Modélisation de la structure génomique tumorale du cancer du sein par analyse factorielle parcimonieuse non paramétrique

Journée GDR-SFB-groupe Biopharma de la SFDS 2019

Sarah-Laure Rincourt[1,2], Stefan Michels[1,2] et Damien Drubay[1,2]

[1] Université Paris-Saclay, Univ. Paris-Sud, UVSQ, CESP, INSERM U1018 Oncostat, F-94805, Villejuif, France

[2] Service de Biostatistique et d'Epidémiologie, Institut Gustave Roussy, F-94805, 114 Rue Edouard Vaillant, 94800 Villejuif, France

09 octobre 2019

Introduction

Contexte - Cancer du sein

Deux chiffres de 2018 en France

- 58 459 nouveaux cas
- 12 146 décès estimés

Données génomiques : données de grande dimension

- Espoir d'amélioration de nouvelles cibles thérapeutiques
- Structures complexes (voies métaboliques, interactions)
- Grandes quantités de gènes (quelques centaines à quelques milliers)
  - Problème d'estimation
  - Problème de structuration de modèle (potentiellement complexité, ex : interaction)

Contexte

Données génomiques : données de grande dimension

- Espoir d'amélioration de nouvelles cibles thérapeutiques
- Structures complexes (voies métaboliques, interactions)
- Grandes quantités de gènes (quelques centaines à quelques milliers)

Idée : Trouver des structures latentes

- Hypothèse : redondance d'information = éléments d'une même voie métabolique (pathway)
- Méthodes les plus classiques : Clustering et ACP
- Moins courante : l'analyse factorielle

Objectif : Établir de la parcimonie dans la modélisation de structures génomiques

Tous les individus ne sont pas reliés à tous les facteurs

$\Rightarrow$ Analyse factorielle parcimonieuse

Exemple d'application : Modélisation du statut des récepteurs à oestrogènes et de l'expression génomique dans le cancer du sein

