Modèles markoviens graphiques pour la fusion de données individuelles et d'interactions : application à la classification de gènes

Matthieu VIGNES

Thèse préparée au sein de l'équipe Mistis (INRIA Rhône-Alpes – LJK – Grenoble),
sous la direction de Florence FORBES et Gilles CELEUX.
Spécialité : mathématiques appliquées.

Maison Jean Kuntzmann, Université Joseph Fourier Grenoble I,
le mardi 30 octobre 2007.

La classification (de gènes), pourquoi ?

But : Organiser des entités statistiques en groupes en tenant compte des données.

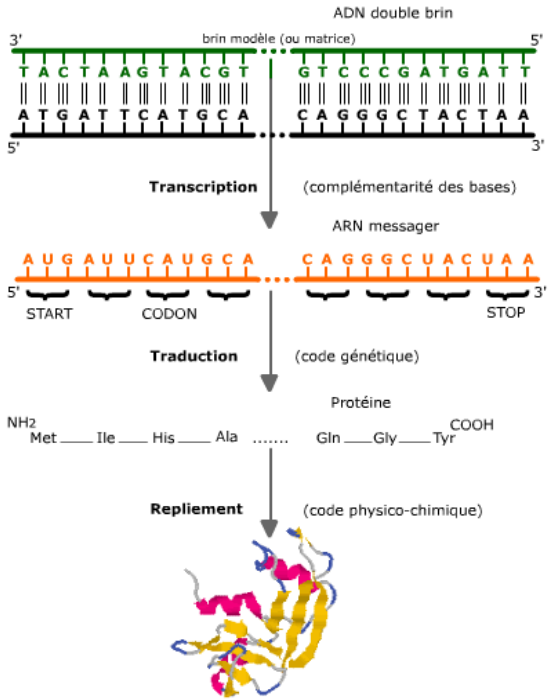
Problématique : Identification de groupes de gènes fonctionnant de façon concertée.

$\rightsquigarrow$ Les données post-génomiques arrivent à point nommé.

Ex : le taux de production de produit de gènes dépend : du type et de l'état de la cellule, des conditions environementales, du stade de développement,...
Intérêt : mesures de l'expression des ARNm dans des tissus atteints de tumeurs et des tissus sains pour un grand nombre d'échantillons à comparer $\longrightarrow$ prédire les gènes et protéines qui pourraient avoir une implication dans le cancer.

L'information génétique et la diversité observée

Simplification : vie = transmission entre générations d'informations qui permettent la construction de machines à survie ou organismes vivants.



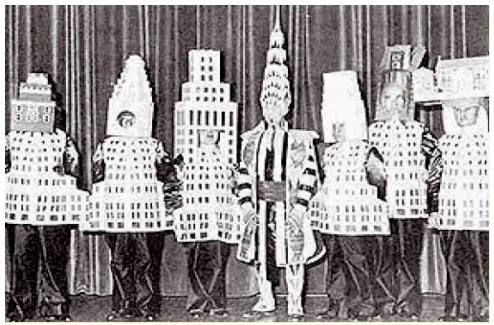
description textuelle (ADN)
design (transcription, ARNp)

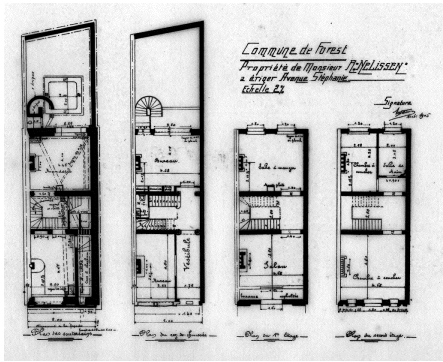
plan (ARNm)
construction (traduction)

produit : "la" cellule

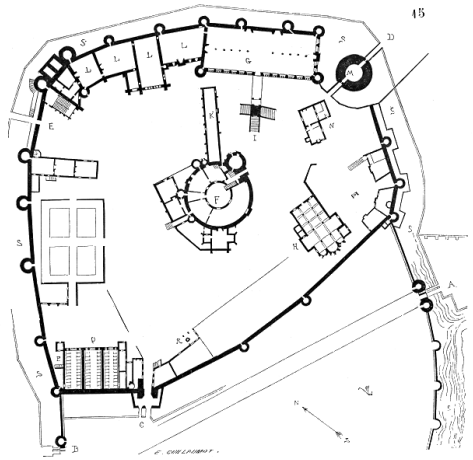
Une première analogie avec le bâtiment

Si on vous décrit : des briques et du bois, des murs avec un toit par dessus, des divisions en pièces et quelques fenêtres, une cave, des combles, *etc.*



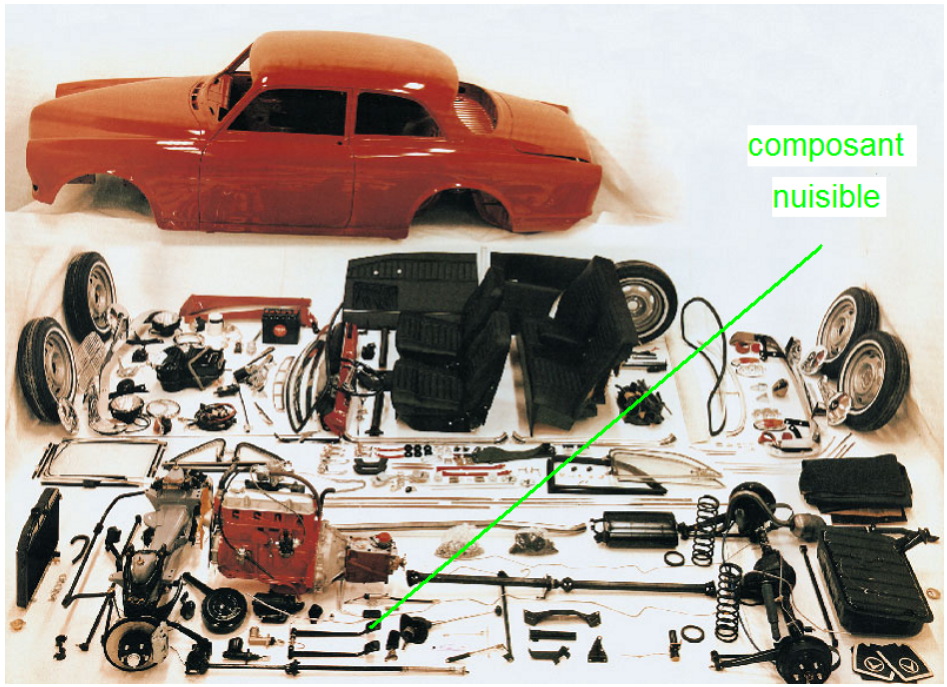


Une construction alternative avec le même cahier des charges



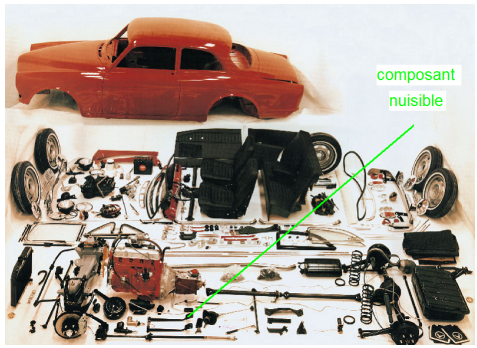
Un biologiste peut-il réparer une radio ? (Lazebnik 2002)

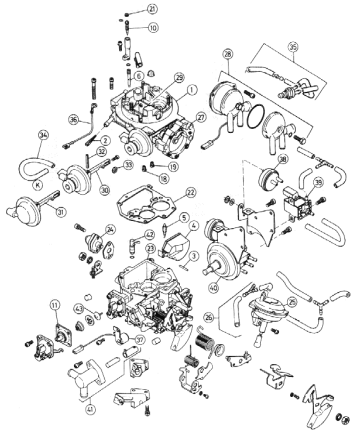
Un biologiste peut-il réparer une radio ? (Lazebnik 2002)

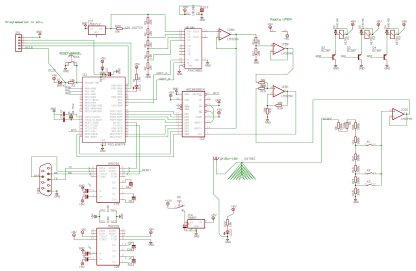


Une application pharmaceutique ?

Un biologiste peut-il réparer une radio ? (Lazebnik 2002)







$\rightsquigarrow$ La nécessité d'intégrer les connaissances de façon rigoureuse.

Dogme simpliste, de nombreuses interactions entre les composants $\longrightarrow$ nécessité de prendre en compte ces dépendances entre les entités biologiques (véhiculant l'information ou agissant).

Notre contexte de travail

- Intégration de données post-génomiques pour une meilleure compréhension des interactions entre les entités d'un système biologique complexe.
- Le type d'approche qui a retenu notre attention : outils probabilistes pour rendre compte de
  1. **mesures individuelles** (*e.g.* expression de gènes) → bruitées et pour lesquelles certaines valeurs peuvent être manquantes,
  2. **interactions** entre les entités biologiques (*e.g.* protein-protein interactions) → traduites dans une structure de réseau.
- Outil proposé : Champs de Markov cachés (*Hidden Markov Random Fields*, HRMF) et algorithmes de type EM pour l'estimation du modèle.

Notre contexte de travail

- Intégration de données post-génomiques pour une meilleure compréhension des interactions entre les entités d'un système biologique complexe.
- Le type d'approche qui a retenu notre attention : outils probabilistes pour rendre compte de
  1. **mesures individuelles** (*e.g.* expression de gènes) → bruitées et pour lesquelles certaines valeurs peuvent être manquantes,
  2. **interactions** entre les entités biologiques (*e.g.* protein-protein interactions) → traduites dans une structure de réseau.
- Outil proposé : Champs de Markov cachés (*Hidden Markov Random Fields*, HRMF) et algorithmes de type EM pour l'estimation du modèle.

Notre contexte de travail

- Intégration de données post-génomiques pour une meilleure compréhension des interactions entre les entités d'un système biologique complexe.
- Le type d'approche qui a retenu notre attention : outils probabilistes pour rendre compte de
  1. **mesures individuelles** (*e.g.* expression de gènes) $\longrightarrow$ bruitées et pour lesquelles certaines valeurs peuvent être manquantes,
  2. **interactions** entre les entités biologiques (*e.g.* protein-protein interactions) $\longrightarrow$ traduites dans une structure de réseau.
- Outil proposé : Champs de Markov cachés (*Hidden Markov Random Fields*, HRMF) et algorithmes de type EM pour l'estimation du modèle.

