Séparation de Sources pour l'Analyse de Données d'Expression

Pierre Chiappetta, Marie-Christine Roubaud, Bruno Torrésani

Séparation de Sources pour l'Analyse de Données d'Expression

Blind Source Separation for Analyzing Expression Data

Pierre CHIAPPETTA[†,‡] Marie-Christine ROUBAUD[‡] Bruno TORRÉSANI[‡]

[†] Centre de Physique Théorique, CNRS-Luminy, case 907, 13288 Marseille Cedex 9
[‡] Laboratoire d'Analyse, Topologie et Probabilités, Centre de Mathématiques et Informatique, 31 rue Joliot-Curie, 13013 Marseille Cedex 13

Courriel : chiapeta@cpt.univ-mrs.fr, {mcroubau,Bruno.Torresani}@cmi.univ-mrs.fr

Résumé

Nous décrivons une nouvelle approche pour l'analyse de données issues du transcriptome, basée sur les techniques de séparation de sources en aveugle. Cette approche fournit en sortie des "profils d'expression élémentaires", ou "sources", éventuellement interprétables comme de possibles voies de régulation. Une analyse plus fine des sources ainsi obtenues montre qu'elles sont généralement caractérisées par une sur-expression (resp. sous-expression) significative d'un petit nombre de gènes, parfois accompagnée d'une sous-expression (resp. sur-expression) d'une famille complémentaire de gènes. Les résultats obtenus sur deux jeux de données d'expression montrent que certaines des familles ainsi détectées correspondent à des familles connues de gènes co-régulés, ce qui valide l'approche.

Mots clés : *données d'expression, séparation de source, analyse en composantes indépendantes, gènes co-régulés.*

Abstract

We present a new approach for analyzing gene expression data, based upon blind source separation techniques. This approach yields "elementary expression patterns", or "sources", which may be interpreted as potential regulation channels. Further analysis of the so-obtained sources show that they are generally characterized by a small numbers of specific co-regulated genes. The results obtained on two datasets show that some of the obtained gene families correspond to well known families of co-regulated genes, which validates our approach.

Keywords: *expression data, source separation, independent component analysis, co-regulated genes.*

1 Introduction

Les expériences de transcriptome fournissent des volumes considérables de données, et soulèvent de nombreux problèmes de nature statistique, que ce soit au niveau de l'acquisition même des données (problèmes de traitement d'images, de quantification, de normalisation) ou au niveau de leur confrontation à des modèles. En particulier, les données d'expression contiennent potentiellement de nombreuses informations sur les interactions gèniques, des voies ou réseaux de régulation,... et c'est sur ce point que se focalise la présente contribution. De nombreuses approches ont été proposées pour inférer de telles interactions à partir de données d'expression (sous les noms de modèles booléens, "reverse engineering", réseaux logiques ou Bayesiens, machines à support vecteur...).

Nous proposons ici une nouvelle approche pour le problème d'analyse de co-régulations de gènes, basée sur l'analyse en composantes indépendantes. Il s'agit essentiellement d'une technique d'identification de modèle linéaire (sur les logarithmes de données d'expression) basée sur les statistiques d'ordres supérieurs. De telles approches ont connu un grand succès ces dernières années dans un contexte de traitement du signal, mais n'ont pas à notre connaissance été exploitées pour l'analyse de données d'expression.

L'approche que nous proposons fournit des candidats pour de possibles voies de régulation. Ces derniers sont souvent caractérisés par l'existence d'une petite famille de gènes qui y sont sur-exprimés ou sous-exprimés. Nous montrons sur un jeu de données de cancer du sein que l'approche proposée est pertinente, dans la mesure où elle met facilement en évidence des groupes de gènes co-régulés (et pas nécessairement fortement corrélés -voir plus bas pour plus de détails) positivement ou négativement, qui ont un sens biologique reconnu.