En se basant sur la structure parcimonieuse

$\Rightarrow$ Régression logistique

Matériel et Méthodes

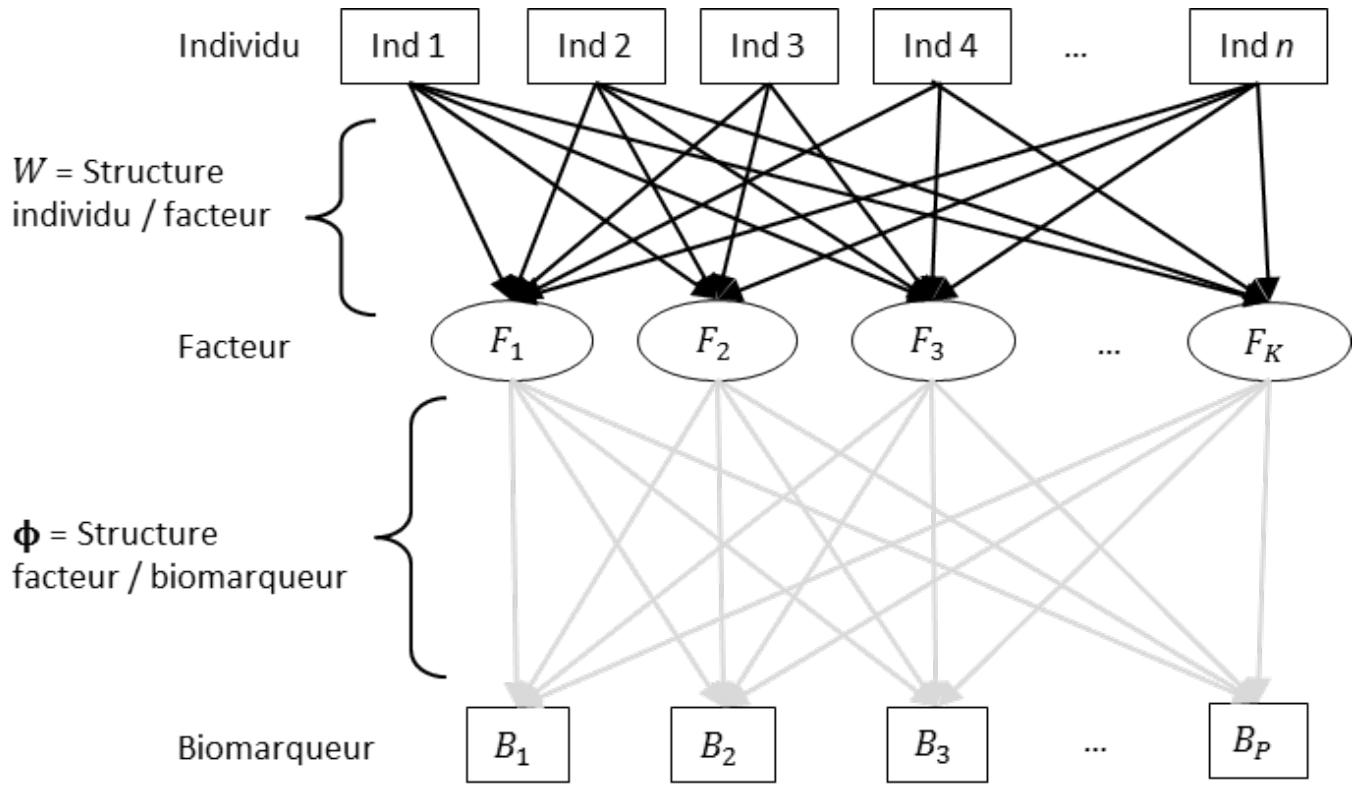
Jeu de données : Cancer du sein

- 523 patientes
- Jeu d'un regroupement d'essais cliniques
  - Traitées par chimiothérapie (anthracycline couplé ou non avec de la taxane)
- Expressions de 1689 gènes
- Variable d'entrée : Présence des récepteurs d'œstrogène (ER)

Projet Gene Expression Omnibus
http ://www.ncbi.nlm.nih.gov/geo

Analyse factorielle classique

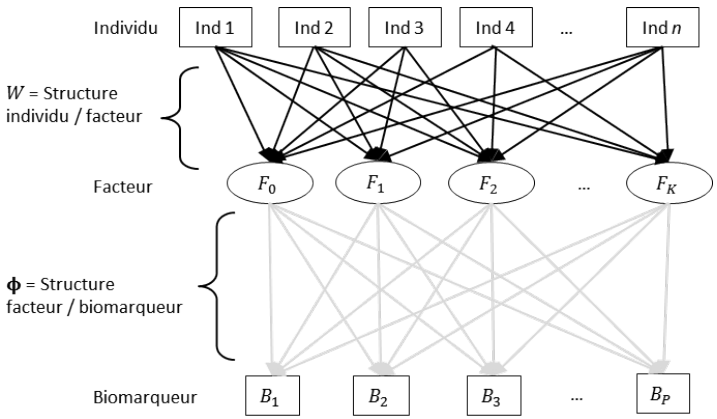
$X = W \phi$



Analyse factorielle classique

Z	F_1	F_2	F_3	...	F_K
Y_1	1	1	1	...	1
Y_2	1	1	1	...	1
Y_3	1	1	1	...	1
Y_4	1	1	1	...	1
...	...	...	...	...	...
Y_n	1	1	1	...	1

$X = (W \circ Z)\,\boldsymbol{\phi}$

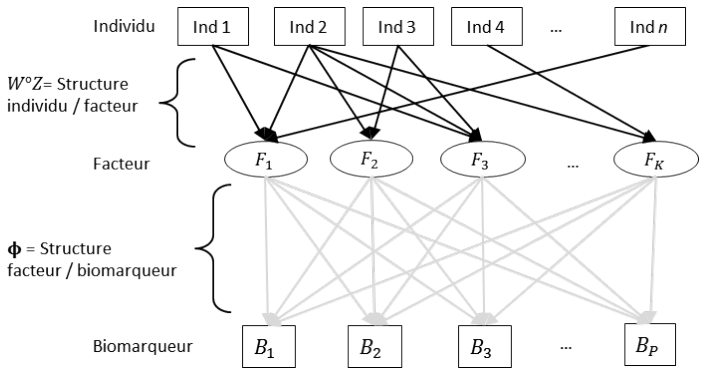


- Tous les individus et toutes les variables sont rattachés à tous les facteurs

Analyse factorielle parcimonieuse

Z	F_1	F_2	F_3	...	F_K
Y_1	1	0	1	...	0
Y_2	1	1	1	...	1
Y_3	0	1	1	...	0
Y_4	0	0	0	...	1
...	...	...	...	...	...
Y_n	1	0	0	...	0

