# Analyse en Composantes Indépendantes et Réseaux Bayésiens

Francis R. BACH[1], Michael I. JORDAN[2],

[1]Computer Science Division
University of California
Berkeley, CA 94720, USA

[2] Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

fbach@cs.berkeley.edu,jordan@cs.berkeley.edu

**Résumé –** Une généralisation de l'analyse en composantes indépendantes (ACI) est introduite: au lieu de déterminer une application linéaire qui rend les composantes indépendantes, nous cherchons une application linéaire qui rend les composantes modélisables par un réseau Bayésien dont la structure est un arbre. Ce nouveau modèle, que nous dénommons TCA (Tree-dependent Component Analysis), permet de relaxer l'hypothèse d'indépendance et d'adapter la structure de dépendance aux observations. En particulier, lorsque l'arbre a plusieurs composantes connexes, notre méthode permet de trouver des « amas » de composantes de telle sorte que les composantes à l'intérieur d'un amas sont dépendantes les unes des autres, mais indépendantes de composantes dans d'autres amas. Notre approche s'applique aussi bien à des sources intemporelles et non Gaussienne qu'à des sources stationaires.

**Abstract –** We present a generalization of independent component analysis (ICA), where instead of looking for a linear transform that makes the data components independent, we look for a transform that makes the data components well fit by a tree-structured graphical model. This *tree-dependent component analysis (TCA)* provides a tractable and flexible approach to weakening the assumption of independence in ICA. In particular, TCA allows the underlying graph to have multiple connected components, and thus the method is able to find "clusters" of components such that components are dependent within a cluster and independent between clusters. Our framework applies equally well for temporally independent non-Gaussian sources and for stationary Gaussian sources.

## 1 Introduction

L'analyse en composantes indépendantes (ACI) est une technique statistique dont l'objectif est de décomposer un signal aléatoire multivarié $x \in \mathbb{R}^m$ en une combinaison linéaire de signaux indépendants, i.e. $x = As$, où $s$ est un signal à composantes indépendantes et $A$ une matrice à coefficients réels. Cette technique est couramment appliquée à des problèmes où les sources peuvent être supposées indépendantes, comme la séparation aveugle de sources sonores ou l'imagerie médicale [16, 10, 8, 14]. Dans le cas où il y a autant de sources $s$ que d'observations $x$, le problème de l'ACI peut être reformulé de la façon suivante : le but est de trouver une matrice carrée $W$ (qui correspond à $A^{-1}$) telle que le vecteur obtenu par action de $W$ sur les observations $x$ aient des composantes les plus indépendantes possible [8].

Dans cet article, nous généralisons cette technique en permettant aux sources d'être dépendantes les unes des autres : nous cherchons une matrice $W$ telle que les composantes de $s = Wx$ peuvent être modélisées par un réseau Bayésien dont la structure est un arbre. La topologie $T$ de l'arbre n'est pas déterminée à l'avance ; en effet, notre algorithme permet de trouver à la fois la matrice $W$ et l'arbre $T$. De plus, lorsque l'arbre comprend plusieurs composantes connexes, le réseau Bayésien permet de modéliser le phénomène d'amas (clusters) de composantes, i.e. les composantes à l'intérieur d'un amas sont dépendantes les unes des autres, mais indépendantes des composantes appartenant à d'autres amas (ce modèle est aussi appelé ACI multidimensionnelle [12, 7]). Contrairement aux méthodes déjà proposées pour ce modèle d'amas [7, 13], notre méthode permet de déterminer de manière automatique le nombre d'amas et leur dimension. Nous dénommons ce nouveau modèle TCA (Tree-dependent Component Analysis).

Tout en conservant un modèle flexible et traitable, nous pouvons relaxer l'hypothèse d'indépendance et notre algorithme peut s'appliquer à des situations plus générales que l'ACI, en particulier dans des situations dans lesquelles les sources ne sont pas toutes indépendantes, par exemple en imagerie médicale (électrocardiogramme d'une mère et de son fœtus [7]), ou pour la séparation de signaux musicaux, où certains instruments sont dépendants les uns des autres.

Nous proposons deux cadres d'applications de nos techniques, le premier pour des signaux temporellement indépendants et non Gaussiens, le deuxième, en vue d'application en traitement du signal audio, pour les processus stationnaires Gaussiens.