2 Position du problème, mise en forme des données, réduction de dimension

On considère des données sous la forme usuelle d'un tableau à double entrée $\mathbf{X} = \{X_g^c,\ g = 1, \dots N_g, c = 1, \dots N_c\}$, où l'indice g numérote les gènes et l'exposant c les conditions. On notera $\mathbf{X}^c$ et $\mathbf{X}_g$ les vecteurs (respectivement colonne et ligne) correspondant respectivement aux conditions et aux gènes. Les valeurs typiques de N_g et N_c sont de l'ordre de quelques milliers et quelques dizaines respectivement.

2.1 Transformation non-linéaire

Les données brutes (même normalisées) ne sont pas nécessairement les bonnes variables à analyser : celles qui ont le plus de sens biologique, ou celles qui se prêtent le mieux à une méthode statistique donnée. On peut changer de variables par une transformation non-linéaire ($\log(x)$, $\log(1 + \alpha x)$, $\sqrt{x}$...). Ces transformations non-linéaires ont pour effet immédiat de réduire le domaine des valeurs prises par les niveaux d'expression (tout du moins en ce qui concerne les grandes valeurs), ce qui est important dans certains cas, la plupart des méthodes statistiques étant mises en difficulté par des valeurs extrêmes. Nous optons ici pour une transformation logarithmique

$$\mathbf{Y} = \log \mathbf{X}\ , \tag{1}$$

qui non seulement diminue l'importance des grandes valeurs, mais aussi donne une plus grande résolution pour les faibles valeurs. Un autre effet essentiel du logarithme est qu'il transforme des produits en sommes, donc d'éventuels effets multiplicatifs en effets linéaires, plus faciles à identifier par des modèles classiques.

2.2 Traitement des valeurs nulles ou faibles

Une telle correction logarithmique pose toutefois le problème des valeurs nulles (le logarithme étant singulier à l'origine). Remarquons qu'un x_g^c nul, ne signifie généralement pas que le gène g ne s'est pas exprimé dans la condition c, mais plutôt que le niveau mesuré est inférieur à un certain seuil de détection (ou de précision).
Il importe donc de remplacer les valeurs nulles (et éventuellement les valeurs très faibles) par des valeurs non nulles, mais suffisamment petites pour préserver l'information qualitative "la valeur mesurée est inférieure à un certain seuil". Plutôt que d'utiliser une valeur constante, nous remplaçons les valeurs nulles par des valeurs aléatoires, dont le domaine de variation est adapté condition par condition (en utilisant la plus petite valeur mesurée non nulle). La distribution des valeurs aléatoires choisie ne semble pas affecter lourdement les résultats obtenus.
Plus précisément, la procédure suivante est employée : pour chaque condition $c = 1, 2, \dots N_c$

1. Sélection de la plus faible valeur x_{min}^c non nulle de $\mathbf{X}^c$.
2. Remplacement de toutes les valeurs nulles [1] X_g^c de $\mathbf{X}^c$ par des valeurs $\tilde{X}_g^c$ pseudo-aléatoires uniformément distribuées entre 0 et αx_{min} où α est une constante (typiquement entre 1 et 3) qui peut être adaptée au protocole expérimental (seuils, normalisation,...). Pour les valeurs $X_g^c \neq 0$, on pose $\tilde{X}_g^c = X_g^c$.

On considère alors les logarithmes modifiés des niveaux d'expression mesurés :

$$Y_g^c = \log \tilde{X}_g^c\ .$$

Un exemple de distribution de logarithmes (modifiés) de niveaux d'expression est montré en FIG. 1. Comme on

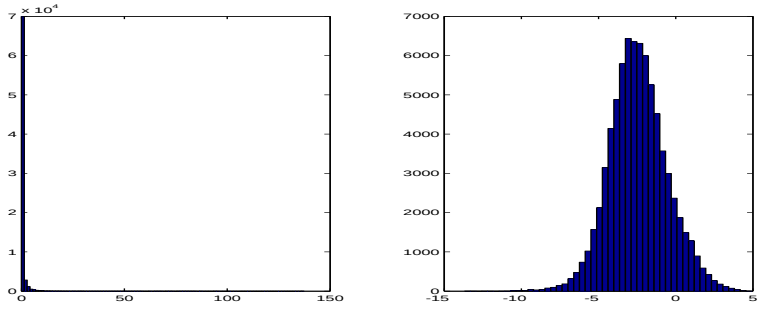

FIG. 1: Histogrammes des niveaux d'expression (à gauche) et de leurs logarithmes corrigés (à droite).

peut le voir, alors que les valeurs originales (histogramme de gauche) sont extrêmement concentrées à l'origine, mais présentent également un certain nombre de grandes valeurs, la distribution des logarithmes est plus étalée [2].