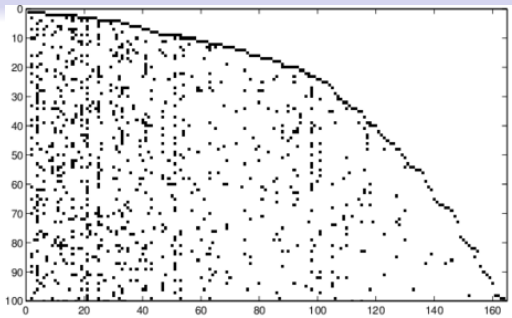
$X = (W \circ Z)\,\boldsymbol{\phi}$



2 difficultés :

- Choix du nombre de facteurs (de 1 à ∞)
- Paramétrisation de la parcimonie (Z)

Analyse factorielle parcimonieuse



2 difficultés :

- Choix du nombre de facteurs (de 1 à ∞)

  Bayésien non-paramétrique : nombre infini de paramètres

- Paramétrisation de la parcimonie (Z)

  $\Rightarrow$ Processus Beta-Bernoulli

  - Modéliser le nombre de facteurs latents : Processus Beta
  - Induire de la parcimonie : Processus Bernoulli

(Chen et al. 2010)

Analyse factorielle - Bayésien non paramétrique

Prior classique de l'analyse factorielle : Gaussien multivarié

$$X \sim \mathcal{N}(\Phi(W \circ Z), diag(\tau_{E_1}^{-1}, ..., \tau_{E_p}^{-1}))$$
$$\Phi_{j,k} \sim \mathcal{N}(0, \tau_{\phi_{j,k}}^{-1})$$
$$W_{k,i} \sim \mathcal{N}(0, \tau_W^{-1} I_K)$$

$$\tau_{E_j} \sim Gamma(g_j, h_j)$$
$$\tau_{\phi_{j,k}} \sim Gamma(c_{j,k}, d_{j,k})$$
$$\tau_W \sim Gamma(e, f)$$

Prior de Z : Processus beta-Bernoulli

$$Z_{k,i} \sim Bernoulli(\pi_k)$$
où $i = 1, ..., n$, $j = 1, ..., p$ et $k = 1, ..., K$

$$\pi_k \sim Beta(\frac{\alpha}{K}, \frac{\beta(K-1)}{K})$$

Inférence Variationnelle (VI)

- Modèle complexe
- Méthodes MCMC : calcul impossible en un temps raisonnable

- VI := dérivé bayésien de l'algorithme EM (+ rapide) (Blei et al. 2017)
- Vrai distribution (nommé p) approximée par une distribution plus simple (nommé q)

$$ELBO(q) = \mathbb{E}(log\ p(x|\theta)) - KL(q(\theta) \parallel p(\theta))$$
$$KL(q(\theta) \parallel p(\theta)) = \sum_{\theta} q(\theta) log \frac{q(\theta)}{p(\theta)}$$

Régression logistique - Prior Horseshoe

$$logit(Y) \sim \mathcal{N}(\beta \widehat{Z}, \lambda_k^2 \tau^2)$$
$$(\beta_k | \lambda_k, \tau) \sim \mathcal{N}(0, \lambda_k^2 \tau^2)$$
$$\lambda_k \sim \mathcal{C}^+(0, 1)$$
$$\tau \sim \mathcal{C}^+(0, 1)$$

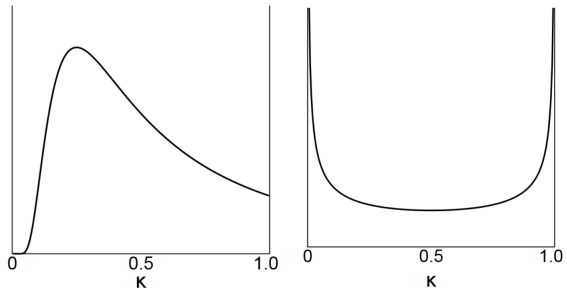


FIGURE 1 – Prior Laplacien VS Prior Horseshoe (Fig2 Carvalho et al.)

