Estimer la proportion de valeurs manquantes complètement aléatoirement dans des jeux de données protéomiques

Quentin Giai Gianetto [1], Yohann Couté [1], Christophe Bruley [1] & Thomas Burger [1,2]

[1] *BIG-BGE, Université Grenoble Alpes, CEA, INSERM, Grenoble, France.*
[2] *BIG-BGE, CNRS, Grenoble, France.*
E-mail : Quentin.GIAI-GIANETTO@cea.fr

Résumé. En protéomique quantitative, les données issues d'analyses par spectrométrie de masse sont caractérisées par un important taux de valeurs manquantes. Parmi ces valeurs manquantes, plusieurs types peuvent coexister : des valeurs manquantes complètement aléatoirement (MCAR) et d'autres manquantes non aléatoirement (MNAR). Dans cette communication, nous proposons une méthode statistique pour estimer la proportion de valeurs MCAR présentes parmi les valeurs manquantes de ces jeux de données. Cette méthode s'appuie sur une modélisation du biais et de la variance asymptotiques d'un estimateur initial à l'aide d'une régression nonlinéaire hétéroscédastique.

Mots-clés. Données manquantes, Modèles de mélange.

Abstract. In quantitative proteomics, mass spectrometry-based data are characterized by a high rate of missing values. Among those missing values, several types can coexist: missing completely-at-random values (MCAR) and other missing not-at-random (MNAR). In this communication, we propose a statistical method to estimate the proportion of MCAR values among the missing values in these datasets. This method is based on a model of the asymptotic bias and variance of an initial estimator using a nonlinear heteroscedastic regression.

Keywords. Missing values, Mixture models.

1 Introduction

En protéomique quantitative, la quantification relative par spectrométrie de masse permet de mettre en évidence des protéines différentiellement abondantes entre au moins deux conditions biologiques. Cette technique d'analyse consiste, dans un premier temps, à répliquer plusieurs échantillons provenant d'une même condition. Ces réplicats sont ensuite digérés par une enzyme, qui va fragmenter les protéines présentes dans l'échantillon en peptides. Chaque réplicat est alors analysé par des techniques de chromatographie liquide et de spectrométrie de masse en tandem [1]. Il en résulte un grand nombre de spectres de masse qui sont traités de façon systématique par des outils bioinformatiques dans le but

d'identifier les peptides présents dans chaque échantillon et fournir une valeur d'intensité à chaque peptide identifié. Ces valeurs d'intensité permettent de déduire si la protéine à laquelle est associé un peptide est plus ou moins présente dans telle ou telle condition biologique. C'est la raison pour laquelle la quantification relative est particulièrement importante pour la recherche de protéines "biomarqueurs" de certaines pathologies, ou pour la détermination de nouvelles cibles thérapeutiques.

Une problématique majeure de cette technique d'analyse est l'absence d'un grand nombre de valeurs. Il est ainsi habituel de constater au minimum 20% de données manquantes dans de tels jeux de données [5]. La raison principale de ces absences est la limite de détection du spectromètre qui engendre des valeurs manquantes car trop petites pour être quantifiées [1][4][5]. Elles peuvent donc être classées comme manquantes non aléatoirement (MNAR) [2]. Toutefois, d'autres raisons liées à la préparation des échantillons, ou à certains aléas techniques dans le processus d'identification et de quantification des peptides, peuvent également déboucher sur d'autres valeurs manquantes [4]. Comme ces raisons impactent aléatoirement les valeurs, celles-ci peuvent être classées comme manquantes aléatoirement (MAR) [2]. En général, ces raisons ne sont pas mesurables, et les valeurs manquantes qu'elles induisent sont dites manquantes complètement aléatoirement (MCAR) [2]. On peut donc considérer que les données manquantes issues de ces analyses sont un mélange de valeurs MNAR et MCAR [4].

