Imputation de données manquantes pour l'inférence de réseau à partir de données RNA-seq

Alyssa Imbert [1], Caroline Le Gall[2], Claudia Armenise[3], Gregory Lefebvre[4], Jorg Hager[4], Armand Valsesia[4], Pierre-Antoine Gourraud[2], Nathalie Viguerie[5] & Nathalie Villa-Vialaneix[1]

[1] *INRA, UR 0875 MIAT, 31326 Castanet Tolosan cedex, France*
{*alyssa.imbert,nathalie.villa*}*@toulouse.inra.fr*
[2] *Methodomics, Toulouse,* {*legall,gourraud*}*@methodomics.com*
[3] *Quartzbio, Geneva, Suisse*
[4] *Nestlé Institute of Health Sciences, Lausanne, Suisse*
[3] *Inserm - UMR1048, Obesity Research Laboratory, Institut of Metabolic and Cardiovascular Diseases (I2MC), Toulouse,* *nathalie.viguerie@inserm.fr*

Résumé. Dans cette proposition de communication, nous nous intéressons au problème de l'inférence de réseaux de gènes à partir de données d'expression mesurées par RNA-seq. Nous présentons une approche qui permet d'intégrer de l'information externe (autre type d'omic par exemple) obtenue sur les mêmes individus ainsi que sur d'autres individus. Notre approche est présentée comme un problème d'imputation que nous résolvons avec des approches de type « hot deck » multiple pour obtenir un réseau plus stable. Nous illustrerons nos résultats sur des données réelles issues d'un programme d'études sur l'obésité. [1]

Mots-clés. réseau, données manquantes, imputation multiple, hot deck

Abstract. In this article, the issue of gene network inference is addressed, in which inference is performed from expression data obtained by RNA-seq sequencing technique. Our proposal aims at integrating external information (another kind of 'omic for instance) measured on the same individuals and on additional individuals. The method is presented as a missing data imputation problem and is solved with multiple hot deck approaches in order to infer a more stable network. Our results will be illustrated on real data coming from a paneuropean project on obesity.

Keywords. network, missing data, multiple imputation, hot deck

1. Ce projet est financé par la société Methodomics http://www.methodomics.com et la région Languedoc-Roussillon Midi-Pyrénées.

1 Introduction

En biologie, les technologies haut débit permettent l'acquisition d'une information riche sur le fonctionnement d'un organisme à divers niveau de l'échelle du vivant. En particulier, les dernières techniques de séquençage (RNA-seq), permettent de mesurer l'expression simultanée de plusieurs milliers de gènes dans un type donné de cellules. L'analyse statistique de ces données comprend plusieurs aspects : pré-traitement, analyse différentielle, classification, inférence de réseau... (voir [4] pour une vision d'ensemble des problématiques liées à ce type de données).

En particulier, l'inférence de réseau a pour but de fournir au biologiste une vision globale des relations de dépendances conditionnelles entre les expressions de gènes (souvent d'une liste restreinte de gènes choisis). Dans ce type d'approches, un graphe est reconstruit à partir de la mesure de l'expression des gènes chez plusieurs individus : les sommets de ce graphe représentent les gènes et les arêtes, les (principales) relations de dépendances conditionnelles entre les expressions de ces gènes.

Cependant, le coût associé à la collecte de ce type de données fait qu'en général, le nombre d'observations collectées (n) est souvent très faible particulièrement devant le nombre de gènes d'intérêt (p). Or, [6] a montré qu'on ne peut espérer une estimation satisfaisante de ce type de réseau à l'aide de modèles gaussiens parcimonieux dans le cadre de l'ultra haute dimension. Par ailleurs, l'inférence de réseau est souvent très sensible à la présence ou non de certaines observations et un certain nombre de travaux se sont intéressés à la question de la recherche des structures stables (au sens de communes à la majorité des observations disponibles) dans l'inférence : [5] propose une approche par ré-échantillonnage qui permet de sélectionner un paramètre de régularisation maximum assurant la stabilité de l'inférence. [2] ont montré que l'inférence de réseau peut être sensible à quelques observations dites « influentes » et ont proposé des mesures d'influence pour filtrer ces observations et stabiliser le réseau inféré.

