Séparation de sources par l'indépendance et la parcimonie

Jean-François Cardoso[a] Dinh-Tuan Pham[b],

[a] LTCI/CNRS,
ENST-TSI, 46 rue Barrault, 75634 Paris, France

[b] LMC/CNRS,
IMAG, B.P. 53X, 38041 Grenoble cedex 9, France
cardoso@tsi.enst.fr, Dinh-Tuan.Pham@imag.fr

Résumé – Dans le problème de séparation de sources indépendantes, la prise en compte de la structure temporelle des sources permet d'utiliser des modèles gaussiens. On peut alors aisément optimiser une vraisemblance en spécifiant la distribution de chaque processus source via une simple fonction (profil de variance ou densité spectrale). Cet article propose une méthode simple et flexible pour estimer cette function à partir des données elles-mêmes comme une fonction constante par morceaux sur une partition dyadique. Dans cette approche, la vraisemblance apparaît à la fois comme un critère d'indépendance (dans sa dépendance au mélange) et comme un critère de parcimonie (dans sa dépendance à la répartition d'énergie des sources).

Abstract – Using the time structure of the source processes allows the source separation problem to be tackled within Gaussian models. It is then possible to build a likelihood function by specifying the source distributions via a simple function (temporal or spectral energy density function). This paper proposes a simple and flexible method for estimating this function from the data themselves as a piecewise constant function over dyadic partition. In this approach, the likelihood appears both as an independence measure (when seen as a function of the mixing matrix) and as a sparseness measure (when seen as a function of the source distributions).

1 Introduction

Nous traitons le classique modèle de séparation de mélanges instantanés non bruités:

$$x(t) = As(t), \quad 1 \leq t \leq n \tag{1}$$

où n échantillons d'une observation m-dimensionnelle $x(t)$ sont modélisés comme résultant du mélange par une matrice A inversible, inconnue, de taille $m \times m$, d'un processus m-dimensionnel $s(t)$. L'hypothèse-clé est que les composantes $s_1(t), \ldots, s_m(t)$ de $s(t)$ sont des processus statistiquement indépendants. Dans cet article, nous suivons une approche au maximum de vraisemblance pour l'estimation du mélange. Il suffit pour cela de construire un modèle de la distribution de chaque processus source puisque si $p_{S_i}(s_i(1), \ldots, s_i(n))$ est la densité de probabilité de la i-ème séquence, alors la log-densité de n observations du modèle (1) est

$$-n \log |\det A| + \sum_{i=1}^{m} \log p_{S_i}(y_i(1), \ldots, y_i(n)) \tag{2}$$

où $y(t) = A^{-1}x(t)$.

L'approche la plus classique pour modéliser les sources suppose (explicitement ou non) que les processus sources sont i.i.d. mais possèdent des lois marginales non gaussiennes. Il est cependant possible d'exploiter une éventuelle structure temporelle des signaux sources pour la séparation aveugle dans un modèle gaussien (ou plus exactement sans tenir compte de la possible non gaussianité des sources). Ceci permet d'obtenir des critères qui sont à la fois 'efficaces' sur le plan statistique et qui se prêtent aisément à l'optimisation.

Par exemple, on peut exploiter une possible non stationnarité des sources via l'estimation de leur densité temporelle de puissance. Nous rappelons cette approche à la section 2; elle utilise un modèle rudimentaire selon lequel la densité de puissance est constante par morceaux sur des intervalles fixes, déterminés à l'avance. Le but de cet article est de montrer comment cette approche peut être améliorée grâce à une estimation plus fine de la densité dans laquelle les intervalles sont estimés à partir des données elles-mêmes: à la section 3, nous explorons la forme de la vraisemblance dans ce cadre et montrons comment elle exprime à la fois l'indépendance entre sources et la parcimonie de leur distribution d'énergie. La section 4 esquisse les aspects algorithmiques: nous montrons comment l'algorithmique de Coifman peut être adapté à la minimisation de notre vraisemblance.