De par leur nombre, ces valeurs manquantes ne peuvent être ignorées et sont souvent imputées. Les méthodes d'imputation habituellement appliquées sont celles développées pour l'étude de l'expression de gènes par puces à ADN. Toutefois, ces méthodes d'imputation supposent que toutes les données manquantes sont du type MAR (ou MCAR), ce qui peut conduire à des résultats biaisés si la plupart des données sont du type MNAR, comme on peut s'y attendre dans le cas de données protéomiques [4]. D'autres approches modélisent explicitement le processus engendrant les valeurs manquantes, notamment à l'aide de modèles de durée de vie : grâce à de telles modélisations, il est possible d'inférer des paramètres d'intérêt, comme la moyenne ou la variance des mesures d'intensité de chaque peptide, en maximisant la vraisemblance du modèle. Par exemple, Taylor et al. [4] propose un modèle distinguant les données MAR/MCAR, des données MNAR. Ce type de modèle nécessite le calcul des probabilités que chaque valeur manquante soit MAR/MCAR, généralement effectué à l'aide de modélisations logit ou probit [4].

Dans cette communication, on propose d'améliorer la connaissance du mécanisme engendrant les valeurs manquantes en estimant, à l'aide d'une approche originale, les proportions de valeurs MCAR et MNAR présentes parmi les valeurs manquantes. À notre connaissance, aucune étude ne s'est jusqu'à présent intéressée à la pertinence d'estimateurs de cette proportion. Or, cette estimation est une étape préliminaire importante pouvant servir à affiner le traitement des valeurs manquantes. Dans le paragraphe 2, nous introduisons quelques notations et hypothèses. Dans le paragraphe 3, nous présentons notre estimateur. Dans le paragraphe 4, des résultats de simulations illustrent la pertinence de notre approche.

2 Notations et hypothèses

Pour estimer la proportion de données MCAR, nous allons faire un certain nombre d'hypothèses spécifiques à notre problématique et auxquelles nous feront référence par la suite. Avant de les introduire, définissons quelques notations : on note x_{ijk} la mesure d'intensité du peptide i dans le réplicat j de la condition k, x_{ijk}^{obs} si x_{ijk} est observé, x_{ijk}^{na} si x_{ijk} est manquante, x_{ijk}^{mcar} si x_{ijk} est MCAR et x_{ijk}^{mnar} si x_{ijk} est MNAR. On a alors le modèle de mélange suivant :

$$F_j(x) = \pi_{na_j} F_j^{na}(x) + (1 - \pi_{na_j}) F_j^{obs}(x) \tag{1}$$

où π_{na_j} est la proportion de valeurs manquantes dans le réplicat j, F_j^{na}, F_j^{obs} et F_j sont, respectivement, les fonctions de répartition des (x_{ijk}^{na}), (x_{ijk}^{obs}) et des données complétées $((x_{ijk}^{na}), (x_{ijk}^{obs}))$ dans le réplicat j. De plus, on suppose :

$$F_j^{na}(x) = \pi_{mcar_j} F_j^{mcar}(x) + (1 - \pi_{mcar_j}) F_j^{mnar}(x) \tag{2}$$

où π_{mcar_j} est la proportion de valeurs MCAR parmi les valeurs manquantes, paramètre que l'on cherche à estimer, et où F_j^{mcar} et F_j^{mnar} sont, respectivement, les fonctions de répartitions des (x_{ijk}^{mcar}) et (x_{ijk}^{mnar}) dans le réplicat j. Puisque les valeurs MCAR se répartissent totalement aléatoirement parmi les valeurs d'intensité, il est important de remarquer que $F_j^{mcar} = F_j$.

Trois autres hypothèses spécifiques aux données protéomiques peuvent être faites. D'une part, à cause du fonctionnement du processus d'ionisation du spectromètre, on peut s'attendre à ce que des peptides différents mais provenant d'une même protéine aient des mesures d'intensité différentes. Cette caractéristique peut se traduire par une hypothèse d'indépendance entre les valeurs d'intensité des peptides présents dans un même réplicat (hypothèse **H1**). D'autre part, la réplication des expériences à partir d'une même condition biologique fait que la valeur manquante d'un peptide dans un réplicat peut être estimée à partir de valeurs mesurées pour ce même peptide dans d'autres réplicats de la condition. Cette autre caractéristique nous permet de déduire de l'information sur la répartition des valeurs manquantes issues d'un réplicat. On va ainsi supposer que $\hat{F}_j^{na} = \tilde{F}_j^{na}$ où $\hat{F}_j^{na}$ et $\tilde{F}_j^{na}$ sont les fonctions de répartitions empiriques des (x_{ijk}^{na}) et de leurs estimations $(\tilde{x}_{ijk}^{na})$ dans le réplicat j, ces estimations étant faites à partir des autres réplicats de la même condition (hypothèse **H2**). Enfin, les données MNAR issues d'un réplicat sont engendrées par la limite de détection du spectromètre, et donc globalement plus petites que les données observées. Cette constatation permet de supposer que $q_j^{mnar}(100\%) < q_j^{obs}(100\%) < +\infty$ où q_j^{mnar} et q_j^{obs} sont les fonctions quantiles des (x_{ijk}^{mnar}) et (x_{ijk}^{obs}) dans le réplicat j de la condition k (hypothèse **H3**). Ces trois hypothèses vont nous permettre d'estimer la proportion de valeurs MCAR présentes parmi les valeurs manquantes d'un réplicat.

