Graphes de représentation minimaux, entropies et divergences: applications

Olivier Michel[◇], *Alfred Hero*[*], *Patrick Flandrin*[⋆]

[◇] Laboratoire d'Astrophysique, UMR 6525,Université de Nice-Sophia Antipolis,
O6108, Nice Cedex 02, France.

[*] Department of EECS, University of Michigan, Ann Arbor, MI 48109-2122, USA

[⋆] Laboratoire de Physique (URA 1325 CNRS), École Normale Supérieure de Lyon,
46 allée d'Italie, 69364 Lyon Cedex 07, France.

e-mail : olivier.michel@ens-lyon.fr, hero@eecs.umich.edu, patrick.flandrin@ens-lyon.fr

October 19, 2000

Catégorie : recherche.

Résumé

Il a été récemment établi que la longueur d'un graphe de représentation minimal (Minimal Spanning Tree, MST) construit sur un ensemble de réalisations d'un processus aléatoire permet d'estimer l'entropie de ce dernier, dans un contexte non paramétrique. Dans cette étude, après avoir rappelé les principales définitions et propriétés des MST, nous en illustrons l'intérêt à travers leur mise en œuvre dans le cadre de problèmes de débruitage, de séparation de mélange statistique ou de détection de trajectoire dans le plan temps-fréquence, dans un contexte d'analyse de signaux non stationnaires.

Mots clés : Graphes de représentation minimaux, - f-divergences - Entropie statistique de Rényi - Séparation de mélange - Débruitage - Temps fréquence.

Abstract

A non parametric approach for entropy estimation was recently proposed by the authors. Based on the statistical properties of minimal spanning trees, the length of these latter is shown to converge toward the Rényi entropy of the underlying process. In this paper, the effectiveness of the MST approach is illustrated on denoising problem, clustering and mixture separation problems. Basic concepts and properties of MST are briefly recalled in a first part, then application examples are given on general denoising or clustering problems, including trajectory detection in the time-frequency plane.

Key words : Minimal spanning Tree - f-divergence - Rényi entropy - Mixture separation - Denoising - Time frequency analysis.

1 Introduction

Dans [18], nous avons établi que le comportement asymptotique d'une classe assez générale de graphes ou de sous graphes minimaux, définis sur un ensemble de points de $\mathbb{R}^d$, vérifiant la propriété de quasi-additivité de Redmond et Yukich [25], permet de construire des estimateurs consistants de l'entropie de Rényi de la distribution de ces points. De tels graphes apparaissent dans

- la résolution de problèmes tels que la construction compétitive (au sens de l'optimisation d'un coût) de réseaux, en télécommunication ou dans le problème de routage de connections en conception de circuits VLSI [9, 24],
- le célèbre problème du voyageur de commerce ou la construction des arbres de Steiner, qui consiste à trouver le trajet minimum permettant de visiter k villes parmi n [8],
- les tests sur la nature aléatoire de champs de données [15],
- dans les problèmes d'optimisation combinatoire.

L'algorithme d'approximation des sous-graphes minimaux contenant k points parmi N ($k < N$) présenté dans [17] généralise l'approche proposée par Ravi et al. [23] dans le cas $d = 2$ et a permis de proposer un estimateur robuste de l'entropie d'une distribution d-dimensionnelle bruitée. Dans le présent article, nous étudions quelques exemples d'utilisation des graphes de représentations minimaux (Minimal Spanning Trees, ou MST) pour l'estimation robuste de l'entropie de Rényi. Dans un premier paragraphe, les principales définitions relatives aux entropies de Rényi et aux divergences associées sont rappelées. La section 3 a pour objet d'introduire les concepts de base sur les MST ainsi que les principaux résultats théoriques obtenus dans les articles [17, 18]. Nous proposons dans les sections suivantes, des applications aux problèmes de séparation de composantes, dans le cas de débruitage ("outlier" rejection) ou de détection de composantes dans le plan temps-fréquence.

2 Entropies et divergence de Rényi

Soit $\mathcal{X}_n = \{x_1, x_2, \ldots, x_n\}$ un ensemble de n réalisation d'un vecteur aléatoire défini sur $\mathbb{R}^d$, de densité de Lebesgue multivariée $f(x)$ dont le support est limité à $[0,1]^d$. L'entropie de Rényi d'ordre ν de ce processus s'exprime [26] par

$$H_\nu(f) = \frac{1}{1-\nu} \ln \int f^\nu(x) dx \tag{1}$$

La divergence d'information (I-divergence) de Rényi entre le processus de densité f et un processus dominé par la densité de Lebesgue f_0, introduite dans le cadre plus général des f-divergences de Csiszàr [10] (voir aussi [1] et les références qui s'y trouvent), prend l'expression suivante :

$$I_\nu(f, f_0) = \frac{-1}{1-\nu} \ln \int \left(\frac{f(x)}{f_0(x)} \right)^\nu f_0(x) dx \tag{2}$$

La quantité $I_\nu(f, f_0)$ apparaît à la fois comme un cas particulier de I-divergence de Chernoff et comme l'entropie conjointe de f et f_0 [3]. Cette I-divergence est minimale (égale à zéro) si et seulement si $f = f_0$ presque partout. La I-divergence de Rényi $I_\nu(f, f_0)$ est égale à l'entropie de Rényi $H_\nu(f)$ lorsque f_0 est la densité uniforme sur $[0, 1]^d$. D'autres divergences peuvent être obtenues en faisant varier le paramètre ν; on mentionnera par exemple le cas $\nu = \frac{1}{2}$, qui conduit à une divergence de Rényi qui est reliée au logarithme de la distance de Hellinger

$$\mathrm{Hel}(f, f_0) = \int \left(\sqrt{f(x)} - \sqrt{f_0(x)} \right)^2 dx = 2 \left(1 - \exp(-\frac{1}{2} I_{\frac{1}{2}}(f, f_0) \right)$$

où

$$I_{\frac{1}{2}}(f, f_0) = -2 \ln \left(\int \sqrt{f(x) f_0(x)} dx \right)$$

et surtout le cas limite $\nu \to 1$ pour lequel la divergence de Rényi tend vers la divergence de Kullback-Leibler, qui appartient elle aussi à la classe des f-divergences de Csiszàr, mais qui est construite à partir de l'entropie de Shannon-Gibbs :

$$\lim_{\nu \to 1} I_\nu(f, f_0) = - \int \ln \frac{f(x)}{f_0(x)} f_0(x) dx.$$

Le problème d'estimation des I-divergences se rencontre dans une très large classe d'applications, par exemple pour la classification de densités de probabilité, à des fins de segmentation ou de séparation de "composantes" dans un mélange (nous développons ce type d'application dans les paragraphes suivants), ou encore dans le contexte de la reconnaissance des formes [3, 10]. Dans ce cadre en effet, un test basé sur l'application de seuils sur les valeurs estimées de $I_\nu(f, f_0)$ est en général la clé de l'algorithme décidant si $f = f_0$. L'estimation de I-divergence apparaît encore dans les problèmes de recalage d'images; dans ce contexte, la I-divergence est directement reliée à l'information mutuelle entre deux images f et f_0 [30]. Pour une revue plus complète sur les problèmes d'estimation d'entropie et de divergence d'information, on pourra se reporter à [7] et [3, 4].