1. Une altenative consiste à parturber toutes les valeurs mesurées en leur ajoutant de petites valeurs aléatoires ; ceci permet de tester la stabilité des résultats obtenus en fin d'analyse.

2. Par comparaison, l'histogramme des logarithmes corrigés en ajoutant une petite constante présenterait un "pic" pour une valeur égale au logarithme de la constante en question, ce qui peut s'avérer problèmatique lorsque le nombre de valeurs nulles est élevé.

2.3 Réduction de dimension

Lorsque le nombre de conditions excède une ou deux dizaines, il devient difficile d'exploiter l'information complète. La dimension "utile" des données peut être réduite en utilisant l'analyse en composantes principales : la diagonalisation de la matrice de covariance "condition-condition" produit généralement un certain nombre de directions dans l'espace N_c-dimensionnel dans lesquelles la variance est négligeable. Ces directions peuvent donc être supprimées par projection des données dans l'espace complémentaire. Il en résulte une diminution de la dimension des données, qui simplifie les calculs numériques ultérieurs.

3 Analyse de données d'expression par séparation de sources

L'analyse en composantes principales (ACP) est un outil d'usage courant pour réduire la dimension d'un ensemble de données, basé sur l'analyse de la matrice des corrélations (ou des covariances), c'est à dire des statistiques d'ordre deux [7]. Lorsque les données sont distribuées suivant une loi (jointe) normale, la matrice des corrélations contient suffisamment d'information pour caractériser la distribution des données. Cependant, tel n'est pas toujours le cas, et les statistiques d'ordre supérieur contiennent souvent une information particulièrement pertinente. Les méthodes de séparation de sources font partie des approches exploitant ces aspects.

3.1 Séparation de sources en aveugle

Le problème de séparation de sources peut être formulé de la façon suivante : étant donnés N "signaux" $\mathbf{Y}^n, n = 1, \dots N$ (en l'occurence, des vecteurs de taille L), mesurés par N "capteurs", on fait l'hypothèse qu'ils sont obtenus par "mélange" de M sources indépendantes inconnues. On représente ceci via le modèle linéaire

$$\mathbf{Y}^n = \sum_{m=1}^{M} A_m^n \mathbf{s}^m \ , \tag{2}$$

où les vecteurs $\mathbf{s}^1, \dots \mathbf{s}^M$, appelés "sources", sont indépendants, dans le sens suivant : les coefficients du développement sur cette base sont statistiquement indépendants.
La séparation de source revient à résoudre le problème inverse, c'est à dire identifier les sources indépendantes $\mathbf{s}^m$ ou, ce qui revient au même, les coefficients de la décomposition A_m^n, à partir des N signaux. Ces coefficients forment une matrice, appelée *matrice de mélange*. Il s'agit d'une problématique classique de traitement du signal. On peut par exemple se référer à [1] ou encore à [4] pour une introduction didactique.
La séparation de source (ou analyse en composantes indépendantes) peut être effectuée suivant diverses approches. L'approche la plus classique consiste toutefois à optimiser un critère d'indépendance : par exemple, on peut définir les sources indépendantes comme les directions $\mathbf{s}^m$ dans l'espace N-dimensionnel qui minimisent l'information mutuelle entre composantes. Cependant, des méthodes basées sur ce type d'approches souffrent parfois d'un relatif manque de robustesse, et on se contente alors d'approximations. C'est l'approche que nous avons suivie ici, basée sur la recherche de directions dans lesquelles la distribution des données s'écarte le plus d'une loi Gaussienne.
Il est facile de voir que la séparation de sources indépendantes est impossible lorsque la distribution (jointe) des données est Gaussienne. Plus encore, il est possible de montrer que la minimisation de l'information mutuelle peut se ramener à la minimisation de l'entropie dans chacune des directions, interprétable comme la recherche de directions "minimalement Gaussiennes". Ainsi, on utilise souvent en pratique [1, 4] des critères de la forme suivante : en notant V la projection des données sur un axe, on cherche à résoudre