3 Estimer la proportion de valeurs MCAR

L'hypothèse **H3** va permettre l'identifiabilité de π_{mcar_j}. En effet, elle implique que

$$\lim_{x \to q_j^{mnar}(100\%)} \frac{1 - F_j^{na}(x)}{1 - F_j(x)} = \lim_{x \to q_j^{mnar}(100\%)} \pi_{mcar_j} + (1 - \pi_{mcar_j}) \frac{1 - F_j^{mnar}(x)}{1 - F_j(x)} = \pi_{mcar_j} \quad (3)$$

Sous l'hypothèse **H1**, si l'on note

$$R(x) = \sum_{i=1}^{n_j^{obs}} \mathbb{1}_{x_{ijk}^{obs} > x} \quad \text{et} \quad S(x) = \sum_{i=1}^{n_j^{na}} \mathbb{1}_{x_{ijk}^{na} > x}$$

où n_j^{obs} est le nombre de valeurs observées dans le réplicat j et n_j^{na} est le nombre de valeurs manquantes dans ce même réplicat, alors $R(x)$ et $S(x)$ sont deux variables indépendantes suivant les lois binomiales respectives $\mathcal{B}(n_j^{obs}, 1 - F_j^{obs}(x))$ et $\mathcal{B}(n_j^{na}, 1 - F_j^{na}(x))$.

On peut alors en déduire l'estimateur du maximum de vraisemblance (MV) de $\pi(x) = (1 - F_j^{na}(x))/(1 - F_j(x))$ lorsque $0 < F_j^{na}(x) < 1$ et $0 < F_j^{obs}(x) < 1$. Il est donné par $\hat{\pi}^{MV}(x) = s(x)/(\pi_{na_j}(s(x) + r(x)))$ où $s(x)$ et $r(x)$ sont les réalisations de $R(x)$ et $S(x)$. Bien entendu, $s(x)$ est inconnu mais peut être estimé par $\tilde{s}(x) = n_j^{na}(1 - \tilde{F}_j^{na}(x))$ où $\tilde{F}_j^{na}$ est définie au niveau de l'hypothèse **H2**. Ainsi, $\tilde{\pi}^{MV}(x) = \tilde{s}(x)/(\pi_{na_j}(\tilde{s}(x) + r(x)))$ constitue un estimateur MV approximatif de $\pi(x)$ qui est égale à l'estimateur MV exact sous l'hypothèse **H2**. Dans ce cadre, pour une proportion de valeurs manquantes π_{na_j} fixée, on peut montrer le résultat asymptotique suivant :

$$\sqrt{n}(\hat{\pi}^{MV}(x) - \pi(x)) \xrightarrow[n \to +\infty]{d} \mathcal{N}\Big(0, \sigma^2(\pi(x), 1 - F_j^{obs}(x))\Big)$$

où la fonction de variance asymptotique s'exprime comme

$$\sigma^2\Big(u, v\Big) = \frac{1 - \pi_{na_j}}{\pi_{na_j}} \frac{h(u,v)}{g(u,v)h(u,v) - k^2(u,v)} \quad (4)$$

avec

$$g(u,v) = \frac{\delta(u)v(1 - \delta(u)v)}{(1 - \pi_{na_j}u)^2} \left(\frac{1}{u} + \frac{v(1 - \pi_{na_j})}{\pi_{na_j}(v-1)u - vu + 1}\right)^2 \quad (5)$$