## 2 Réseaux Bayésiens

Les réseaux Bayésiens (Bayesian networks, graphical models) forment un ensemble de modèles probabilistes pour de larges collections de variables aléatoires où une représentation

parcimonieuse est nécessaire, à la fois pour des raisons numériques (afin d'éviter la manipulation de tableaux trop larges) et statistiques (afin de limiter le nombre de paramètres à estimer). Ils sont couramment utilisés en intelligence artificielle et en apprentissage automatique (machine learning) [18, 15].

## 2.1 Arbres

Soient $s = (s_1, \ldots, s_m)$, $m$ variables aléatoires réelles avec densité $p(s)$[1]. Un réseau Bayésien permet de définir une densité conjointe à partir de densités marginales ou conditionelles locales. L'implication des variables dans ces lois locales est représentée à l'aide d'un graphe : deux variables sont reliées par une arête si et seulement si elles apparaissent dans une même loi locale. Dans le cas d'un arbre $T$, il existe une formule fermée pour exprimer la loi conjointe en fonction des lois marginales (en particulier, il n'y pas de constante de normalisation) :

$$p(s) = \prod_{(u,v)\in T} \frac{p(s_u, s_v)}{p(s_u)p(s_v)} \prod_{u=1}^{m} p(s_u) \qquad (1)$$

($p(s_u)$, $p(s_u, s_v)$ sont les densités marginales). On dit qu'une distribution se factorise dans le graphe $T$, si et seulement si elle vérifie l'équation (1). Une distribution qui se factorise dans un arbre trivial (i.e. sans arêtes) a des composantes indépendantes et la loi conjointe est simplement le produit des lois marginales (ce qui implique que l'ACI est un sous-modèle de TCA). Tout au long de l'article, nous montrerons comment les principales technique couramment appliquées à l'ACI se généralisent à TCA.

Il existe une caractérisation des réseaux Bayésiens à l'aide de relations d'indépendance conditionnelle. Dans le cas des arbres, une loi de probabilité se factorise dans $T$, si et seulement si, deux variables appartenant à deux composantes connexes différentes sont indépendantes et deux variables appartenant à la même composante sont conditionnellement indépendantes sachant toutes les autres variables de cette composante.

## 2.2 Forêts et amas

Un arbre non couvrant est un arbre qui comprend plus d'une composante connexe, et est communément appelé une « forêt », comme montré en Figure 1. Nous proposons de modéliser le phénomène d'amas par une forêt. Soient $C_1, \ldots, C_k$ les $k$ composantes connexes de la forêt $T$. Chacune de ces composantes connexes $C_1, \ldots, C_k$ modélise un amas. L'indépendance entre les amas est modélisée exactement, alors que la dépendance au sein d'un amas est modélisée de manière approchée par une distribution qui se factorise dans un arbre couvrant (i.e. avec une seule composante connexe). Ces distributions sont couramment utilisées en traitement du signal [21] et permettent de modéliser une large classe de signaux multivariés tout en bénéficiant d'algorithmes d'inférence et d'estimation de complexité linéaire par rapport au nombre de variables.

[1]Les réseaux Bayésiens peuvent être aussi bien définis pour des variables discrètes, en remplaçant la densité par la loi de probabilité.

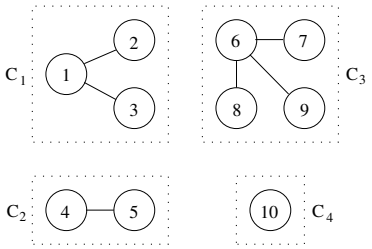


FIG. 1 – Une forêt avec 10 variables et 4 amas.

# 3 Vraisemblance semi-paramétrique

Dans le reste de l'article, nous utiliserons les notations suivantes : $x$ est une variable aléatoire à valeurs dans $\mathbb{R}^m$ ; $p(x_u, x_v)$ et $p(x_u)$ sont les lois marginales. La divergence de Kullback-Leibler (KL) entre deux distributions $p(x)$ et $q(x)$ est définie par $D(p \,||\, q) = E_{p(x)} \log \frac{p(x)}{q(x)}$. Nous utiliserons aussi les informations mutuelles entre les deux variables $x_u$ et $x_v$, définies par

$$I(x_u, x_v) = D(p(x_u, x_v) \,||\, p(x_u)p(x_v))$$

ainsi qu'entre les $m$ variables $x_1, \ldots, x_m$, définies par

$$I(x_1, \ldots, x_m) = D(p(x) \,||\, p(x_1) \cdots p(x_m)).$$

Ces informations mutuelles peuvent s'exprimer en termes d'entropie, i.e. $I(x_u, x_v) = H(x_u) + H(x_v) - H(x_u, x_v)$.

