSC. Droite : Perte de variance expliquée par les 3 axes principaux lorsque des données sont masquées. Les lignes pointillées correspondent aux pourcentages* 5% *et* 20%. *Les boîtes correspondent aux valeurs obtenues par bootstrap à partir du jeu de données original.*

méthode de Josse et Husson (2016) (Figure 1). Le diagramme des variances a permis de mettre en évidence trois axes principaux dans les données. Dans la suite de l'étude de simulation, nous avons utilisé $K = 3$ facteurs pour l'ensemble des méthodes reposant sur la factorisation matricielle (méthodes *pem, pgnmf, wnmf, mimca*). Afin de quantifier la perte d'information liée à la reconstruction automatique du questionnaire, nous avons calculé la corrélation entre les premiers axes principaux de la matrice disjonctive originale et ceux des tableaux de probabilités des matrices reconstruites à partir de données masquées (Josse et Husson, 2016). Les valeurs obtenues pour des répliques par bootstrap du jeu de données montrent que la variabilité est faible. Pour un questionnaire de type HSOPSC, les résultats indiquent qu'il est possible de restituer une part importante de l'information manquante lorsque de grands nombres d'items sont aléatoirement supprimés. Ce résultat justifie les valeurs élevées des taux de suppression considérés dans le reste de notre étude.

Selon le critère RMSE, les valeurs de performance varient entre 0.950 (*missf*) pour un taux de suppression de 10% et 1.210 (*locf*) pour un taux de suppression de 80% (Table 3). Les méthodes naïves présentent des performances plus faibles que les méthodes factorielles. La méthode *wnmf* obtient les meilleures performances pour des taux de suppression situés entre 20% et 30% (valeurs de RMSE situées entre 0.974 et 0.982). L'algorithme *pgnmf* obtient les meilleurs résultats lorsque le taux de suppression est supérieur à 40% (valeurs de RMSE situées entre 0.987 et 1.022, Table 3 et Figure 2).

Selon le critère de divergence de Kullback-Leibler, les valeurs de performance varient entre 0.0003 (*missf*) pour un taux de suppression de 10% et 0.696 (*med*) pour un taux de suppression de 80% (Table 4). Les méthodes *missf* et *wnmf* obtiennent les meilleures performances pour l'ensemble des taux de suppression (valeurs de divergence situées entre 0.0003 et 0.021). L'algorithme *pgnmf* obtient des résultats similaires aux autres méthodes factorielles lorsque le taux de suppression est inférieur à 50% (valeurs de divergence situées entre 0.002 et 0.033, Figure 3).

 ISSN: 2102-6238

TABLE 3. *Questionnaire HSOPSC. Erreur de reconstruction RMSE en fonction du taux de données supprimées. Abbrévations utilisées : missf : missForest, pem : Poisson-EM, pgnmf : Poisson-Gamma NMF, wnmf : Weighted NMF, mimca : Multiple Imputation using Multiple Correspondence Analysis.*

	10%	20%	30%	40%	50%	60%	70%	80%
locf	1.3117	1.3066	1.3083	1.3057	1.3076	1.3144	1.3268	1.3354
moy	1.0522	1.0443	1.0494	1.0475	1.0508	1.0477	1.0483	1.0482
med	1.0073	1.0136	1.0180	1.0208	1.0239	1.0185	1.0171	1.0220
mod	1.0902	1.0993	1.1063	1.0984	1.1098	1.1037	1.0982	1.0856
missf	**0.9500**	0.9819	1.0090	1.0223	1.0551	1.0971	1.1374	1.2034
pem	0.9750	0.9819	0.9881	0.9874	0.9918	0.9994	1.0094	1.0343
pgmnf	0.9746	0.9815	0.9874	**0.9866**	**0.9908**	**0.9973**	**1.0052**	**1.0220**
wnmf	0.9607	**0.9737**	**0.9829**	0.9891	1.0009	1.0272	1.0586	1.1234
mimca	0.9807	0.9880	0.9956	0.9974	1.0001	1.0031	1.0093	1.0255