- Variable explicative : espérance de la matrice $\widehat{Z}$ de l'analyse factorielle
- Prior horseshoe sur les coefficients β (Carvalho et al. 2009)
- Odd Ratio $(OR_k) = e^{\beta_k}$

Monte Carlo Hamiltonienne (HMC)

Utilisation de la méthode HMC car le modèle plus simple

- Basé sur les méthodes de Monte Carlo par Chaîne de Markov (MCMC)
- Utilise les dérivées de la fonction de densité pour générer des transitions efficaces pour couvrir le posterior (Duane et al. 1987)

Implémentation et Optimisation

Analyse factorielle : Inférence Variationnelle

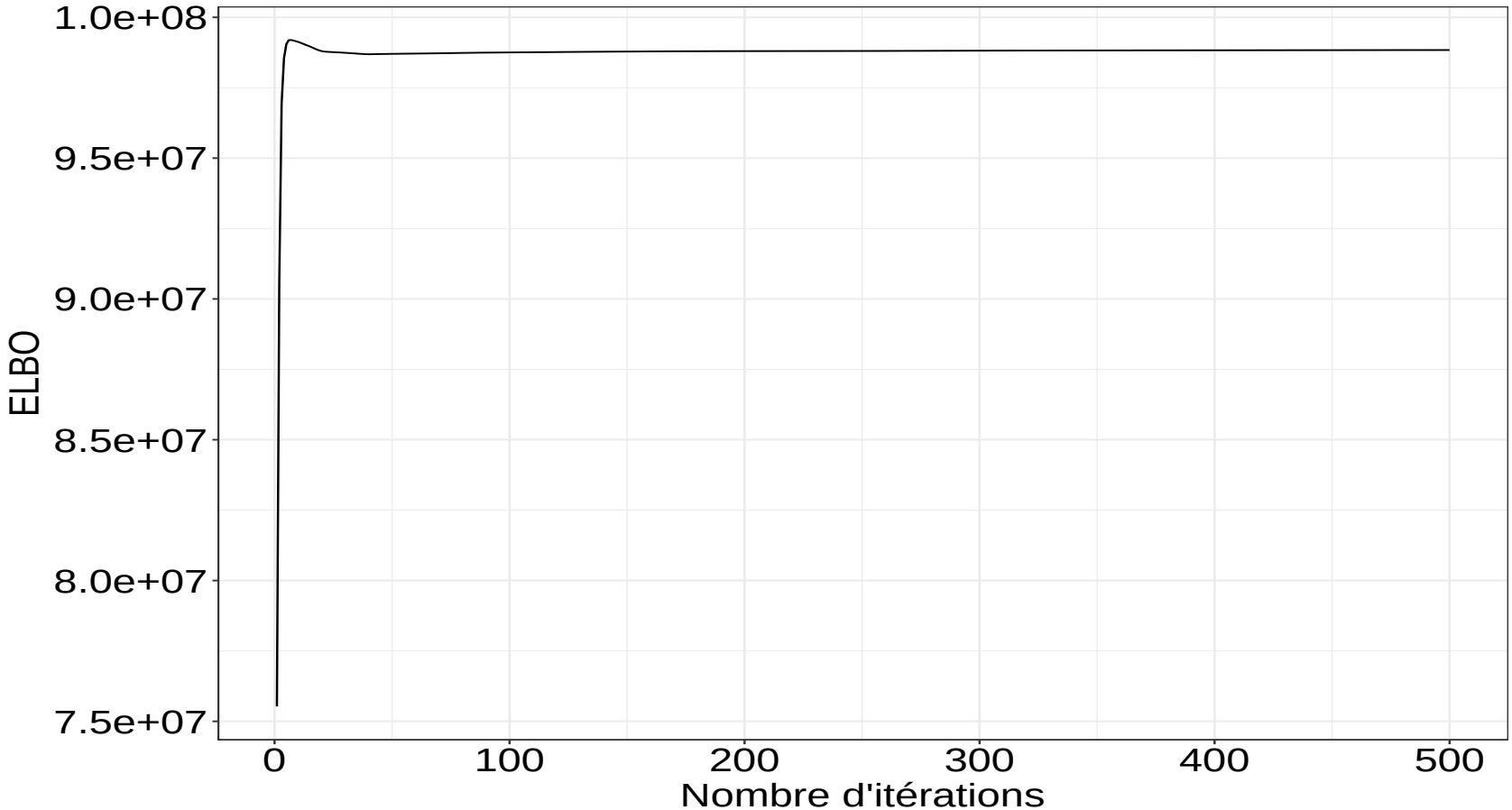
- Critère d'arrêt : décision à 500 itérations
- Vérification de la convergence par l'ELBO
- Ecriture matricielle sur le logiciel R

