Modèle Bernoulli-Gaussien pour l'analyse génétique

Cécile BAZOT[1], Nicolas DOBIGEON[1], Jean-Yves TOURNERET[1], Alfred O. HERO III[2]

[1]Université de Toulouse, IRIT/INP-ENSEEIHT, Toulouse, France
[2]University of Michigan, EECS Department, Ann Arbor, USA
{cecile.bazot, nicolas.dobigeon, jean-yves.tourneret}@enseeiht.fr, hero@umich.edu

Résumé – Cet article présente un modèle Bayésien hiérarchique d'analyse factorielle, appliqué à des données génétiques. Chaque échantillon observé est la combinaison d'un certain nombre de signatures génétiques (ou *facteurs*) parmi une bibliothèque de facteurs, suivant un modèle de mélange linéaire. La particularité de la méthode proposée est de prendre en compte la parcimonie des coefficients du mélange (ou *scores*) de chaque signature génétique dans l'échantillon étudié en choisissant une loi Bernoulli-Gaussienne tronquée comme loi *a priori* pour ces coefficients. Cette loi permet également de satisfaire les contraintes physiques de positivité et d'additivité de ces coefficients. Des simulations conduites sur des données synthétiques, en comparaison avec d'autres méthodes d'analyse factorielle, ont montré les performances du modèle proposé. Cette méthode a également été appliquée sur des données génétiques réelles.

Abstract – This paper investigates a hierarchical Bayesian algorithm for gene factor analysis. Each observed sample is decomposed as a linear combination of gene signatures (also referred to as *factors*) following a linear mixing model. To enforce the sparsity of the relative contribution (called *factor score*) of each gene signature to a specific sample, constrained Bernoulli-Gaussian distributions are elected as prior distributions for these factor scores. This distribution allows the positivity and full-additivity constraints to be ensured. Simulations on synthetic data, in comparison with other factor analysis algorithms, and real data illustrate the proposed method.

1 Introduction et position du problème

Dans le cadre de l'analyse génétique, les méthodes d'analyse factorielle permettent de décomposer une matrice $\mathbf{Y} \in \mathbb{R}^{G\times N}$ dont les lignes (respectivement les colonnes) correspondent aux différents gènes (resp. échantillons), avec $N << G$. Chaque échantillon observé $\mathbf{y}_i$ $(i = 1, \ldots, N)$ se décompose suivant le modèle de mélange linéaire

$$\mathbf{y}_i = \sum_{r=1}^{R} \mathbf{m}_r a_{i,r} + \mathbf{n}_i \tag{1}$$

où $\mathbf{m}_r = [m_{r,1}, \ldots, m_{r,G}]^T$ désigne la $r^{\text{ème}}$ signature génétique (ou *facteur*), $a_{i,r}$ est la proportion (ou *facteur score*) du $r^{\text{ème}}$ facteur dans le $i^{\text{ème}}$ échantillon et $\mathbf{n}_i$ correspond à une erreur résiduelle. Dans cet article, les R facteurs $\{\mathbf{m}_r\}_{r=1,\ldots,R}$ sont supposés appartenir à une bibliothèque $\mathbf{M} \in \mathbb{R}^{G\times K}$ de K signatures possibles, avec $K > R$. En considérant N échantillons, le modèle se réécrit sous la forme matricielle

$$\mathbf{Y} = \mathbf{MA} + \mathbf{N}$$

où $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$, $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N]$ représente la matrice des scores, $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_K]$ celle des signatures génétiques et $\mathbf{N} = [\mathbf{n}_1, \ldots, \mathbf{n}_N]$.