Notre contexte de travail

- Intégration de données post-génomiques pour une meilleure compréhension des interactions entre les entités d'un système biologique complexe.
- Le type d'approche qui a retenu notre attention : outils probabilistes pour rendre compte de
  1. **mesures individuelles** (*e.g.* expression de gènes) $\longrightarrow$ bruitées et pour lesquelles certaines valeurs peuvent être manquantes,
  2. **interactions** entre les entités biologiques (*e.g.* protein-protein interactions) $\longrightarrow$ traduites dans une structure de réseau.
- Outil proposé : Champs de Markov cachés (*Hidden Markov Random Fields*, HRMF) et algorithmes de type EM pour l'estimation du modèle.

Notre contexte de travail

- Intégration de données post-génomiques pour une meilleure compréhension des interactions entre les entités d'un système biologique complexe.
- Le type d'approche qui a retenu notre attention : outils probabilistes pour rendre compte de
  1. **mesures individuelles** (*e.g.* expression de gènes) $\longrightarrow$ bruitées et pour lesquelles certaines valeurs peuvent être manquantes,
  2. **interactions** entre les entités biologiques (*e.g.* protein-protein interactions) $\longrightarrow$ traduites dans une structure de réseau.
- Outil proposé : Champs de Markov cachés (*Hidden Markov Random Fields*, HRMF) et algorithmes de type EM pour l'estimation du modèle.

Travaux antérieurs

- Classification non supervisée de données d'expression : nombreuses approches envisagées (voir Kim *et al*, 2005 : classification hiérarchique, k-means, SOM, modèle de mélange, bi-clustering, ...).
- Observations manquantes : Little & Rubin 2002 (cadre statistique), Troyanskaya *et al* 2001 et raffinements (imputation des expressions manquantes) et Kim *et al* 2007 (estimation alternée avec classification).

Limitations majeures des approches existantes :

- la structure de réseau n'est pas prise en compte (exception : Yamanishi *et al*, 2003 mais dans un cadre supervisé),
- la nature stochastique des données est ignorée (exception : modèle de mélange, voir Yeung *et al* 2001).
- imputation préalable des valeurs manquantes (exception : Kim *et al*, 2007 mais aucune dépendance entre les gènes),

Travaux antérieurs

- Classification non supervisée de données d'expression : nombreuses approches envisagées (voir Kim *et al*, 2005 : classification hiérarchique, k-means, SOM, modèle de mélange, bi-clustering, ...).
- Observations manquantes : Little & Rubin 2002 (cadre statistique), Troyanskaya *et al* 2001 et raffinements (imputation des expressions manquantes) et Kim *et al* 2007 (estimation alternée avec classification).

Limitations majeures des approches existantes :

- la structure de réseau n'est pas prise en compte (exception : Yamanishi *et al*, 2003 mais dans un cadre supervisé),
- la nature stochastique des données est ignorée (exception : modèle de mélange, voir Yeung *et al* 2001).
- imputation préalable des valeurs manquantes (exception : Kim *et al*, 2007 mais aucune dépendance entre les gènes),

Travaux antérieurs

- Classification non supervisée de données d'expression : nombreuses approches envisagées (voir Kim *et al*, 2005 : classification hiérarchique, k-means, SOM, modèle de mélange, bi-clustering, ...).
- Observations manquantes : Little & Rubin 2002 (cadre statistique), Troyanskaya *et al* 2001 et raffinements (imputation des expressions manquantes) et Kim *et al* 2007 (estimation alternée avec classification).

Limitations majeures des approches existantes :

- la structure de réseau n'est pas prise en compte (exception : Yamanishi *et al*, 2003 mais dans un cadre supervisé),
- la nature stochastique des données est ignorée (exception : modèle de mélange, voir Yeung *et al* 2001).
- imputation préalable des valeurs manquantes (exception : Kim *et al*, 2007 mais aucune dépendance entre les gènes),

Travaux antérieurs

- Classification non supervisée de données d'expression : nombreuses approches envisagées (voir Kim *et al*, 2005 : classification hiérarchique, k-means, SOM, modèle de mélange, bi-clustering, ...).
- Observations manquantes : Little & Rubin 2002 (cadre statistique), Troyanskaya *et al* 2001 et raffinements (imputation des expressions manquantes) et Kim *et al* 2007 (estimation alternée avec classification).

Limitations majeures des approches existantes :

- la structure de réseau n'est pas prise en compte (exception : Yamanishi *et al*, 2003 mais dans un cadre supervisé),
- la nature stochastique des données est ignorée (exception : modèle de mélange, voir Yeung *et al* 2001).
- imputation préalable des valeurs manquantes (exception : Kim *et al*, 2007 mais aucune dépendance entre les gènes),

Travaux antérieurs

- Classification non supervisée de données d'expression : nombreuses approches envisagées (voir Kim *et al*, 2005 : classification hiérarchique, k-means, SOM, modèle de mélange, bi-clustering, ...).
- Observations manquantes : Little & Rubin 2002 (cadre statistique), Troyanskaya *et al* 2001 et raffinements (imputation des expressions manquantes) et Kim *et al* 2007 (estimation alternée avec classification).

Limitations majeures des approches existantes :

- la structure de réseau n'est pas prise en compte (exception : Yamanishi *et al*, 2003 mais dans un cadre supervisé),
- la nature stochastique des données est ignorée (exception : modèle de mélange, voir Yeung *et al* 2001).
- imputation préalable des valeurs manquantes (exception : Kim *et al*, 2007 mais aucune dépendance entre les gènes),

plan

Plan détaillé ; c'est quand qu'on va où ?

Notations utilisées

- x_i : mesure individuelle associée à l'objet (gène) i, $\forall i = 1, ..., N$ $(x_i \in \mathbf{R}^D)$,
- z_i : étiquette (ou classe) de i,
- Loi des données observées :

$$P(x_i|\underbrace{\Psi}_{:=(\Delta,\theta)}) = \sum_{k=1}^{K} P(Z_i = k|\underbrace{\Delta}_{\text{param. loi a priori}})f(x_i|\underbrace{\theta_k}_{\text{param. classe } k}),$$

- *But* : Affecter chaque objet à une des K classes sachant $\mathbf{x}$ et le réseau.
  $\rightsquigarrow$ *Maximum A Posteriori* (MAP)
  $\mathbf{z}^{MAP} = \arg\max P(\mathbf{Z} = \mathbf{z}|\mathbf{x}, \Psi)$.
- Le gène i interagit avec son voisinage $\nu(i)$ *via* un champ de Markov (*Markov Random Field*, MRF) qui est...

Notations utilisées

- x_i : mesure individuelle associée à l'objet (gène) i, $\forall i = 1, ..., N$ $(x_i \in \mathbf{R}^D)$,
- z_i : étiquette (ou classe) de i,
- Loi des données observées :

$$P(x_i|\underbrace{\Psi}_{:=(\Delta,\theta)}) = \sum_{k=1}^{K} P(Z_i = k|\underbrace{\Delta}_{\text{param. loi a priori}}) f(x_i|\underbrace{\theta_k}_{\text{param. classe } k}),$$

- *But* : Affecter chaque objet à une des K classes sachant $\mathbf{x}$ et le réseau.
  $\rightsquigarrow$ *Maximum A Posteriori* (MAP)
  $\mathbf{z}^{MAP} = \arg\max P(\mathbf{Z} = \mathbf{z}|\mathbf{x}, \Psi)$.
- Le gène i interagit avec son voisinage $\nu(i)$ *via* un champ de Markov (*Markov Random Field*, MRF) qui est...

Notations utilisées

- x_i : mesure individuelle associée à l'objet (gène) i, $\forall i = 1, ..., N$ $(x_i \in \mathbf{R}^D)$,
- z_i : étiquette (ou classe) de i,
- Loi des données observées :

$$P(x_i|\underbrace{\Psi}_{:=(\Delta,\theta)}) = \sum_{k=1}^{K} P(Z_i = k|\underbrace{\Delta}_{\text{param. loi a priori}}) f(x_i|\underbrace{\theta_k}_{\text{param. classe } k}),$$

- *But* : Affecter chaque objet à une des K classes sachant $\mathbf{x}$ et le réseau.
  $\rightsquigarrow$ *Maximum A Posteriori* (MAP)
  $\mathbf{z}^{MAP} = \arg\max P(\mathbf{Z} = \mathbf{z}|\mathbf{x}, \Psi)$.
- Le gène i interagit avec son voisinage $\nu(i)$ *via* un champ de Markov (*Markov Random Field*, MRF) qui est...

Notations utilisées

- x_i : mesure individuelle associée à l'objet (gène) i, $\forall i = 1, ..., N$ $(x_i \in \mathbf{R}^D)$,
- z_i : étiquette (ou classe) de i,
- Loi des données observées :

$$P(x_i|\underbrace{\Psi}_{:=(\Delta,\theta)}) = \sum_{k=1}^{K} P(Z_i = k|\underbrace{\Delta}_{\text{param. loi a priori}})f(x_i|\underbrace{\theta_k}_{\text{param. classe } k}),$$

- *But* : Affecter chaque objet à une des K classes sachant $\mathbf{x}$ et le réseau.
  $\rightsquigarrow$ *Maximum A Posteriori* (MAP)
  $\mathbf{z}^{MAP} = \arg\max P(\mathbf{Z} = \mathbf{z}|\mathbf{x}, \Psi)$.
- Le gène i interagit avec son voisinage $\nu(i)$ *via* un champ de Markov (*Markov Random Field*, MRF) qui est...

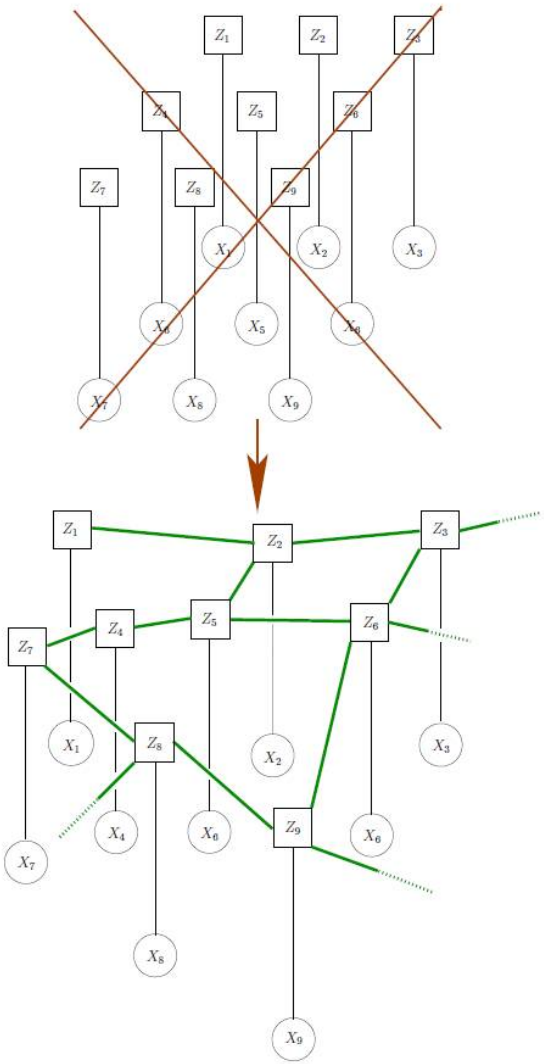
Notations utilisées

- x_i : mesure individuelle associée à l'objet (gène) i, $\forall i = 1, ..., N$ $(x_i \in \mathbf{R}^D)$,
- z_i : étiquette (ou classe) de i,
- Loi des données observées :

$$P(x_i|\underbrace{\Psi}_{:=(\Delta,\theta)}) = \sum_{k=1}^{K} P(Z_i = k|\underbrace{\Delta}_{\text{param. loi a priori}}) f(x_i|\underbrace{\theta_k}_{\text{param. classe } k}),$$

- *But* : Affecter chaque objet à une des K classes sachant $\mathbf{x}$ et le réseau.
  $\rightsquigarrow$ *Maximum A Posteriori* (MAP)
  $\mathbf{z}^{MAP} = \arg\max P(\mathbf{Z} = \mathbf{z}|\mathbf{x}, \Psi)$.
- Le gène i interagit avec son voisinage $\nu(i)$ *via* un champ de Markov (*Markov Random Field*, MRF) qui est...