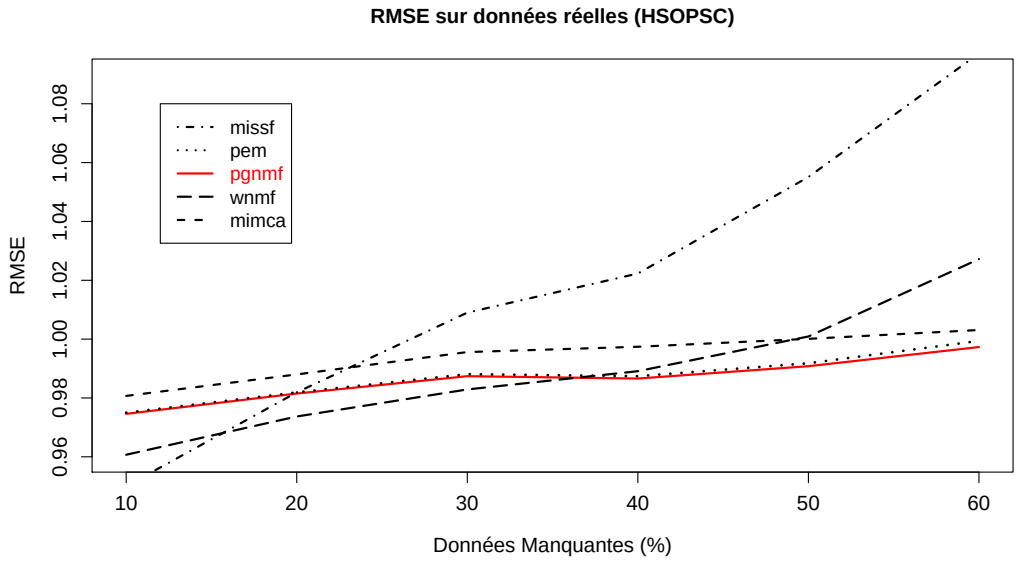

FIGURE 2. *Questionnaire HSOPSC. Représentation graphique de la performance des méthodes de reconstruction automatique (RMSE) en fonction du taux de données supprimées. Abbrévations utilisées : missf : missForest, pem : Poisson-EM, pgnmf : Poisson-Gamma NMF, wnmf : Weighted NMF, mimca : Multiple Imputation using Multiple Correspondence Analysis.*

 ISSN: 2102-6238

TABLE 4. *Questionnaire HSOPSC. Divergence de Kullback-Leibler en fonction du taux de données supprimées. Abbrévations utilisées : missf : missForest, pem : Poisson-EM, pgnmf : Poisson-Gamma NMF, wnmf : Weighted NMF, mimca : Multiple Imputation using Multiple Correspondence Analysis.*

	10%	20%	30%	40%	50%	60%	70%	80%
locf	0.0024	0.0075	0.0143	0.0228	0.0321	0.0432	0.0550	0.0693
moy	0.0046	0.0174	0.0404	0.0722	0.1163	0.1776	0.2656	0.4107
med	0.0081	0.0313	0.0714	0.1262	0.2022	0.3086	0.4641	0.6964
mod	0.0075	0.0292	0.0660	0.1169	0.1930	0.2951	0.4293	0.6659
missf	**0.0003**	**0.0012**	**0.0026**	**0.0044**	**0.0061**	**0.0089**	**0.0131**	**0.0214**
pem	0.0021	0.0085	0.0197	0.0340	0.0549	0.0764	0.1014	0.1228
pgmnf	0.0021	0.0085	0.0196	0.0337	0.0548	0.0763	0.1015	0.1266
wnmf	0.0015	0.0057	0.0126	0.0217	0.0331	0.0405	0.0492	0.0473
mimca	0.0017	0.0068	0.0155	0.0269	0.0441	0.0635	0.0880	0.1143

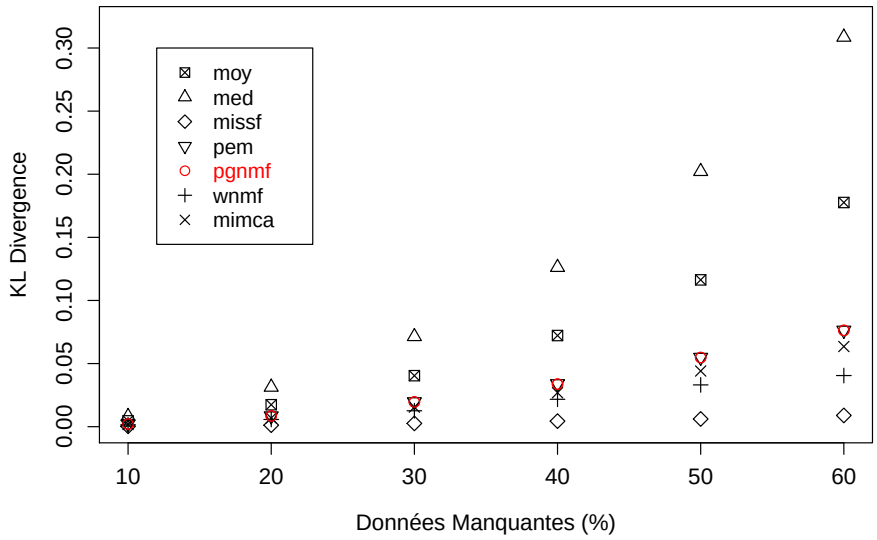

FIGURE 3. *Questionnaire HSOPSC. Représentation graphique de la performance des méthodes de reconstruction automatique (Divergence de Kullback-Leibler) en fonction du taux de données supprimées. Abbrévations utilisées : missf : missForest, pem : Poisson-EM, pgnmf : Poisson-Gamma NMF, wnmf : Weighted NMF, mimca : Multiple Imputation using Multiple Correspondence Analysis.*