$$\max \left(\mathbb{E}\left\{\Phi(V)\right\} - \mathbb{E}\left\{\Phi(\gamma)\right\} \right)^2 \ , \tag{3}$$

où le maximum est pris sur toutes les directions de l'espace considéré. Ici, γ est une variable aléatoire Gaussienne centrée réduite, Φ est une fonction (non quadratique), et $\mathbb{E}\left\{X\right\}$ représente l'espérance d'une variable aléatoire X. Les axes ainsi obtenus fournissent les directions indépendantes estimées.
Remarque : Il faut noter qu'à la différence de l'ACP (qui fournit également des décompositions telles que (2)), l'analyse en composantes indépendantes fournit une base $\{\mathbf{s}^1, \dots \mathbf{s}^M\}$ qui n'est pas nécessairement orthogonale.
Remarque : Il est aussi important de signaler que les sources sont obtenues à un facteur multiplicatif près (qui peut être éliminé par une normalisation appropriée), et à un signe près. De plus, elles sont obtenues dans un ordre arbitraire.

3.2 Application aux données d'expression

Nous cherchons à interpréter les profils d'expression à l'aide d'un modèle linéaire

$$\mathbf{Y}^c = \sum_{m=1}^{M} A_m^c \mathbf{s}^m \tag{4}$$

où les Y_g^c sont les logarithmes des données, et les sources $\mathbf{s}^m$ sont "maximalement indépendantes". Ces sources représentent des "pseudo profils d'expression", et fournissent des candidats pour des mécanismes de régulation indépendants. L'approche décrite plus haut fournit une estimation pour ces sources, ainsi que pour la matrice de mélange $\{A_m^c, c = 1, \dots N_c, m = 1, \dots M\}$.
Les algorithmes numériques que nous utilisons sont basés sur le logiciel libre `FastICA` [4] (utilisant le langage de programmation `Matlab` ; signalons également l'existence d'une implémentation utilisant le logiciel libre de statistiques `R` [6]). La recherche de sources indépendantes est précédée d'une réduction de dimension, et d'une normalisation des conditions dans les dimensions restantes (afin que l'algorithme se focalise sur les statistiques d'ordre supérieur). Plusieurs critères d'écart à la Gaussianité ont été testés, et fournissent des résultats concordants. Les résultats décrits dans la section 4 utilisent la fonction $\Phi(x) = x^4$.
Le nombre de sources indépendantes (inférieur par construction au nombre de conditions considérées) est un paramètre libre de la méthode, et doit être choisi à l'avance. On observe que les composantes indépendantes obtenues, et en particulier les gènes qui les caractérisent (voir ci dessous) sont assez stables lorsque l'on fait varier le nombre de sources. L'algorithme d'optimisation du critère (3) est basé sur une méthode de type "descente", et est donc susceptible de converger vers un minimum local. Par conséquent, une "inspection a posteriori" des résultats obtenus est nécessaire.

Analyse de la matrice de mélange : Pour chacune des sources m obtenues le vecteur $\{A_m^c, c = 1, \dots N_c\}$ fournit les coefficients des projections des différentes conditions sur la source m (en d'autres termes, le "poids" de la source m dans la condition c). La distribution des coefficients A_m^c peut alors être particulièrement instructive. On peut notamment observer dans certains cas une distribution très nettement bimodale, signe que la source considérée a un pouvoir discriminant significatif sur les conditions. Un exemple spectaculaire est donné ci-dessous dans le cas de données de cancer du sein (voir FIG. 2).
Analyse des sources : Il s'avère que chacune des sources estimées se caractérise par un (petit) nombre de gènes significativement sur-exprimés ou sous-exprimés (par rapport au comportement d'ensemble des gènes). Les ensembles de gènes qui se signalent ainsi ont souvent une cohérence biologique claire, ce qui en fait des candidats pour de possibles voies de régulation.
Il faut également signaler que la méthode permet de mettre en évidence simultanément des ensembles de gènes co-régulés, et anti-régulés (contrairement à ce que ferait naturellement un algorithme de classification, qui aurait tendance à regroupes des profils d'expression semblables). Là encore, un exemple est présenté ci-dessous (voir FIG. 3).