Dans cette proposition de communication, nous nous étudions une approche différente : il est, en effet, commun de mesurer, simultanément aux données de RNA-seq, d'autres types de données 'omiques. D'un coût moins élevé, ces données sont fréquemment disponibles pour un nombre plus grand d'individus. Nous présentons ici une approche utilisant cette information supplémentaire au travers de méthodes d'imputation d'individus manquants et étudions sa pertinence en terme d'inférence de réseau. La proposition de communication est organisée comme suit : dans la section 2, nous décrivons une approche d'imputation que nous mettons en œuvre sur ces données dans le cadre de l'inférence de réseau. Puis, dans la section 3, nous présentons les données du projet DiOGenes sur lesquelles nous travaillons. Enfin, dans la section 4, nous présentons quelques résultats préliminaires montrant l'influence des individus utilisés pour l'inférence sur le réseau lui-même et les réseaux obtenus à partir de l'imputation multiple.

2 Présentation de la méthode d'imputation

L'approche que nous proposons se place dans le cadre où deux types de données sont observés : le premier type de données correspond aux données d'expression d'intérêt qui sont contenues dans une matrice $\mathbf{X}$ de dimension $n_1 \times p$ (n_1 individus et p gènes) dont les entrées sont des données de comptages (entiers). La second type de données est celui sur lequel nous pouvons nous appuyer pour l'inférence : il s'agit d'une matrice $\mathbf{Y}$ de mesures numériques de dimension $n \times q$ où $n > n_1$ est le nombre d'individus sur lequels les mesures ont été effectuées et q est le nombre de variables observées. Tous les individus pour lesquels des données d'expression sont disponibles sont observés dans $\mathbf{Y}$ et on peut, sans perte de généralité, supposer qu'ils correspondent aux n_1 premières lignes de cette matrice et que $\forall\, i = 1, \ldots, n_1$, $\mathbf{x}_i$ (i-ème ligne de la matrice $\mathbf{X}$) correspond au même individu que $\mathbf{y}_i$ (i-ème ligne de la matrice $\mathbf{Y}$).

Ce problème peut être compris comme un problème d'imputation de valeurs manquantes dans une matrice $[\widetilde{\mathbf{X}},\ \mathbf{Y}]$ de dimension $n \times (p+q)$ dans laquelle, $\tilde{\mathbf{x}}_i = \mathbf{x}_i$, $\forall\, i = 1, \ldots, n_1$ et $\tilde{\mathbf{x}}_i$ est manquant $\forall\, i \geq n_1 + 1$. De manière similaire à [7], afin de conserver la structure de corrélation entre les variables imputées, des lignes entières sont imputées à partir d'individus existant par une méthode de type « hot deck ». Cette procédure est appliquée selon une approche d'imputation multiple qui permet d'observer la stabilité des arêtes inférées.

De manière plus précise, nous mettons en œuvre une méthode d'imputation multiple de type « hot deck » comme décrite dans [3]. Celle-ci repose sur la création, pour un individu $i \in [\![n_1 + 1, n]\!]$, d'un ensemble de « donneurs », $\mathcal{D}(i)$, qui sont des observations $j \in [\![1, n_1]\!]$. Cet ensemble peut être déterminé :

— soit par le calcul d'un score dit d'affinité, calculé pour tous les individus $j \in [\![1, n_1]\!]$. Ce score est de la forme :

$$s(i,j) = \frac{1}{q} \sum_{k=1}^{q} \mathbb{I}_{\{|\mathbf{y}_{ik} - \mathbf{y}_{jk}| < \sigma\}}$$

où σ est un seuil fixé et $\mathcal{D}(i)$ est défini comme l'ensemble $\{j :\ s(i,j) = \max_{l \neq i} s(i,l)\}$;

— soit par une méthode de plus proches voisins : dans ce cas, l'ensemble $\mathcal{D}(i)$ est défini comme l'ensemble des K (où K est un entier fixé) plus proches voisins au sens de la métrique euclidienne usuelle : $d(i,j) = \sum_{k=1}^{q} (\mathbf{y}_{ik} - \mathbf{y}_{jk})^2$.

Contrairement à [3], qui impute les variables $k = 1, \ldots, p$ indépendamment les unes des autres, l'intégralité de la ligne $\tilde{\mathbf{x}}_i$ est imputée par la ligne $\tilde{\mathbf{x}}_j$ en prenant au hasard une observation $j \in \mathcal{D}(i)$. Cette opération est effectuée B fois pour toutes les observations $i = 1, \ldots, n_1$ pour obtenir B matrices imputées différentes $\mathbf{X}^{*,b}$ ($b = 1, \ldots, B$) de tailles $n \times p$. Les B tableaux de données imputées sont ensuite utilisés pour l'inférence d'un réseau de gènes. Pour l'inférence, nous utilisons un modèle log-linéaire de Poisson, adapté

aux données de comptage, comme décrit dans [1]. Enfin, le réseau final retenu correspond au réseau dans lequel les arêtes les plus stables sont conservées.