La méthode proposée permet d'estimer, de manière totalement non supervisée, les proportions $\{\mathbf{a}_i\}_{i=1,\ldots,N}$, avec $\mathbf{a}_i = [a_{i,1}, \ldots, a_{i,R}]^T$, et les signatures génétiques $\{\mathbf{m}_r\}_{r=1,\ldots,R}$, à partir des échantillons observées $\{\mathbf{y}_i\}_{i=1,\ldots,N}$. Le nombre de facteurs est ainsi déterminé directement à partir des données. L'avantage de cette méthode, en comparaison avec d'autres méthodes d'analyse factorielle comme l'anayse factorielle non paramétrique (NPBFA) [1] ou le modèle BFRM [2], est qu'elle permet de satisfaire aux contraintes de positivité et d'additivité liées à la physique du modèle

$$\begin{aligned} &a_{i,k} \geq 0,\ \sum_{k=1}^{K} a_{i,k} = 1,\ i = 1, \ldots, N,\ k = 1, \ldots, K, \\ &m_{k,l} \geq 0,\ k = 1, \ldots, K,\ l = 1, \ldots, L. \end{aligned} \tag{2}$$

Remarquons que de telles contraintes ont déjà été imposées dans [3] pour l'imagerie hyperspectrale. Cependant l'approche adoptée ici diffère de [3] dans le sens où une contrainte de parcimonie est en plus imposée pour les vecteurs des facteurs scores.

Comme dans de nombreuses méthodes d'analyse factorielle Bayésienne, le vecteur $\mathbf{n}_i = [n_{i,1}, \ldots, n_{i,G}]^T$ est une séquence de bruit additif que l'on supposera indépendante et identiquement distribuée (i.i.d.) suivant une loi normale centrée et de matrice de covariance $\mathbf{\Sigma} = \sigma^2\mathbf{I}_G$

$$\mathbf{n}_i|\sigma^2 \sim \mathcal{N}\left(\mathbf{0}_G, \sigma^2\mathbf{I}_G\right) \tag{3}$$

où $\mathbf{I}_G$ est la matrice identité de dimension $G \times G$.

La modélisation Bayésienne proposée dans cet article est un moyen efficace de prendre en compte toutes les contraintes citées précédemment, notamment en choisissant des lois *a priori* adéquates pour les paramètres inconnus. La détermination des paramètres inconnus du modèle Bayésien proposé dans cet article est réalisée par un algorithme de Gibbs qui génère des données distribuées asymptotiquement suivant la loi *a posteriori* des paramètres inconnus. Des résultats de simulations conduites sur des données synthétiques et réelles permettent d'illustrer l'intérêt de la méthode.

2 Modèle Bayésien hiérarchique

Le modèle Bayésien utilisé est basé sur la vraisemblance des observations et sur la définition de lois *a priori* adéquates pour les paramètres, notamment les vecteurs des facteurs scores $\{\mathbf{a}_i\}_{i=1,\ldots,N}$ et des signatures génétiques $\{\mathbf{m}_k\}_{k=1,\ldots,K}$.

2.1 Fonction de vraisemblance

Le modèle de mélange linéaire défini par (1) ainsi que les propriétés statistiques du vecteur de bruit $\mathbf{n}_i$ (3) permettent d'écrire $\mathbf{y}_i|\mathbf{M},\mathbf{a}_i,\sigma^2 \sim \mathcal{N}\left(\mathbf{M}\mathbf{a}_i,\sigma^2\mathbf{I}_G\right)$. Ainsi, la vraisemblance des observations $\mathbf{Y}$ est

$$f(\mathbf{Y}|\mathbf{M},\mathbf{A},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{GN/2}} \exp\left[-\frac{\sum_{i=1}^{N}\|\mathbf{y}_i-\mathbf{M}\mathbf{a}_i\|^2}{2\sigma^2}\right] \tag{4}$$

où $\|\cdot\|$ est la norme l_2.