## 3.1 Contraste pour TCA

Dans une approche semi-paramétrique de l'ACI, où le paramètre d'intérêt est la matrice de démixage $W$ et les densités des sources sont considérées comme des paramètres de nuisance, maximiser la vraisemblance est équivalent à minimiser l'information mutuelle $I(s_1, \ldots, s_m)$ entre les estimations $s = Wx$ des sources [8]. Nous pouvons généraliser ce résultat au contexte de TCA, les paramètres d'intérêt étant maintenant la matrice $W$ et l'arbre $T$ :

**Théorème 1** *Etant donnée une variable aléatoire $x$ à valeurs dans $\mathbb{R}^m$, minimiser la vraisemblance semi-paramétrique du modèle TCA est équivalent à minimiser l'expression suivante par rapport à la matrice de démixage $W$ et l'arbre $T$ ($s = Wx$ est l'estimation des $m$ sources) :*

$$J(W, T) = I(s_1, \ldots, s_m) - \sum_{(u,v)\in T} I(s_u, s_v) \qquad (2)$$

Le résultat précédent, prouvé dans [3, 4], généralise à la fois le résultat classique pour le modèle d'ACI—qui est équivalent à TCA avec un arbre sans arête—et le résultat de Chow et Liu [9]—équivalent à TCA sans mixage. L'algorithme que nous proposons est simplement une minimisation du contraste $J(W, T)$ défini par l'équation (2), par rapport aux variables $W$ et $T$, et est présenté dans la section 4.2.

Comme pour l'ACI, dans le cas intemporel, le modèle n'est identifiable que pour des sources non Gaussiennes [10]. Cependant cette condition n'est pas toujours suffisante pour le modèle TCA [3].

### 3.2 Distribution a priori et arbres

Afin d'obtenir des arbres non couvrants, nous imposons une distribution a priori $p(T)$ pour l'arbre $T$, i.e. le contraste à minimiser est la somme du contraste défini à l'équation (2) et de $f(T) = -\log p(T)$. Les distributions a priori que notre algorithme peut utiliser incluent toutes les distributions $p(T)$ telles que $f(T) = -\log p(T)$ est une fonction concave du nombre d'arêtes de $T$. Le contraste que nous minimisons est alors

$$F(W,T) = I(s_1,\ldots,s_m) - \sum_{(u,v)\in T} I(s_u,s_v) + f(T). \quad (3)$$

## 4 Estimation et minimisation

Comme dans le cadre de l'ACI, afin de minimiser le contraste défini à la section précédente, il faut tout d'abord pouvoir l'estimer à partir d'un échantillon fini des signaux $x$.

### 4.1 Estimation pour des sources intemporelles

Dans le cadre de sources temporellement indépendantes et non Gaussiennes, il est possible d'étendre certaines techniques de l'ACI à TCA.

Comme l'entropie conjointe de $s = Wx$ peut se décomposer en $H(s) = H(x) + \log|\det W|$, il est seulement nécessaire d'estimer les entropies unidimensionelles et bidimensionnelles, Pour la première méthode, nous estimons $m(m-1)/2$ densités en utilisant l'estimateur de Parzen [20] et nous utilisons ces densités estimées pour calculer les entropies.

Dans une deuxième approche, nous étendons le contraste présenté dans [2], qui est basé sur la « variance généralisée à noyaux » (kernel generalized variance, KGV), une approximation de l'information mutuelle utilisant des espaces de Hilbert à noyaux reproductifs [19].

Enfin, dans une troisième approche, nous utilisons des expansions de *Gram-Charlier* pour les entropies unidimensionnelles et bidimensionnelles [1]. La fonction de contraste est calculée à partir de cumulants d'ordre au plus quatre et est donc plus facile à calculer et à optimiser. Même si elle ne permet pas d'arriver toujours à un résultat satisfaisant, elle permet de trouver une initialisation rapide pour les deux autres approches.

### 4.2 Optimisation

Le contraste dépend de la matrice $W$ et de l'arbre $T$. Nous utilisons une procédure de minimisation alternée.