4 Applications à des données d'expression.

La méthode a été appliquée à deux jeux de données d'expression à savoir des données de cancer du sein [5] et de Bacillus subtilis [8]. Nous discutons ici uniquement des résultats obtenus à partir du premier jeu en référant à [2] pour une présentation plus complète.

Données de cancer du sein. Nous cherchons à interpréter les profils d'expression obtenus dans une expérience de transcriptome effectuée par l'équipe TAGC du CIML de Marseille [5]. Les ADNc sont disposés sur une membrane de nylon et hybridés avec des ARN extraits de tumeurs du cancer du sein marqués radioactivement (provenant de l'Institut Paoli-Calmettes de Marseille). On dispose de 1045 gènes, de 12 lignées cellulaires et de 55 tumeurs. La transformation logarithmique modifiée présentée dans les Sections 2.1 et 2.2 a été effectuée (elle était nécessaire dans ce cas compte tenu du grand nombre de valeurs nulles présentes dans les données).
L'analyse en composantes indépendantes a été est réalisée sur l'ensemble complet de données d'une part, et sur les tumeurs seules dans un second temps. L'analyse sur l'ensemble complet de données fournit un nombre significatif de sources permettant de discriminer nettement les tumeurs des lignées cellulaires.
Une analyse plus fine montre que pour un certain nombre de sources, les niveaux d'expression correspondants prennent (après normalisation) des valeurs "raisonnables" (généralement compatibles avec une distribution normale centrée réduite), à l'exception d'un certain nombre d'entre eux, manifestement sur-exprimés ou sous-exprimés. Comme nous l'avons signalé plus haut, le signe global affecté à une source est dépourvu de signification ; par contre, on observe souvent, pour un groupe de gènes sur-exprimés dans une source donnée, un groupe complémentaire de gènes significativement sous-exprimés, ce qui est interprétable comme une co-régulation négative : pour la source en question, une sur-expression significative du premier groupe de gènes va de pair avec une sous-expression du second.
C'est le cas de la composante indépendante caractérisée par l'`ESR1` (récepteur estrogène). Celui-ci est co-régulé positivement avec les gènes `GATA3, XBP1, THBS1, MYB, CRABP2, KRT19, ILF1` et `PLAT` et co-régulé négativement avec notamment `GSPT1` et `WNT6`. Le premier groupe (`ESR1, GATA3, XBP1`...) corrobore les résultats précédemment obtenus sur le même jeu de données [5], ainsi que d'autres résultats publiés (voir par exemple [3]). Il est à noter que les relations entre logarithmes de niveaux d'expression sont proches de relations

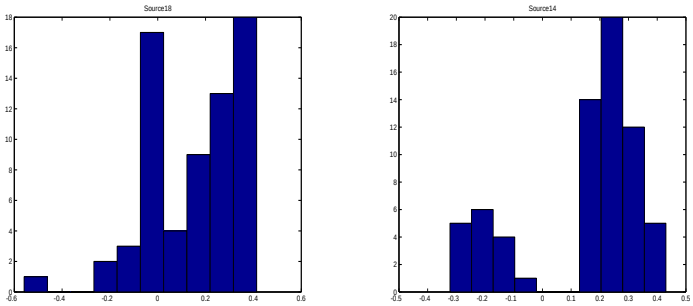

FIG. 2: Coefficients de la matrice de mélange correspondant à deux sources estimées : la composante GATA3−GSTP1 (à gauche) et la composante CIDE A (à droite).