2 Modèles gaussiens

Deux méthodes particulièrement simples ont été développées pour la modélisation gaussienne des sources. Dans la première, chaque signal source est modélisé comme

$$s_i(t) = \sigma_i(t) v_i(t) \tag{3}$$

où $v_i(t)$ est une séquence i.i.d. gaussienne de variance unité et $\sigma_i(t)$ est un profil d'amplitude, déterministe, supposé à 'variation lente' [2]. Il s'agit donc d'un modèle très

simple de non-stationnarité dans lequel chaque terme de la somme (2) se réduit à

$$-\frac{1}{2}\sum_{t=1}^{n}\frac{y_i^2(t)}{\sigma_i^2(t)}+\log\sigma_i^2(t)+\text{cst} \tag{4}$$

La seconde méthode est en quelque sorte le 'dual de Fourier' de la première. Elle utilise l'approximation de Whittle de la vraisemblance, basée sur les propriétés asymptotiques de la transformée de Fourier discrète (TFD) d'un signal stationnaire: les coefficients de Fourier sont approximativement gaussiens et décorrélés avec une variance proportionnelle au spectre du processus. Autrement dit, la séquence des coefficients de Fourier se comporte essentiellement comme (3) en remplaçant le temps discret par les fréquences discrètes de la TFD et en remplaçant le profil de variance $\sigma_i^2(t)$ par la densité spectrale de puissance.

Dans chacun de ces deux modèles, une simple fonction décrit la distribution de chaque source: son profil de variance (densité temporelle de puissance) ou son spectre (densité spectrale de puissance). Le premier modèle permet donc d'exploiter la non stationnarité, tandis que le second permet d'exploiter la corrélation temporelle. Dans la suite de cet article, nous nous contentons de décrire la méthode non stationnaire, la méthode exploitant la corrélation temporelle s'en déduisant par transformée de Fourier.

Partitionnement Dans le contexte de séparation aveugle, il convient d'estimer la distribution des sources à partir des données elles-mêmes. Dans les deux modèles gaussiens esquissés ci-dessus, il suffit d'estimer une densité temporelle ou fréquentielle. Un point important est que l'on peut se contenter d'une estimation très grossière de cette densité. Par exemple, dans [2, 3], la densité temporelle de puissance est représentée comme étant constante par morceaux sur un nombre fini de sous-intervalles fixés a priori et (typiquement) de même longueur. Autrement dit, l'intervalle d'observation $\mathcal{I}=[1,n]$ est partitionné en Q sous-intervalles: $\mathcal{I}=\cup_{q=1}^{Q}\mathcal{I}_q$ et l'on suppose la variance constante sur chaque sous-intervalle: $\sigma_i^2(t)=\sigma_{iq}^2$ si $t\in\mathcal{I}_q$. On notera $\mathcal{P}$ une telle partition.

Dans ce modèle, en utilisant (4) et le fait que $\sigma_i(t)$ est constant sur chaque sous-intervalle, la log-vraisemblance (2) devient

$$-n\log|\det A|-\frac{1}{2}\sum_{i=1}^{m}\sum_{q=1}^{Q}n_q\left(\frac{\hat\sigma_{iq}^2}{\sigma_{iq}^2}+\log\sigma_{iq}^2\right)+\text{cst} \tag{5}$$

où n_q est le nombre de points dans le q-ième intervalle de la partition $\mathcal{P}$ et $\hat\sigma_{iq}^2$ est la variance empirique de $y_i(t)$ sur cet intervalle:

$$\hat\sigma_{iq}^2=\frac{1}{n_q}\sum_{t\in\mathcal{I}_q}(A^{-1}x(t))_i^2 \tag{6}$$

La vraisemblance dépend de la matrice de mélange A et des variances locales σ_{iq}^2 mais il est immédiat de la maximiser par rapport à ces derniers paramètres: pour A fixé, l'expression (5) est maximale pour $\sigma_{iq}^2=\hat\sigma_{iq}^2$ et chaque terme entre parenthèses se réduit en ce point à $\log\hat\sigma_{iq}^2+\text{cst}$. Par conséquent, la maximisation de la vraisemblance se ramène à la minimisation de la fonction

$$\phi(A,\mathcal{P})=2n\log|\det A|+\sum_{i=1}^{m}\sum_{q=1}^{Q}n_q\log\hat\sigma_{iq}^2 \tag{7}$$

qui, à une constante et à un facteur -2 près, est le maximum de la log-vraisemblance par rapport aux paramètres de nuisance σ_{iq}.