2.2 Lois *a priori* des paramètres

2.2.1 Scores

Considérons tout d'abord une loi normale de moyenne nulle et de variance α^2, tronquée sur l'intervalle $]0,\mu^+]$, et notée $\mathcal{N}_{]0,\mu^+]}\left(0,\alpha^2\right)$. Sa densité de probabilité s'écrit [4]

$$\varphi_{]0,\mu^+]}(x) = \frac{C}{\sqrt{2\pi\alpha^2}} \exp\left(-\frac{x^2}{2\alpha^2}\right) \mathbf{1}_{]0,\mu^+]}(x) \tag{5}$$

où $\mathbf{1}_{\mathbb{E}}(x)$ est la fonction indicatrice définie sur $\mathbb{E}$ (i.e., $\mathbf{1}_{\mathbb{E}}(x) = 1$ si $x \in \mathbb{E}$ et 0 sinon). Dans (5), $C = \left[\Phi\left(\frac{\mu^+}{\alpha}\right) - \frac{1}{2}\right]^{-1}$ est une constante de normalisation où Φ est la fonction de répartition de la loi normale centrée réduite. La génération d'échantillons distribués suivant la loi normale tronquée (5) peut être effectuée avec une stratégie similaire à celle décrite dans [5].

En général, un nombre réduit de facteurs (noté $R < K$) appartenant à la bibliothèque (notée $\mathbf{M}$) participent au mélange (1), ce qui se traduit par de nombreux coefficients $a_{i,k}$ égaux à 0. Nous proposons l'utilisation d'une loi *a priori* exploitant cette propriété de parcimonie des vecteurs $\mathbf{a}_i$ ($i = 1,\ldots,N$). En suivant l'approche décrite dans [6], il semble intéressant d'utiliser une loi *a priori* définie par un mélange d'une masse à l'origine et d'une loi normale tronquée. Ainsi, si nous notons $\mathbf{a}_{i,1:k-1}$ ($k = 2,\ldots,K-1$) le vecteur constitué des $k-1$ premiers éléments du vecteur $\mathbf{a}_i$, la loi *a priori* choisie pour les scores est la loi tronquée Bernoulli-Gaussienne suivante

$$\begin{array}{ll} a_{i,1} & \sim (1-w_i)\,\delta(a_{i,1}) + w_i\,\mathcal{N}_{]0,1]}(0,\alpha^2), \\ a_{i,k}|\mathbf{a}_{i,1:k-1} & \sim (1-w_i)\,\delta(a_{i,k}) + w_i\,\mathcal{N}_{]0,\mu^+_{i,k}]}(0,\alpha^2), \end{array} \tag{6}$$

où $\delta(\cdot)$ est une masse en zéro et w_i est un hyperparamètre inconnu renseignant sur la probabilité *a priori* d'avoir un coefficient non-nul. De plus, pour respecter la contrainte d'additivité, la loi normale associée aux termes non nuls du mélange est tronquée à droite par $\mu^+_{i,k} = 1 - \sum_{j=1}^{k-1} a_{i,j}$ ($k = 2,\ldots,K-1$) et le dernier élément du vecteur des abondances est fixé à $a_{i,K} = \mu^+_{i,K} \triangleq 1 - \sum_{k=1}^{K-1} a_{i,k}$. Ainsi, la distribution *a priori* pour le vecteur des proportions $\mathbf{a}_i$ ($i = 1,\ldots,N$) dont le dernier élément $a_{i,K}$ est fixé à $\mu^+_{i,K}$ s'écrit

$$f(\mathbf{a}_i) = f(a_{i,1})\left[\prod_{k=2}^{K-1} f(a_{i,k}|\mathbf{a}_{i,1:k-1})\right]\delta\left(a_{i,K}-\mu^+_{i,K}\right).$$

En supposant que les vecteurs de proportions $\mathbf{a}_i$ ($i = 1,\ldots,N$) sont *a priori* indépendants, on obtient la loi jointe *a priori* suivante pour la matrice des scores $\mathbf{A}$: $f(\mathbf{A}) = \prod_{i=1}^{N} f(\mathbf{a}_i)$.