linéaires (par exemple, les coefficients de corrélation entre ESR1, GATA3, XBP1 et MYB sont compris entre $r = 0.6$ et 0.85). Par contre, la mise en évidence de la co-régulation négative avec le groupe GSTP1−WNT6 sur ce jeu de données semble être un résultat nouveau. Cette relation spectaculaire est en fait presque linéaire, comme on peut le voir sur la FIG. 3, où sont représentés les profils d'expression comparés de GSTP1 et GATA3 (à gauche) et de WNT6 et XBP1 (à droite). Les coefficients de corrélation correspondants sont compris entre $r = -0.5$ et $r = -0.6$.
La distribution des coefficients de la matrice de mélange (voir FIG. 2, figure de gauche) présente un aspect assez nettement bimodal. Il s'avère de plus que cette distribution se corrèle de façon quasi-parfaite aux données cliniques de taux de récepteur estrogène présent dans les tumeurs. En fixant un seuil s aux alentours de $s \approx 0.14$, on observe que parmi les 24 patients c tels que $A^c_{ESR} < s$, 23 se trouvent dans la classe ER-, un seul est ER+. Parallèlement, parmi les 31 patients c tels que $A^c_{ESR} > s$, 30 sont ER+ et un seul ER-.

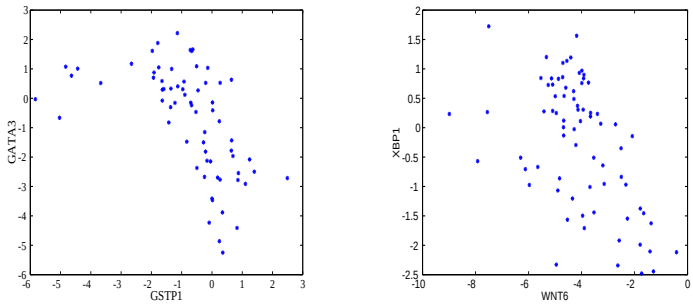

FIG. 3: Comparaison de gènes anti-corrélés : GATA3−GSTP1 (à gauche) et WNT6−XBP1 (à droite).

Une seconde composante se caractérise par une distribution des coefficients de matrice de mélange très nettement bimodale (voir la FIG. 2) : il s'agit d'une source indépendante impliquant un certain nombre de gènes, notamment le récepteur prolactine (PRLR), ainsi que CIDEA, CDH15, CDKN3, MLANA, CNTFR, IGFBP1, BCL2 (transcrit alternatif)... qui ont déjà été identifiés [5] pour leur rôle dans le cancer du sein (notamment impliqués la résistance à la chimiothérapie) : une surexpression ce ce groupe de gènes est associée à un pronostic défavorable. Une analyse comparée de l'expression de ces gènes permet de retrouver le comportement d'ensemble détecté sur la Fig. 2, à savoir une partition claire des conditions en deux groupes. Les valeurs des coefficients de corrélation sont sensiblement plus faibles que dans le cas précédent, ce qui indique une relation pas nécessairement linéaire (comme on peut le voir clairement sur la FIG. 4, figure de droite).
On peut noter que ce groupe de gènes ne semble pas s'accompagner d'un groupe significatif de gènes anti-régulés. C'est également le cas d'une source impliquant un grand nombre de gènes liés à la fonction immunitaire (notamment IGHM, IGHA1, IGKV1D, GATA1, GATA2, GATA4, GATA6, IL2RG, CSF1, NFYB, SUI1, RELA, ...). Cette source sépare nettement les lignées cellulaires des tumeurs. La distribution des coefficients de mélange correspondants $\{A^c_m, c = 1, \ldots N_c\}$ est proche d'une distribution uniforme (voir [2] pour plus de détails).

5 Conclusions

Nous avons décrit les grandes lignes d'une méthode d'analyse exploratoire de données d'expression basée sur les techniques de séparation de sources en aveugle, que nous appliquons aux logarithmes (corrigés) de profils d'expression. Cette méthode fournit en sortie des "profils d'expression élémentaires", éventuellement interprétables comme de possibles voies de régulation. Ces dernières sont généralement catactérisées par une sur-expression

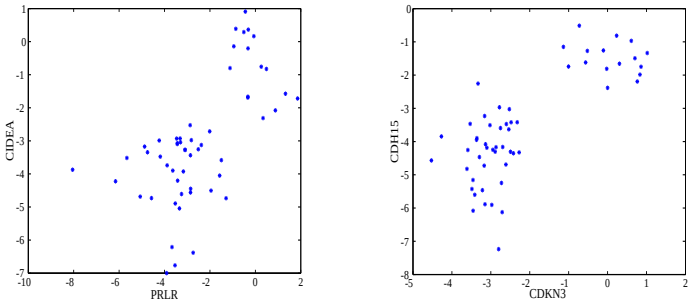

FIG. 4: Gènes de la classe `PRLR, CIDEA, ...` : `PRLR-CIDEA` (à gauche) et `CDH15-CDKN3` (à droite).