3 Modèle gaussien adaptatif

Les méthodes gaussiennes citées plus haut reposent sur le choix *a priori* d'une partition $\mathcal{P}$ de l'axe des temps ou des fréquences en sous-intervalles de même longueur (il est possible de travailler avec des sous-intervalles de longueurs arbitraires mais la partition reste toujours fixée a priori). Autrement dit, ces méthodes ne considèrent la maximisation de $\phi(A,\mathcal{P})$ que par rapport au mélange A et non par rapport à la partition $\mathcal{P}$.

Pourtant, le choix d'un bon partitionnement doit dépendre de la nature des signaux traités. Dans le cas non stationnaire, par exemple, il faut choisir une partition suffisamment fine pour saisir les variations locales d'énergie des sources (par exemple, pour séparer des signaux de parole continue, on devrait partitionner à l'échelle du phonème). Cependant, un partitionnement trop fin conduit à une forte variabilité statistique de l'estimation de variance, qui se répercute sur la qualité de la séparation. Dans ce travail, nous développons une méthode de détermination automatique du partitionnement de l'axe des temps (ou des fréquences) et ceci en fonction des données elles-mêmes. Ainsi, les sous-intervalles de la partition seront courts dans les zones où la variance varie rapidement et longs dans les zones où elle varie lentement.

Nous proposons une méthode de séparation fondée sur la minimisation par rapport au mélange A et à la partition $\mathcal{P}$ d'une version pénalisée du critère de vraisemblance (7). Algorithmiquement, nous proposons une minimisation alternée par rapport à ces deux paramètres, comme le suggère les reformulations (ci-après) du critère de vraisemblance (7).

Dépendance par rapport au mélange Pour une partition $\mathcal{P}$ fixée, définissons pour chaque sous-intervalle $\mathcal{I}_q$ la matrice $\hat R_q$ de covariance empirique des observations sur cet intervalle:

$$\hat R_q=\frac{1}{n_q}\sum_{t\in\mathcal{I}_q}x(t)x(t)^\dagger. \tag{8}$$

Puisque les $\hat\sigma_{iq}^2$ sont les éléments diagonaux de $A^{-1}\hat R_qA^{-\dagger}$, on a $\sum_i\log\hat\sigma_{iq}^2=\log\det\operatorname{diag}(A^{-1}\hat R_qA^{-\dagger})$. Par ailleurs, on a $\log\det(A^{-1}\hat R_qA^{-\dagger})=-2\log\det|A|+\log\det\hat R_q$. Par conséquent, le critère de vraisemblance s'écrit aussi

$$\phi(A,\mathcal{P})=\sum_{q=1}^{Q}n_q\ \operatorname{off}(A^{-1}\hat R_qA^{-\dagger})+f_1(\mathcal{P}) \tag{9}$$

où $f_1(\mathcal{P}) = \sum_q n_q \log\det \hat{R}_q$ ne dépendant pas de A et où

$$\mathrm{off}(M) = \log\det \mathrm{diag} M - \log\det M \qquad (10)$$

est une mesure de l'écart d'une matrice positive M à la diagonalité. Pour une partition donnée, commune à toutes les sources, le critère de vraisemblance est donc un critère de diagonalisation conjointe. Ce critère est aussi [2, 3] la mesure de l'information mutuelle entre les composantes de $A^{-1}x(t)$ dans notre modèle gaussien non-stationnaire.

