MST et divergences informationelles : applications

Olivier Michel♦, Alfred Hero□, Patrick Flandrin♦

♦ Laboratoire de Physique (URA 1325 CNRS), École Normale Supérieure de Lyon,
46 allée d'Italie, 69364 Lyon Cedex 07, France ; tel : 04 72 72 85 59 - fax : 04 72 72 80 80

□ Department of EECS, University of Michigan, Ann Arbor, MI 48109-2122, USA

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

February 11, 2000

Catégorie : recherche.

Résumé

Mots clés : Minimal spanning Tree - f-divergences - Entropie statistique de Rényi - Séparation de mélange - Débruitage - Temps fréquence.

Abstract

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

1 Introduction

Dans [13], 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 R^d, vérifiant la propriété de quasi-additivité de Redmond et Yukich [18], 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 [7, 17], dans le célèbre problème du voyageur de commerce ou dans la construction des arbres de Steiner, qui consiste à trouver le trajet minimum permettant de visiter k villes parmi n [?], les tests sur la nature aléatoire de champs de données [10], et de manière générale dans les problèmes d'optimisation combinatoire. L'algorithme d'approximation des sous graphes minimaux contenant k points parmi N ($k < N$) présenté dans [12] généralise l'approche proposée par Ravi et al. [16] dans le cas $d = 2$ et a permis de proposer un estimateur robuste de l'entropie d'une distribution d-dimensionelle 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 2 a pour objet d'introduire les concepts de bases sur les MST ainsi que les principaux résultats théoriques obtenus dans les articles [12, 13]. Nous proposons 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. Dans une dernière partie, nous montrons comment ces estimateurs permettent, à partir de transformations opérées sur les données observées, d'en estimer la divergence informationnelle par rapport à une fonction de densité de probabilité (fdp) multidimensionnelle de référence, sans recourir à l'estimation de la fdp observée et dans un cadre non paramètrique.

2 Entropies et divergence de Rényi

Soit $X_n = \{x_1, x_2, \ldots, x_n\}$ une réalisation d'un processus aléatoire i.i.d. défini sur R^d, de densité de Lebesgue multivariée $f(x_i)$ dont le support est limité à $[0,1]^d$. L'entropie de Rényi d'ordre ν de ce processus s'exprime [19]

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

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 [8] (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 Cherno 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 (p.p.). 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 proportionelle au logarithme de la distance de Hellinger

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

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 :

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

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, 8]. 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 estimation apparaît encore dans les problèmes de recalage d'image; dans ce contexte, la I-divergence est directement reliée à l'information mutuelle entre deux images f et f_0 [21]. Pour une revue complète sur les problèmes d'estimation d'entropie et de divergence d'information, on pourra se reporter à [6] 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) est un graphe (ou arbre) T_n connectant l'ensemble des réalisations $X_n = \{x_1, x_2, \ldots, x_n\}$ d'un processus ponctuel défini dans 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 :

$$L_{n,\gamma} = \sum_{e_{i,j} \in T_n} |e_{i,j}|^\gamma$$

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

$$T_n^* = \mathrm{Arg}\min_{T_n} L_{n,\gamma}$$

Ce dernier peut être calculé de façon exacte à partir d'algorithmes dont le coût varie comme $n \log n$. Cette définition est étendue aux sous graphes ne connectant qu'un sous ensemble de points dans 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 $X_{n,k}$ de 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 :

$$X_{n,k}^* = \mathrm{Arg}\min_{i_1,\ldots,i_k} \mathrm{Arg}\min_{T_n} L_{n,k,\gamma}$$

