Imputation multiple de type hot-deck pour l'inférence de réseaux à partir de données RNA-seq

Alyssa Imbert
Encadrée par Nathalie Villa-Vialaneix et Nathalie Viguerie

Séminaire des doctorants

28/04/2017

Sommaire

Motivation

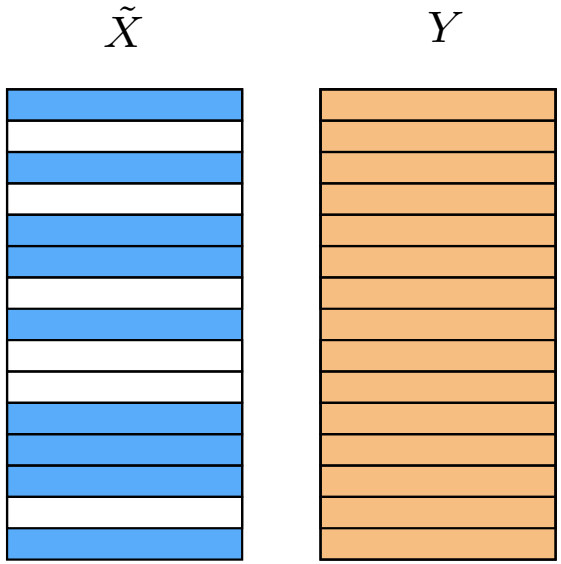
- Données RNA-seq : généralement, peu d'échantillons
  $\hookrightarrow$ inférence de réseau difficile
- L'inférence de réseau est sensible aux observations influentes, *Bar-Hen A., 2016*.

- Objectif : Trouver une solution pour limiter la perte d'information

- Données annexes : apport d'information
  $\hookrightarrow$ utiliser cette information supplémentaire pour améliorer l'inférence de réseau

Notation

- Matrice $\tilde{X}$ de dimension $n_1 \times p$ comprenant les mesures d'expression d'intérêt (données RNA-seq) ;
- matrice Y de dimension $n \times q$ de mesures numériques (métabolome, données phénotypiques, expression qPCR,...) ;
- n_1 échantillons (individus) en commun entre $\tilde{X}$ et Y ;
- présence de données manquantes $\longrightarrow$ raisons expérimentales



Problématique

Recherche d'une méthode d'imputation qui permet de :

- préserver le lien existant entre les variables
  $\rightarrow$ imputer les individus manquants **en entier**

- Prendre en compte l'incertitude liée à l'imputation des données manquantes

Objectif : augmenter la qualité de l'inférence en utilisant de l'information externe (important car données avec n petit)

Sommaire

Modèle graphique log-linéaire de Poisson

Allen G.I. et Liu Z., 2012

- Transformation puissance des données : $x_{ij} \rightarrow x_{ij}^{\alpha}$, $\alpha \in]0, 1]$
- Soit $z_j = (x_{1j}^{\alpha}, ..., x_{nj}^{\alpha})$ le vecteur des valeurs d'expression transformées pour le gène j

$$p(Z_{ij}|z_{i(-j)}) \sim \mathcal{P}(\mu_j) \text{ avec } \log(\mu_j) = \sum_{j' \neq j} \beta_{jj'} \tilde{z}_{ij'}$$

où $\tilde{z}$ correspond aux données log-transformées et réduites.

- arête entre gènes j et j' $\Leftrightarrow$ $\beta_{jj'}\beta_{j'j} \neq 0$
- modèle parcimonieux $\rightarrow$ ajout d'une pénalisation lasso avec paramètre de régularisation λ à la log-vraisemblance
- Choix de λ par une procédure de rééchantillonnage : critère StARS [1]

1. Stability Approach to Regularization Selection *Liu H. et al., 2010*

Sommaire

Définitions

Imputation hot-deck [2]