Par rapport à $T$, l'objectif $J(W,T)$ est une somme indexée par les arêtes de l'arbre $T$, plus une fonction concave du nombre d'arêtes. Ainsi, une fois que les informations mutuelles $I(s_u,s_v)$ sont calculées pour chaque paire, nous faisons face à une simple modification d'un problème d'arbre couvrant de poids minimal (minimum weight spanning tree [11]), que nous pouvons résoudre exactement en temps polynomial par un « algorithme gourmand » (greedy algorithm), présenté en Figure 2, que nous appliquons avec $w_{uv} = -I(s_u,s_v)$. Cet algorithme suit un principe très simple—tant que l'on peut ajouter une arête sans

**Input** : Poids $\{w_{uv}, u,v \in \{1,\ldots,m\}\}$
$t_{max} \geqslant 0$, fonction concave $w(t)$
**Algorithme** :
1. Initialisation : $T = \varnothing$, $t = 0$
$E = \{1,\ldots,m\} \times \{1,\ldots,m\}$
2. Tant que $E \neq \varnothing$ et $t < t_{max}$
a. Calculer $w_{u_0v_0} = \max_{(u,v)\in E} w_{uv}$
b. Si $w_{u_0v_0} + f(t+1) - f(t) > 0$
$T \leftarrow T \cup (u_0,v_0)$, $\quad t \leftarrow t+1$
$E \leftarrow \{e \in E, T \cup \{e\}$ n'a pas de cycle$\}$
sinon $E = \varnothing$
**Output** : arbre $T$

FIG. 2 – Algorithme gourmand pour trouver l'arbre (non-couvrant) de poids maximal avec au plus $t_{max}$ arêtes.

**Input** : Observations $\{x\} = \{x^1,\ldots,x^N\}, \forall n,\ x^n \in \mathbb{R}^m$
**Algorithme** :
1. Initialisation : $T = \varnothing$, $W$ random
2. Pour $i = 0,\ldots,m-1$
a. Tant que $F(W,T)$ décroît
1. Fixer $W$ et calculer l'arbre (non-couvrant) $T$ avec au plus $i$ arêtes
2. Calculer le gradient of $F$ par rapport à $W$
3. Rechercher exhaustivement le minimum de $F$ par rapport à $W$, dans la direction du gradient
b. $W_i = W$, $T_i = T$, $J_i = J$
3. Calculer $i^* = \arg\max_i J_i$
**Output** : matrice de démixage $W = W_{i^*}$, arbre $T = T_{i^*}$

FIG. 3 – l'algorithme TCA

créer de cycle, on ajoute l'arête qui génère le gain immédiat maximal—et détermine toujours un arbre optimal.

Enfin, pour minimiser par rapport à la matrice de démixage $W$, nous utilisons une méthode de descente de gradient, pour laquelle nous avons développé plusieurs heuristiques afin d'éviter certains minimums locaux (pour plus de détails et simulations numériques, voir [4]). Une d'entre elles est de commencer par des arbres sans arêtes, i.e. de commencer par un algorithme d'ACI, et ensuite d'augmenter progressiment le nombre maximal d'arêtes. L'algorithme final est présenté en Figure 3

### 4.3 Généralisation aux séries temporelles

Dans les cas des source stationnaires Gaussiennes, nous pouvons estimer les différentes informations mutuelles, à l'aide de la densité spectrale multidimensionnelle, généralisant la fonction de contraste de Pham [17] au modèle TCA. Ce contraste correspond à une généralisation de la notion de réseaux Bayésiens aux séries temporelles [6] et permet d'appliquer notre méthode à des signaux où la dépendance temporelle est essentielle. Pour plus de détails, voir [4].

TAB. 1 – Résultats de simulation pour des sources intemporelles non Gaussiennes.

| $m$ | Jade | FastICA | TCA-Cum | TCA-Kde | TCA-Kgv |
|---|---|---|---|---|---|
| 4 | 0.6 | 0.65 | 0.25 | 0.15 | 0.14 |
| 6 | 1.3 | 1.2 | 0.7 | 0.51 | 0.5 |
| 8 | 2.4 | 2.5 | 1.1 | 0.9 | 0.9 |

# 5 Simulations

Dans toutes les simulations que nous présentons, nous avons généré des données à partir de modèles d'amas connus (i.e. la matrice $W$ et les amas $C_1, \ldots, C_k$ sont connus), afin de pouvoir comparer la performance de notre algorithme avec des algorithmes dédiés à l'ACI, i.e. dont le but n'est pas d'obtenir des amas, mais des composantes unidimensionnelles (pour une description plus précise du cadre expérimental, voir [4]). Le but de ces simulations est de montrer que pour des examples simples d'amas, utiliser l'ACI n'est pas suffisant pour déterminer la matrice de démixage et des techniques comme celle que nous avons présentée sont donc nécessaires.