Régression logistique : HMC

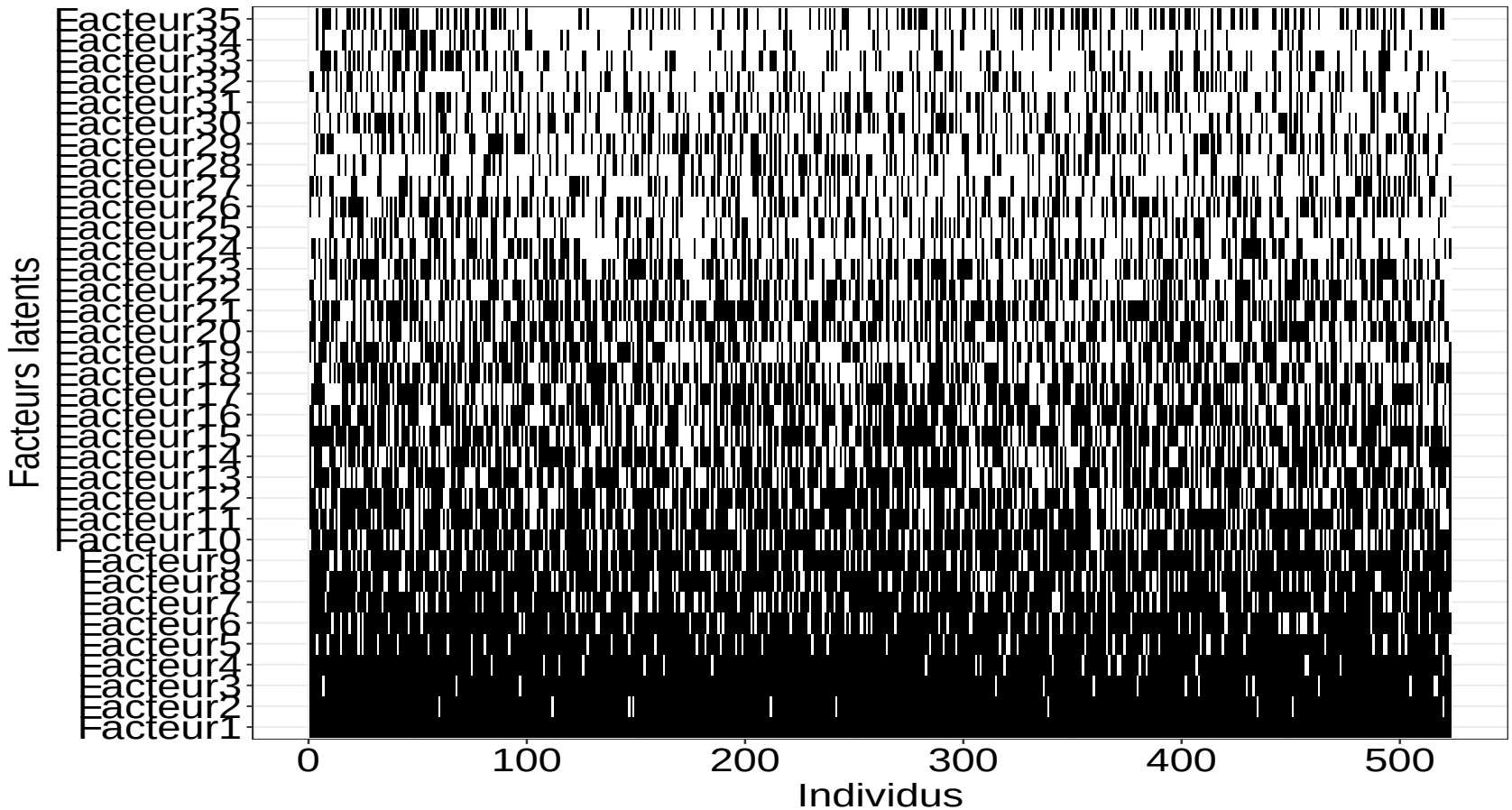
- 2 000 itérations dont 500 de chauffe
- 4 chaînes
- logiciel R package RStan (Carpenter et al. 2017)