$$h(u,v) = \frac{\pi_{na_j}^{-1} - 1}{v(1-v)} + \delta(u)v(1 - \delta(u)v)\left(\frac{1}{v} + \frac{u(1 - \pi_{na_j})}{\pi_{na_j}(v-1)u - vu + 1}\right)^2 \quad (6)$$

$$k(u,v) = \frac{(1 - \delta(u)v)(1 - \pi_{na_j})}{(\pi_{na_j}(v-1)u - vu + 1)^2} \quad (7)$$

où $\delta(u) = \frac{(\pi_{na_j} - 1)u}{(\pi_{na_j}u - 1)}$.

En conséquence, sous les hypothèses **H1**, **H2** et **H3**, $\tilde{\pi}^{MV}(x)$ est un estimateur asymptotiquement sans biais de π_{mcar_j} lorsque $n \to +\infty$ et $x \in [q_j^{mnar}(1), m_j[$, où $m_j = \min(\max_i(\tilde{x}_{ijk}^{na}), \max_i(x_{ijk}^{obs}))$. Malheureusement, il peut être montré que

$$\sigma^2\Big(\pi(x), 1 - F_j^{obs}(x)\Big) \to +\infty$$

lorsque $x \to q_j^{obs}(1)$, où σ^2 est définie en (4). Ainsi, bien que $\tilde{\pi}^{MV}(x)$ soit asymptotiquement sans biais, on peut s'attendre à ce que sa variance diverge à proximité de m_j. Pour affiner l'estimation de π_{mcar_j}, l'hypothèse d'une loi paramétrique pour la répartition des valeurs MNAR va permettre de modéliser le biais de $\tilde{\pi}^{MV}(x)$. Un choix particulièrement adapté à notre cas est celui d'une loi de Weibull translatée $F_j^{mnar}(x) = 1 - \exp(-(\frac{1}{\lambda}\frac{x-l_j}{m_j-l_j})^d)$ où $d > 0$, $\lambda > 0$ et $l_j = \min(\min_i(\tilde{x}_{ijk}^{na}), \min_i(x_{ijk}^{obs}))$ est le minimum des données complétées dans le réplicat j. Nous discuterons plus amplement de l'adéquation de cette loi à notre problématique lors de notre communication. Cette hypothèse conduit à supposer un modèle représentant le biais et la variance asymptotiques de $\tilde{\pi}^{MV}(x)$:

$$\tilde{\pi}^{MV}(x) = \kappa + \frac{1-\kappa}{1-\tilde{F}_j(x)} \exp(-\alpha(\frac{x-l_j}{m_j-l_j})^d) + \epsilon(x) \tag{8}$$

où $\kappa \in [0,1]$, $\alpha > 0$, $d > 0$, $\tilde{F}_j(x) = \pi_{na_j}\tilde{F}_j^{na}(x) + (1-\pi_{na_j})\hat{F}_j^{obs}(x)$ est la fonction de répartition empirique des intensités complétées dans le réplicat j, et $\epsilon(x)$ est indépendamment et identiquement distribué suivant une loi normale centrée de variance $\sigma^2_{\epsilon(x)} = \sigma^2(\tilde{\pi}^{MV}(x), 1 - \hat{F}_j^{obs}(x))$ où σ^2 est définie en (4). À partir de ce modèle, des estimateurs $\hat{\kappa}$, $\hat{\alpha}$ et $\hat{d}$ des paramètres du modèle (8) sont déterminés en minimisant :

$$L(\kappa, \alpha, d) = \sum_{g=1}^{G-1} \frac{1}{\sigma^2_{\epsilon(y_g)}} \Big[\tilde{\pi}^{MV}(y_g) - \kappa - \frac{1-\kappa}{1-\tilde{F}_j(y_g)} \exp(-\alpha \times (\frac{y_g - l_j}{m_j - l_j})^d)\Big]^2 \tag{9}$$

où $y_g = l_j + g\frac{m_j - l_j}{G}$ et G est un nombre fixé de sous-intervalle de $[l_j, m_j]$ de longueurs égales. Un algorithme de type quasi-Newton peut être utilisé pour minimiser (9) en tenant compte des contraintes suivantes : $\kappa \in [0,1]$, $\alpha > 0$ et $d > 0$.