 ISSN: 2102-6238

5. Discussion

En utilisant les résultats d'une enquête HSOPSC réalisée au centre hospitalier universitaire de Grenoble auprès de travailleurs médicaux, nous avons envisagé de réduire aléatoirement le questionnaire destiné à évaluer la culture de la sécurité des patients afin d'augmenter son acceptabilité auprès du personnel hospitalier. Nous avons estimé qu'une part importante de l'information contenue dans le questionnaire était redondante et pouvait être restituée automatiquement. Nous avons délibérément masqué une partie des réponses des participants, et comparé plusieurs méthodes d'imputation de données manquantes pour reconstituer l'information masquée.

Parmi les approches testées, nous avons exploré les propriétés de nouveaux algorithmes s'appuyant sur des lois de Poisson. En particulier, nous avons décrit un modèle bayésien de factorisation matricielle pour l'imputation des données discrètes. Concernant l'estimation des paramètres de ce modèle, nous avons proposé un nouvel algorithme d'échantillonnage de Gibbs, implémenté une approche variationnelle et un algorithme EM. La procédure EM effectue directement l'inférence des matrices latentes à partir de la vraisemblance, alors que l'algorithme bayésien utilise des a priori "parcimonieux" sur les facteurs latents.

Concernant l'enquête HSOPSC de Grenoble, nous avons obtenu les résultats suivants. Les méthodes d'imputation naïves, telles que l'imputation par la moyenne ou par la médiane des valeurs observées, présentent de grandes erreurs relatives par rapport aux méthodes de factorisation matricielle ou d'apprentissage automatique. En particulier, nous proposons que les méthodes d'imputation par la moyenne, la médiane ou le mode ne soient pas utilisées pour la reconstruction des histogrammes de réponses aux items. Les méthodes factorielles présentent des résultats proches. Les performances de la méthode bayésienne *pgnmf* sont comparables à d'autres méthodes de factorisation matricielle proposées récemment. Les performances de la méthode *pgnmf* sont parmi les meilleures selon le critère RMSE. Pour ce même critère, les méthodes d'apprentissage par forêts aléatoires (*missf*) ou par factorisation non-négative pondérée (*wnmf*) démontrent des performances souvent inférieures à *pgnmf*. Les performances de telles méthodes sont néanmoins supérieures en ce qui concerne la reconstruction des lois marginales. Ce résultat est important pour l'interprétation des résultats des enquêtes portant sur la culture de sécurité des patients en santé publique. Le fait marquant est que plusieurs méthodes produisent des résultats satisfaisants, supérieurs aux méthodes traditionnellement utilisées dans ce domaine. Si l'on personnalise le questionnaire HSOPSC comme il est suggéré dans cette étude, il paraît souhaitable de combiner les résultats de reconstruction de plusieurs méthodes présentées.

En conclusion, les méthodes de factorisation matricielles et la méthode d'apprentissage automatique *missf* permettent de reconstruire de grandes quantités de données supprimées de manière efficace. Leur utilisation lors d'enquêtes médicales portant sur la culture de sécurité des patients permettrait d'envisager une gestion optimisée des questionnaires hospitaliers. En effet, nos résultats suggèrent que des enquêtes médicales similaires à celle effectuée à Grenoble pourraient être réalisées en réduisant substantiellement le nombre de questions posées à chaque travailleur médical avec une perte limitée des interprétations de l'enquête.

 ISSN: 2102-6238

Remerciements

Les auteurs sont reconnaissants envers deux rapporteurs anonymes pour leurs commentaires constructifs ayant permis d'améliorer la présentation de ce travail. Ils remercient l'éditeur-en-chef du Journal de la Société Française de Statistique, Gilles Celeux, pour sa relecture attentive de l'article et ses commentaires très utiles. DBD a reçu un financement de la fondation SIMONS, projet NGALA, et une aide du LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) pour un séjour au laboratoire TIMC-IMAG de l'Université Grenoble-Alpes.

Références

AUDIGIER, V., HUSSON, F. et JOSSE, J. (2017). MIMCA : multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing*, 27(2):501–518.

BEN OTHMAN AMROUSSI, L. (2011). *Conception et validation d'une méthode de complétion des valeurs manquantes fondée sur leurs modèles d ?apparition*. Thèse de doctorat, Université de Caen.

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, USA.

BREIMAN, L. (2001). Random Forests. *Machine Learning*, 45:5–32.