Dans les sections suivantes, nous proposons de nouvelles méthodes d'estimation robuste de l'entropie de Rényi $H_\nu(f)$ de processus de densité f inconnue, et de la I-divergence de Rényi $I_\nu(f, f_0)$, entre une densité f inconnue dominée par une densité f_0 arbitraire.

3 MST et k-MST

3.1 Définitions

Un graphe acyclique minimal (MST, pour "Minimal Spanning Tree") est un graphe (ou arbre) $\mathcal{T}_n$ connectant l'ensemble des réalisations $\mathcal{X}_n = \{x_1, x_2, \ldots, x_n\}$ d'un processus vectoriel défini dans $\mathbb{R}^d$. C'est donc une liste de sommets (les points x_i) et de connections $e_{i,j}$ entre ces sommets. La longueur totale d'ordre γ du graphe est la somme des longueurs (norme euclidienne) pondérées en loi de puissance d'ordre $\gamma \in]0, d[$, de l'ensemble des connections :

$$L_{n,\gamma} = \sum_{e_{i,j} \in \mathcal{T}_n} |e_{i,j}|^\gamma \tag{3}$$

Le MST est, parmi tous les graphes acycliques totalement connectés qu'il est possible de construire, le graphe dont la longueur est minimale :

$$\mathcal{T}_n^\star = \operatorname{Arg}\min_{\mathcal{T}_n} L_{n,\gamma} \tag{4}$$

Le graphe défini par les équations (3,4) peut être calculé de façon exacte à partir d'algorithmes dont le coût varie comme $n \log n$ [27]. Cette définition est étendue aux sous-graphes ne connectant qu'un sous-ensemble de points dans $\mathbb{R}^d$: les k-MST.

Un k-MST est un graphe minimal ne connectant que k points parmi n. C'est aussi le MST associé au sous-ensemble $\mathcal{X}_{n,k}$ de $\mathcal{X}_n$ ne contenant que ces k points. La minimisation porte alors à la fois sur la détermination de ce sous-ensemble et sur la longueur du MST connectant les points du sous-ensemble :

$$\mathcal{X}_{n,k}^\star = \operatorname{Arg}\min_{i_1,\ldots,i_k} \operatorname{Arg}\min_{\mathcal{T}_k} L_{k,\gamma}$$

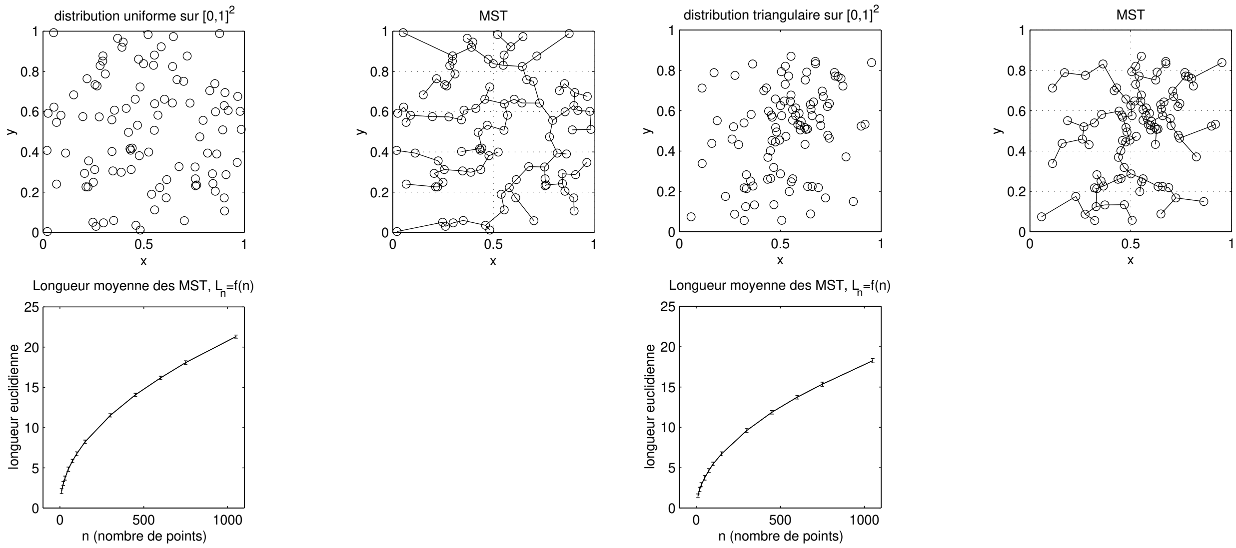
où $\mathcal{X}_{n,k} = \{x_{i_1}, \ldots, x_{i_n}\}$. En pratique, la double minimisation est conduite conjointement; c'est le cas en particulier des algorithmes que nous avons développés [17, 18]. Il a été démontré que le problème d'estimation d'un k-MST dans $\mathbb{R}^2$ est un problème NP-complet [23, 31]. Ravi *et al* ont proposé un algorithme d'approximation à coût polynômial dans le cas de distributions bidimensionnelles. Dans [18], nous avons étendu ce travail et proposé un algorithme d'approximation des k-MST dans le cas plus général d-dimensionnel, fournissant une solution dont le rapport d'approximation est majoré en $O(k^{(1-1/d)^2})$. Le détail de l'algorithme de calcul approché des k-MST, sa robustesse calculée à partir des courbes d'influence, et des éléments de preuve de sa convergence asymptotique sont donnés dans l'article [18]. Cette partie , très technique, ne sera pas développée dans cet article.

3.2 Exemples

Les figures 1-a et 1-b présentent un exemple qui illustre l'intérêt des MST dans le cadre de l'analyse d'une distribution à partir d'un ensemble fini (souvent faible) de réalisations du processus étudié.

Les deux distributions considérées sont définies sur $[0, 1]^d$. En 1-a, la distribution est uniforme alors qu'en figure 1-b, nous avons utilisé une distribution séparable triangulaire, maximale en $(0.5, 0.5)$. Sur chaque figure sont portés trois graphiques : un exemple de distribution obtenu pour 100 réalisations de la variable aléatoire bidimensionnelle considérée, le MST construit sur cette distribution, et la représentation de l'évolution de la longueur moyenne des MST obtenus en fonction du nombre de réalisations considéré. Ce dernier graphe est calculé en reproduisant 256 constructions de MST pour chaque valeur n étudiée. La longueur utilisée ici est la longueur euclidienne, soit pour $\gamma = 1$.

Il apparaît que l'arbre de représentation minimal obtenu dans le cas de la distribution uniforme s'étend sans caractéristique géométrique particulière, sur la totalité du support considéré. Le graphe construit sur les réalisations de distribution triangulaire montre nettement une structure dans laquelle il existe une forte densité de segments courts au centre du support, où la fonction de densité de probabilité est maximale. De plus, de très nombreux segments sont orientés selon des lignes radiales dont le centre correspond au point où la densité de probabilité est maximale. Ainsi, il apparaît qualitativement pour l'instant que les MST contiennent la signature de caractéristiques importantes de la distribution de leurs sommets.



(a) (b)

Figure 1: Distribution pour $n = 100$ réalisations de la variable aléatoire, MST, et longueur des MST en fonction de n, dans le cas -(a)- d'une distribution uniforme, -(b)- d'une distribution triangulaire.