Fonction d'énergie H d'un champ de Markov



Définition : $\mathbf{Z}$ est un champ de Markov ssi $P(z_i|\mathbf{z} \setminus \{z_i\}) = P(z_i|\mathbf{z}_{\nu(i)})$ et $P > 0$.
Et (indépendance conditionnelle)
$P(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^{N} P(x_i|z_i) \rightarrow (\mathbf{X}, \mathbf{Z})$ est un champ de Markov caché.

Fonction d'énergie et restriction du modèle

Théorème (Hammersley-Clifford)

P est un champ de Markov $\Longleftrightarrow$
P s'écrit comme une distribution gibbsienne :
$P_G(\mathbf{z}|\Delta) = W(\Delta)^{-1} \exp\left(-H(\mathbf{z}, \Delta)\right)$ *avec*
$H(\mathbf{z}, \Delta) = \sum_{c \in C} V_c(\mathbf{z}_c)$.

On se limite aux interactions de paires :

$$H(\mathbf{z}, \Delta) = \sum_{i \sim j} V_{ij}(z_i, z_j)$$

Fonction d'énergie et restriction du modèle

Théorème (Hammersley-Clifford)

P est un champ de Markov $\Longleftrightarrow$
P s'écrit comme une distribution gibbsienne :
$P_G(\mathbf{z}|\Delta) = W(\Delta)^{-1} \exp\left(-H(\mathbf{z}, \Delta)\right)$ *avec*
$H(\mathbf{z}, \Delta) = \sum_{c \in C} V_c(\mathbf{z}_c)$.

On se limite aux interactions de paires :

$$H(\mathbf{z}, \Delta) = \sum_{i \sim j} V_{ij}(z_i, z_j)$$

Fonction d'énergie et restriction du modèle

Théorème (Hammersley-Clifford)

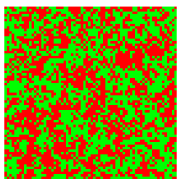
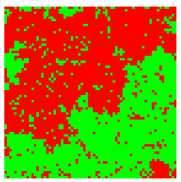
P est un champ de Markov $\Longleftrightarrow$
P s'écrit comme une distribution gibbsienne :
$P_G(\mathbf{z}|\Delta) = W(\Delta)^{-1} \exp\left(-H(\mathbf{z}, \Delta)\right)$ *avec*
$H(\mathbf{z}, \Delta) = \sum_{c \in C} V_c(\mathbf{z}_c)$.

Dans le cas d'un champ homogène et isotrope :

$$H(\mathbf{z}) = \sum_{i \sim j} V(z_i, z_j)$$

Cas particulier : le modèle de Potts

$$H(\mathbf{z}) = -\beta \sum_{i \sim j} \mathbb{1}_{z_i = z_j}$$

Fonction d'énergie et restriction du modèle

Théorème (Hammersley-Clifford)

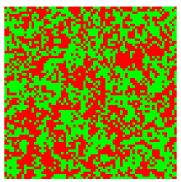
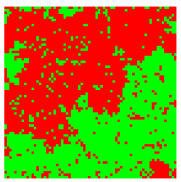
P est un champ de Markov $\Longleftrightarrow$
P s'écrit comme une distribution gibbsienne :
$P_G(\mathbf{z}|\Delta) = W(\Delta)^{-1} \exp\left(-H(\mathbf{z}, \Delta)\right)$ *avec*
$H(\mathbf{z}, \Delta) = \sum_{c \in C} V_c(\mathbf{z}_c)$.

Dans le cas d'un champ homogène et isotrope :

$$H(\mathbf{z}) = \sum_{i \sim j} V(z_i, z_j)$$

Cas particulier : le modèle de Potts

$$H(\mathbf{z}) = -\beta \sum_{i \sim j} \mathbb{1}_{z_i = z_j}$$

Bons résultats de l'approximation de type champ moyen

Ψ doit être estimé $\rightsquigarrow$ algorithme EM qui permet en outre de remplir l'objectif de classification.

Difficultés : calcul de $W(\Delta)$ et des probabilités *a posteriori* $\rightsquigarrow$ approximations de type champ moyen (Celeux *et al* 2003) qui se ramènent à un système de variables indépendantes :
$P_G(\mathbf{z}) \approx \prod_i P(z_i|\tilde{z}_{\nu(i)})$.
Approche en champ simulé (*simulated-field*, SF) la plus performante.

$\rightsquigarrow$ article dans *IEEE Transactions on Computational Biology and Bioinformatics* contenant des applications à des données simulées et à des données de programme de sporulation associées à un graphe métabolique chez la levure.

Bons résultats de l'approximation de type champ moyen

Ψ doit être estimé $\rightsquigarrow$ algorithme EM qui permet en outre de remplir l'objectif de classification.

Difficultés : calcul de $W(\Delta)$ et des probabilités *a posteriori* $\rightsquigarrow$ approximations de type champ moyen (Celeux *et al* 2003) qui se ramènent à un système de variables indépendantes :
$P_G(\mathbf{z}) \approx \prod_i P(z_i|\tilde{z}_{\nu(i)})$.
Approche en champ simulé (*simulated-field*, SF) la plus performante.

$\rightsquigarrow$ article dans *IEEE Transactions on Computational Biology and Bioinformatics* contenant des applications à des données simulées et à des données de programme de sporulation associées à un graphe métabolique chez la levure.

Bons résultats de l'approximation de type champ moyen

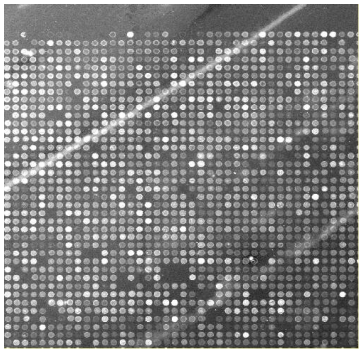
Ψ doit être estimé $\rightsquigarrow$ algorithme EM qui permet en outre de remplir l'objectif de classification.

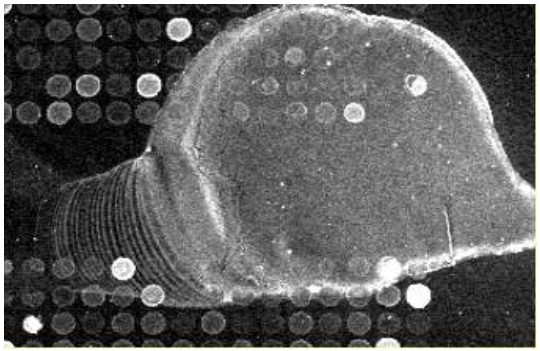
Difficultés : calcul de $W(\Delta)$ et des probabilités *a posteriori* $\rightsquigarrow$ approximations de type champ moyen (Celeux *et al* 2003) qui se ramènent à un système de variables indépendantes :
$P_G(\mathbf{z}) \approx \prod_i P(z_i|\tilde{z}_{\nu(i)})$.
Approche en champ simulé (*simulated-field*, SF) la plus performante.

$\rightsquigarrow$ article dans *IEEE Transactions on Computational Biology and Bioinformatics* contenant des applications à des données simulées et à des données de programme de sporulation associées à un graphe métabolique chez la levure.

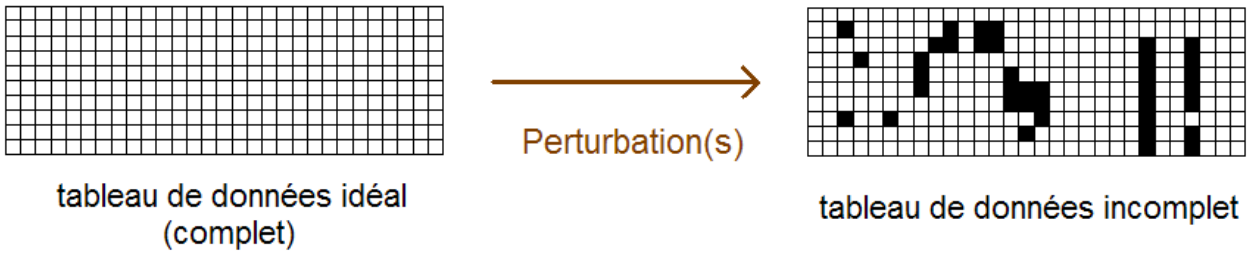
Plan détaillé ; c'est quand qu'on va où ?

Pourquoi des données viennent à manquer ?





Formalisation : R $(N \times D)$ mécanisme d'absence de paramètre(s) ρ ; $R_{ij} = 0$ ou 1 selon que x_{ij} est manquant ou observé.



Notations utilisées

- $o_i \subset \{1, \ldots, D\}$: indices des valeurs observées et m_i des valeurs manquantes pour le gène i ($o_i \cup m_i = \{1, \ldots, D\}$).
- Notations résultantes : $x_i^{o_i} = \{x_{id}, d \in o_i\}$ et $\mathbf{x}^o = \{x_i^{o_i}, i \in \{1, \ldots, N\}\}$. De même pour $x_i^{m_i}$ et $\mathbf{x}^m$.