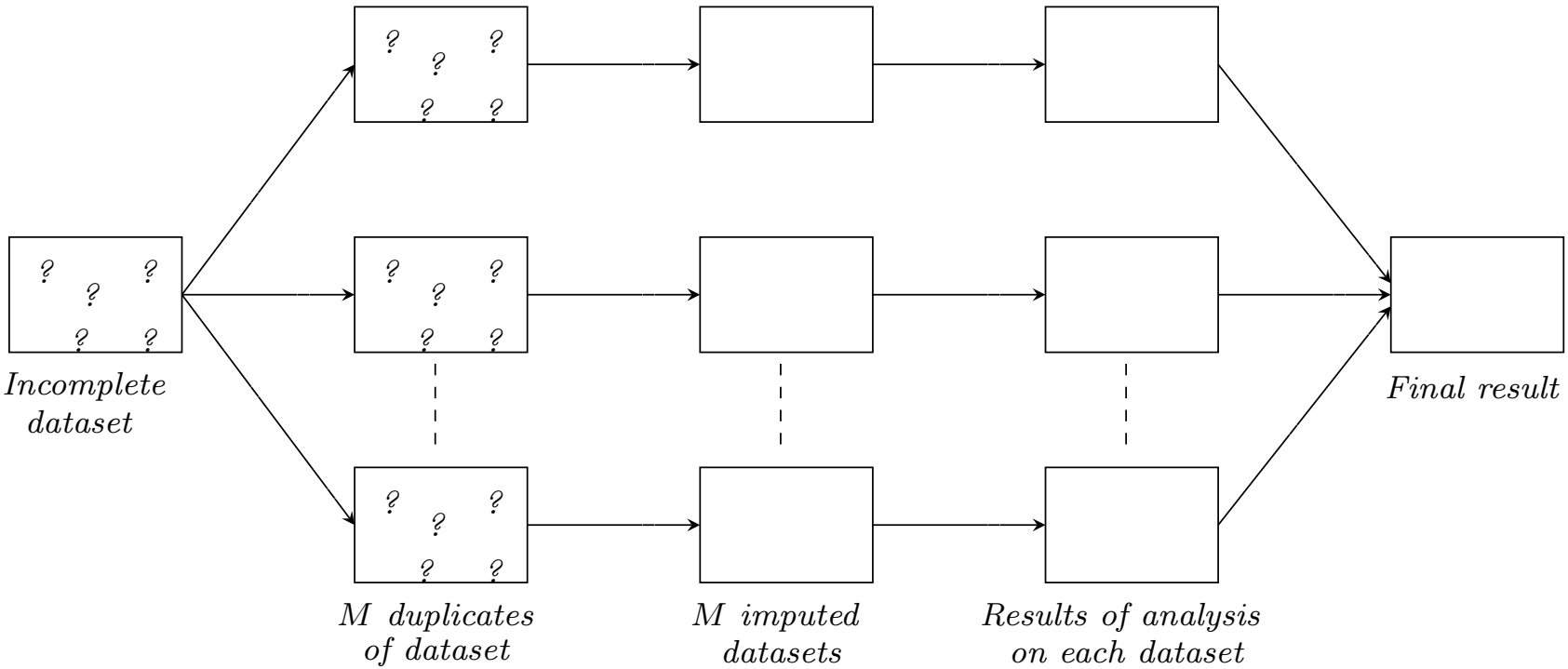
- Regroupe un ensemble de méthodes basées sur le concept de "donneurs"

- Pour chaque variable manquante,
  - rechercher des individus similaires à partir des données observées $\rightarrow$ groupe de donneurs
  - sélectionner un donneur au hasard
  - imputer la valeur de ce donneur pour la donnée manquante

2. Revue : *Andridge R. R. et Little R. J. A., 2010*

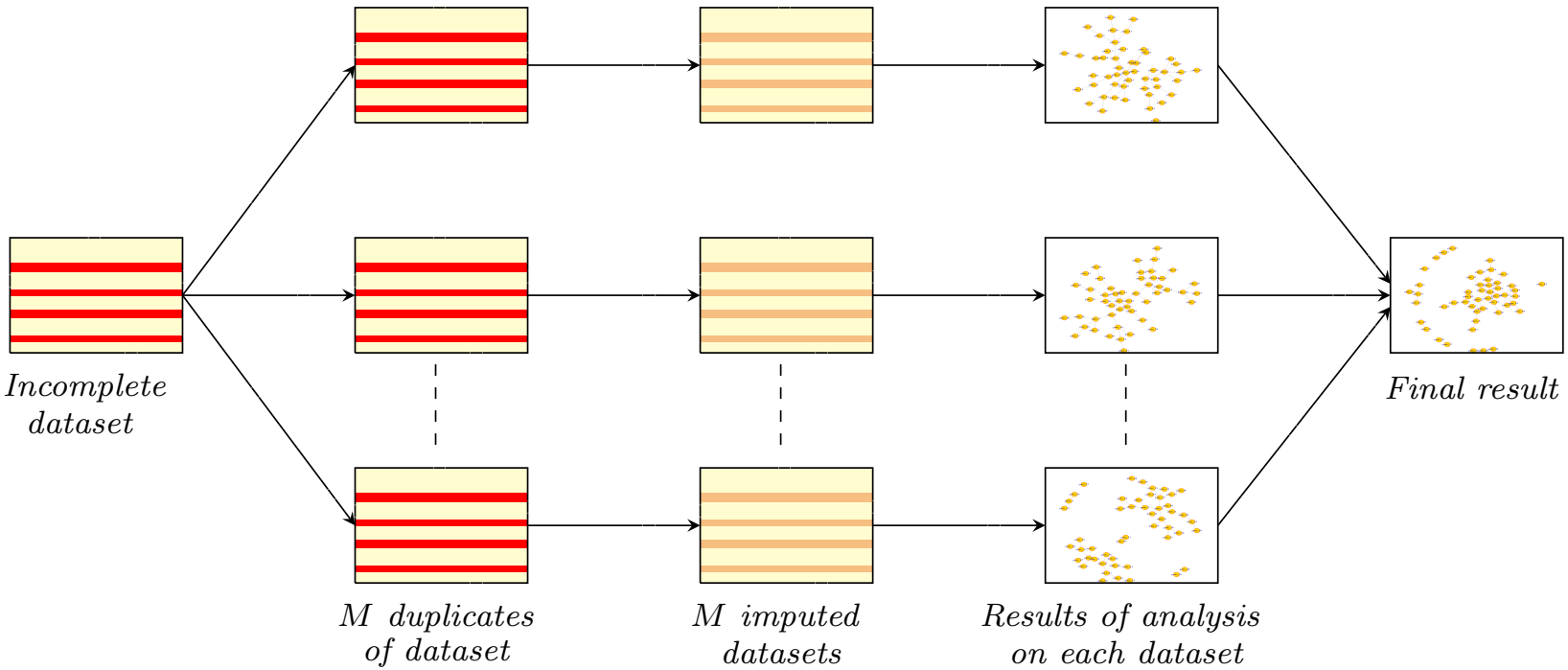
Définitions

Imputation multiple



Imputation multiple hot-deck (hd-MI)