Afin de comparer les résultats, nous avons construit une métrique qui est invariante par rapport aux invariances connues du problème (i.e. seuls les sous-espaces sont potentiellement identifiables [7]). Cette métrique permet de mesurer la performance d'une matrice $\hat{W}$ sachant que la matrice optimale est $W$ et la décomposition en amas est connue. Le problème est invariant par permutation des composantes et notre métrique est basée sur la minimisation par rapport à tous les assignements possibles ; cette minimization est rendue possible par l'utilisation de l'« algorithme Hongrois » [5]. Cette métrique est toujours comprise entre 0 et $m$, et égale à 0 pour un démixage parfait (voir [4] pour une définition précise).

Dans le Tableau 1, nous comparons notre approche avec deux algorithmes d'ACI, Jade [8] et FastICA [14] : l'ACI toute seule ne permet pas de retrouver parfaitement la matrice de démixage correcte, alors que notre approche permet de la retrouver plus efficacement.

# 6 Conclusion

Nous avons présenté un algorithme qui généralise l'analyse en composantes indépendantes en modélisant les sources par un réseau Bayésien à structure d'arbre. Cet arbre $T$ et la matrice de démixage sont obtenus par la minimisation d'une fonction de contraste, dans une approche semi-paramétrique. En particulier, lorsque l'arbre est non couvrant, notre algorithme permet de trouver des amas. La détermination de l'arbre (non couvrant) optimal nous permet de trouver à la fois le nombre et la taille de ces amas, ce qui n'était pas possible avec les approches existantes [7, 13].

Enfin, bien que dans cet article nous nous limitons aux arbres, il est possible de généraliser notre approche à des réseaux Bayésiens plus complexes afin de considérer des dépendances plus riches.

# Références


[1] S. Akaho, Y. Kiuchi, et S. Umeyama. MICA : Multimodal independent component analysis. In *Proc. of the International Joint Conf. on Neural Networks*, 1999.

[2] F. R. Bach et M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3 :1–48, 2002.

[3] F. R. Bach et M. I. Jordan. Tree-dependent component analysis. In *Proc. of UAI 2002*, 2002.

[4] F. R. Bach et M. I. Jordan. Finding clusters in independent component analysis. In *Fourth Int. Symp on ICA and BSS*, 2003.

[5] D. Bertsimas et J. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientic, 1997.

[6] D. Brillinger. Remarks concerning graphical models for time series and point processes. *Revista de Econometria*, 16 :1–23, 1996.

[7] J.-F. Cardoso. Multidimensional independent component analysis. In *Proc. of ICASSP*, 1998.

[8] J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11(1) :157–192, 1999.

[9] C. K. Chow et C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Information Theory*, 14 :462–467, 1968.

[10] P. Comon. Independent component analysis, a new concept ? *Signal Processing*, 36(3) :287–314, 1994.

[11] T. H. Cormen, C. E. Leiserson, et R. L. Rivest. *Introduction to Algorithms*. MIT Press, 1989.

[12] L. De Lathauwer, B. De Moor, et J. Vandewalle. Fetal electrocardiogram extraction by blind source subspace separation. *IEEE Trans. Biomed. Eng*, 47(5) :567–572, 2000.

[13] A. Hyvärinen et P. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7) :1705–1720, 2000.

[14] A. Hyvärinen, J. Karhunen, et E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[15] M. I. Jordan. Graphical models. *Statistical Science (Special Issue on Bayesian Statistics)*, 2002. In press.

[16] C. Jutten et J. Hérault. Blind separation of sources, Part I : An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1) :1–10, 1991.

[17] D. T. Pham. Mutual information approach to blind separation of stationary sources. *IEEE Trans. Information Theory*, 48(7) :1935–1946, 2002.

[18] S. Russell et P. Norvig. *Artificial Intelligence : A Modern Approach*. Prentice Hall, 2002.

[19] B. Schölkopf et A. J. Smola. *Learning with Kernels*. MIT Press, 2001.

[20] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1985.

[21] A. S. Willsky. Multiresolution Markov models for signal and image processing. *Proc. IEEE*, 90(8) :1396–1458, 2002.