La comparaison entre les résultats obtenus pour ces deux distributions est présentée sur la figure 2. Le premier graphe reprend en partie les courbes des figures précédentes. Le graphe de droite reproduit ces mêmes courbes, normalisées par $\sqrt{n}$, et transformées par la fonction $-2\log(.)$. Il apparaît clairement que ces valeurs transformées de la longueur des MSTs convergent vers des constantes différentes pour chacune des distributions. En fait, comme nous l'avons indiqué dans [16], les valeurs asymptotiques sont égales aux valeurs de l'entropie de Rényi d'ordre $\frac{1}{2}$ de chaque distribution. Ce dernier point sera développé dans le paragraphe suivant.

L'utilisation des MST dans la cadre plus général de la discrimination entre distributions n'est pas nouvelle; on peut par exemple citer les travaux de Hoffmann et Jain [15], qui proposent d'utiliser les MST pour tester la nature aléatoire d'observations dans $\mathbb{R}^2$, ou ceux de Dussert et Rasigni [11] qui appliquent une démarche similaire dans des tests d'ordre ou de désordre en physique de la matière condensée. La prise en compte dans la cadre des problèmes de discrimination de densité, des propriétés des MST comme estimateurs d'entropie ou de divergence informationnelle a été abordée dans [19] et ne sera pas développée dans cet article.

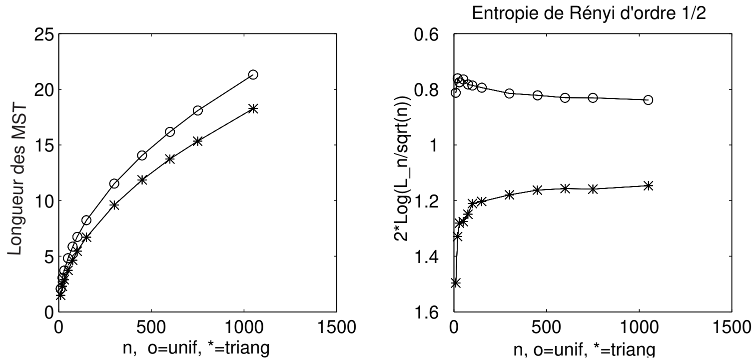


Figure 2: Évolution de la longueur des MST pour des distributions uniformes ou triangulaires, en fonction du nombre de réalisations considéré. À gauche : longueur euclidienne; à droite : entropie de Rényi d'ordre $\frac{1}{2}$.

3.3 Propriétés, estimation d'entropies

Soient $L_{n,\gamma}$ définie comme en (3), fonction quasi-additive euclidienne d'ordre γ, et $\mathcal{X}_n$ un ensemble de réalisations indépendantes du processus aléatoire de densité de Lebesgue $f(x)$, défini sur $\mathbb{R}^d$. Steele [29] a démontré le théorème suivant, en généralisant un résultat établi par Beardwood, Halton et Hammersley [6] :

$$\lim_{n\to\infty} \frac{L_{n,\gamma}(\mathcal{X}_n)}{n^{\frac{d-\gamma}{d}}} = \beta(\gamma, d) \int_{\mathbb{R}^d} f(x)^{\frac{d-\gamma}{d}} dx \qquad (p.s.)$$

où $\beta(\gamma, d)$ est une fonction indépendante de la densité f. A partir de ce résultat, nous avons établi dans [18] la propriété suivante, pour les k-MST :

Soient

- $\nu \in]0,1[$ défini par l'égalité $\nu = (d-\gamma)/d$, $\gamma \in]0,d[$, et

$$\hat{H}_\nu(\mathcal{X}^*_{n,k}) = \frac{1}{1-\nu} \ln \left(n^{-\nu} L_{k,\gamma}(\mathcal{X}^*_{n,k}) \right) + \tilde{\beta}(\nu, d) \tag{5}$$

la statistique construite à partir de la longueur $L_{k,\gamma}(\mathcal{X}^*_{n,k}) = \sum_{e_{i,j} \in \mathcal{T}^\star_{n,k}} |e_{i,j}|^{(1-\nu)d}$, associée au k-MST $\mathcal{T}^\star_{n,k}$.

- $\hat{L}_{k,\gamma}(\mathcal{X}^*_{n,k})$ la valeur approchée de $L_{k,\gamma}(\mathcal{X}^*_{n,k})$, obtenue par l'algorithme d'estimation des k-MST [18]. $\tilde{\beta}(\nu, d)$ s'exprime aisément à partir de $\beta(\gamma, d)$.

Si $k = \lfloor \alpha n \rfloor$, $\alpha \in [0, 1]$, en substituant $\hat{L}_{k,\gamma}(\mathcal{X}^*_{n,k})$ à $L_{k,\gamma}(\mathcal{X}^*_{n,k})$ dans (5), on obtient un estimateur consistant et robuste de l'entropie de Rényi de la distribution de densité f :

$$\lim_{n\to\infty} \hat{H}_\nu(\mathcal{X}^*_{n,k}) = \min_{A:P(A)\geq\alpha} \frac{1}{1-\nu} \ln \int_A f^\nu(x)dx \quad (p.s.) \tag{6}$$

Dans cette expression, la minimisation est conduite sur tous les sous-ensembles boréliens A définis sur $[0, 1]^d$, dont la probabilité $P(A)$ vérifie l'inégalité $P(A) = \int_A f(x)dx \geq \alpha$.

Un certain nombre de propriétés remarquables peuvent être déduites de l'équation (6) :

- La valeur de β dans l'expression (5), est égale à l'entropie de Rényi d'une distribution de densité uniforme sur $[0, 1]^d$. Le terme β n'est par conséquent fonction que de ν et d.

- La variable k qui fixe la taille (en terme de nombre de sommets connectés) du graphe minimal cherché, joue un rôle identique au rôle tenu par le paramètre α dans les estimateurs de moyenne α-tronquée : en présence de points de bruit (outliers), k peut être ajusté de sorte à assurer une certaine robustesse à l'estimateur d'entropie [17, 18].

- L'estimateur de l'entropie de Rényi construit sur les k-MST est un estimateur direct et ne requiert donc pas de devoir estimer la densité f, ce qui est toujours difficile.

- L'estimation de l'entropie de Rényi d'ordre ν quelconque sur l'intervalle $]0, 1[$ s'obtient par modification du paramètre γ, ce dernier pouvant varier continûment sur $]0, d[$.

- La méthode proposée s'étend sans difficulté au problème d'estimation d'autre type d'entropies et donc de I-divergences, par exemple l'entropie structurelle de Havrda-Charvàt, non additive, qui généralise l'entropie de Rényi [14] :

  $$HC_\alpha(f) = \frac{1}{1-\nu}\left[\int f^\nu(x)dx - 1\right]$$

  (Comme l'entropie de Rényi, l'entropie Havrda-Charvàt est concave pour $0 < \nu < 1$ et tend vers l'entropie de Shannon quand $\nu \to 1$.

Il est important de souligner la nature non paramétrique de l'estimateur d'entropie décrit par les équations (5,6). De plus, ces estimateurs ne requièrent pas la définition de paramètres extérieurs (nombre de bins pour une estimation via un histogramme par exemple). Dans la cas de distributions possédant des queues lourdes, l'estimation d'entropie par développement tronqué en moments ou cumulants ne peut être satisfaisante (la contribution des moments d'ordres supérieurs négligés est importante). Au contraire, l'estimation directe de l'entropie par les MST prend en compte, par nature, la totalité des moments de la distribution.