L'estimateur final de π_{mcar_j} est défini par $\hat{\kappa}$.

4 Simulations

On simule des jeux de données protéomiques par $x_{ijk} \underset{i.i.d}{\sim} \mathcal{N}(\mu_{ik}, 0.2)$ avec $\mu_{ik} \underset{i.i.d}{\sim} \mathcal{N}(25, 2)$ où $i \in [1, \ldots, n]$. D'un côté, les valeurs MCAR sont déterminées par un tirage aléatoire sans remise effectué uniformément parmi la liste des peptides $[1, \ldots, n]$. D'un autre côté, les valeurs MNAR sont générées par un tirage aléatoire sans remise parmi les peptides restant en respectant :

$$P(x_{ijk} \text{ is MNAR} | x_{ijk}) = \frac{1}{b} f_{\beta(1,b)}(\frac{x_{ijk} - min_{i\in[1,n]} x_{ijk}}{\max_{i\in[1,n]} x_{ijk} - \min_{i\in[1,n]} x_{ijk}}) \tag{10}$$

où $f_{\beta(1,b)}$ correspond à la densité de la loi Beta $\beta(1,b)$ où $b \geq 1$ permet d'ajuster la distribution des valeurs MNAR de façon plus ou moins proche de celle des valeurs MCAR.

		π_j^{mcar}				
		0.05	0.10	0.20	0.30	0.50
Notre estimateur $\hat{\kappa}$	Biais	3.64%	3.28%	2.80%	2.17%	1.31%
	EQM	(0.73%)	(0.88%)	(1.06%)	(1.12%)	(1.19%)
Estimateur logit	Biais	5.40%	1.30%	−6.44%	−13.25%	−22.68%
	EQM	(1.17%)	(0.83%)	(1.08%)	(2.33%)	(5.56%)
Estimateur de [3]	Biais	7.43%	5.01%	−1.06%	−8.44%	−28.43%
	EQM	(0.85%)	(0.56%)	(3.51%)	(3.84%)	(9.52%)

Tableau 1: Exemple de résultats empiriques obtenus pour $n = 2000$, $b = 4$, $\pi_{na_j} = 30\%$ et 5 réplicats présents dans $K = 200$ conditions (1000 réplicats simulés). *(Biais, EQM = Ecart moyen, erreur quadratique moyenne entre l'estimateur et la vraie valeur.)*

Les résultats des simulations montrent que, quelque soit la valeur de π_j^{mcar}, notre méthode permet d'obtenir une estimation de la proportion de valeurs MCAR peu biaisée par rapport à d'autres approches telles qu'une méthode utilisant une modélisation logit du mécanisme des valeurs manquantes, ou bien l'estimateur heuristique de [3] appliqué à notre problématique (Tableau 1). Au travers de ces simulations, il apparait également que, plus la répartition des valeurs MNAR est décalée à gauche, et plus le biais de l'estimateur est faible (non reporté ici). De plus, des résultats obtenus sur des jeux de données protéomiques réels ont montré que la proportion de valeurs MCAR parmi les valeurs manquantes est en général évaluée entre 10% et 20% (non reporté ici).

Bibliographie

[1] Karpievitch, Y. V., Dabney, A. R., & Smith, R. D. (2012). Normalization and missing value imputation for label-free LC-MS analysis. BMC bioinformatics, 13(Suppl 16), S5.

[2] Little, R. J., & Rubin, D. B. (2014). Statistical analysis with missing data. John Wiley & Sons.

[3] Patra, R. K., & Sen, B. (2015). Estimation of a two-component mixture model with applications to multiple testing. Journal of the Royal Statistical Society: Series B (Statistical Methodology).

[4] Taylor, S. L., Leiserowitz, G. S., & Kim, K. (2013). Accounting for undetected compounds in statistical analyses of mass spectrometry 'omic studies. Statistical applications in genetics and molecular biology, 12(6), 703-722.

[5] Webb-Robertson, B. J. M., Wiberg, H. K., Matzke, M. M., Brown, J. N., Wang, J., McDermott, J. E., ... & Waters, K. M. (2015). Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. Journal of proteome research, 14(5), 1993-2001.