2.2.2 Signatures génétiques

L'ensemble des échantillons observés respectant les contraintes de positivité et d'additivité appartiennent à un polytope convexe de $\mathbb{R}^L$ dont les sommets sont les K signatures génétiques à estimer. Les données peuvent ainsi être représentées dans un sous-espace $\mathcal{V}_{K-1}$ de dimension $K-1$, avec $R \leq K \leq G$. Ce sous-espace peut être estimé préalablement par une méthode de réduction de dimensionalité, comme l'analyse en composantes principales (ACP) par exemple. Notons que ce genre d'approche a été utilisé avec succès pour l'imagerie hyperspectrale dans [3] et permet de réduire considérablement les degrés de liberté des paramètres à estimer. Ainsi, au lieu d'estimer directement les signatures génétiques $\mathbf{m}_k$ ($k = 1,\ldots,K$), nous proposons d'estimer leurs projections $\mathbf{t}_k$ sur ce sous-espace $\mathcal{V}_{K-1}$ de dimension réduite. Les signatures génétiques $\mathbf{m}_k$ et leurs projections $\mathbf{t}_k$ sur les composantes principales pertinentes issues de l'ACP sont reliées par l'équation $\mathbf{t}_k = \mathbf{P}(\mathbf{m}_k - \bar{\mathbf{y}})$, où $\mathbf{P}$ est la matrice de projection (de taille $K-1 \times G$) et $\bar{\mathbf{y}}$ la moyenne empirique des échantillons.

La loi *a priori* choisie pour les signatures projetées $\mathbf{t}_k$ est alors une loi normale multivariée $\mathcal{N}_{\mathcal{T}_k}\left(\mathbf{e}_k, s_k^2\mathbf{I}_{k-1}\right)$ tronquée sur $\mathcal{T}_k$

$$\mathbf{t}_k|\mathbf{e}_k, s_k^2 \sim \mathcal{N}_{\mathcal{T}_k}\left(\mathbf{e}_k, s_k^2\mathbf{I}_{k-1}\right). \tag{7}$$

La troncature sur l'ensemble $\mathcal{T}_k$ assure la positivité des coefficients des signatures définie dans (2)

$$\{m_{k,g} \geq 0,\ \forall g = 1,\ldots,G\} \quad \Leftrightarrow \quad \{\mathbf{t}_k \in \mathcal{T}_k\}. \tag{8}$$

Les vecteurs moyennes $\mathbf{e}_k$ sont fixés comme les solutions d'un algorithme d'extraction de pôles de mélange dédié à l'imagerie hyperspectrale, par exemple VCA [7]. Les variances s_k sont fixées à de grandes valeurs. En supposant que les facteurs projetés sont *a priori* indépendants, la loi jointe *a priori* pour la matrice des facteurs projetés $\mathbf{T} = [\mathbf{t}_1,\ldots,\mathbf{t}_K]$ est $f(\mathbf{T}) = \prod_{k=1}^{K} f(\mathbf{t}_k)$.

2.2.3 Variance du bruit

La loi *a priori* de la variance σ^2 est une loi conjuguée inverse-Gamma de paramètres $\nu/2$ et $\gamma/2$, i.e., $\sigma^2|\nu,\gamma \sim \mathcal{IG}\left(\frac{\nu}{2},\frac{\gamma}{2}\right)$. L'hyperparamètre ν sera fixé à $\nu = 2$ alors que l'hyperparamètre γ sera ajustable (comme dans [3, 6]).

2.3 Lois *a priori* des hyperparamètres

Notons $\mathbf{\Psi} = \{\boldsymbol{w},\gamma\}$ le vecteur d'hyperparamètres associés au modèle défini précédemment, avec $\boldsymbol{w} = [w_1,\ldots,w_N]^T$.

L'algorithme proposé ici estime ces hyperparamètres à partir des données en définissant des lois *a priori* adéquates pour ces hyperparamètres. Plus précisément, une loi uniforme sur l'ensemble $[0,1]$ est choisie comme loi *a priori* pour la proportion moyenne des facteurs scores non-nuls, i.e., $w_i \sim \mathcal{U}([0,1])$. Enfin, en suivant l'approche décrite dans [3, 6], une loi *a priori* non-informative de Jeffreys est choisie pour l'hyperparamètre γ, i.e., $f(\gamma) \propto \frac{1}{\gamma}\mathbf{1}_{\mathbb{R}^+}(\gamma)$.