Hypothèse : le mécanisme d'absence est Missing At Random (MAR) *i.e.* ne dépend que des valeurs observées, pas des manquantes (pas de censure).
$\longrightarrow$ Séparation de l'estimation des paramètres du mécanisme d'absence des autres :

$$\begin{aligned} P(\mathbf{X}, \mathbf{Z}, R|\Psi, \rho) &= P(R|\mathbf{X}, \mathbf{Z}, \rho) P(\mathbf{X}, \mathbf{Z}|\Psi) \\ &= P(R|\mathbf{x}^o, \rho) P(\mathbf{X}, \mathbf{Z}|\Psi) \end{aligned}$$

Notations utilisées

- $o_i \subset \{1, \ldots, D\}$: indices des valeurs observées et m_i des valeurs manquantes pour le gène i ($o_i \cup m_i = \{1, \ldots, D\}$).
- Notations résultantes : $x_i^{o_i} = \{x_{id}, d \in o_i\}$ et $\mathbf{x}^o = \{x_i^{o_i}, i \in \{1, \ldots, N\}\}$. De même pour $x_i^{m_i}$ et $\mathbf{x}^m$.

Hypothèse : le mécanisme d'absence est Missing At Random (MAR) *i.e.* ne dépend que des valeurs observées, pas des manquantes (pas de censure).

$\longrightarrow$ Séparation de l'estimation des paramètres du mécanisme d'absence des autres :

$$\begin{aligned} P(\mathbf{X}, \mathbf{Z}, R|\Psi, \rho) &= P(R|\mathbf{X}, \mathbf{Z}, \rho) P(\mathbf{X}, \mathbf{Z}|\Psi) \\ &= P(R|\mathbf{x}^o, \rho) P(\mathbf{X}, \mathbf{Z}|\Psi) \end{aligned}$$

Notations utilisées

- $o_i \subset \{1, \ldots, D\}$: indices des valeurs observées et m_i des valeurs manquantes pour le gène i ($o_i \cup m_i = \{1, \ldots, D\}$).
- Notations résultantes : $x_i^{o_i} = \{x_{id}, d \in o_i\}$ et $\mathbf{x}^o = \{x_i^{o_i}, i \in \{1, \ldots, N\}\}$. De même pour $x_i^{m_i}$ et $\mathbf{x}^m$.

Hypothèse : le mécanisme d'absence est Missing At Random (MAR) *i.e.* ne dépend que des valeurs observées, pas des manquantes (pas de censure).
$\longrightarrow$ Séparation de l'estimation des paramètres du mécanisme d'absence des autres :

$$\begin{aligned} P(\mathbf{X}, \mathbf{Z}, R|\Psi, \rho) &= P(R|\mathbf{X}, \mathbf{Z}, \rho) P(\mathbf{X}, \mathbf{Z}|\Psi) \\ &= P(R|\mathbf{x}^o, \rho) P(\mathbf{X}, \mathbf{Z}|\Psi) \end{aligned}$$

Paramètres du modèle de champ de Markov caché

Paramètres à estimer :

- θ_k, paramètre(s) de la distribution $f(x_i|Z_i = k)$ (*e.g.* vecteur moyen μ_k et matrice de covariance Σ_k pour f gaussienne multivariée), $k = 1, \ldots, K$ et
- Δ, paramètres (spatiaux) de P_G.

$\rightarrow \Psi = \{\theta_1, \ldots, \theta_K, \Delta\}$ est l'ensemble des paramètres du modèle...

...Estimés selon $\tilde{\Psi} = \arg\max P(\mathbf{x}^o|\Psi)$,
et dans le cadre de l'algorithme EM : $\tilde{\Psi} = \arg\max P(\mathbf{x}^o, \mathbf{x}^m, \mathbf{z}|\Psi)$.

Paramètres du modèle de champ de Markov caché

Paramètres à estimer :

- θ_k, paramètre(s) de la distribution $f(x_i|Z_i = k)$ (*e.g.* vecteur moyen μ_k et matrice de covariance Σ_k pour f gaussienne multivariée), $k = 1, \ldots, K$ et
- Δ, paramètres (spatiaux) de P_G.

$\rightarrow \Psi = \{\theta_1, \ldots, \theta_K, \Delta\}$ est l'ensemble des paramètres du modèle...

...Estimés selon $\tilde{\Psi} = \arg\max P(\mathbf{x}^o|\Psi)$,
et dans le cadre de l'algorithme EM : $\tilde{\Psi} = \arg\max P(\mathbf{x}^o, \mathbf{x}^m, \mathbf{z}|\Psi)$.

Paramètres du modèle de champ de Markov caché

Paramètres à estimer :

- θ_k, paramètre(s) de la distribution $f(x_i|Z_i = k)$ (*e.g.* vecteur moyen μ_k et matrice de covariance Σ_k pour f gaussienne multivariée), $k = 1, \ldots, K$ et
- Δ, paramètres (spatiaux) de P_G.

$\rightarrow \Psi = \{\theta_1, \ldots, \theta_K, \Delta\}$ est l'ensemble des paramètres du modèle...

...Estimés selon $\tilde{\Psi} = \arg\max P(\mathbf{x}^o|\Psi)$,
et dans le cadre de l'algorithme EM : $\tilde{\Psi} = \arg\max P(\mathbf{x}^o, \mathbf{x}^m, \mathbf{z}|\Psi)$.

Mise en place de l'algorithme EM

- L'algorithme EM (*Expectation-Maximisation*) maximise l'espérance de la (log-)vraisemblance complète du modèle conditionnellement aux données observées et à l'état courant des paramètres :

$$\begin{aligned} Q(\Psi|\Psi^{(q)}) &:= \mathbb{E}\left[\log P(\mathbf{x}^{Obs}, \mathbf{X}^{Man}, \mathbf{Z}|\Psi)|\mathbf{x}^{Obs}, \Psi^{(q)}\right] \\ &= \underbrace{\mathbb{E}\left[\log P(\mathbf{x}^{Obs}, \mathbf{X}^{Man}|\mathbf{Z}, \theta)|\mathbf{x}^{Obs}, \Psi^{(q)}\right]}_{:=Q_\theta(\theta|\Psi^{(q)})} \\ &\quad + \underbrace{\mathbb{E}\left[\log P(\mathbf{Z}|\Delta)|\mathbf{x}^{Obs}, \Psi^{(q)}\right]}_{:=Q_\Delta(\Delta|\Psi^{(q)})} \end{aligned}$$

- Difficulté (supplémentaire par rapport au cas complet) : $\mathbf{X}^{Man}$ est une nouvelle variable latente.

Mise en place de l'algorithme EM

- L'algorithme EM (*Expectation-Maximisation*) maximise l'espérance de la (log-)vraisemblance complète du modèle conditionnellement aux données observées et à l'état courant des paramètres :

$$\begin{aligned} Q(\Psi|\Psi^{(q)}) &:= \mathbb{E}\left[\log P(\mathbf{x}^{Obs}, \mathbf{X}^{Man}, \mathbf{Z}|\Psi)|\mathbf{x}^{Obs}, \Psi^{(q)}\right] \\ &= \underbrace{\mathbb{E}\left[\log P(\mathbf{x}^{Obs}, \mathbf{X}^{Man}|\mathbf{Z}, \theta)|\mathbf{x}^{Obs}, \Psi^{(q)}\right]}_{:=Q_\theta(\theta|\Psi^{(q)})} \\ &\quad + \underbrace{\mathbb{E}\left[\log P(\mathbf{Z}|\Delta)|\mathbf{x}^{Obs}, \Psi^{(q)}\right]}_{:=Q_\Delta(\Delta|\Psi^{(q)})} \end{aligned}$$

- Difficulté (supplémentaire par rapport au cas complet) : $\mathbf{X}^{Man}$ est une nouvelle variable latente.

Mise en place de l'algorithme EM

- L'algorithme EM (*Expectation-Maximisation*) maximise l'espérance de la (log-)vraisemblance complète du modèle conditionnellement aux données observées et à l'état courant des paramètres :

$$\begin{aligned} Q(\Psi|\Psi^{(q)}) &:= \mathbb{E}\left[\log P(\mathbf{x}^{Obs}, \mathbf{X}^{Man}, \mathbf{Z}|\Psi)|\mathbf{x}^{Obs}, \Psi^{(q)}\right] \\ &= \underbrace{\mathbb{E}\left[\log P(\mathbf{x}^{Obs}, \mathbf{X}^{Man}|\mathbf{Z}, \theta)|\mathbf{x}^{Obs}, \Psi^{(q)}\right]}_{:=Q_\theta(\theta|\Psi^{(q)})} \\ &\quad + \underbrace{\mathbb{E}\left[\log P(\mathbf{Z}|\Delta)|\mathbf{x}^{Obs}, \Psi^{(q)}\right]}_{:=Q_\Delta(\Delta|\Psi^{(q)})} \end{aligned}$$

- Difficulté (supplémentaire par rapport au cas complet) : $\mathbf{X}^{Man}$ est une nouvelle variable latente.

Algorithme de type champ moyen

Init. stratégie : plusieurs tirages.

étape NR Créer (simuler selon $P(\mathbf{Z}|\mathbf{x}^o, \Psi^{(q)})$ pour SF $\longrightarrow$ algorithme SFmiss) $\tilde{z}_i^{(q+1)}$ à partir de $x_i^{o_i}$ et $\Psi^{(q)}$ en remplaçant $P_G(\mathbf{z})$ par $\prod_{i=1}^{N} P_G(z_i|\tilde{\mathbf{z}}_{\nu(i)}^{(q+1)})$.

étape EM Appliquer une itération de EM pour mettre à jour $\Psi^{(q)}$ en $\Psi^{(q+1)}$ *i.e.* :

- calcul de $\tilde{t}_{ik}^{(q+1)} = P_G(Z_i = k|x_i^{o_i}, \Psi^{(q)}, \tilde{\mathbf{z}}_{N_i}^{(q+1)})$ (étape E) et
- étape M : mise à jour de $\Psi = (\theta, \Delta)$ qui maximise séparément $Q_\theta(\theta|\Psi^{(q)})$ et $Q_\Delta(\Delta|\Psi^{(q)})$ (identique au cas des données complètes pour les paramètres spatiaux).

$\longrightarrow$ La structure markovienne des données est prise en compte tout en se ramenant au cas d'une distribution factorisée traitable par EM.

Algorithme de type champ moyen

Init. stratégie : plusieurs tirages.

étape NR Créer (simuler selon $P(\mathbf{Z}|\mathbf{x}^o, \Psi^{(q)})$ pour SF $\longrightarrow$ algorithme SFmiss) $\tilde{z}_i^{(q+1)}$ à partir de $x_i^{o_i}$ et $\Psi^{(q)}$ en remplaçant $P_G(\mathbf{z})$ par $\prod_{i=1}^{N} P_G(z_i|\tilde{\mathbf{z}}_{\nu(i)}^{(q+1)})$.

étape EM Appliquer une itération de EM pour mettre à jour $\Psi^{(q)}$ en $\Psi^{(q+1)}$ *i.e.* :

- calcul de $\tilde{t}_{ik}^{(q+1)} = P_G(Z_i = k|x_i^{o_i}, \Psi^{(q)}, \tilde{\mathbf{z}}_{N_i}^{(q+1)})$ (étape E) et
- étape M : mise à jour de $\Psi = (\theta, \Delta)$ qui maximise séparément $Q_\theta(\theta|\Psi^{(q)})$ et $Q_\Delta(\Delta|\Psi^{(q)})$ (identique au cas des données complètes pour les paramètres spatiaux).