Schéma général



Imputation
hot-deck

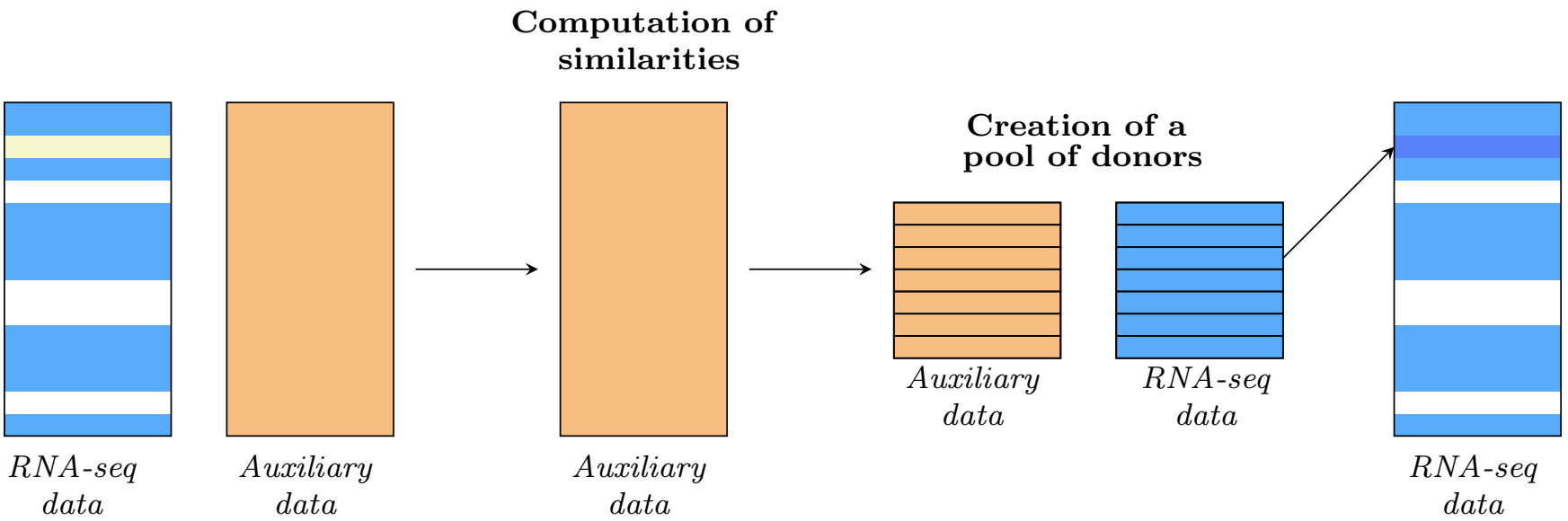
Network inference
llgm + StARS

"Pool"
study of frequency of appearance of edges

llgm = log-linear Poisson graphical model (*Allen et Liu, 2012*)

Imputation multiple hot-deck (hd-MI)

Première étape : imputation hot-deck



Imputation multiple hot-deck

Tester 2 approches :

- avec **un score d'affinité** [3] (package R *hot.deck*) :

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

où σ = seuil fixé et $\mathcal{D}(i) = \{j : \ s(i,j) = \max_{l\neq i} s(i,l)\}$

- avec k **plus proches voisins** (knn), au sens métrique euclidienne :

$$d(i,j) = \sum_{k=1}^{q}(y_{ik} - y_{jk})^2$$

3. *Cranmer S.J. and Gill J., 2012*

Choix du seuil σ

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

Critère : Etude de l'inertie moyenne intra-$\mathcal{D}(i)$:

$$V_{intra} = \frac{\sum_i \frac{\sum_{\text{d : donor of i}} (x_i - x_d)^2}{D_i}}{n}$$

où n est le nombre d'individus manquants et D_i le nombre de donneurs pour l'individu i.