Nous avons développé dans [17] une application reposant sur ces propriétés des MST pour la résolution d'un problème de séparation de mélange de densités du type $f = (1-\varepsilon)f_1 + \varepsilon f_0$ dans le cas où f_0 est une densité uniforme. Nous décrirons rapidement cette étude dans le paragraphe suivant.

4 Deux exemples d'application

4.1 Débruitage - séparation de mélange

Nous avons évoqué, dans les paragraphes précédents, la construction d'estimateurs consistant d'entropies à l'aide des MST. Dans cette application, nous mettons en évidence la sensibilité des MST au bruit. Le bruit, dans ce contexte, se manifeste par la présence de points d'observation à répartition uniforme de densité f_0, se superposant à l'ensemble des observations de densité inconnue $f_1(x)$. On considère donc la densité de mélange suivante :

$$f(x) = (1-\epsilon)f_1(x) + \epsilon f_0(x) \tag{7}$$

Sur l'exemple considéré ici, f_1 est une densité définie sur $\mathbb{R}^2$, associée à une distribution en anneau, et prend la forme :

$$f_1 = c\exp(-\frac{225}{2}(||x-(0.4,0.4)||-0.25)^2)$$

où c est une constante de normalisation, $||x||^2 = ||(x_1,x_2)||^2 = x_1^2 + x_2^2$. Un ensemble de 50 observations associées à cette densité, , contaminé par la présence de 50 points de distribution uniforme, est représenté sur les figures 3 et 4. Le MST (noté 100-MST) construit sur ce mélange de distributions est représenté sur le graphique inférieur droit de la figure 4. Alors que le MST associé au seul ensemble des observations de densité f_1 peut être utilisé pour l'estimation de l'entropie de f_1, le MST construit sur le mélange est sévèrement influencé par la présence de bruit... Les branches du graphe connectant des points associés à f_0 apparaîssent, sur cet exemple, nettement plus longues que les branches connectant entre elles deux réalisations du processus en anneau, de densité f_1. La longueur de MST résultant est largement affectée par la présence des réalisations de densité f_0. L'entropie estimée par cette méthode est l'entropie du mélange, très différente de celle de f_1 seule. D'autre part, cet exemple illustre que lorsque le nombre total N de réalisations (du mélange) est faible, l'importance relative des branches connectant des points de bruit peut être très importante. Nous développons dans la suite quelques idées, dont l'utilisation des k-MSTs, pour rendre robustes les estimateurs précédents.

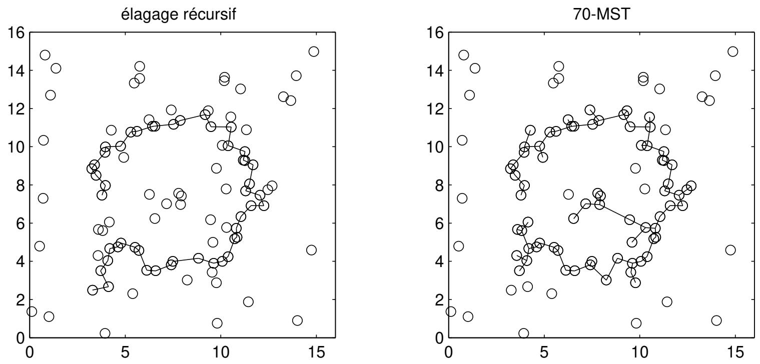


Figure 3: Comparaison des résultats de réjection de bruit (densité uniforme) pour -à gauche : l'algorithme de type "Banks", -à droite : l'algorithme basé sur les k-MST.

Une première solution à ce problème de sensibilité au bruit a été développée par Banks *et al.* [2] dans le contexte de régression non linéaire non paramétrique. Les auteurs y suggèrent de couper les plus grandes branches de l'arbre construit sur le mélange, jusqu'à dégager un "tronc", associé au signal recherché. Présentée par ces auteurs sans justification autre que son efficacité, cette méthode peut aisément être justifiée par nos études, qui replace cette démarche dans le cadre de l'identification de sous-ensembles d'entropie minimale. Sur la partie gauche de la figure 3, nous présentons le résultat obtenu pour le problème de séparation de f_1 et f_0, pour un algorithme dérivé de l'algorithme de Banks, basé sur une approche itérative de type "*cut and merge*" sur le MST du mélange [20].

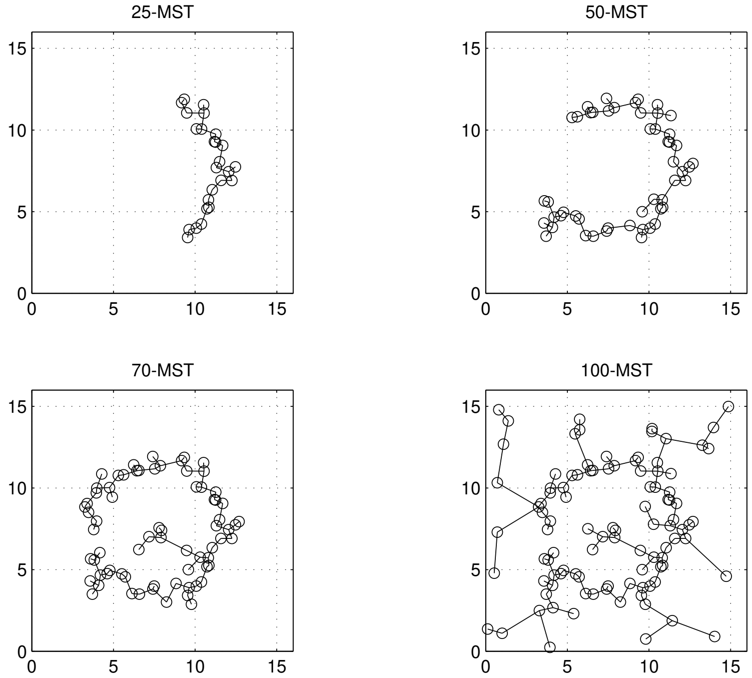


Figure 4: k-MST estimés pour différentes valeurs de k, sur le mélange de densité annulaire - uniforme.

Sur la figure 4 sont présentés les k-MST calculés pour différentes valeurs du paramètre k. Il est évident que lorsque k augmente, de plus en plus de points "appartenant" à la distribution f_1 sont pris en compte. Cependant, si pour $k \geq 50$, une grande part des points de f_1 semble avoir été détectée (au sens où ces points sont des sommets du k-MST estimé), il est difficile de déterminer la valeur de k à choisir pour optimiser la réjection des points de bruit[1]. Un critère simple peut-être construit à partir de la longueur estimée $\hat{L}_{k,\gamma}(\mathcal{X}_{n,k})$ des k-MSTs, en fonction de k. Sur la figure 5, le graphique supérieur gauche représente l'évolution d'une telle fonction pour le mélange de distributions annulaire - uniforme précédent. Pour les faibles valeurs de k, $\hat{L}(\mathcal{X}_{n,k})$ croît presque linéairement avec k : les segments sont tous de longueur faible, presque uniforme (voir figure 6), tant que seuls les points associés à f_1 sont agrégés par le k-MST. Il est intéressant de souligner ici que le comportement linéaire de $\hat{L}(\mathcal{X}_{n,k})$ en fonction de k est démontré dans le cas où n et k tendent vers l'infini (avec $\frac{n}{k} = \alpha$) pour des distributions uniformes sur leur support [18]. L'approximation proposée ici correspond donc à assimiler la distribution f_1 à une distribution uniforme sur un support annulaire, et n grand.