$\longrightarrow$ La structure markovienne des données est prise en compte tout en se ramenant au cas d'une distribution factorisée traitable par EM.

Algorithme de type champ moyen

Init. stratégie : plusieurs tirages.

étape NR Créer (simuler selon $P(\mathbf{Z}|\mathbf{x}^o, \Psi^{(q)})$ pour SF $\longrightarrow$ algorithme SFmiss) $\tilde{z}_i^{(q+1)}$ à partir de $x_i^{o_i}$ et $\Psi^{(q)}$ en remplaçant $P_G(\mathbf{z})$ par $\prod_{i=1}^{N} P_G(z_i|\tilde{\mathbf{z}}_{\nu(i)}^{(q+1)})$.

étape EM Appliquer une itération de EM pour mettre à jour $\Psi^{(q)}$ en $\Psi^{(q+1)}$ *i.e.* :

- calcul de $\tilde{t}_{ik}^{(q+1)} = P_G(Z_i = k|x_i^{o_i}, \Psi^{(q)}, \tilde{\mathbf{z}}_{N_i}^{(q+1)})$ (étape E) et
- étape M : mise à jour de $\Psi = (\theta, \Delta)$ qui maximise séparément $Q_\theta(\theta|\Psi^{(q)})$ et $Q_\Delta(\Delta|\Psi^{(q)})$ (identique au cas des données complètes pour les paramètres spatiaux).

$\longrightarrow$ La structure markovienne des données est prise en compte tout en se ramenant au cas d'une distribution factorisée traitable par EM.

Algorithme de type champ moyen

Init. stratégie : plusieurs tirages.

étape NR Créer (simuler selon $P(\mathbf{Z}|\mathbf{x}^o, \Psi^{(q)})$ pour SF $\longrightarrow$ algorithme SFmiss) $\tilde{z}_i^{(q+1)}$ à partir de $x_i^{o_i}$ et $\Psi^{(q)}$ en remplaçant $P_G(\mathbf{z})$ par $\prod_{i=1}^N P_G(z_i|\tilde{\mathbf{z}}_{\nu(i)}^{(q+1)})$.

étape EM Appliquer une itération de EM pour mettre à jour $\Psi^{(q)}$ en $\Psi^{(q+1)}$ *i.e.* :

- calcul de $\tilde{t}_{ik}^{(q+1)} = P_G(Z_i = k|x_i^{o_i}, \Psi^{(q)}, \tilde{\mathbf{z}}_{N_i}^{(q+1)})$ (étape E) et
- étape M : mise à jour de $\Psi = (\theta, \Delta)$ qui maximise séparément $Q_\theta(\theta|\Psi^{(q)})$ et $Q_\Delta(\Delta|\Psi^{(q)})$ (identique au cas des données complètes pour les paramètres spatiaux).

$\longrightarrow$ La structure markovienne des données est prise en compte tout en se ramenant au cas d'une distribution factorisée traitable par EM.

Algorithme de type champ moyen

Init. stratégie : plusieurs tirages.

étape NR Créer (simuler selon $P(\mathbf{Z}|\mathbf{x}^o, \Psi^{(q)})$ pour SF $\longrightarrow$ algorithme SFmiss) $\tilde{z}_i^{(q+1)}$ à partir de $x_i^{o_i}$ et $\Psi^{(q)}$ en remplaçant $P_G(\mathbf{z})$ par $\prod_{i=1}^{N} P_G(z_i|\tilde{\mathbf{z}}_{\nu(i)}^{(q+1)})$.

étape EM Appliquer une itération de EM pour mettre à jour $\Psi^{(q)}$ en $\Psi^{(q+1)}$ *i.e.* :

- calcul de $\tilde{t}_{ik}^{(q+1)} = P_G(Z_i = k|x_i^{o_i}, \Psi^{(q)}, \tilde{\mathbf{z}}_{N_i}^{(q+1)})$ (étape E) et
- étape M : mise à jour de $\Psi = (\theta, \Delta)$ qui maximise séparément $Q_\theta(\theta|\Psi^{(q)})$ et $Q_\Delta(\Delta|\Psi^{(q)})$ (identique au cas des données complètes pour les paramètres spatiaux).

$\longrightarrow$ La structure markovienne des données est prise en compte tout en se ramenant au cas d'une distribution factorisée traitable par EM.

Calculs explicites

Étape E : $\tilde{t}_{ik}^{(q+1)} := P(Z_i = k | \tilde{\mathbf{z}}_{\nu(i)}, x_i^{o_i}, \Psi^{(q)}) = \frac{\tilde{\pi}_{ik}^{(q)} f(x_i^{o_i} | \theta_k^{(q)})}{\sum_l \tilde{\pi}_{il}^{(q)} f(x_i^{o_i} | \theta_l^{(q)})}$

Étape M : (Δ *idem* complet)
$\theta_k^{(q+1)} = \arg\max \sum_i \tilde{t}_{ik}^{(q+1)} \mathbb{E}[\log f(x_i^{o_i}, X_i^{m_i} | \theta_k) | x_i^{o_i}, \theta_k^{(q)}]$
$\longrightarrow$ moyenne et matrice de covariance pour une distribution normale.

Calculs explicites

Étape E : $\tilde{t}_{ik}^{(q+1)} := P(Z_i = k | \tilde{\mathbf{z}}_{\nu(i)}, x_i^{o_i}, \Psi^{(q)}) = \frac{\tilde{\pi}_{ik}^{(q)} f(x_i^{o_i} | \theta_k^{(q)})}{\sum_l \tilde{\pi}_{il}^{(q)} f(x_i^{o_i} | \theta_l^{(q)})}$

Étape M : (Δ *idem* complet)
$\theta_k^{(q+1)} = \arg\max \sum_i \tilde{t}_{ik}^{(q+1)} \mathbb{E}[\log f(x_i^{o_i}, X_i^{m_i} | \theta_k) | x_i^{o_i}, \theta_k^{(q)}]$
$\longrightarrow$ moyenne et matrice de covariance pour une distribution normale.

Plan détaillé ; c'est quand qu'on va où ?

Aperçu des données simulées

♣♢ 4 jeux de données synthétiques ♡♠

- 2 jeux de données MCAR, (i) $D = 1$ (image "triangle-carrés-disque" 128×128, $K = 4$) et (ii) $D = 10$ (simulation d'un modèle de Potts, $\beta = 0.4$ sur une grille 100×100 avec $K = 2$ groupes), bruit gaussien diagonal.

- 2 jeux de données NMAR (censure) (iii) $D = 1$ et (iv) $D = 4$ sur une image (128×128) composée de carrés ($K = 4$), bruit gaussien non diagonal.

Aperçu des données simulées

♣♢ 4 jeux de données synthétiques ♡♠

- 2 jeux de données MCAR, (i) $D = 1$ (image "triangle-carrés-disque" 128×128, $K = 4$) et (ii) $D = 10$ (simulation d'un modèle de Potts, $\beta = 0.4$ sur une grille 100×100 avec $K = 2$ groupes), bruit gaussien diagonal.
- 2 jeux de données NMAR (censure) (iii) $D = 1$ et (iv) $D = 4$ sur une image (128×128) composée de carrés ($K = 4$), bruit gaussien non diagonal.

Aperçu des données simulées

♣♢ 4 jeux de données synthétiques ♡♠

- 2 jeux de données MCAR, (i) $D = 1$ (image "triangle-carrés-disque" 128×128, $K = 4$) et (ii) $D = 10$ (simulation d'un modèle de Potts, $\beta = 0.4$ sur une grille 100×100 avec $K = 2$ groupes), bruit gaussien diagonal.

- 2 jeux de données NMAR (censure) (iii) $D = 1$ et (iv) $D = 4$ sur une image (128×128) composée de carrés ($K = 4$), bruit gaussien non diagonal.

(i) Performance de SFmiss dans le cas MCAR - $D = 1$

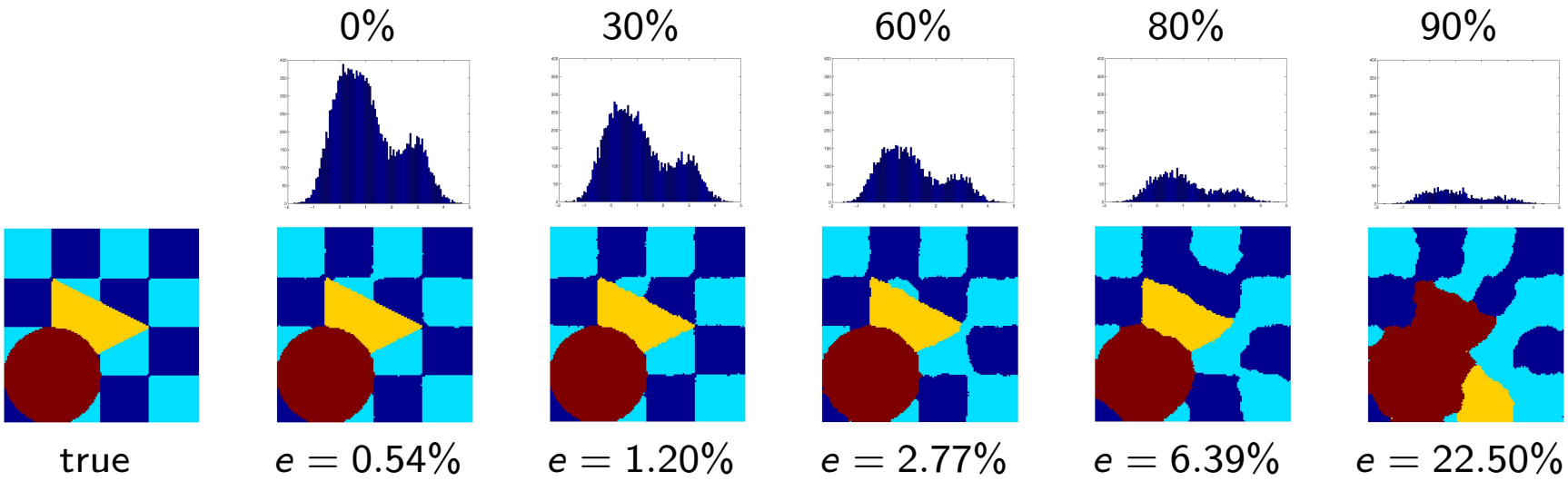


FIG.: Données synthétiques $1D$: histogramme des données (ligne 1) et taux d'erreur de classification e

(iii) Comparaisons des performances de plusieurs algorithmes dans le cas MCAR - $D = 10$