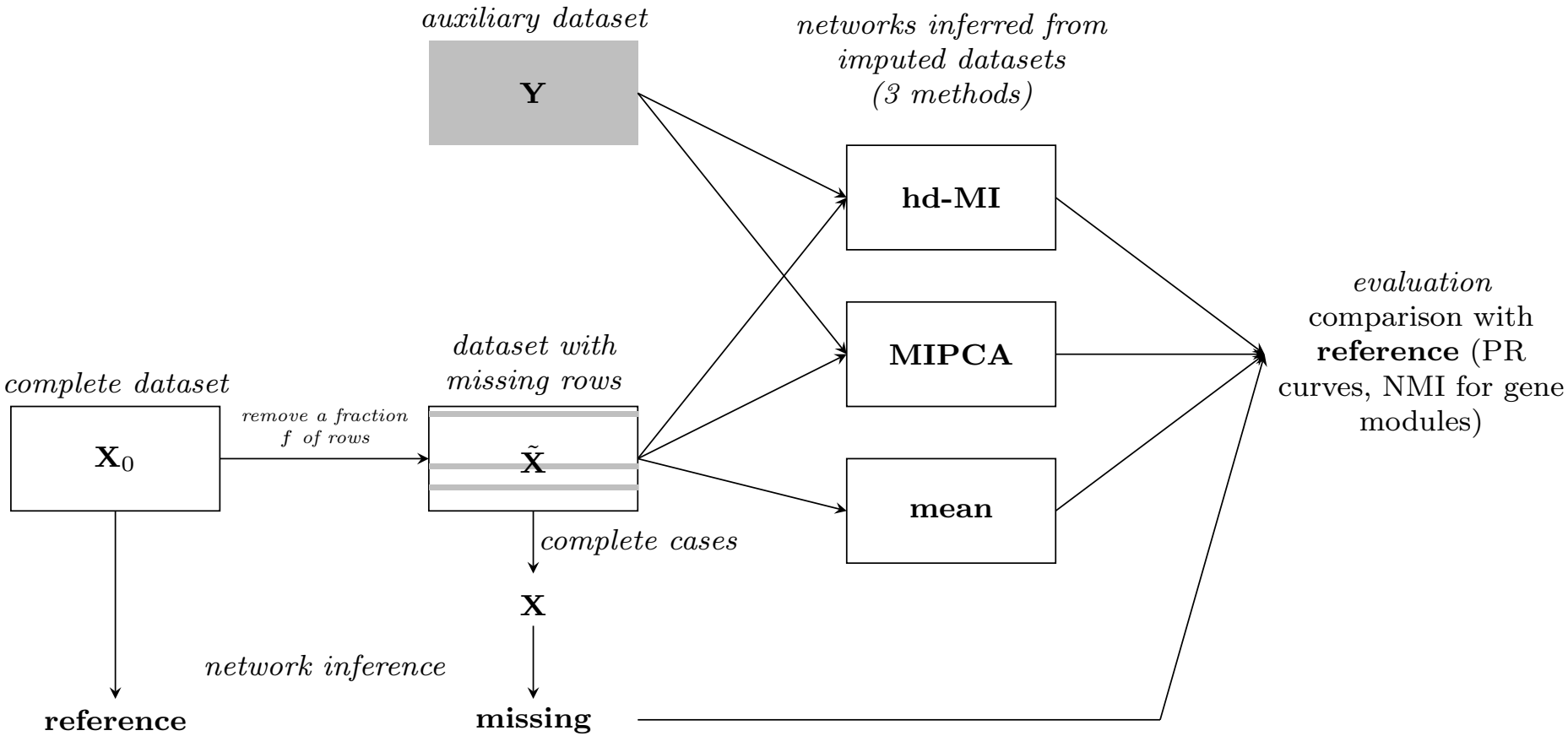
Sommaire

Cadre

- Test sur données réelles issus de deux projets :
  - GTEx : Genotype-Tissue Expression [4],
  - DiOGenes [5],
- 3 méthodes d'imputation :
  - imputation simple et naïve : imputation par la moyenne
  - imputation multiple par ACP : MIPCA, *Josse et al., 2011*
  - notre méthode : hd-MI
- 10%, 20%, 30%, 40% et 50% d'individus manquants

4. *Lonsdale et al., 2013*
5. *Larsen et al., 2010*

Schéma du processus d'évaluation



Sommaire

Présentation des données

GTEx

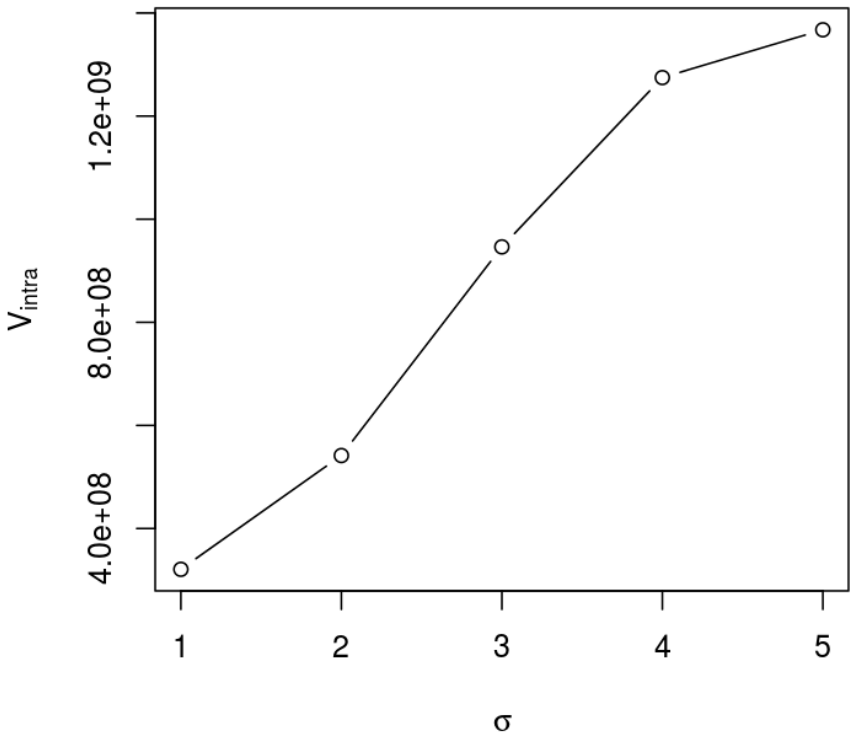
- Données RNA-seq sur plus de 30 tissus humains ;
- **choix des tissus** : deux tissus dont le profil d'expression est proche [6]
  - X : lung,
  - Y : thyroid,
- normalisation des données RNA-seq : TMM package edgeR ;
- description des jeux de données :
  - 320 échantillons pour X,
  - 323 échantillons pour Y,
- évaluation : garder seulement les **221 échantillons en commun** ;
- **sélection des gènes les plus variables** (ceux avec les variances les plus élevées) :
  - pour X : p= 100,
  - pour Y : q = 50,
  - 36 gènes en commun.
- Résultat pour **20%** d'observations manquantes