(resp. sous-expression) significative d'un petit nombre de gènes, parfois accompagnée d'une sous-expression (resp. sur-expression) d'une famille complémentaire.
Bien que l'approche proposée n'ait pas pour objectif premier la classification de gènes, les groupes de gènes caractérisant les sources indépendantes ont généralement une cohérence biologique ; la possibilité de mettre en évidence de groupes complémentaires de gènes anti-régulés semble être un plus significatif.
Rappelons que l'identification de sources par ACI est précédée d'une "standardisation" des données, qui réduit la matrice de variance-covariance à l'unité, et permet donc de focaliser l'analyse sur le rôle des moments d'ordre supérieur. En tant que telle, elle est complémentaire des techniques classiques d'analyse en composantes principales (ACP) couramment utilisées, qui se basent sur l'étude des moments d'ordre deux.

Signalons toutefois que cette approche est encore loin d'une procédure systématique, dans la mesure où la méthode repose sur un certain nombre de paramètres dont l'optimisation n'est pas nécessairement aisée. Par exemple :

- Le critère de non-Gaussianité : il est possible de montrer que pour une distribution(non-gaussienne) donnée, il existe un choix optimal, qui minimise la variance de l'estimateur. Les distributions des sources étant généralement inconnues, on se contente de choix "passe-partout".
- Le nombre de sources indépendantes recherchées est lui aussi inconnu, et doit être spécifié à l'avance.
- L'algorithme d'optimisation utilisé est basé sur des méthodes de type "méthode de gradient", et converge généralement vers des optima locaux. Ce point mérite d'être approfondi.

Néanmoins, les résultats obtenus (notamment sur les données de cancer du sein, décrits dans cette contribution) sont remarquablement stables et semblent ne dépendre que faiblement de ces paramètres.

Références

[1] J.F. Cardoso (1998) : Blind signal separation : statistical principles. *Proc. of the IEEE* **9** :10, pp. 2009-2025.

[2] P. Chiappetta, M.C. Roubaud et B. Torrésani (2002) : Blind source separation for analyzing microarray data, En préparation.

[3] S. Gruvberger, M. Ringnér, Y. Chen, S. Panavally, L.H. Saal, A. Borg, M. Fernö, C. Peterson and P. Meltzer (2001) : Estrogen Receptor Status in Breast Cancer Is Associated with Remarkably Distinct Gene Expression Patterns. *Cancer Research* **61**, pp 5979-5984.

[4] A. Hyvärinen, E. Oja (2000) : Independent Component Analysis : Algorithms and Applications. *Neural Networks*, **13**, pp. 411-430, Voir aussi "Independent Component Analysis : a tutorial", disponible sur le site `http ://www.cis.hut.fi/projects/ica`

[5] R. Houlgatte et al (2001) : Unpublished results.
Voir aussi F. Bertucci, R. Houlgatte et al (2000) : Gene expression profiling of primary breast carcinomas using arrays of candidate genes. *Human Molecular Genetics* **9** :20, pp. 2981-2991.

[6] J.L. Marchini et C. Heaton (2001) : An `R` and `C` implementation of the `FastICA` package. Disponible sur le site `http ://cran.r-project.org/`.

[7] G. Saporta (1990) : *Probabilité, Analyse de Données et Statistique*, Editions Technip, Paris.

[8] A. Sekowska, S. Robin, J. J. Daudin, A. Hénaut, A. Danchin (2001) : Extracting biologival information from DNA arrays : an unexpected link between arginine and methionine metabolism in Bacillus subtilis. *GenomeBiology* **2**, pp. 19.1-19.12.