CEMGIL, A. T. (2009). Bayesian Inference for Nonnegative Matrix Factorisation Models. *Computational Intelligence and Neuroscience*, 2009:785152.

DEMISSIE, S., LAVALLEY, M. P., HORTON, N. J., GLYNN, R. J. et CUPPLES, L. A. (2003). Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. *Statistic in Medicine*, 22(4):545–557.

DEMPSTER, A. P., LAIRD, N. M. et RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38.

ETCHEGARAY, J. M. et THOMAS, E. J. (2012). Comparing two safety culture surveys : safety attitudes questionnaire and hospital survey on patient safety. *BMJ Quality and Safety*, 21(6):490–498.

GHAHRAMANI, Z. et BEAL, M. (2000). Propagation Algorithms for Variational Bayesian Learning. *In Advances in Neural Information Processing Systems*, pages 507–513, Cambridge, Massachusetts, USA. MIT Press.

GHOMRAWI, H. M., MANDL, L. A., RUTLEDGE, J., ALEXIADES, M. M. et MAZUMDAR, M. (2011). Is there a role for expectation maximization imputation in addressing missing data in research using WOMAC questionnaire ? Comparison to the standard mean approach and a tutorial. *BMC Musculoskeletal Disorders*, 12:109.

GOPALAN, P., CHARLIN, L. et BLEI, D. (2014). Content-based recommendations with Poisson factorization. *In Advances in Neural Information Processing Systems 27*, pages 3176–3184.

GOPALAN, P., HOFMAN, J. M. et BLEI, D. M. (2015). Scalable recommendation with hierarchical Poisson factorization. *In UAI'15 Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 326–335, Amsterdam, Netherlands.

JOSEPH, L., BELISLE, P., TAMIM, H. et SAMPALIS, J. S. (2004). Selection bias found in interpreting analyses with missing data for the prehospital index for trauma. *Journal of Clinical Epidemiology*, 57(2):147–153.

JOSSE, J., CHAVENT, M., LIQUET, B. et HUSSON, F. (2010). Handling missing values with Regularized Iterative Multiple Correspondence Analysis. *Journal of Classification*, 29(1):91–116.

JOSSE, J. et HUSSON, F. (2016). missMDA : a package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31.

KEMP, K., WARREN, S., CHAN, N., MCCORMACK, B., SANTANA, M. et QUAN, H. (2016). Qualitative complaints and their relation to overall hospital rating using an H-CAHPS-derived instrument. *BMJ Quality and Safety*, 25(10):770–777.

KIM, H. et PARK, H. (2008). Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30:713–730.

KIM, Y. D. et CHOI, S. (2009). Weighted nonnegative matrix factorization. *In IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1541–1544, Taipei. IEEE.

LEE, D. D. et SEUNG, H. S. (2001). Algorithms for Non-negative Matrix Factorization. *In Advances in neural information processing systems*, pages 556–562. MIT Press.

 ISSN: 2102-6238

LITTLE, R. J., D'AGOSTINO, R., COHEN, M. L. et AL (2012). The Prevention and Treatment of Missing Data in Clinical Trials. *New England Journal of Medicine*, 367:1355–1360.

LITTLE, R. J. A. et RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

LITTLE, R. J. A. et RUBIN, D. B. (2002). *Statistical Analysis with Missing Data. 2nd ed*. Wiley, New York.

OCCELLI, P., QUENON, J.-L., KRET, M. et AL (2013). Validation of the French version of the Hospital Survey on Patient Safety Culture questionnaire. *International Journal for Quality in Health Care*, 25(4):459–468.

ROTNITZKY, A. et WYPIJ, D. (1994). A note on the bias of estimators with missing data. *Biometrics*, 50(4):1163–1170.

RUBIN, D. B. (1976). Inference in Missing Data. *Biometrika*, 63:581–592.

SHRIVE, F. M., STUART, H., QUAN, H. et GHALI, W. A. (2006). Dealing with missing data in a multi-question depression scale : a comparison of imputation methods. *BMC Medical Research Methodology*, 6:57.

SORRA, J. S. et NIEVA, V. F. (2004). *Hospital Survey on Patient Safety Culture. (Prepared by Westat, under Contract No. 290-96-0004)*. AHRQ Publication No. 04-0041, Rockville, MD : Agency for Healthcare Research and Quality.

STEKHOVEN, D. J. et BÜHLMANN, P. (2011). Missforest-nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28:112–118.

STREINER, D. L. (2008). Missing data and the trouble with LOCF. *Evidence Based Mental Health*, 11(1):3–5.

VAN BUUREN, S. et OUDSHOORN, K. (1999). Flexible Multivariate Imputation by mice. *TNO Prevention Center, Leiden, The Netherlands*, pages 1–20.

WALJEE, A. K., MUKHERJEE, A., SINGAL, A. G. et AL (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, 3.

 ISSN: 2102-6238