6. *Melé et al., 2015*

Choix de sigma et distribution d'apparition des arêtes

GTEx, 20% d'observations manquantes

Choix de sigma

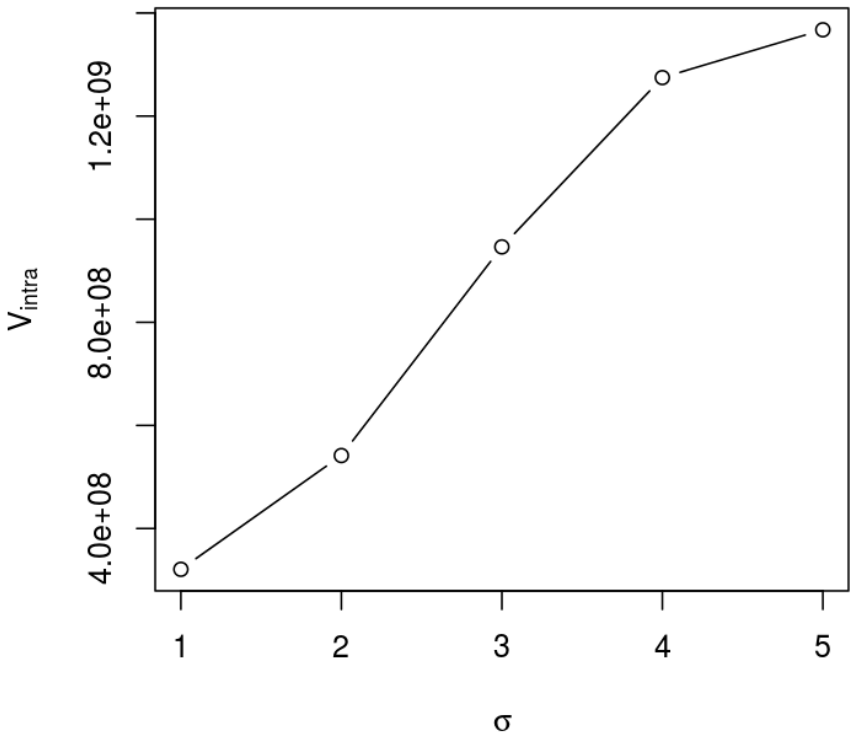


Choix : $\sigma = 2$

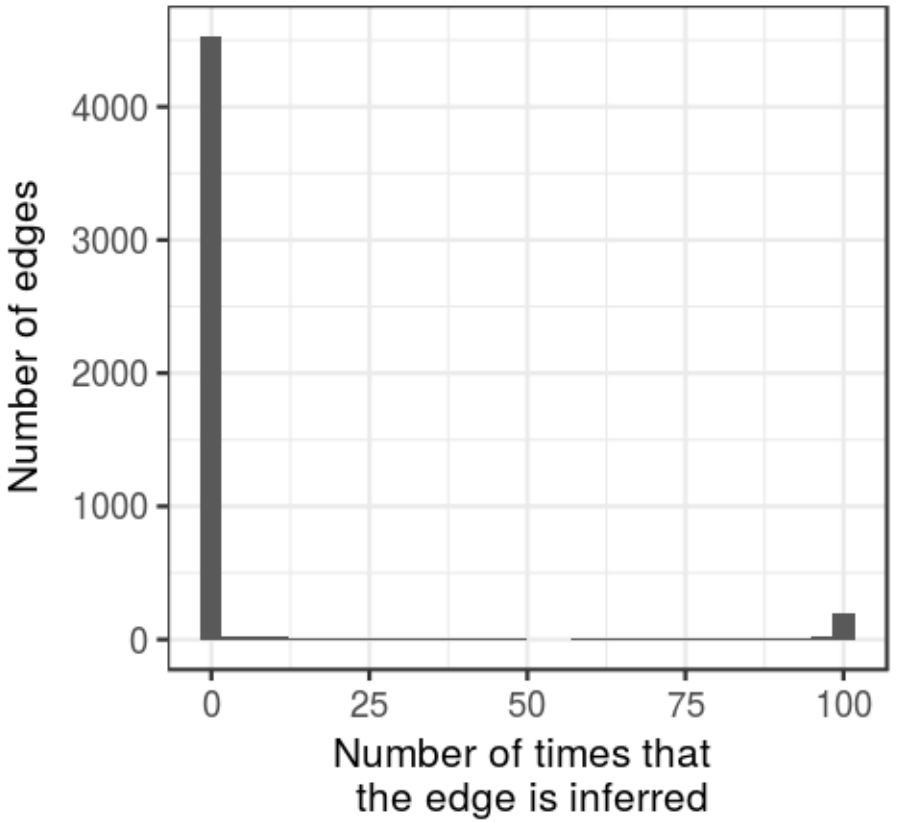
Choix de sigma et distribution d'apparition des arêtes