Résultat

ELBO


1.0e+08
9.5e+07
9.0e+07
8.5e+07
8.0e+07
7.5e+07
ELBO
0
100
200
300
400
500
Nombre d'itérations

Représentation de la matrice binaire Z



35 facteurs latents (non nuls) observés

Résultat de la régression logistique

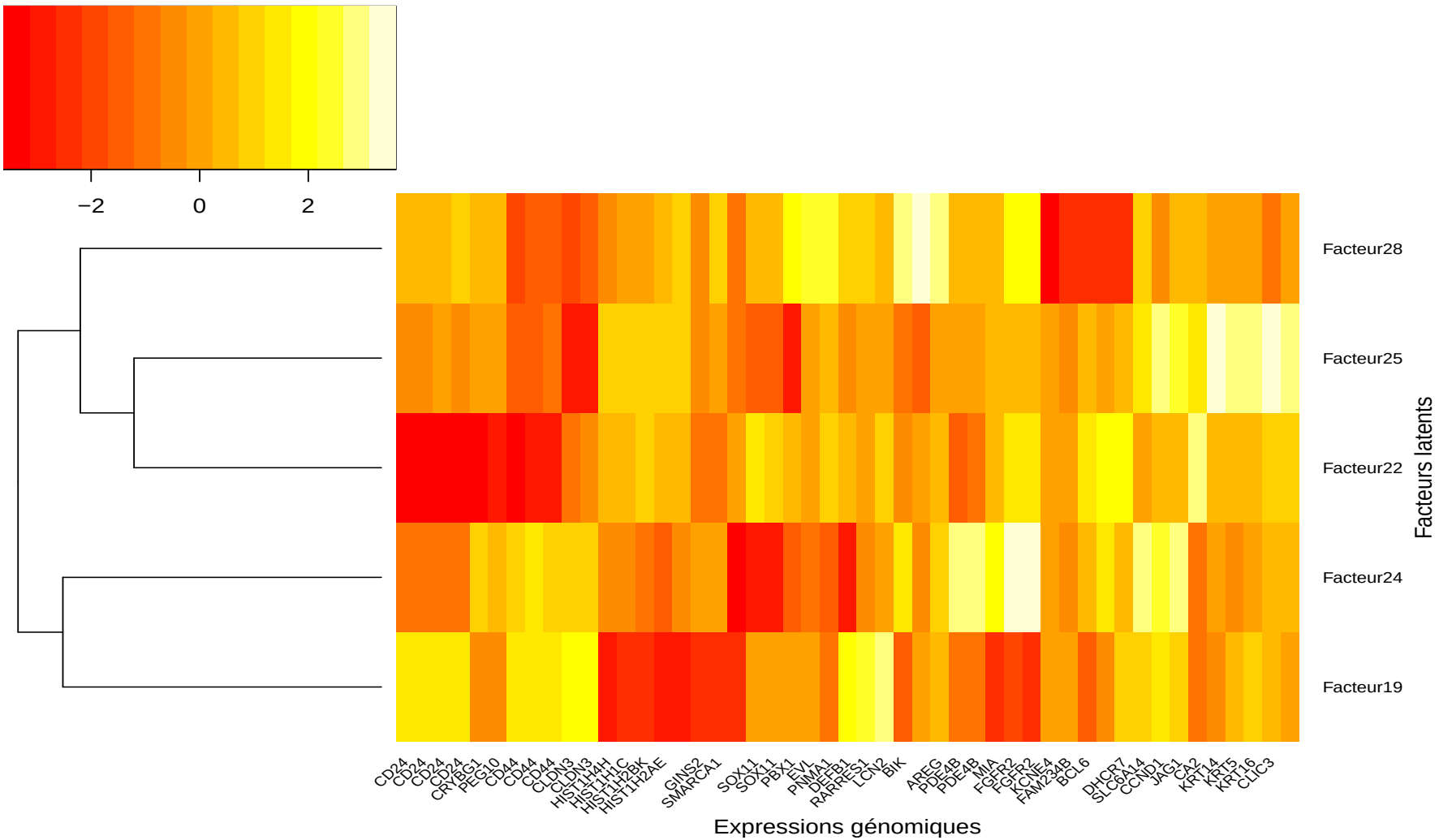
Odds Ratio des différents facteurs latents sélectionnés