Dépendance par rapport à la partition La dépendance du critère de vraisemblance par rapport à la seule partition prend aussi une forme très particulière et suggestive. Notons $\mathcal{P}_0$ la partition de $[1, n]$ en un unique intervalle et $\hat{\sigma}_i^2$ la variance empirique moyenne sur cet intervalle. Pour toute partition $\mathcal{P}$, on a $\hat{\sigma}_i^2 = n^{-1}\sum_q n_q \hat{\sigma}_{iq}^2$ ce qui permet de ré-écrire le critère de vraisemblance (7) comme

$$\phi(A, \mathcal{P}) = \phi(A, \mathcal{P}_0) - \sum_i \sum_q n_q \, h\left(\frac{\hat{\sigma}_{iq}^2}{\hat{\sigma}_i^2}\right) \qquad (11)$$

où la fonction $h(\cdot)$ est définie comme $h(u) = u - 1 - \log u$. Puisque $h(u) \geq 0$ avec égalité seulement si $u = 1$, on voit que le critère, vu comme fonction de $\mathcal{P}$, mesure la somme sur toutes les sources des écarts entre les variances empiriques locales et leurs valeurs moyennes sur tout l'échantillon.

Par conséquent, minimiser le critère par rapport à $\mathcal{P}$, c'est encore trouver les intervalles faisant apparaître la plus grande diversité de variances locales au sens de la mesure $\sum_q n_q \, h(\hat{\sigma}_{iq}^2/\hat{\sigma}_i^2)$. En ce sens, le critère apparaît comme une mesure de la parcimonie de la distribution d'énergie des sources.

Pénalisation Il n'est pas possible d'utiliser le critère (7) dans la minimisation par rapport à $\mathcal{P}$ sans pénaliser la complexité de la partition. En effet, tout raffinement d'une partition par découpage d'un intervalle quelconque $\mathcal{I}_0$ de taille n_0 en deux sous-intervalles $\mathcal{I}_1, \mathcal{I}_2$ diminue le critère si les variances empiriques σ_1^2 et σ_2^2 sur $\mathcal{I}_1$ et $\mathcal{I}_2$ sont différentes. On vérifie en effet aisément que si $n_0 = n_1 + n_2$ et $n_0\sigma_0^2 = n_1\sigma_1^2 + n_2\sigma_2^2$, alors $n_0 \log\sigma_0^2 - n_1 \log\sigma_1^2 - n_2 \log\sigma_2^2 = n_1 h(\sigma_1^2/\sigma_0^2) + n_2 h(\sigma_2^2/\sigma_0^2)$ qui est strictement positif sauf si $\sigma_1 = \sigma_2$.

Pour se faire une idée du type de pénalisation de la complexité à apporter au critère (7), examinons la différence

$$\delta = \sum_{q=1}^{Q} n_q \log\hat{\sigma}_q^2 - \sum_{q=1}^{Q} n_q \log\sigma_q^2 \qquad (12)$$

quand le modèle tient, c'est-à-dire, pour T échantillons i.i.d. distribués selon une partition $\mathcal{P}$ de $[1, T]$ en Q sous-intervalles de variances $\sigma_1^2, \ldots, \sigma_Q^2$. Le premier terme de (12) est en fait un estimateur biaisé du second; un calcul asymptotique (pour n_i grand) fournit le biais au premier ordre: $\mathrm{E}\delta = -Q$. Ceci suggère de pénaliser le critère (7) en lui ajoutant, pour chaque source, un terme proportionnel au nombre d'intervalles de la partition. De manière plus générale, nous proposons le critère pénalisé

$$\phi_\star(A, \mathcal{P}) = \phi(A, \mathcal{P}) + m \sum_q \rho(n_q) \qquad (13)$$

avec une fonction $\rho(\cdot)$ convenablement choisie. En prenant pour ρ la fonction constante égale à 1, nous compensons exactement le biais asymptotique. En prenant pour ρ une fonction décroissante de la taille de l'intervalle, nous pénalisons les découpages trop fins qui sont dommageables à l'estimation des variances.

Le critère $\phi(A, \mathcal{P})$ dérivant de la vraisemblance, une pénalisation additive peut être comprise comme un terme bayesien résultant d'une probabilité *a priori* sur les partitions. La forme (13) correspond à une probabilité $p(\mathcal{P}) = \exp\left(-m/2 \sum_q \rho(n_q)\right)/Z$ où Z est une constante de normalisation.