GTEx, 20% d'observations manquantes

Choix de sigma



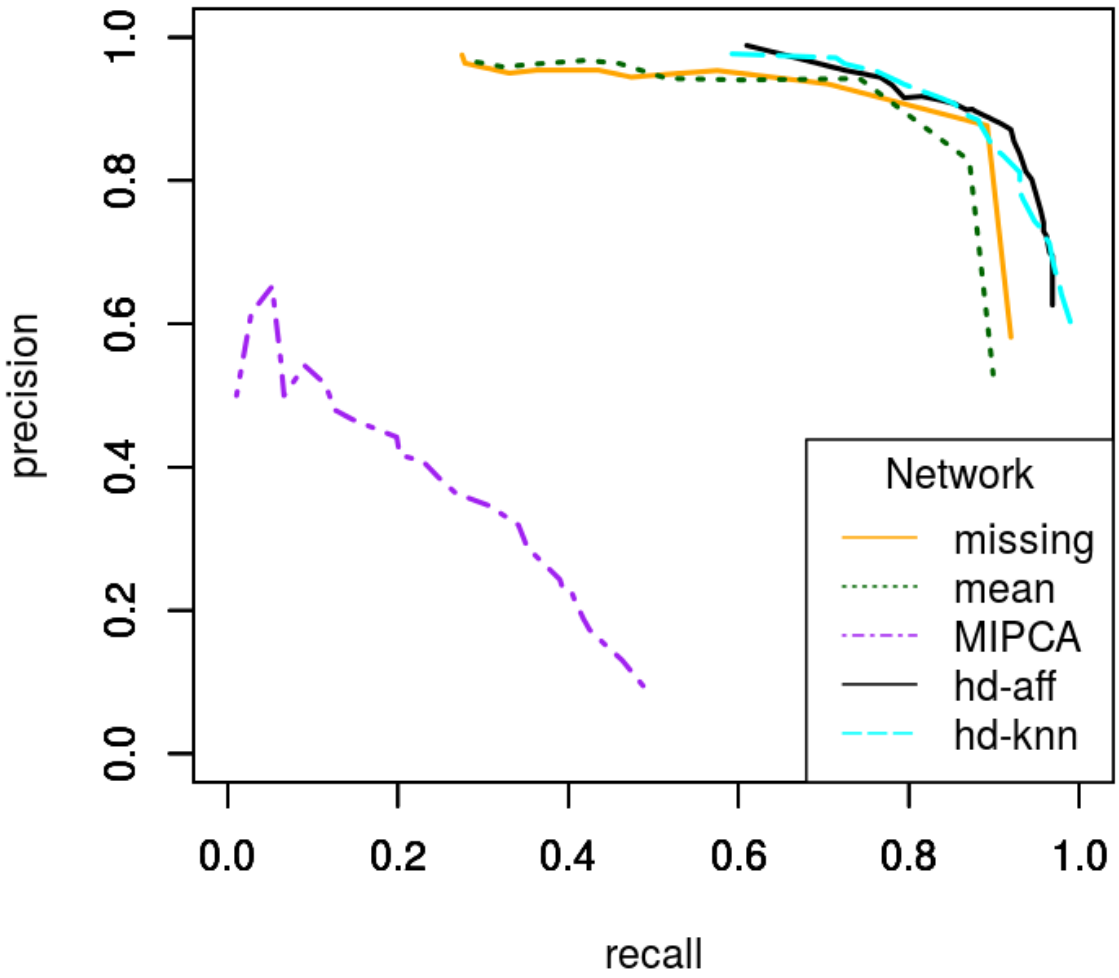
Distribution d'apparition d'une arête



Choix : $\sigma = 2$

Courbe précision/rappel

GTEx, 20% d'observations manquantes



Comparaison des modules de gènes

GTEx, 20% d'observations manquantes

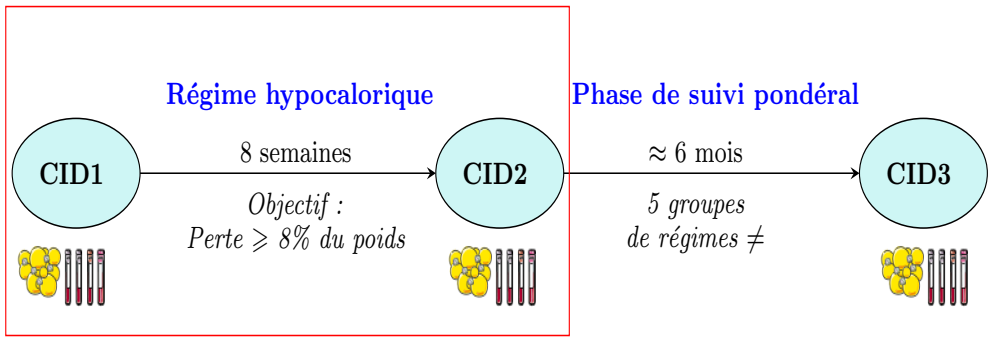
- Recherche de modules de gènes sur la plus grande composante du réseau : classification de sommets
  $\hookrightarrow$ fonction *spinglass_community()*
- comparaison des modules de gènes avec ceux du réseau de référence : NMI[7]
  $\hookrightarrow$ NMI compris entre $[0, 1]$
  $\hookrightarrow$ NMI $= 1$: les modules entre les deux réseaux sont identiques