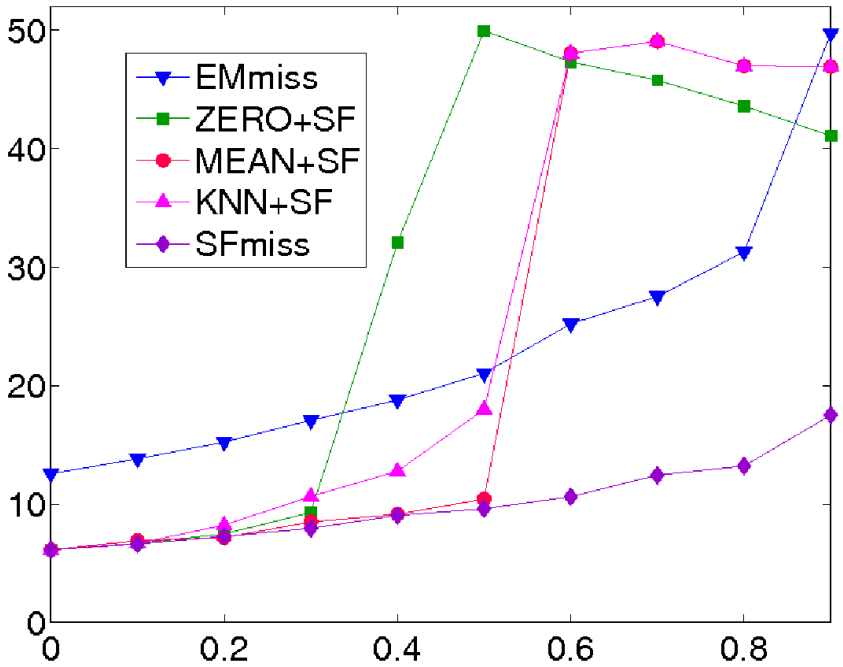


FIG.: Pourcentage de sites mal classés pour une simulation d'un modèle de Potts sur une image 100×100 avec $K = 2$ groupes et un mécanisme d'absence MCAR.

(iii) Performance de SFmiss dans le cas NMAR - $D = 1$

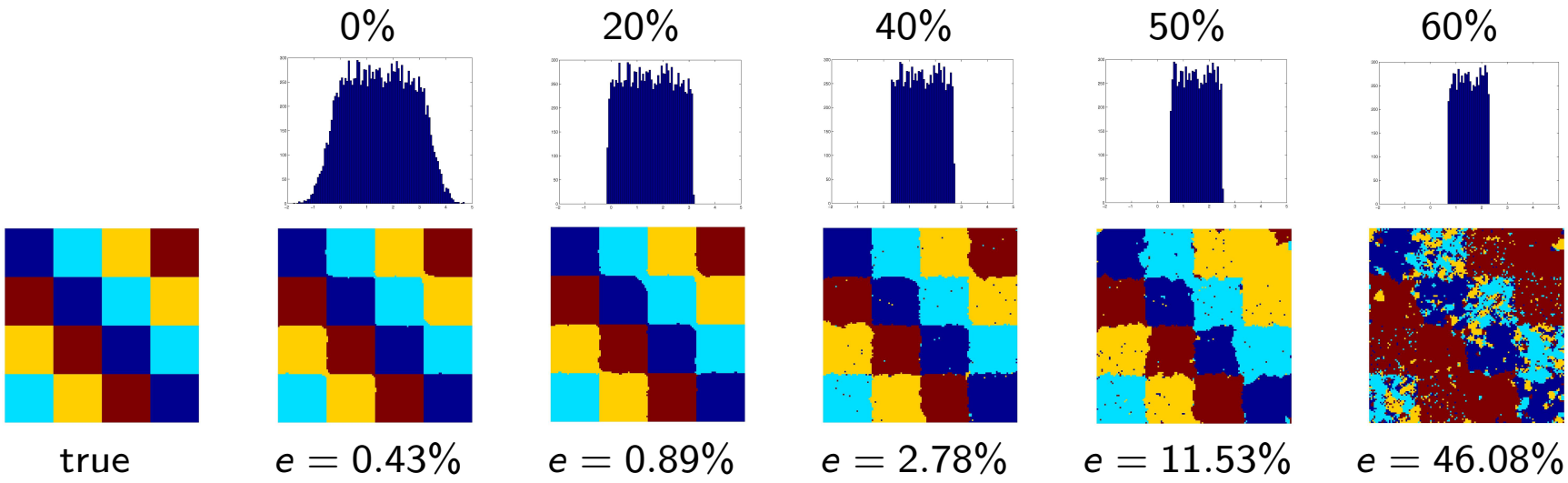


FIG.: Données synthétiques $1D$: histogramme des données (ligne 1) et taux d'erreur de classification e (ligne 2)

(iv) Comparaisons des performances de différents algorithmes dans le cas NMAR - $D = 4$

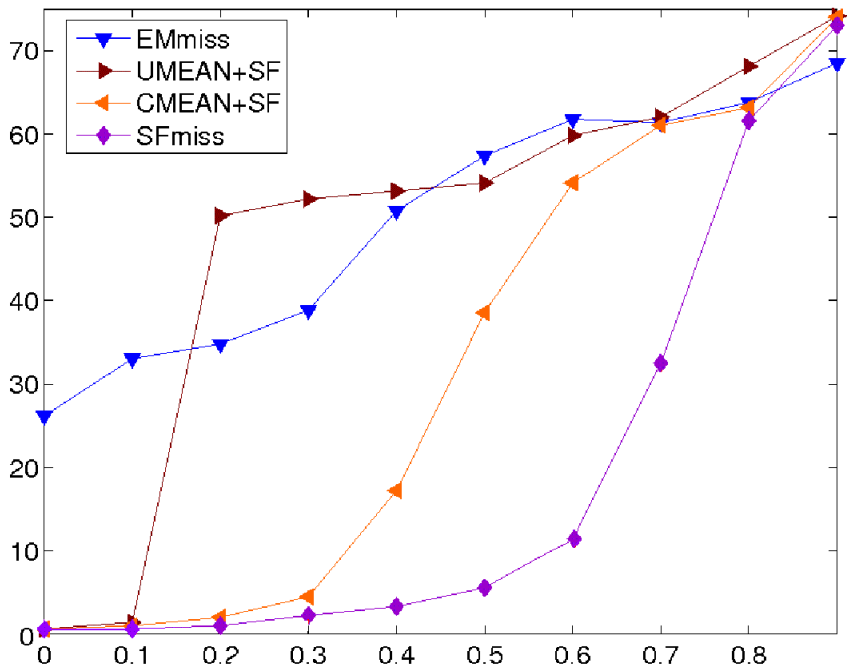


FIG.: Pourcentage de sites mal classés pour une image 128×128 à $K = 4$ groupes dans un cas de censure à gauche et à droite (NMAR).

(iv) Performance de SFmiss dans le cas NMAR (suite)

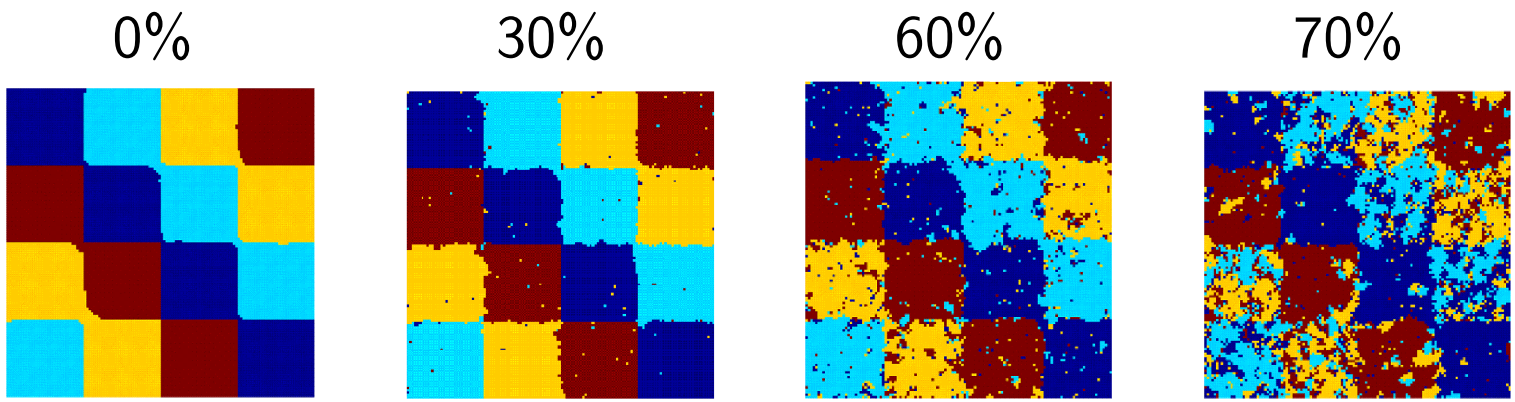
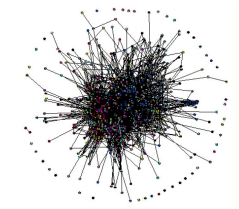


FIG.: Données synthétiques $D = 4$: taux d'erreur de classification e

Données d'expression de cycle cellulaire combinées avec des interactions protéine-protéine

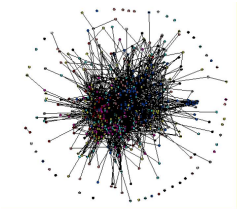
- Expérience *cdc*28 de Cho *et al*, 1998 ($D = 17$) en ne considérant que les 800 gènes identifiés comme mis en cause dans le cycle cellulaire par Spellman *et al*, 1998.
- Données d'interactions (vérifiées et prédites) issues de STRING (`http://string.embl.de`)



- Le jeu de données : 612 gènes reliés par 3530 interactions. "Seulement" 5% d'observations manquantes mais 80% des gènes affectés.
- Pas de consensus de validation pour une telle classification non supervisée (Handl *el al* 2005) $\rightarrow$ Étude de quelques propriétés. $K = 9$ est choisi par BIC.

Données d'expression de cycle cellulaire combinées avec des interactions protéine-protéine

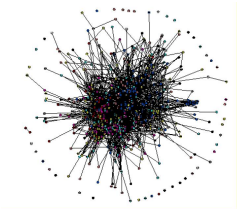
- Expérience *cdc*28 de Cho *et al*, 1998 ($D = 17$) en ne considérant que les 800 gènes identifiés comme mis en cause dans le cycle cellulaire par Spellman *et al*, 1998.
- Données d'interactions (vérifiées et prédites) issues de STRING (`http://string.embl.de`)



- Le jeu de données : 612 gènes reliés par 3530 interactions. "Seulement" 5% d'observations manquantes mais 80% des gènes affectés.
- Pas de consensus de validation pour une telle classification non supervisée (Handl *el al* 2005) $\rightarrow$ Étude de quelques propriétés. $K = 9$ est choisi par BIC.