En supposant que tous les hyperparamètres de ce modèle Bayésien sont *a priori* indépendants, la loi du vecteur d'hyperparamètres est

$$f(\mathbf{\Psi}) = f(\boldsymbol{w})f(\gamma) \propto \frac{1}{\gamma}\prod_{i=1}^{N}\mathbf{1}_{[0,1]}(w_i)\,\mathbf{1}_{\mathbb{R}^+}(\gamma)\,. \tag{9}$$

où $\propto$ signifie "proportionnel à".

2.4 Loi *a posteriori*

La loi *a posteriori* jointe des vecteurs des paramètres inconnus $\mathbf{\Theta} = \{\mathbf{T}, \mathbf{A}, \sigma^2\}$ et hyperparamètres $\mathbf{\Psi} = \{\boldsymbol{w}, \gamma\}$ s'écrit

$$f(\mathbf{\Theta}, \mathbf{\Psi}|\mathbf{Y}) \propto f(\mathbf{Y}|\mathbf{\Theta})f(\mathbf{\Theta}|\mathbf{\Psi})f(\mathbf{\Psi}) \tag{10}$$

où $f(\mathbf{Y}|\mathbf{\Theta})$ et $f(\mathbf{\Psi})$ ont été respectivement définis dans (4) et (9). Sous l'hypothèse que les paramètres sont *a priori* indépendants, on obtient

$$f(\mathbf{\Theta}|\mathbf{\Psi}) = f(\mathbf{T})f(\mathbf{A}|\boldsymbol{w}, \alpha^2)f(\sigma^2|\nu, \gamma). \tag{11}$$

3 Echantillonneur de Gibbs

Cette section présente l'algorithme de Gibbs utilisé pour générer aléatoirement des échantillons asymptotiquement distribués suivant la loi *a posteriori* d'intérêt. Cette méthode MCMC est plus particulièrement détaillée dans [3, 6]. Elle consiste à simuler suivant les lois conditionnelles de $f\left(\mathbf{T}, \mathbf{A}, \sigma^2, \boldsymbol{w}|\mathbf{Y}\right)$. Plus précisément, les lois conditionnelles utilisées sont mesurées ci-dessous.