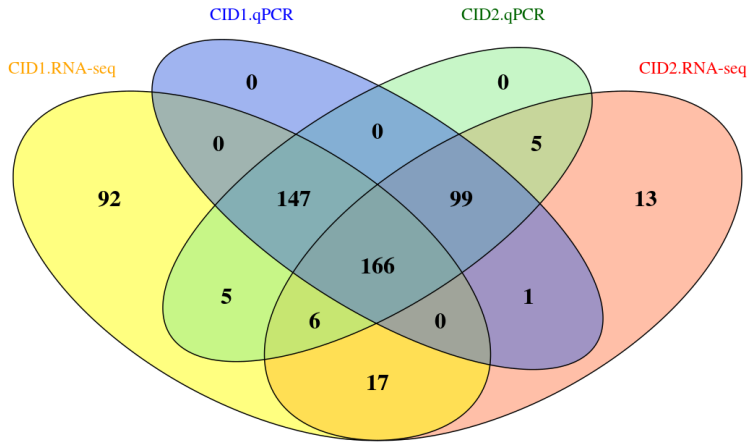
graph	reference	missing	mean	MIPCA	hd-aff	hd-knn
$\sharp$ modules	7	7	7	1	8	8
NMI		0.557	0.573	1[8]	**0.667**	0.603

7. normalized mutual information measure, *Danon L. and al (2005)*
8. 3 gènes seulement dans la plus grande composante

Présentation des données

DiOGenes





RNA-seq :

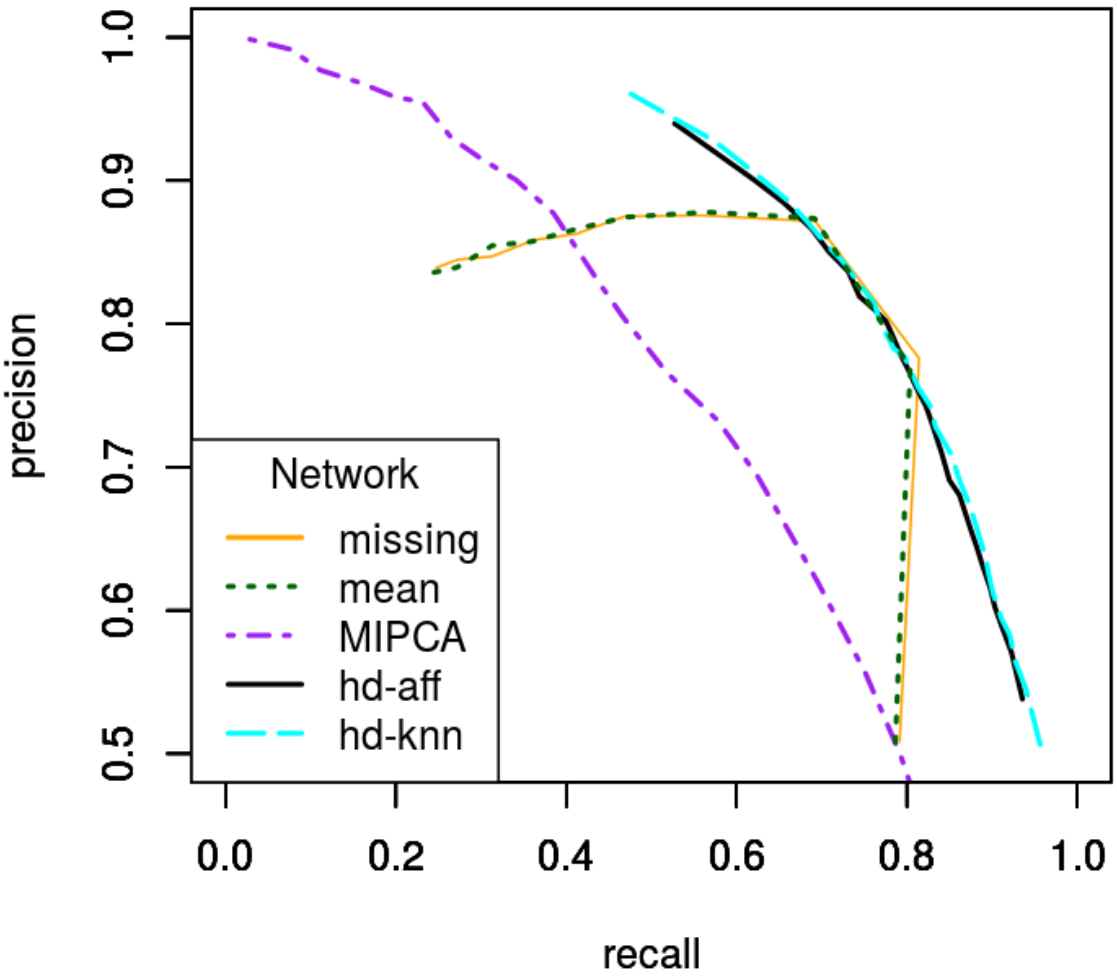
- 433 individus en CID1,
- 307 individus en CID2,
- **189 individus en commun**,
- 317 gènes

Jeu suppport : RT-qPCR :

- 166 individus pour CID1,
- 172 individus pour CID2,
- 284 gènes.