Données d'expression de cycle cellulaire combinées avec des interactions protéine-protéine

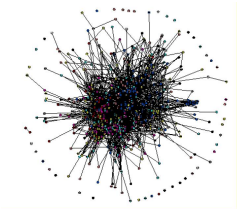
- Expérience *cdc*28 de Cho *et al*, 1998 ($D = 17$) en ne considérant que les 800 gènes identifiés comme mis en cause dans le cycle cellulaire par Spellman *et al*, 1998.
- Données d'interactions (vérifiées et prédites) issues de STRING (`http://string.embl.de`)



- Le jeu de données : 612 gènes reliés par 3530 interactions. "Seulement" 5% d'observations manquantes mais 80% des gènes affectés.
- Pas de consensus de validation pour une telle classification non supervisée (Handl *el al* 2005) $\rightarrow$ Étude de quelques propriétés. $K = 9$ est choisi par BIC.

Données d'expression de cycle cellulaire combinées avec des interactions protéine-protéine

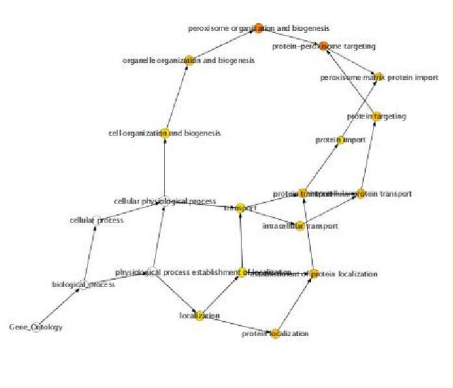
- Expérience *cdc*28 de Cho *et al*, 1998 ($D = 17$) en ne considérant que les 800 gènes identifiés comme mis en cause dans le cycle cellulaire par Spellman *et al*, 1998.
- Données d'interactions (vérifiées et prédites) issues de STRING (`http://string.embl.de`)



- Le jeu de données : 612 gènes reliés par 3530 interactions. "Seulement" 5% d'observations manquantes mais 80% des gènes affectés.
- Pas de consensus de validation pour une telle classification non supervisée (Handl *el al* 2005) $\rightarrow$ Étude de quelques propriétés. $K = 9$ est choisi par BIC.

Étude de l'enrichissement en termes GO de quelques groupes

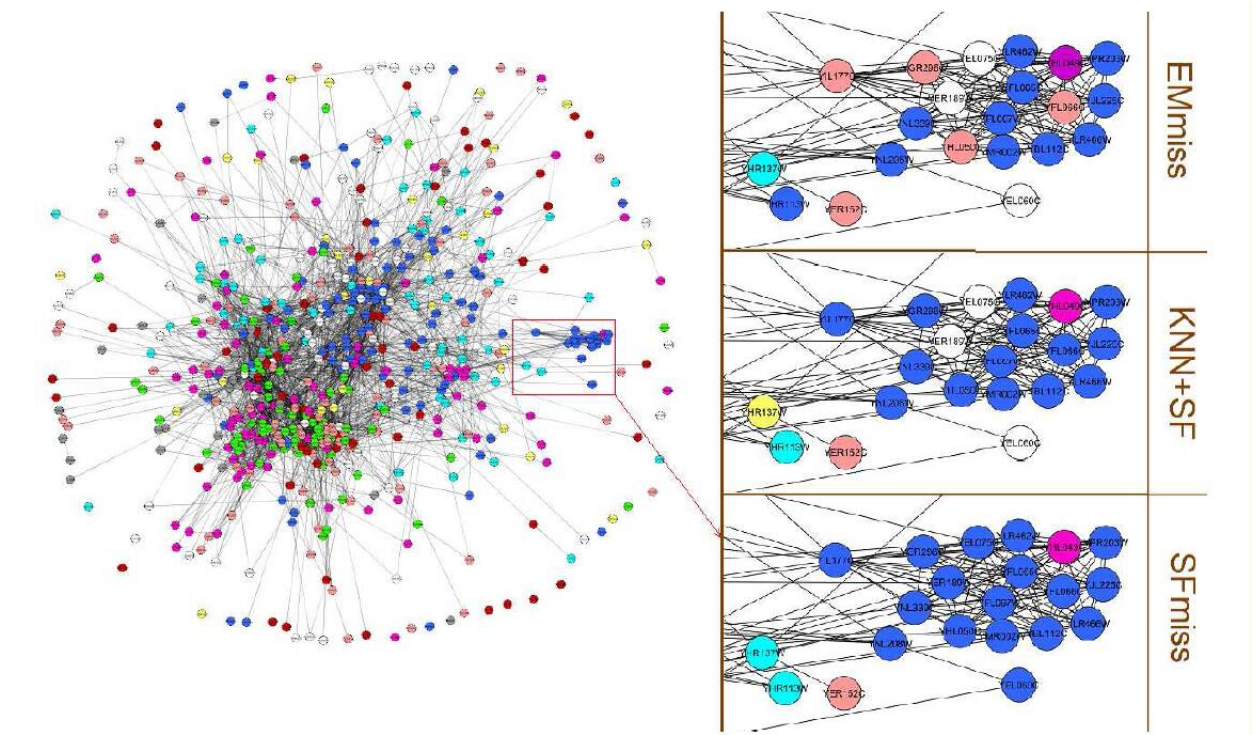
Groupe SFmiss	Terme GO & P-val. corresp.	Meilleure P-val. parmi les gr. EMmiss	Meilleure P-val. parmi les gr. KNN+SF
k=3	GO :0006732, coenzyme met. process		
	$1.1\,10^{-2}$	> 0.1	> 0.1
k=4	GO :0005819, spindle		
	$4.6\,10^{-9}$	$6.7\,10^{-7}$	$2.0\,10^{-6}$
	GO :0006790, sulf. met. process		
	$1.1\,10^{-4}$	$2.4\,10^{-4}$	$8.7\,10^{-4}$
	GO :0000278, mitotic cell cycle		
	$2.2\,10^{-3}$	$7.7\,10^{-3}$	> 0.1
	GO :0030472, mit. spin. org. & biogen. in nucleus		
	$5.2\,10^{-3}$	$8.8\,10^{-3}$	$2.0\,10^{-2}$
k=8	GO :0042555, MCM cplx		
	$3.4\,10^{-4}$	$8.3\,10^{-4}$	$4.0\,10^{-4}$
	GO :0008026, ATP-dep. helicase act.		
	$5.5\,10^{-4}$	$1.3\,10^{-3}$	$4.5\,10^{-4}$
	GO :0006268, DNA unwind. replic.		
	$2.8\,10^{-3}$	$6.7\,10^{-3}$	$1.1\,10^{-3}$
	GO :0042623, ATPase act. coupl.		
	$4.4\,10^{-3}$	$1.5\,10^{-2}$	$4.3\,10^{-2}$



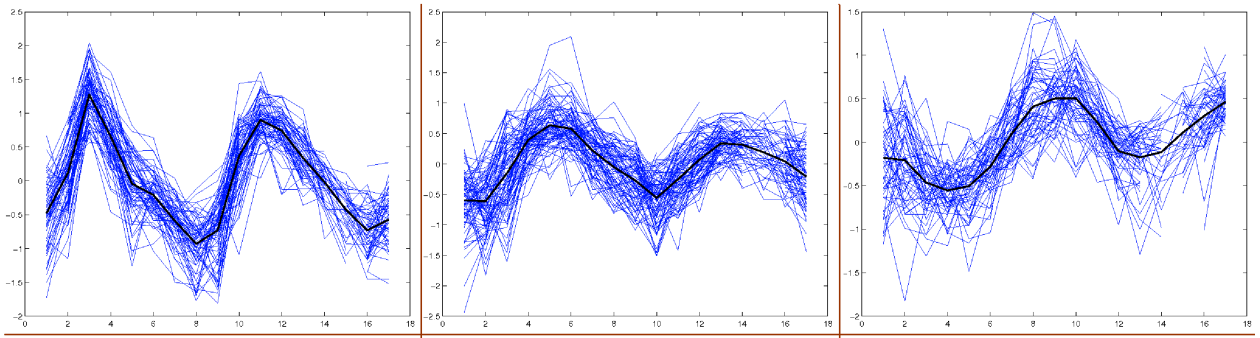
Exemple de hiérarchie GO avec quelques termes significativement surepresentés.

P-valeur corrigée pour contrôler le FDR de Benjamini & Hochberg, 1995.

Mise en évidence de module(s) ?

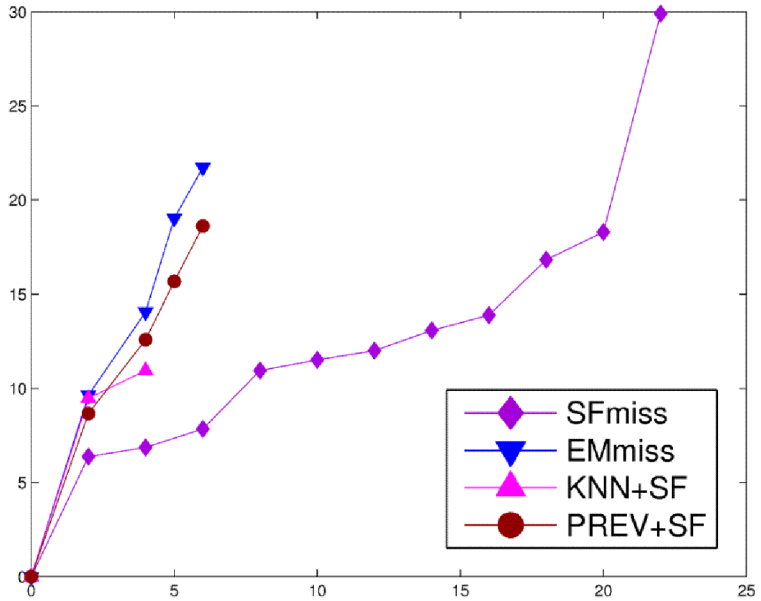


Interprétation des profils d'expression



Profils de quelques groupes SFmiss qui permettent une interprétation de la classification en phases du cycle cellulaire.

Stabilité de la classification vis-à-vis de la quantité d'observations manquantes



Taux d'erreur de différents algorithmes vs. proportion d'observations manquantes ajoutées (MCAR).

Plan détaillé ; c'est quand qu'on va où ?

Faisons le point

- Cadre statistique markovien pour l'intégration d'observations individuelles (éventuellement manquantes) et de données de paires.
  + Communication vers la communauté biologique.

- Résultats prometteurs de l'approche aussi bien sur données simulées que sur données de la levure.

Faisons le point

- Cadre statistique markovien pour l'intégration d'observations individuelles (éventuellement manquantes) et de données de paires.
  + Communication vers la communauté biologique.

- Résultats prometteurs de l'approche aussi bien sur données simulées que sur données de la levure.