La prise en compte de nouveaux sommets, de probabilité très faible au sens de la densité f_1, implique la présence dans le k-MST de branches de grande longueur; $\hat{L}(\mathcal{X}_{n,k})$ augmente alors beaucoup plus rapidement

[1]L'optimisation est comprise ici au sens où le nombre de points de f_1 est maximal et celui de f_0 minimal dans la liste des sommets du k-MST.

en fonction de k. Nous proposons d'utiliser un critère de sélection de k construit sur la détection de cette rupture de pente de la fonction $\hat{L}_{k',\gamma}(\mathcal{X}_{n,k'}) = h(k'), \;\; k \leq k'$.

Les arguments développés dans le paragraphe précédent conduisent à estimer la valeur de k permettant la meilleure séparation du mélange décrit par l'équation (7), par la valeur maximale k_s telle que pour $k' \leq k_s$, l'hypothèse de linéarité de $h(k)$ soit vérifiée[2]. La difficulté de détection d'une valeur optimale k_s par cette approche est illustrée par l'absence d'une rupture franche dans la courbe $\hat{L}_{k,\gamma=1}$ en fonction de k, presentée dans la vignette supérieure gauche de la figure 5.

Soit l'équation de régression linéaire suivante :

$$L_{k',1} = A(k).k' + l(k) \;, \;\; 2 \leq k' \leq k-1 \tag{8}$$

où $A(k)$ et $l(k)$ minimisent l'erreur quadratique moyenne de régression linéaire jusqu'au $(k-1)$ième point

$$\hat{\sigma}^2_{k-1} = \frac{1}{k-2} \sum_{k'=2}^{k'=k-1} \varepsilon^2(k') \tag{9}$$

dans laquelle

$$\varepsilon(k') = \hat{L}_{k',1} - L_{k',1} \;, \;\; k' \leq k-1.$$

Soit

$$\hat{\varepsilon}(k) = \hat{L}_{k,1} - A(k).k - l(k) \tag{10}$$

l'erreur de prédiction linéaire de $\hat{L}_{k,1}$, à partir du modèle décrit par l'équation (8), identifié sur les valeurs de k' inférieures à k. La vignette supérieure droite de la figure 5 représente l'évolution de la variance de l'erreur de régression définie par l'équation (9) en fonction de k; La variation du carré de l'erreur de prédiction linéaire $\hat{\varepsilon}(k)$ définie par l'équation (10) est illustrée sur la vignette inférieure gauche.

Le graphique de la vignette inférieure droite (figure 5) représente l'évolution du rapport $r(k) = \frac{\hat{\varepsilon}(k)^2}{\hat{\sigma}^2_{k-1}}$ en fonction de k. Ce rapport est classique en statistique, pour les problèmes de test d'hypothèse linéaire généralisée, pour les problèmes de comparaisons multiples [12] ou encore dans les problèmes de détection de rupture [28],[5]. Notons H_0, l'hypothèse selon laquelle le modèle linéaire (8) est vérifié. Sous H_0, la quantité $r(k)$ suit une distribution $\mathcal{F}_{k,1}$ [12] (rapport de deux lois de χ^2, à k et 1 degrés de liberté respectivement). La mise en œuvre, l'étude et la discussion du test de rupture du modèle linéaire ne seront pas abordés dans cet article. On ne retiendra que les heuristiques suivantes sur $r(k)$:

Deux raisons peuvent conduire à de grandes valeurs de $r(k)$

- H_0 est vérifiée et la probabilité a priori d'observer la valeur obtenue de $\varepsilon(k)$ était très faible

[2] l'hypothèse de linéarité de $h(k')$ n'est pas restrictive : un raisonnement identique peut être formulé avec tout autre modèle. Dans le cas où la distribution cherchée est uniforme sur son support, cette hypothèse est naturellement vérifiée.

- H_0 n'est pas vérifiée

Cette deuxième interprétation est retenue, et conduit à estimer $k_s = 70$ sur l'exemple présenté ici.

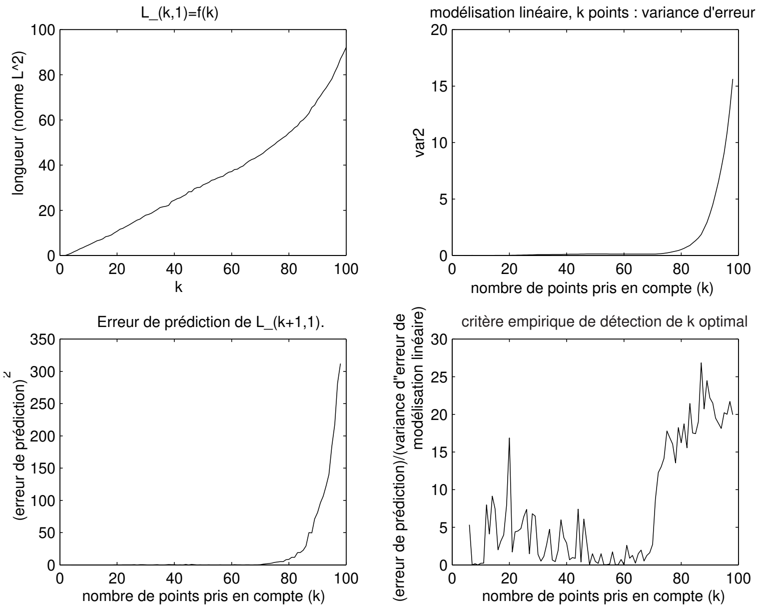


Figure 5: En haut -à gauche : $\hat{L}_{k,\gamma}(\mathcal{X}_{n,k})$ en fonction de k; -à droite : erreur de régression linéaire sur $L(\mathcal{X}_{n,k})$ en fonction de k. En bas-à gauche : variance d'erreur de prédiction de $L(\mathcal{X}_{n,k'})$, pour $k' = k + 1$, sous l'hypothèse de linéarité de $L(\mathcal{X}_{n,k'})$, $k' \leq k$; -à droite : critère de détection empirique de sélection de k (se reporter au texte).

La figure 6 complète cette étude en présentant les histogrammes des longueurs de segments des k-MST pour des valeurs de k identiques à celles utilisées pour la figure 4 : pour $k \geq 70$, les histogrammes mettent nettement en évidence l'existence d'une queue très épaisse dans la distribution des longueurs de segment. Le mode principal correspond aux segments de faible longueur, caractéristiques des connections entre des points de la distribution recherchée. Le second mode, qui compte les segments de longueur plus élevée, souligne la présence dans le k-MST considéré de branches longues connectant la structure en arbre à un point éloigné qui, par conséquent, sera considéré comme du bruit.