Courbes PR, CID1

DiOGenes, 20% d'observations manquantes



Modules de gènes, CID1

DiOGenes, 20% d'observations manquantes

- Recherche de modules de gènes sur la plus grande composante du réseau : fonction *spinglass_community()*
- comparaison des modules de gènes avec ceux du réseau de référence : NMI

graph	reference	missing	mean	MIPCA	hd-aff	hd-knn
♯ modules	7	7	7	10	8	8
NMI		0.526	0.612	0.346	0.493	0.492
NMI avec CID2	0.423	0.421	0.424	0.341	0.38	0.383

Conclusion

- Pour une précision élevée, meilleur recall avec l'imputation multiple hot-deck
- moins de faux positifs avec hd-MI
- GTEx : meilleur nmi (en comparant avec le réseau référence) avec hd-MI
  $\longrightarrow$ préserve les modules de gènes
- Au delà de 30% d'observations manquantes, les résultats se dégradent :
  $\longrightarrow$ la courbe PR pour hd-MI est en dessous de celle construite avec des observations manquantes

Perspectives

- Article : fin rédaction
- Package R : RNAseqNet

Merci pour votre attention

StARS

Choix de λ avec le critère StARS :

- création d'un vecteur Λ avec des valeurs décroissantes de λ
- sous-échantillons de X
- inférer un graphe pour chaque sous-échantillon et paramètre de régularisation du vecteur Λ
- $\lambda_{opt} = argmin_{\lambda} \left\{ min_{0 \leq \rho \leq \lambda} \left[\sum_{j<k} 2\bar{A}_{jk}(\rho)(1 - \bar{A}_{jk}(\rho)) / \binom{p}{2} \right] \leq \beta \right\}$ où $\bar{A}_{jk}(\lambda) = \frac{1}{B} \sum_{b=1}^{B} A_{jk}^{(b)}$, $\beta = 0.05$ par défaut

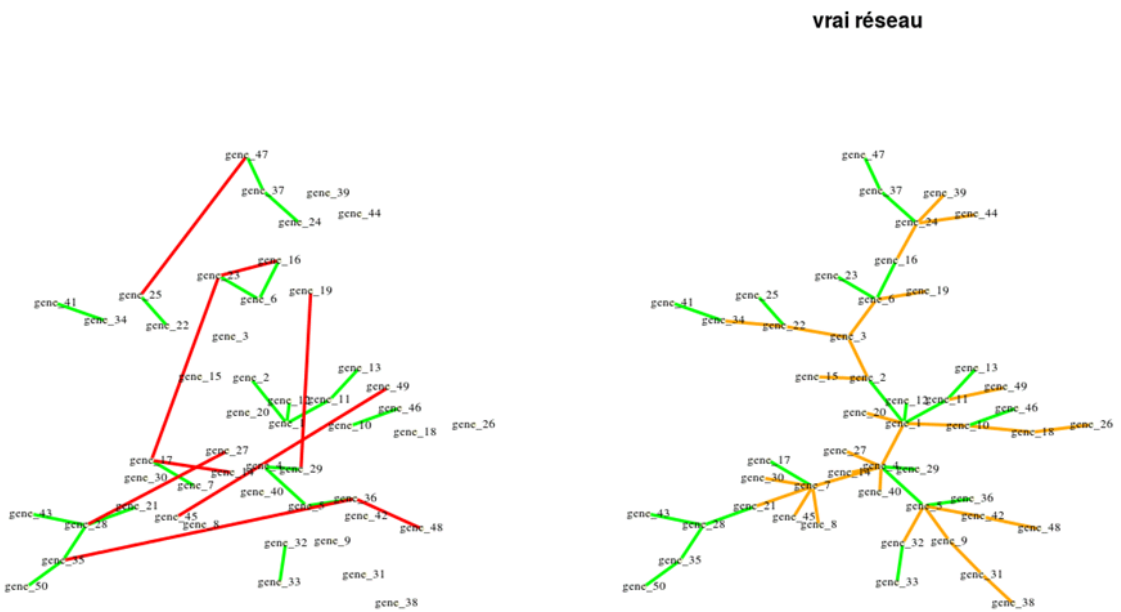
Précision/rappel

- Précision : $\mathrm{Pr} = VP/(VP + FP)$

$$\frac{\text{nombre d'arêtes } \textbf{prédites} \text{ présentes dans le graphe "vrai"}}{\text{nombre total d'arêtes prédites}}$$

- Rappel : $\mathrm{R} = VP/(VP + FN)$

$$\frac{\text{nombre d'arêtes } \textbf{prédites} \text{ présentes dans le graphe "vrai"}}{\text{nombre d'arêtes dans le graphe "vrai"}}$$



NMI

$$NMI(A,B) = \frac{-2\sum_{i=1}^{C_a}\sum_{j=1}^{C_b} N_{ij} log(\frac{N_{ij}N}{N_{i.}N_{.j}})}{\sum_{i=1}^{C_a} N_{i.} log(\frac{N_{i.}}{N}) + \sum_{j=1}^{C_b} N_{.j} log(\frac{N_{.j}}{N})}$$

- N : matrice de confusion, lignes = communautés "réelles", colonnes = communautés "trouvées"
- C_a = nombre de vraies communautés ; C_b : nombre de communautés trouvées
- N_{ij} = nombre de noeuds dans la vraie communauté i qui apparaît dans la communauté trouvée j
- $N_{i.}$ = somme sur la ligne i de la matrice N_{ij}, $N_{.j}$ = somme sur la colonne j