Faisons le point

- Cadre statistique markovien pour l'intégration d'observations individuelles (éventuellement manquantes) et de données de paires.
  + Communication vers la communauté biologique.

- Résultats prometteurs de l'approche aussi bien sur données simulées que sur données de la levure.

Suites envisagées

- [en cours] :
  - Reconstruction des observations manquantes par estimation par moyenne conditionnelle en tenant compte de la classification.
  - Analyse sur un jeu de données complet, *e.g.* tout le réseau de la levure.

  Recquiert : beaucoup de temps et/ou un partenariat complet pour une expertise biologique.
- [court-moyen-long terme] :
  - Utilisation de pondérations fixées relatives à une distance ou une confiance en les arêtes du graphe (déjà implémenté) : $V_{ij}(z_i, z_j) = \omega_{ij} V(z_i, z_j)$.
  - Effets de la structure du réseau (grille régulière, réseau aléatoire, réseau *small world* ou exponentiel,...),
  - Étude de la robustesse de l'absence de certaines arêtes : graphe non fixé, incomplet, partiellement erroné.
  - modèle avec classification empiétante (dans l'esprit de Gash *et al* 2002 et Battle *et al* 2004) : aspects informatiques et expériences.
  - Prise en compte du mécanisme d'absence des observations pour améliorer les performances dans le cas des données NMAR (virtuellement tous les jeux de données réels).

Suites envisagées

- [en cours] :
  - Reconstruction des observations manquantes par estimation par moyenne conditionnelle en tenant compte de la classification.
  - Analyse sur un jeu de données complet, *e.g.* tout le réseau de la levure.

    Recquiert : beaucoup de temps et/ou un partenariat complet pour une expertise biologique.
- [court-moyen-long terme] :
  - Utilisation de pondérations fixées relatives à une distance ou une confiance en les arêtes du graphe (déjà implémenté) : $V_{ij}(z_i, z_j) = \omega_{ij} V(z_i, z_j)$.
  - Effets de la structure du réseau (grille régulière, réseau aléatoire, réseau *small world* ou exponentiel,...),
  - Étude de la robustesse de l'absence de certaines arêtes : graphe non fixé, incomplet, partiellement erroné.
  - modèle avec classification empiétante (dans l'esprit de Gash *et al* 2002 et Battle *et al* 2004) : aspects informatiques et expériences.
  - Prise en compte du mécanisme d'absence des observations pour améliorer les performances dans le cas des données NMAR (virtuellement tous les jeux de données réels).

Bibliographie sélective

Gideon Schwarz.

Estimating the dimension of a model.
The Annals of Statistics, 6(2) :131–134, 1978.

Yoav Benjamini and Yosef Hochberg.

Controlling the false dicovery rate – a practical and powerful approach to multiple testing.
Journal of the Royal Statistical Society B, 57(1) :289–300, 1995.

Ka-Yee Yeung, Chris Fraley Alejandro Murua, Adrian E. Raftery and Walter L. Ruzzo.

Model-based clustering and data transformations for gene expression data.
Bioinformatics, 17(10) :977–987, 2001

Roderick J.A. Little and Donald B. Rubin.
Statistical analysis with missing data.
emphWiley, 2002.

Yuri Lazebnik

Can a biologist fix a radio ? – Or what I learned while studying apoptosis.
Cancer Cell 2(3) :179–182, 2002.

Gilles Celeux, Florence Forbes and Nathalie Peyrard.

EM procedures using mean-field like approximations for Markov-model based image segmentation.
Pattern recognition, 36(1) :131–144, 2003.

Florence Forbes and Nathalie Peyrard.

Hidden Markov random field model selection criteria based on mean field-like approximations.
IEEE Transansactions on Pattern Analysis and Machine Intelligence, 25(9) :1089-1101, 2003.

Bibliographie sélective (suite)

Yoshihiro Yamanishi, Jean-Philippe Vert, Akihiro Nakaya and Minoru Kanehisa.

Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis.
Bioinformatics, 19(Suppl.1) :i323–i330, 2003.

Dae-Won Kim, Kwang H. Lee and Doheon Lee.

Detecting clusters of different geometrical shapes in microarray data.
Bioinformatics, 21(9) :1927–1934, 2005.

Julia Handl and Joshua Knowles and Douglas B. Kell.

Computational cluster validation in the post-genomic data analysis.
Bioinformatics, 21(15) :3201–3212, 2005.

Alexis Battle and Eran Segal and Daphne Koller.

Probabilistic discovery of overlapping cellular processes and their regulation.
Proceedings of the 8th Annual International Conference on Computational Molecular Biology, 167–176, ACM, 2004.

Dae-Won Kim and Ki-Young Lee and Kwang H. Lee and Doheon Lee.

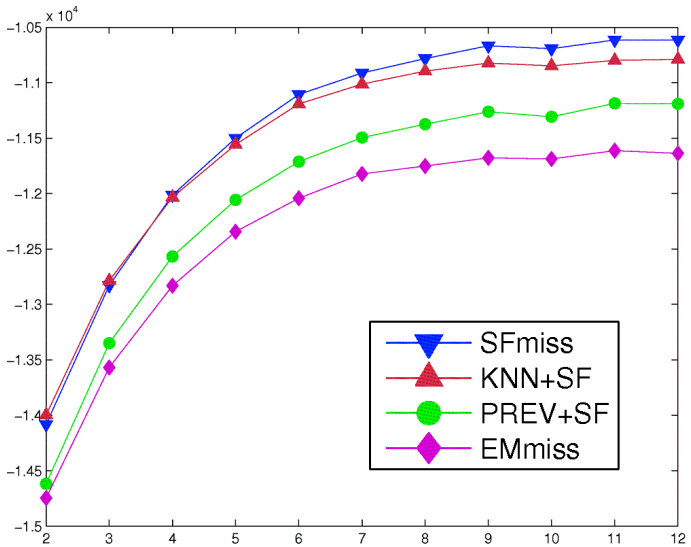
Towards clustering of incomplete microarray data without the use of imputation
Bioinformatics, 23(1) :107–113, 2007.

Merci pour votre écoute.

Maintenant les questions ?

Sélection du nombre de classes

Calcul de BIC (Schwarz 1977) pour $K = 2$ à $K = 12$ pour différents modèles. Approximation de type champ moyen (Forbes & Peyrard 2003) dans le cadre de champ de Markov cachés.
$K = 9$ est sélectionné (maximum) et SFmiss est le modèle retenu.



Calculs explicites

Étape E : $\tilde{t}_{ik}^{(q+1)} := P(Z_i = k|\tilde{\mathbf{z}}_{\nu(i)}, x_i^{o_i}, \Psi^{(q)}) = \frac{\tilde{\pi}_{ik}^{(q)} f(x_i^{o_i}|\theta_k^{(q)})}{\sum_l \tilde{\pi}_{il}^{(q)} f(x_i^{o_i}|\theta_l^{(q)})}$

Étape M : (Δ *idem* complet)

$\theta_k^{(q+1)} = \arg\max \sum_i \tilde{t}_{ik}^{(q+1)} \mathbb{E}[\log f(x_i^{o_i}, X_i^{m_i}|\theta_k)|x_i^{o_i}, \theta_k^{(q)}]$

moyennes :

$\mu_k^{(q+1)} = \frac{\sum_i t_{ik}^{(q)}[r_i \otimes x_i + (1-r_i) \otimes \eta_{ik}^{(q)}]}{\sum_i t_{ik}^{(q)}}$

matrice de covariance

$\Sigma_k^{(q)} = \frac{\sum_i t_{ik}^{(q)} S_{ik}^{(q)}}{\sum_i t_{ik}^{(q)}}$,

$S_{ik}^{(q)} = r_i.r_i^t \otimes (x_i - \mu_k^{(q)}).(x_i - \mu_k^{(q)})^t + \ldots$

$\ldots + \Gamma_{i,k}^{(q)}$

où $f(x_i^{o_i}|\theta_k) \sim \mathcal{N}(\mu_k^{o_i}, \Sigma_k^{o_i})$ et $f(x_i^{m_i}|x_i^{o_i}, \theta_k) \sim \mathcal{N}(\eta_{ik}, \Gamma_{ik})$ (et $\eta_{i,k} = \mu_k^{Man_i} + \Sigma_k^{Man_i, Obs_i}(\Sigma_k^{Obs_i})^{-1}(x_i^{Obs_i} - \mu_k^{Obs_i})$ ainsi que $\Gamma_{i,k} = \Sigma_k^{Man_i} - \Sigma_k^{Man_i, Obs_i}(\Sigma_k^{Obs_i})^{-1}\Sigma_k^{Obs_i, Man_i}$).

Calculs explicites

Étape E : $\tilde{t}_{ik}^{(q+1)} := P(Z_i = k|\tilde{\mathbf{z}}_{\nu(i)}, x_i^{o_i}, \Psi^{(q)}) = \frac{\tilde{\pi}_{ik}^{(q)} f(x_i^{o_i}|\theta_k^{(q)})}{\sum_l \tilde{\pi}_{il}^{(q)} f(x_i^{o_i}|\theta_l^{(q)})}$

Étape M : (Δ *idem* complet)
$\theta_k^{(q+1)} = \arg\max \sum_i \tilde{t}_{ik}^{(q+1)} I\!E[\log f(x_i^{o_i}, X_i^{m_i}|\theta_k)|x_i^{o_i}, \theta_k^{(q)}]$

moyennes :
$\mu_k^{(q+1)} = \frac{\sum_i t_{ik}^{(q)}[r_i \otimes x_i + (1-r_i) \otimes \eta_{ik}^{(q)}]}{\sum_i t_{ik}^{(q)}}$

matrice de covariance
$\Sigma_k^{(q)} = \frac{\sum_i t_{ik}^{(q)} S_{ik}^{(q)}}{\sum_i t_{ik}^{(q)}}$,
$S_{ik}^{(q)} = r_i.r_i^t \otimes (x_i - \mu_k^{(q)}).(x_i - \mu_k^{(q)})^t + \ldots$
$\ldots + \Gamma_{i,k}^{(q)}$

où $f(x_i^{o_i}|\theta_k) \sim \mathcal{N}(\mu_k^{o_i}, \Sigma_k^{o_i})$ et $f(x_i^{m_i}|x_i^{o_i}, \theta_k) \sim \mathcal{N}(\eta_{ik}, \Gamma_{ik})$ (et $\eta_{i,k} = \mu_k^{Man_i} + \Sigma_k^{Man_i, Obs_i}(\Sigma_k^{Obs_i})^{-1}(x_i^{Obs_i} - \mu_k^{Obs_i})$ ainsi que $\Gamma_{i,k} = \Sigma_k^{Man_i} - \Sigma_k^{Man_i, Obs_i}(\Sigma_k^{Obs_i})^{-1}\Sigma_k^{Obs_i, Man_i}$).