3 Une étude sur l'impact de la restriction alimentaire chez les obèses : le projet DiOGenes

Les données sur lequelles ont été mises en œuvre les méthodes décrites dans la section 2 sont issues du projet pan-européen DiOGenes. Ce projet est une étude d'intervention diététique contrôlée sur des personnes obèses, réalisée dans 8 pays européens. Les sujets inclus dans cette étude suivent un régime très faible calories de 8 semaines avec pour objectif de perdre au moins 8% de leur poids initial. De nombreuses mesures biologiques ont été réalisées avant le début du régime (CID1) et à l'issue de celui-ci (CID2). En particulier, des mesures transcriptomiques ont été obtenues à partir de biopsie de tissus adipeux selon trois types de techniques de séquençage : qPCR, puces à ADN, et RNA-seq. Dans la suite, nous nous intéresserons à l'inférence du réseau obtenu à partir d'une sélection de 317 gènes dont l'expression a été mesurée par la technique RNA-seq en CID1 et CID2. Le but final de l'étude est la compréhension de l'impact de la restriction alimentaire sur la structure de dépendance dans cet ensemble de gènes. Le nombre d'observations disponible correspond à 433 individus en CID1 et 307 individus en CID2. Pour des raisons techniques, seuls 189 observations correspondent à des individus en commun entre les deux expériences.

4 Premiers résultats

L'objectif étant de comparer les réseaux obtenus à partir des données collectées en CID1 et en CID2, une première étude a consisté à comparer les réseaux obtenus à partir de l'ensemble des individus disponibles aux deux pas de temps (resp. 433 et 307 observations) aux réseaux obtenus à partir des individus communs aux deux observations (189 individus) : les premiers ont a priori une fiabilité globale plus importante (inférence effectuée sur $1,5$ à $2,5$ fois plus d'individus) mais les seconds semblent plus pertinents dans un but de comparaison des deux pas de temps (inférence effectuée à partir d'individus communs). L'inférence est effectuée à l'aide d'un modèle log-linéaire de Poisson comme décrit dans [1] et le critère de sélection du nombre d'arête est le critère StARS introduit dans [5].

Les résultats de cette première analyse, sont résumés dans le tableau 1 en terme de nombre d'arêtes communes entre les différents réseaux inférés. Ceux-ci montrent une forte influence du groupe d'individus considérés qui modifie le réseau inféré avec de l'ordre de 1/3 des arêtes qui diffèrent.

L'objectif du travail présenté lors de cette communication sera de mettre en œuvre

dataset	CID1 (tous)	CID1 (communs)	CID2 (tous)	CID2 (communs)
CID1 (tous)	1663			
CID1 (communs)	1065	1540		
CID2 (tous)	835	718	1576	
CID2 (communs)	750	659	1093	1486

TABLE 1 – Nombre d'arêtes en commun dans les deux réseaux inférés. La diagonale indique le nombre total d'arêtes du réseau correspondant.

dataset	CID1 (tous)	CID1 (communs)	CID1 imp1	CID1 imp2	CID1 imp3	CID1 imp4	CID1 imp5
CID1 (tous)	1663						
CID1 (communs)	1065	1540					
CID1 imp1	1521	1167	2739				
CID1 imp2	1512	1182	2330	2757			
CID1 imp3	1509	1186	2314	2330	2756		
CID1 imp4	1515	1169	2320	2364	2351	2752	
CID1 imp5	1507	1172	2355	2349	2335	2320	2745

TABLE 2 – Nombre d'arêtes en commun dans les deux réseaux inférés à partir des données initiales et les réseaux inférés à partir des cinq tables imputées. La diagonale indique le nombre total d'arêtes du réseau correspondant.

l'approche d'imputation décrite dans la section 2 en utilisant des données externes pour simuler des individus communs manquant dans l'une ou l'autre des deux conditions. Pour ce faire, nous utilisons des données d'expression mesurées en qPCR qui sont disponibles pour tous les individus aux deux pas de temps : cette approche permet d'améliorer l'inférence en augmentant le nombre d'observations, tout en conservant une structure de corrélation réaliste entre les gènes. Enfin, l'imputation multiple permettra de dériver une mesure de confiance et de stabilité de l'inférence.