	OR	$IC^*_{95\%}$
Facteur19	0.59	[0.38 ; 0.95]
Facteur22	1.57	[1.00 ; 2.49]
Facteur24	0.38	[0.24 ; 0.58]
Facteur25	0.42	[0.26 ; 0.70]
Facteur28	2.64	[1.65 ; 4.24]

(*) : Intervalle de crédibilité à 95 % (percentiles des distributions des itérations convergées des 4 chaînes)

Résultat de la régression logistique

Heatmap des facteurs latents sélectionnés en fonction des gènes (Φ)



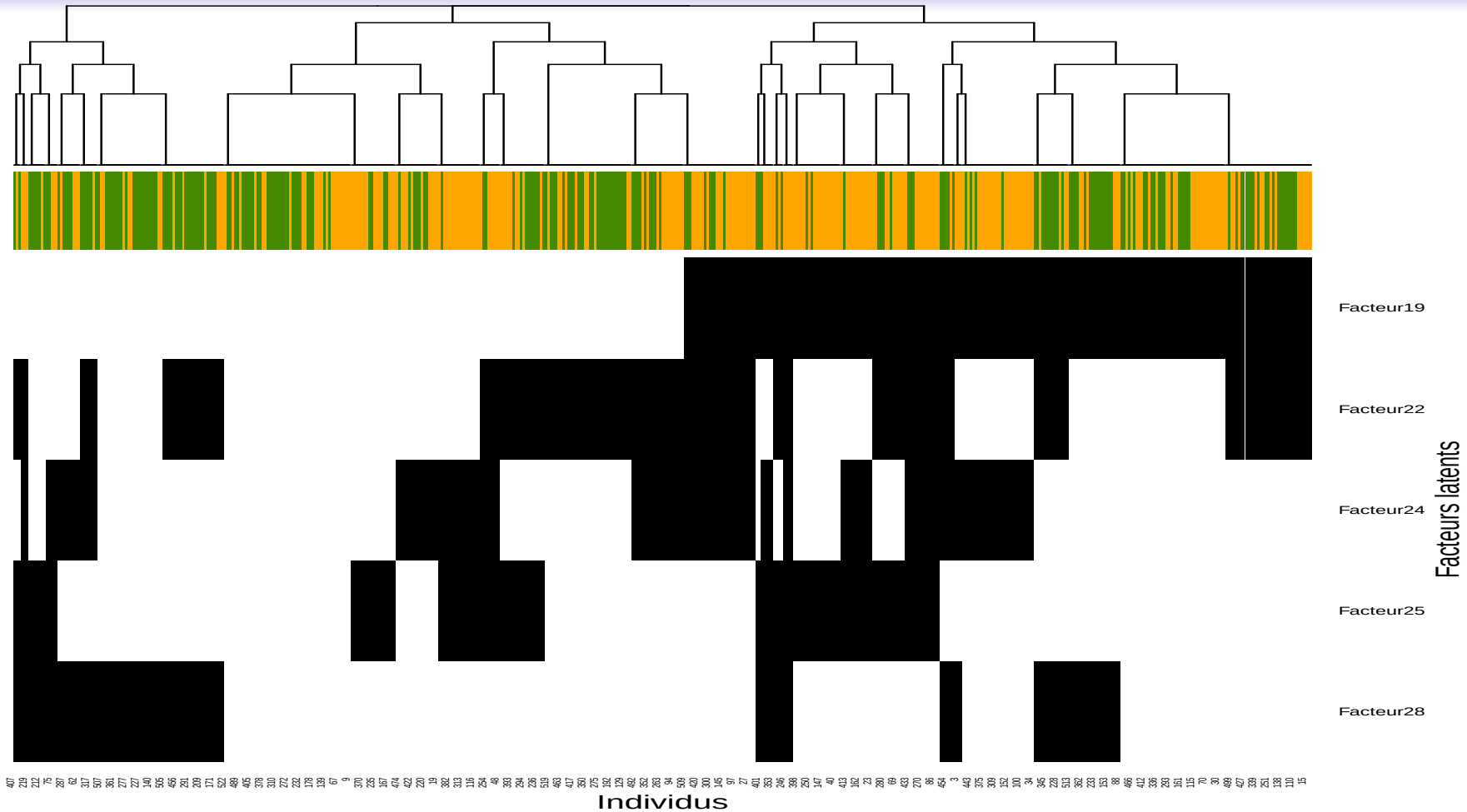
Résultat de la régression logistique

Tableaux des gènes reliés par leur fonction aux facteurs latents

Facteur19	Facteur22	Facteur24	Facteur25	Facteur28
HIST1H1C	CD44	FGFR2	KRT14	AREG
HIST1H2BK	CD44	FGFR2	PBX1	EVL
HIST1H4H	CD44	PDE4B	KRT16	BCL6
HIST1H2AE	CA2	PDE4B	KRT5	DHCR7
LCN2	CD24	DEFB1	CCND1	KCNE4
GINS2	CD24	JAG1	CLDN3	BIK
NA	CD24	SLC6A14	CLDN3	PNMA1
RARRES1	CD24	SOX11	JAG1	NA
MIA	PEG10	SOX11	CLIC3	NA
SMARCA1	CRYBG1	NA	NA	FAM234B

Résultat de la régression logistique

Profils obtenus des individus



Avec 51 individus sans facteur

Vert : statut ER+ ; orange statut ER-

Conclusion

Conclusion

Mise en place d'une modélisation en 2 étapes

1. Analyse factorielle parcimonieuse
2. Régression logistique horseshoe pour la modélisation du statut des récepteurs

Conclusion

Mise en place d'une modélisation en 2 étapes

1. Analyse factorielle parcimonieuse
2. Régression logistique horseshoe pour la modélisation du statut des récepteurs

Dans notre exemple :

Analyse factorielle

- Espérance du nombre de facteur latent