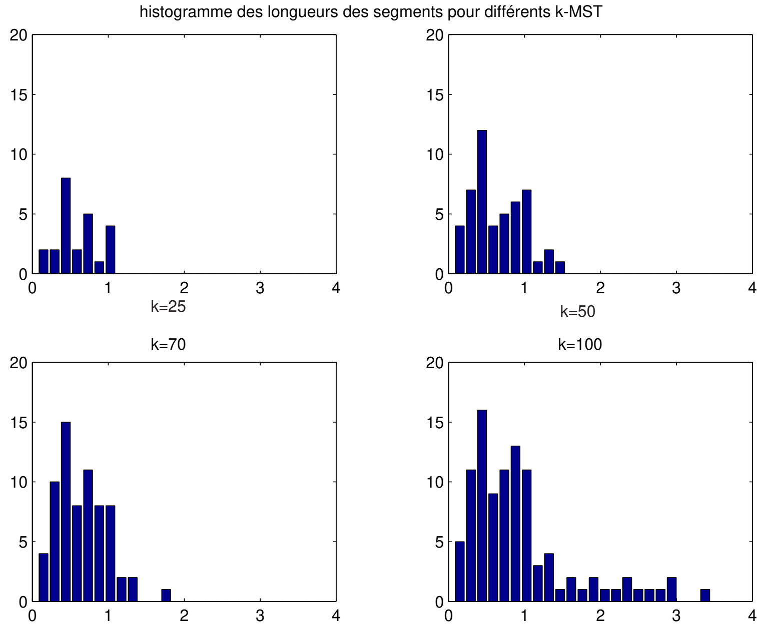


Figure 6: Histogrammes des longueurs de segments des k-MST, pour différentes valeurs de k; les k-MST sont estimés sur le mélange de distributions annulaire - uniforme.

Discussion

- L'exemple précédent a été choisi à dessein : les barycentres des distributions f_1 et f_0 sont en effet confondus, tous deux au centre de l'image présentée. Dans ce cas, les méthodes de clustering usuelles (e.g. la méthode des k-means [?]) qui repose sur l'existence de composantes dont les centres de masses sont distincts, ne peut donner de résultats satisfaisants.

- Dans l'exemple choisi, les supports des distributions f_1 et f_0 se recouvrent; la méthode proposée s'applique dans la mesure où la distance moyenne avec le plus proche voisin chez les points de f_1 est significativement inférieure à la distance moyenne entre un point de f_1 et un point de f_0, ainsi qu'à la distance moyenne entre deux points de f_0. Il s'agit d'une limitation intrinsèque de l'approche proposée. Cependant, lorsque la densité des points de bruit f_0 est comparable à celle des points de signal f_1, le problème de séparation dans le cas de supports non disjoints est mal posé en l'absence de toute autre information a priori sur les distributions. Dans [19], nous proposons une solution à ce problème, dans le cas où l'une des deux distributions f_1 ou f_0 est connue.

4.2 Extraction de trajectoire dans le plan temps-fréquence

Dans cette partie, nous illustrons l'intérêt des approches recourant aux MST ou k-MST pour l'analyse de composantes dans le cas des représentations temps-fréquence. Le point de départ (l'observation) est la donnée d'une représentation temps-fréquence. L'ensemble des maxima relatifs de la distribution est identifié dans une première Ètape. Chacun de ces maxima relatifs peut être considÈrÈ comme une rÈalisation d'un processus alÈatoire tridimensionnel; les variables considérées sont du type $x = [t, \omega, E(t, \omega)]$, où $E(t, \omega) \in \mathbb{R}$ est la valeur prise par la distribution temps-fréquence à la date $t \in \Delta T$ et à la fréquence $\omega \in \Delta\Omega$. Le problËme de dÈtection des composantes est alors reformulÈ comme un problËme de sÈparation de mÈlange $f = (1-\varepsilon)f_1 + \varepsilon f_2$, dans lequel $f_1 = g(x/Bruit)$, $f_2 = g(x/Signal)$ où $g(x/.)$ est la fonction de distribution des maxima de la distribution, conditionnellement à la présence de bruit ou de signal respectivement.

Un problème crucial rencontré dans ce contexte réside dans la définition nécessaire d'une norme dans l'espace $\Delta T \times \Delta\Omega \times \mathbb{R}$. La définition d'une norme dans le seul plan temps-fréquence (TF) doit conduire à une notion de distance qui soit indépendante de l'échantillonnage dans ce dernier : la distance entre deux "paquets" d'énergie ne devrait pas dépendre de la fréquence d'échantillonnage F_e de la série temporelle, ni du nombre de canaux fréquentiels N_b utilisés dans l'estimation de la distribution TF. Cette propriété peut être obtenue en introduisant deux constantes K et K' (toutes deux dimensionnellement homogènes à un temps) et la définition de distance suivante entre deux points $P_i = (t_i, \omega_i)$, $(i \in \{1, 2\})$ du plan TF :

$$D(P_1, P_2) = \sqrt{\left(\frac{t_1 - t_2}{KF_e}\right)^2 + \left(\frac{K'\pi F_e}{N_b}\frac{(\omega_1 - \omega_2)}{2\pi}\right)^2}$$

où $t_i \in [0, N-1]$ et $\omega_i \in [0, N_b - 1]$ sont les indices respectivement en temps et en fréquence des points considérés. Dans la suite, nous nous limiterons au cas $F_e = 1$, $K = 1$, $K' = N_b$ et $N_b = \frac{N}{2}$. La distance $D(P_1, P_2)$ possède alors une interprétation simple : c'est la distance en pixels entre P_1 et P_2 dans une image $N \times N_b$ où chaque pixel représente la valeur de la distribution TF à la date t_i/F_e, à la fréquence $\omega_i.F_e/N_b$. Il est important de noter que cette normalisation est arbitraire, au sens où les poids relatifs des axes temps, fréquence et énergie (exprimés entre autre par les valeurs de K et K') ne prennent de signification réelle qu'à travers la nécessité de rendre lisibles les variations qui nous intéressent dans la distribution temps-fréquence étudiée. Le type de normalisation adopté dépend donc essentiellement de l'application envisagée. La dynamique de la troisième variable, homogène à une énergie, est totalement arbitraire dans la représentation TF. Le problème général de définition d'une norme dans le plan TF est un problème largement ouvert et ne sera pas étudié ici.

Deux approches sont discutÈes sur cet exemple:

• la premiËre est une transposition directe des mÈthodes proposÈes dans des travaux antÈrieurs, en dimension 2 [20]. Cette méthode repose sur un algorithme d'élagage récursif du MST construit dans le plan TF, sur la distribution des maximas relatifs les plus forts. Comme cela a été brièvement décrit au paragraphe précédent, l'algorithme d'élagage utilisé a été proposé par Banks [2] dans un contexte de régression non paramétrique. L'ensemble des maxima les plus forts est déterminé par seuillage sur l'énergie. La recherche deux ensembles de maximas relatifs se traduit dans l'étude de la fonction de distribution de probabilité (fdp) des maximas relatifs par la recherche de deux modes. Le seuil recherché doit séparer au mieux ces deux modes, et correspondre à un minimum relatif de la fdp, ou de manière équivalente à un palier de dérivée nulle dans la fonction de distribution cumulative (fdc). Nous avons recherché en pratique le palier le plus long dans la fdc pour determiner le seuil de réjection. La fdc est estimée numériquement à partir de l'intégration numérique de l'histogramme de la distribution des hauteurs, calculé sur N_q quanta, où $N_q \simeq N_{max}/2$, N_{max} étant le nombre de maxima relatifs détectés (voir figure 7). Bien que donnant des résultats satisfaisants, cette approche frustre requiert la détermination d'un grand nombre de paramètres de réglages, et invite à proposer une approche alternative, décrite dans le paragraphe suivant.