- **Echantillonnage suivant** $f(\mathbf{T}|\mathbf{A}, \sigma^2, \mathbf{Y})$
  Soit $\mathbf{T}_{\backslash k}$ la matrice $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_K]$ privée de sa $k^{\text{ème}}$ colonne. La loi *a posteriori* de $\mathbf{t}_k$ est une loi Gaussienne multivariée tronquée
  $$\mathbf{t}_k|\mathbf{T}_{\backslash k}, \mathbf{a}_k, \sigma^2, \mathbf{Y} \sim \mathcal{N}_{\mathcal{T}_k}(\boldsymbol{\tau}_k, \mathbf{\Gamma}_k) \tag{12}$$
  où
  $$\begin{cases} \mathbf{\Gamma}_k &= \left[\sum_{i=1}^{N} a_{i,k}^2 \mathbf{P}\mathbf{\Sigma}^{-1}\mathbf{P}^T + \frac{1}{s_k^2}\mathbf{I}_K\right]^{-1}, \\ \boldsymbol{\tau}_k &= \mathbf{\Gamma}_k\left[\sum_{i=1}^{N} a_{i,k}\mathbf{P}\mathbf{\Sigma}^{-1}\boldsymbol{\epsilon}_{i,k} + \frac{1}{s_k^2}\mathbf{e}_k\right], \\ \boldsymbol{\epsilon}_{i,k} &= \mathbf{y}_i - a_{i,k}\bar{\mathbf{y}} - \sum_{j\neq k} a_{i,k}\mathbf{m}_j. \end{cases} \tag{13}$$
- **Echantillonnage suivant** $f(\mathbf{A}|\boldsymbol{w}, \sigma^2, \mathbf{Y})$
  La loi *a posteriori* des scores $a_{i,k}$ est une loi Bernoulli-Gaussienne tronquée de paramètres $(\widetilde{w}_{i,k}, \mu_{i,k}, \eta_{i,k}^2, \mu_{i,k}^+)$, i.e.
  $$a_{i,k}|w_i, \sigma^2, \mathbf{a}_{i,\backslash k}, \mathbf{y}_i \sim (1-\widetilde{w}_{i,k})\delta(a_{i,k}) + \widetilde{w}_{i,k}\mathcal{N}_{]0,\mu_{i,k}^+[}(\mu_{i,k}, \eta_{i,k}^2) \tag{14}$$
  où $\mathbf{a}_{i,\backslash k}$ correspond au vecteur $\mathbf{a}_i$ privé de sa $k^{\text{ème}}$ composante et
  $$\begin{cases} \widetilde{w}_{i,k} &= \frac{u_{i,k}}{u_{i,k}+(1-w_i)}, \\ u_{i,k} &= w_i\frac{\eta_{i,k}}{\alpha}\exp\left(\frac{\mu_{i,k}^2}{2\eta_{i,k}^2}\right)\left[\Phi\left(\frac{\mu_{i,k}^+-\mu_{i,k}}{\eta_{i,k}}\right) - \Phi\left(\frac{-\mu_{i,k}}{\eta_{i,k}}\right)\right], \\ \eta_{i,k}^2 &= \left(\frac{\|\mathbf{m}_k\|^2}{\sigma^2} + \frac{1}{\alpha^2}\right)^{-1}, \\ \mu_{i,k} &= \eta_{i,k}^2\left(\frac{\mathbf{m}_k^T\boldsymbol{\epsilon}_{\backslash k}}{\sigma^2}\right), \\ \boldsymbol{\epsilon}_{\backslash k} &= \mathbf{y}_i - \sum_{j=1,j\neq k}^{K}\mathbf{m}_j a_i j. \end{cases} \tag{15}$$
  Notons que cette loi *a posteriori* Bernoulli-Gaussienne tronquée sur l'ensemble $]0, \mu_{i,k}^+[$ permet de respecter les contraintes d'additivité et de positivité (2).
- **Echantillonnage suivant** $f(\boldsymbol{w}|\mathbf{A})$
  La génération d'échantillons distribués suivant $f(\boldsymbol{w}|\mathbf{A})$ se fait selon la loi Beta suivante (pour $i = 1, \ldots, N$)
  $$w_i|\mathbf{a}_i \sim \mathcal{B}(1+n_{1,i}, 1+n_{0,i}), \tag{16}$$
  avec $n_{1,i} = \sharp\{k|a_{i,k} \neq 0\}$ et $n_{0,i} = K - n_{1,i}$.
- **Echantillonnage suivant** $f(\sigma^2|\mathbf{M}, \mathbf{A}, \mathbf{Y})$
  La loi *a posteriori* de $\sigma^2|\mathbf{M}, \mathbf{A}, \mathbf{Y}$ est une loi inverse-Gamma de paramètres $\frac{LN}{2}$ et $\frac{1}{2}\sum_{i=1}^{N}\|\mathbf{y}_i - \mathbf{M}\mathbf{a}_i\|^2$.

4 Résultats de simulation

4.1 Données synthétiques

Les données synthétiques générées sont constituées de $N = 128$ échantillons, chacun composé exactement de $R = 4$ facteurs, parmi une bibliothèque inconnue $\mathbf{M}$ de $K = 9$ signatures possibles, et $G = 256$ gènes. Les scores ont été générés aléatoirement suivant une distribution de Dirichlet $\mathcal{D}(1, \ldots, 1)$ et les échantillons sont bruités avec un rapport signal-à-bruit fixé à SNR $= 20$ dB. Les vecteurs moyennes $\mathbf{e}_k$ $(k = 1, \ldots, K)$ nécessaires pour évaluer la loi *a priori* des signatures génétiques sont choisis comme les projections des signatures identifiées par une analyse VCA [7]. Les échantillons générés suivant l'échantillonneur de Gibbs permettent d'approcher les estimateurs MMSE et MAP des scores $\mathbf{a}_i$ $(i = 1, \ldots, N)$ et des facteurs projetés $\mathbf{t}_k$ $(k = 1, \ldots, K)$. Les erreurs quadratiques moyennes MSE des scores et des signatures sont définies par $(r = 1, \ldots, R)$