Indépendance, parcimonie, complexité Pour conclure, notons qu'il est possible —et probablement souvent souhaitable— de généraliser notre modèle en partitionnant indépendamment les profils de chaque source. Le modèle est alors spécifié par A et m partitions $\mathcal{P}_1, \ldots, \mathcal{P}_m$ de tailles $Q_1, \ldots, Q_m$ et le critère de vraisemblance pénalisé devient

$$\phi_\star(A, \mathcal{P}_1, \ldots, \mathcal{P}_m) = n \log\det|A| + \sum_{i=1}^{m} \sum_{q=1}^{Q_m} n_{iq} \log\hat{\sigma}_{iq}^2 + \rho(n_{iq}). \qquad (14)$$

Comme vu plus haut, la dépendance $\phi_\star$ en A est une mesure d'indépendance, la dépendance en chaque $\mathcal{P}_i$ est une mesure de la parcimonie de la distribution temporelle d'énergie de chaque source et la fonction ρ permet d'assigner à chaque partition $\mathcal{P}_i$ une probabilité *a priori* proportionnelle à $\exp(-1/2 \sum_q \rho(n_{iq}))$.

Il est aussi intéressant de considérer la forme (11) du critère de vraisemblance en fonction de A. Sous cette forme, le critère dépend de A à travers la parcimonie des profils de variance $\hat{\sigma}_{iq}^2$ mais aussi via $\phi(A, \mathcal{P}_0)$. Or $\phi(A, \mathcal{P}_0) = \mathrm{off}(A^{-1}\hat{R}A^{-\dagger})$ où $\hat{R} = n^{-1}\sum_q n_q \hat{R}_q = n^{-1}\sum_t x(t)x(t)^\dagger$, ce qui montre que ce dernier terme est une mesure de la corrélation globale des sources. Ainsi, il apparaît que la mesure de dépendance (9) est la somme d'une mesure de corrélation et d'une mesure de la parcimonie du profil d'énergie de chacune des sources (voir[4] pour une perspective plus générale).

4 Algorithmes

Nous proposons de minimiser le critère de vraisemblance pénalisé par des minimisations alternées par rapport au mélange A et aux partitions $\mathcal{P}_i$.

Optimisation du mélange

Deux cas se présentent pour la minimisation de (11) par rapport à A.

Le cas le plus favorable est celui où une même partition est utilisée pour toutes les sources: $\mathcal{P}_i = \mathcal{P}$. Alors, le critère prend la forme (9) et sa minimisation est donc obtenue par la diagonalisation conjointe des matrices de covariance empiriques estimées sur $\mathcal{P}$. Un algorithme très efficace existe à cet effet [2, 3].

Un modèle plus riche et plus réaliste n'impose pas un partitionnement identique pour toutes les sources. Dans ce cas, le critère de vraisemblance pénalisé n'est plus équivalent à (9) et il faut recourir à une optimisation *ad hoc*. On peut néanmoins aisément construire un algorithme de type quasi-Newton car il est facile, comme dans [2], de calculer le gradient et le hessien approximatif du critère par rapport au mélange.

Optimisation des partitions

L'optimisation du critère par rapport à toutes les partitions possibles souffre d'une complexité combinatoire mais elle peut être réalisée très rapidement si l'on accepte de restreindre la recherche à un ensemble bien structuré de partitions. Nous proposons ici d'adapter l'algorithme de Coifman [1] (pour la recherche des bases d'entropie minimale) dont l'idée essentielle est d'exploiter l'additivité du critère *et* de restreindre la recherche à un ensemble de *partitions dyadiques*.

Nous décrivons brièvement l'adaptation de cet algorithme à notre problème pour la recherche de la meilleure partition dyadique du profil de variance d'une seule source. Dans la version la plus simple, on suppose que l'intervalle d'observation est partitionné en 2^J 'atomes' (intervalles élémentaires) constituant la partition la plus fine. Dans une première étape, chaque paire d'intervalles consécutifs est fusionnée (ou non) selon que cette fusion diminue (ou non) le critère. Dans les étapes suivantes, on poursuit cette fusion conditionnelle de la façon suivante. Soient deux intervalles adjacents $\mathcal{I}_1$ et $\mathcal{I}_2$ de longueurs n_i, de variance moyenne v_i, partitionnés selon $\mathcal{P}_i$, contribuant un coût ϕ_i au critère ($i = 1, 2$). On considère deux possibilités pour partitionner $\mathcal{I}_0 = \mathcal{I}_1 \cup \mathcal{I}_2$ selon $\mathcal{P}_0$: soit $\mathcal{P}_0$ est la partition triviale de $\mathcal{I}_0$ en un seul intervalle (fusion), soit $\mathcal{P}_0$ est l'union des deux partitions $\mathcal{P}_1$ et $\mathcal{P}_2$ (concaténation). Le choix entre fusion et concaténation se fait selon le critère, c'est-à-dire selon le signe de