• La seconde approche est appliquée directement en trois dimensions. L'énergie est normalisée de sorte que les dynamiques sur chacun des axes temps, fréquence et énergie soient numériquement identiques. Pratiquement, $K = N$, $K' = \frac{N_b}{N}$. Dans le problème de débruitage, présenté dans la section précédente, nous avons proposé et étudié une solution qui repose directement sur le calcul des k-MST, pour toute valeur possible de k. La figure 8 montre que dans le cas présent, ni les longueurs des k-MST [3] en fonction de k, ni l'entropie de Rényi estimée à partir de ces derniers ne permettent de définir facilement un critère de sélection d'une valeur de k optimale.

Soit $\mathcal{T}^{\star}_{N_{max}}$ le MST construit sur la distribution des maxima relatifs de la distribution TF et $\{e_{i,j}\}$ l'ensemble de ses segments. N_{max} est le nombre de maxima relatifs détectés dans la distribution TF. Soit c une coupure sur $\mathcal{T}^{\star}_{N_{max}}$, définissant deux sous-ensembles de points S_1,S_2. On cherche c tel que

$$c = \mathrm{Arg} \min_{e_{i,j}} \max\{H(S_1), H(S_2)\} \tag{11}$$

où H est une fonction de coût. Si c est la coupure à appliquer pour obtenir deux distributions, sous contraintes de minimalité de l'entropie maximale des distributions résultantes, on choisit pour H l'entropie de Rényi, estimée par les MST. Cette approche reformule le problème de détection de composantes dans le plan TF comme un problème de 'clustering' sur l'ensemble des maxima relatifs. Le critère (11) est largement utilisé en reconnaissance des formes [13], dans les problèmes de séparation de "clusters" de tailles différentes. La quantité c désigne le lien dans $\mathcal{T}^{\star}_{N_{max}}$ qui doit être supprimé pour définir deux sous-ensembles ("clusters"). Celle-ci étant déterminée, les MST bidimensionnels sont alors construits sur chacun de ces sous-ensembles (voir figure 9), ce qui permet une meilleure représentation de la trajectoire temps-fréquence de la composante

[3] L'algorithme d'estimation des k-MST, décrit dans [18] fournit une approximation (le degré d'approximation étant borné)de la solution exacte, dont la détermination est un problème NP complet. Nécessairement $L_{k,\gamma} < L_{k+1,\gamma}$, mais l'existence d'erreurs d'estimation $(\hat{L}_{k,\gamma} - L_{k,\gamma})$ montre que cette inégalité peut n'être pas vérifiée pour les k-MST estimés.

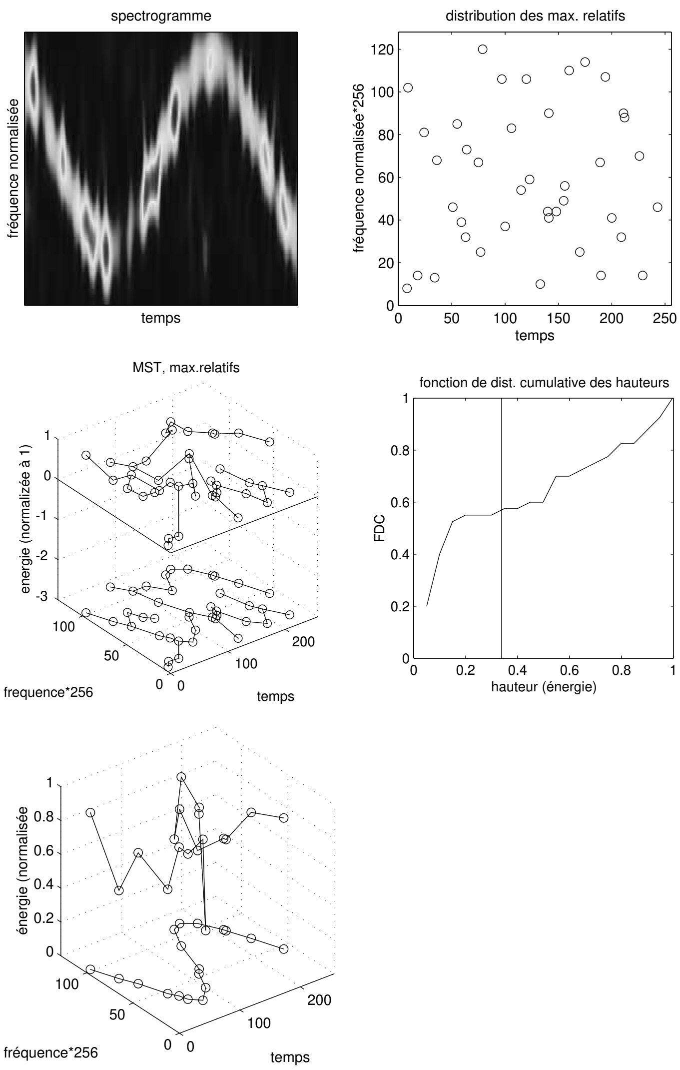


Figure 7: Extraction automatique d'une modulation sinusoïdale de fréquence (RSB=5dB); En haut, à gauche: spectrogramme. En haut, à droite : carte des maxima locaux. Centre, à gauche : MST 3-D de la carte des maxima locaux et projection 2-D (dans le plan z=-3). Centre, à droite: Fonction de distribution cumulative (fdc) des valeurs des maxima locaux et position du seuil détecté. En bas, à gauche : Structure extraite par seuillage de la fdc: MST 3-D et MST 2-D (dans le plan z=0) après élagage.

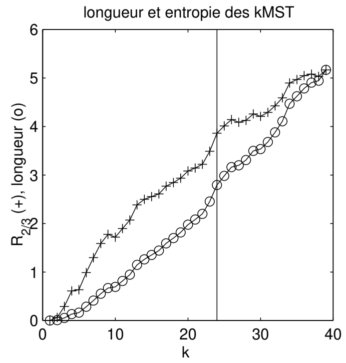


Figure 8: Sur ce graphique sont superposées deux courbes : la fonction cumulative de la distribution des hauteurs des maxima relatifs (o)(l'échelle est modifiée à des fins de représentation), et la longueur des k-MST estimés en fonction de k (+), pour la distribution des maxima relatifs considérés comme des variables définies dans $\mathbb{R}^3$. Aucune de ces courbes ne permet de définir aisément un seuil. (Le seuil obtenu sur la fdc des hauteurs de maxima relatifs est représenté pour information)

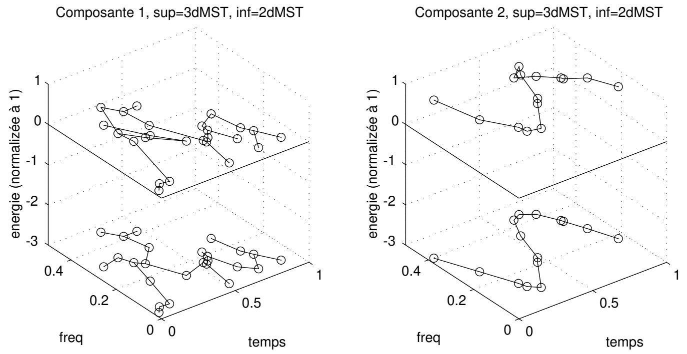


Figure 9: Graphiques supérieurs : représentation dans $\mathbb{R}^3$ des composantes identifiées par le critère de séparation entropique défini dans le texte. Graphiques inférieurs : MST calculés dans le seul plan temps-fréquence (la composante 'energie' est négligée), pour chacune des composantes identifiées.

cherchée. L'extension de cette approche à la séparation de composantes multiples est obtenu par la recherche de plusieurs coupures sur $\mathcal{T}^{\star}_{N_{max}}$ [22]. Notons que dans le cas présenté, l'utilisation de la norme euclidienne usuelle ($\gamma = 1$, cf sections précédentes) dans un espace de dimension $d = 3$ conduit à utiliser comme fonction de coût H, l'entropie de Rényi d'ordre $2/3$. La détermination du meilleur ordre à utiliser dans les problèmes de discrimination est, à notre connaissance, un problème totalement ouvert. Les performances de cette méthode dépendent du seul choix de la dynamique sur les trois axes de la représentation TF. Contrairement à la première démarche proposée, peu de paramètres de réglages sont nécessaires.