$$\text{GMSE}_r^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{a}_{i,r} - a_{i,r}\right)^2,\ \text{MSE}_r^2 = \|\widehat{\mathbf{m}}_r - \mathbf{m}_r\|^2. \tag{17}$$

Les résultats de simulations sont reportés dans le Tableau 1 où la méthode BeG proposée est comparée avec la méthode non-paramétrique (NPBFA) [1], le modèle BFRM proposé par Carvalho *et al.* [2], un algorithme de factorisation en matrices non-négatives (NMF) [8] et l'analyse en composantes principales (ACP). Les méthodes ACP et NMF sont appliquées en fixant le nombre de facteurs à trouver à $R = 4$, alors que les autres méthodes permettent d'estimer ce nombre de facteurs. En revanche, puisque l'algorithme proposé prend en compte explicitement les contraintes d'additivité et de positivité des scores,

TAB. 1 – Comparaison des performances d'estimation entre les algorithmes NPBFA, BFRM, NMF, ACP et l'approche proposée.

		BeG	NPBFA	BFRM	NMF	PCA
MSE^2 $(\times 10^3)$	Facteur 1	**0.372**	1.352	1.827	3.438	2.186
	Facteur 2	**0.288**	1.558	1.434	3.303	0.364
	Facteur 3	**0.016**	1.237	1.946	4.281	0.413
	Facteur 4	**0.012**	2.645	N/A	6.580	0.381
$GMSE^2$	Facteur 1	6.955	40.013	198.735	19.709	**4.191**
	Facteur 2	8.065	35.282	183.638	34.913	**0.022**
	Facteur 3	**5.410**	23.687	191.699	17.254	7.069
	Facteur 4	14.293	26.766	N/A	18.802	**5.119**

il fournit une décomposition unique à une permutation de facteurs près, alors que les autres méthodes nécessitent une mise à l'échelle. Ces résultats illustrent la précision de la méthode BeG proposée.

4.2 Données réelles

L'algorithme proposé a également été évalué sur les données génétiques décrites dans [9], correspondant aux niveaux d'expression des gènes de $N = 108$ échantillons collectés sur six sujets, à cinq instants : 0, 1, 2, 4 et 12 heures après que les sujets aient bu une boisson particulière (alcool, jus de raisin, eau ou vin rouge). La méthode BeG est appliquée sur ces données avec $K = 9$ facteurs. La Figure 1 représente les 3 premières signatures biologiques, après avoir réorganisé les indices des $G = 22283$ gènes de manière à regrouper les facteurs entre eux. Plus précisément, le $k^{\text{ème}}$ pic de la figure correspond au gène le plus dominant du $k^{\text{ème}}$ facteur et les gènes compris entre ce pic et le $(k+1)^{\text{ème}}$ sont aussi dominants pour ce facteur mais à des degrés moins importants. Par exemple, la figure représentant la première signature biologique (haut) montre que le facteur correspondant est dominant pour les 10000 premiers gènes. Les scores sont représentés dans la Figure 2 sous forme d'images, dont les colonnes (resp. lignes) correspondent aux 6 sujets (resp. 5 instants sous 4 expériences différentes). On peut remarquer que certains facteurs sont clairement associés à des individus, par exemple, le facteur 5 (sujet 5), facteur 6 (sujet 2), facteur 7 (sujets 4, 6), et facteur 9 (sujets 1, 3).