Un travail préliminaire a été réalisée inférant 100 réseaux pour des tableaux de données imputées à CID1 en utilisant l'approche par affinité. Par étude de l'évolution d'un critère d'inertie moyenne intra-$\mathcal{D}(i)$ en fonction de σ, nous avons fixé σ à 3. Les résultats sont résumés dans le tableau 2 en terme de nombre d'arêtes communes entre les différents réseaux obtenus à partir des jeux imputés et dans la figure 1 qui représente la distribution de l'apparition d'une arête sur les différents réseaux. Ces résultats préliminaires montrent une assez grande stabilité de l'inférence à partir des données imputées (avec 1892 apparaissant plus de 90% du temps et 1376 arêtes apparaissant 100 fois). Ces premiers résultats sont encourageants pour la méthode proposée. Des résultats plus complets seront présentés lors de la conférence.

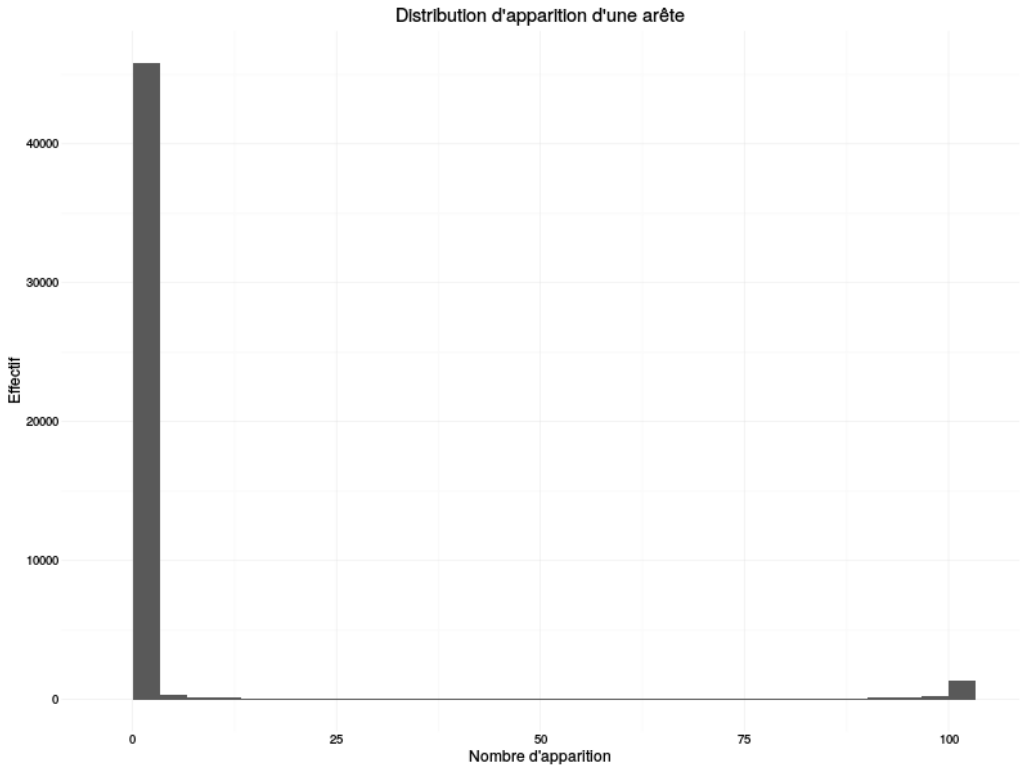

FIGURE 1 – Distribution d'apparition d'une arête dans les réseaux inférés à partir de cinq jeux de données imputés (CID1)

Références

[1] G. Allen and Z. Liu. A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2012.

[2] A. Bar-Hen and J.M. Poggi. Influence measures and stability for graphical models. *Journal of Multivariate Analysis*, 147 :145–154, 2016.

[3] S.J. Cranmer and J. Gill. We have to be discrete about this : a non-parametric imputation technique for missing categorical data. *British Journal of Political Science*, 43 :425–449, 2012.

[4] M. Gallopin. *Classification et inférence de réseaux pour les données RNA-seq.* Thèse de doctorat, Université Paris-Sud, Paris, France, 2015.

[5] H. Liu, K. Roeber, and L. Wasserman. Stability approach to regularization selection (StARS) for high dimensional graphical models. In *Proceedings of Neural Information Processing Systems (NIPS 2010)*, volume 23, pages 1432–1440, Vancouver, Canada, 2010.

[6] N. Verzelen. Minimax risks for sparse regressions : ultra-high-dimensional phenomenons. *Electronic Journal of Statistics*, 6 :38–90, 2012.

[7] V. Voillet, I. Gonzáles, L. Liaubet, P. Besse, and M. San Cristobal. Recovering missing individual block information in a multiblock multiple factor analysis. In *missData 2015*, Rennes, France, 2015.