- Parcimonie de notre analyse factorielle
- Profil des individus hétérogènes

Régression logistique

Relation entre les facteurs latents et les récepteurs à oestrogènes

$\Rightarrow$ Forte importance de la prolifération

Conclusion

Perspectives

- Modélisation jointe de l'analyse factorielle avec une réponse :
  - $\Rightarrow$ Binaire ou de Survie
- Mettre en place un modèle plus interprétable : mise en place d'un profil moyen

Merci de votre attention

Bibliographie

Blei,D., Kucukelbir A, McAuliffe, J. (2017) Variational Inference : A Review for Statisticians. Journal of the American Statistical Association, 112 :518, 859-877. DOI : 10.1080/01621459.2017.1285773

Carvalho, C.M., Polson, N.G. , Scott, J.G.. (2009). Handling Sparsity via the Horseshoe. Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, in PMLR 5 :73-80

Chen, B., Chen, M., Paisley, J., Zaas, A., Woods, C., Ginsburg, G. S., ... Carin, L. (2010). Bayesian inference of the number of factors in gene-expression analysis : application to human virus challenge studies. BMC bioinformatics, 11, 552. doi:10.1186/1471-2105-11-552

Duane, S., Kennedy, A.D.,Pendleton B. J. , Roweth D. (1987). Hybrid Monte Carlo. Phys. Lett. B, 195, pp. 216-222. DOI 10.1016/0370-2693(87)91197-X

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, ... Allen Riddell. (2017). Stan : A probabilistic programming language. Journal of Statistical Software 76(1). DOI 10.18637/jss.v076.i01