5 Conclusions et perspectives

Ce papier présente un algorithme d'estimation Bayésienne pour l'analyse génétique. Une loi Bernoulli-Gaussienne tronquée est choisie comme loi *a priori* pour les scores afin d'assurer les contraintes de positivité et de somme à un. Un échantillonneur de Gibbs est proposé ici pour générer des échantillons distribués asymptotiquement suivant la loi *a posteriori* d'intérêt. Ces échantillons sont ensuite utilisés afin de déterminer des estimateurs MMSE et MAP. Des simulations ont été conduites sur données synthétiques et réelles afin d'illustrer le modèle Bayésien proposé.

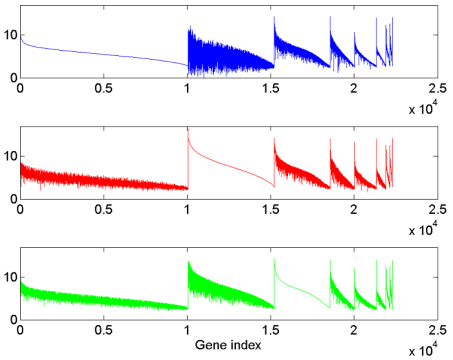

FIG. 1 – Exemples de signatures biologiques rangées par dominance décroissante.

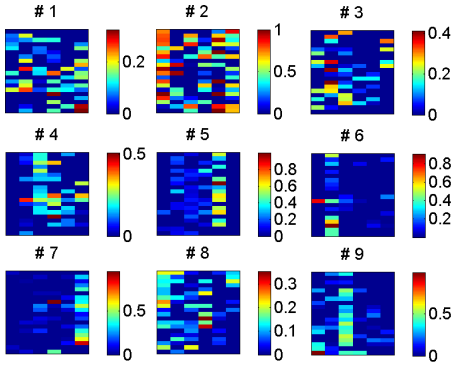

FIG. 2 – Facteur scores pour chacun des $K = 9$ facteurs.

Références

[1] B. Chen, M. Chen, J. Paisley, A. Zaas, C. Woods, G. S. Ginsburg, A. O. Hero, J. Lucas, D. Dunson, and L. Carin, "Bayesian inference of the number of factors in gene-expression analysis : application to human virus challenge studies," *BMC Bioinformatics*, vol. 11, no. 1, p. 552, 2010.

[2] C. M. Carvalho, J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West, "High-dimensional sparse factor modelling : Applications in gene expression genomics," *J. Amer. Stat. Assoc.*, vol. 103, no. 484, pp. 1438–1456, December 2008.

[3] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, and A. O. Hero, "Joint Bayesian endmember extraction and linear unmixing for hyperspectral imagery," *IEEE Trans. Signal Processing*, vol. 57, no. 11, pp. 4355–4368, Nov. 2009.

[4] C. P. Robert, "Simulation of truncated normal variables," *Statistics and Computing*, vol. 5, no. 2, pp. 121–125, June 1995.

[5] V. Mazet, D. Brie, and J. Idier, "Simuler une distribution normale à support positif à partir de plusieurs lois candidates," in *Actes 20^e coll. GRETSI*, vol. 2, Sept. 2005, pp. 1077–1080.

[6] N. Dobigeon, A. O. Hero, and J.-Y. Tourneret, "Hierarchical Bayesian sparse image reconstruction with application to MRFM," *IEEE Trans. Image Processing*, vol. 18, no. 9, pp. 2059–2070, Sept. 2009.

[7] J. M. Nascimento and J. M. Bioucas-Dias, "Vertex component analysis : A fast algorithm to unmix hyperspectral data," *IEEE Trans. Geosci. and Remote Sensing*, vol. 43, no. 4, pp. 898–910, April 2005.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Proc. of Neural Info. Process. Syst.*, 2000.

[9] F. Baty, M. Facompre, J. Wiegand, J. Schwager, and M. Brutsche, "Analysis with respect to instrumental variables for the exploration of microarray data structures," *BMC Bioinformatics*, vol. 7, no. 1, p. 422, 2006.