5 Conclusion

Nous avons montré comment les estimateurs d'entropie et de divergence informationnelle construits sur la base d'outils issus de la théorie des graphes, ici les MST, permettent de proposer des méthodes robustes de débruitage, de segmentation ou de séparation de mélange. L'intérêt majeur de cette approche tient à la fois à l'existence d'algorithme rapides d'estimation ou d'approximation des MST et k-MST, et au fait qu'il n'est fait usage d'aucun modèle paramètrique. Les exemples présentés montrent comment ces méthodes permettent de proposer une alternative aux outils issus du traitement d'image dans le cadre de la détection de composante ou de trajectoire dans le plan temps fréquence, s'appuyant sur la mise en oeuvre de fonctions de coût entropique. D'autres extensions de ces approches fondées sur la théorie des graphes doivent être envisagées, tant pour aborder des problèmes plutôt théoriques d'estimations de distances ou de divergence informationnelles entre distributions de probabilités, que pour s'interesser à des applications plus concrètes, telles que le recalage d'images par exemple.

References

[1] A.E.Badel, O.Michel, A.O.Hero : "Comparaisons de systèmes et arbres de régression", *Traitement du Signal*, **15-2** 1998, pp 103-118.

[2] D. Banks, M. Lavine, and H. J. Newton, "The minimal spanning tree for nonparametric regression and structure discovery," in *Computing Science and Statistics. Proceedings of the 24th Symposium on the Interface*, H. J. Newton, editor, pp. 370–374, 1992.

[3] M.Basseville,"Distances measures for signal processing and pattern recognition", *Signal Processing*, vol. 18, 1989, pp. 349-369.

[4] M.Basseville, "Information : entropies, divergences et moyennes,"IRISA, publication interne no. 1020, mai 1996.

[5] M.Bassevile, I.V.Nikivirov, "Detection of abrupt changes - Theory and applications.", Prentice Hall information and System Sciences Series, 1993.

[6] J. Beardwood, J. H. Halton, and J. M. Hammersley, "The shortest path through many points," *Proc. Cambridge Philosophical Society*, vol. 55, pp. 299–327, 1959.

[7] J. Beirlant, E. J. Dudewica, L. Györfi, and E. van der Meulen, "Nonparametric entropy estimation: an overview," *Intern. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.

[8] F. Chapeau-Blondeau, F. Janez, J.L. Ferrier, "A dynamic adaptative relaxation scheme applied to the euclidean Steiner minimal tree problem.", *SIAM J. Optim.*, vol. 7, no. 4, pp. 1037–1053, nov. 1997.

[9] C. Chiang, M. Sarrafzadeh, and C. K. Wong, "Powerful global router: Based on Steiner min-max trees," in *IEEE International Conference on Computer-Aided Design*, pp. 2–5, Santa Clara, CA, 1989.

[10] I. Csiszár and J. Korner, *Information theory: coding theorems for discrete memoryless systems*, Academic Press, Orlando FL, 1981.

[11] C. Dussert, G. Rasigni, J. Palmari, and A. Llebaria, "Minimal spanning tree: a new approach for studying order and disorder," *Phys. Rev. B*, vol. 34, no. 5, pp. 3528–3531, 1986.

[12] T.S. Ferguson, "Mathematical Statistics - A decision theoretic approach.", Academic Press, Orlando FL, 1967.

[13] K. Fukunaga, "Statistical pattern recognition.", Academic Press, San Diego CA, 1990.

[14] J. Havrda, F. Chàrvat, "Quantification methods of classification processes," *KIBERNETIKA CISLO 1, ROCNIK 3* pp. 30–34, 1967.

[15] R. Hoffman and A. K. Jain, "A test of randomness based on the minimal spanning tree," *Pattern Recognition Letters*, vol. 1, pp. 175–180, 1983.

[16] A.O.Hero, O.Michel : "Robust estimation of point process intensity features using k-minimal spanning tree.", proc. of ISIT, International Symposium on Information theory, 1997, Ulm, Germany, pp.74.

[17] A.O.Hero, O.Michel :"Robust entropy estimation strategies based on edge weighted random graphs (with connections).", SPIE International Symposium on Optical Science, Engineering and Instrumentation, July 1998, San Diego, USA.

[18] A.O.Hero, O.Michel : "Asymptotic theory of greedy approximations to minimal K-point random graphs.", *IEEE Trans. on Information theory*, vol.45, No.6, pp.1921-1938, septembre 1999.

[19] A.O.Hero, O.Michel : "Estimation of Rényi information divergence via pruned minimal spanning trees.", proc. of IEEE Workshop on HOS, Caesarea, Israel, PP.264-268, 1999.

[20] O.Michel, A.O.Hero : "Pruned MST's for entropy estimation and outlier rejection.", IEEE-IT workshop on DECI, Detection, Classification and Imaging, Santa-Fe, NM, USA., Feb 99. **Communication invitée**.

[21] O.Michel, P.Flandrin and A.O.Hero :"Détection de structures dans le plantemps-fréquence à l'aide de graphes minimaux.", Gretsi.99, Vannes, France, pp.713-716.

[22] O.Michel, P.Flandrin and A.O.Hero :"Automatic extraction of time-frequency skeletons with minimal spanning trees.", accepté ICASSP'2000, Istambul, Turquie.

[23] R. Ravi, M. Marathe, D. Rosenkrantz, and S. Ravi, "Spanning trees short or small," in *Proc. 5th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 546–555, Arlington, VA, 1994.

[24] R. Ravi, M. Marathe, D. Rosenkrantz, and S. Ravi, "Spanning trees – short or small," *SIAM Journal on Discrete Math*, vol. 9, pp. 178–200, 1996.

[25] C. Redmond and J. E. Yukich, "Limit theorems and rates of convergence for Euclidean functionals," *Ann. Applied Probab.*, vol. 4, no. 4, pp. 1057–1073, 1994.

[26] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, volume 1, pp. 547–561, 1961.

[27] M.I. Shamos, D. Hoey, "Closest-point problems," in *Proc. 16th Annual Symp. on Foundations of computer sciences*, pp. 151–162, 1975.

[28] D. Siegmund, "Sequential analysis: tests and confidence intervals.", Springer Verlag , New-York, 1985.

[29] J. M. Steele, "Growth rates of euclidean minimal spanning trees with power weighted edges," *Ann. Probab.*, vol. 16, pp. 1767–1787, 1988.

[30] P. Viola and W. Wells, "Alignment by maximization of mutual information," in *Proc. of 5th Int. Conf. on Computer Vision, MIT*, volume 1, pp. 16–23, 1995.

[31] A. A. Zelikovsky and D. D. Lozevanu, "Minimal and bounded trees," in *Proc. of Tezele Congres XVIII Acad. Romano-Americaine*, pp. 25–26, Kishinev, 1993.