$$n_0 \log v_0 + \rho(n_0) - \phi_1 - \phi_2 \qquad (15)$$

avec $n_0 = n_1 + n_2$ et $v_0 = (n_1 v_1 + n_2 v_2)/(n_1 + n_2)$. En cas de fusion, on pose $\phi_0 = n_0 \log v_0 + \rho(n_0)$; en cas de concaténation, on pose $\phi_0 = \phi_1 + \phi_2$.

En procédant de la sorte, niveau après niveau, on propage $(n_i, v_i, \mathcal{P}_i, \phi_i)$ le long des branches d'un arbre binaire. On montre aisément que cette technique permet de trouver la meilleure partition dyadique au sens du critère.

Deux remarques concernant le point initial. La taille des 'atomes' détermine la partition la plus fine. Il semble raisonnable de choisir de atomes de longueur au moins égale à m de telle sorte que —dans le cas où l'on alterne avec une étape de diagonalisation conjointe— les matrices de covariances empiriques soient (probablement) de rang plein. Par ailleurs, le nombre d'échantillons n'est généralement pas un multiple d'une puissance de 2. Une possibilité est de choisir des atomes de tailles variables pour en avoir un nombre exactement égal à 2^J. Une autre possibilité est de choisir des atomes de tailles identiques et de gérer un arbre sur-dimensionné avec, à une extrémité, des atomes vides, lesquels échappe évidemment au processus de fusion conditionnelle à partir d'un certain niveau.

Conclusion

Nous avons vu que, dans une classe de modèles gaussiens pour la séparation de sources, le critère de vraisemblance est lié à l'information mutuelle et à la parcimonie de la distribution de densité de puissance (temporelle ou spectrale) des sources. Remarquons qu'il est tout-à-fait possible d'invoquer l'idée de parcimonie pour arriver au critère (7) sans faire référence à un modèle gaussien. Ce critère en effet favorise l'émergence de sources parcimonieuses en ceci que leurs distributions d'énergie (dans le domaine temporel ou fréquentiel) sont concentrées dans des régions de faible étendue.

Cette approche peut améliorer les algorithmes existants car elle permet une adaptation plus souple à la distribution des sources. De plus cette adaptation est 'automatique' même si le choix d'une partition dépend encore de la pénalisation de la vraisemblance.

L'algorithme proposé pour mettre en oeuvre ce programme alterne entre optimisation par rapport au mélange A (éventuellement via une diagonalisation conjointe) et par rapport aux partitions dyadiques (via l'algorithme de Coifman).

La modélisation par partitions dyadiques permet une recherche très rapide au prix d'une contrainte forte. Néanmoins la classe des partitions dyadiques reste très vaste et nous pensons qu'elle peut convenir à de nombreuses situations pratiques. Lors de la conférence, nous présenterons des résultats expérimentaux illustrant l'application de notre approche à divers types de signaux.

References

[1] Ronald Raphael Coifman and Mladen Victor Wickerhauser. Entropy based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 32:712–718, March 1992.

[2] Dinh-Tuan Pham and Jean-François Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. on Sig. Proc.*, 49(9):1837–1848, September 2001.

[3] D.T. Pham. Blind separation of instantaneous mixture of sources via the Gaussian mutual information criterion. *Signal Processing*, (4):855–870, 2001.

[4] Jean-François Cardoso. The three easy routes to independent component analysis; contrasts and geometry. In *Proc. ICA 2001, San Diego*, 2001.