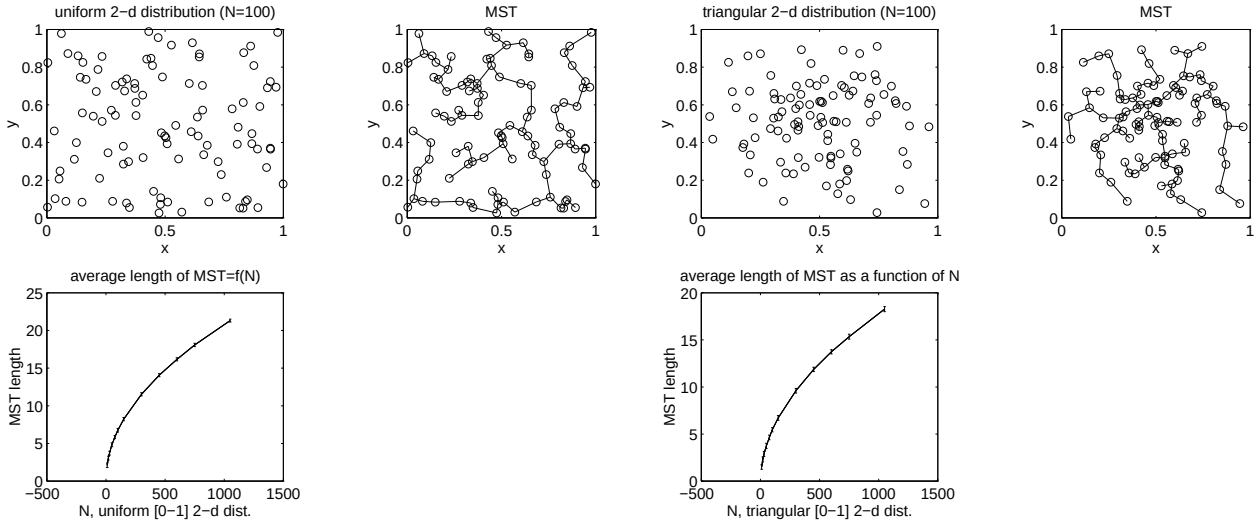
où $X_{n,k} = \{x_{i_1}, \ldots, x_{i_n}\}$. En pratique, la double minimisation est conduite conjointement; c'est le cas en particulier des algorithmes que nous avons développés [12, 13]. Il a été démontré que le problème

d'estimation d'un k-MST dans R^2 est un problème NP-complet [16, 22]. Ravi et al ont proposé un algorithme d'approximations à coût polynômial dans le cas de distributions bidimensionnelles. Dans [13], 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 $O(k^{\frac{d-1}{d}^2})$. 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 [13] proposé en annexe. Cette partie , très technique ne sera pas développée dans cet article.

Exemples

Les figures 1-a et 1-b représentent un exemple qui illustre l'intérêt des MST dans la cadre des problèmes de discrimination entre deux distributions. L'utilisation des MST dans le contexte de discrimination entre distributions n'est pas nouvelle; on peut par exemple citer les travaux de Ho□mann et Jain [10], qui proposent d'utiliser les MST pour tester la nature aléatoire d'observations dans R^2, ou ceux de Dussert et Rasigni [9] qui appliquent une démarche similaire dans des tests d'ordre ou de désordre en physique de la matière condensée.

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 MSTs 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 $\nu = \frac{1}{2}$.



(a) (b)

Figure 1: Distribution pour N = 100 réalisations de la variable aléatoire, MST, et longueur des MSTs 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 gauche 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 [11], les valeurs asymptotiques sont égales aux valeurs de l'entropie de Rényi d'ordre $\frac{1}{2}$ de chaque distribution.

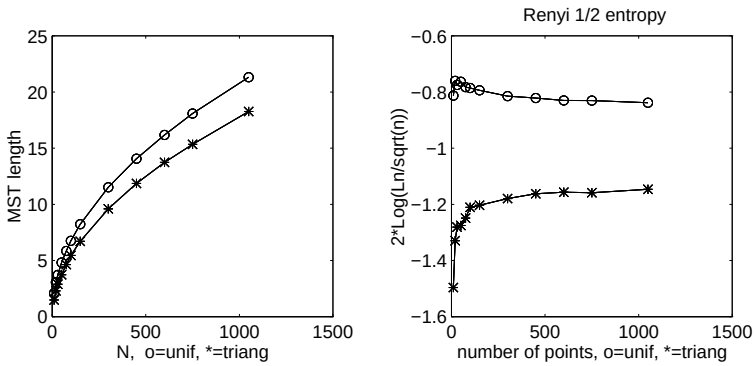


Figure 2: Évolution de la longueur des MSTs pour des distributions uniformes ou triangulaires, en fonction du nombre de réalisations considéré. À gauche : longueur euclidienne; à droite : entropie de Rényi d'ordre $\frac{1}{2}$.

3.2 Propriétés, estimation d'entropies

Soient L définie auparavant, fonction quasi-additive euclidienne d'ordre γ, et X_n un ensemble de réalisation indépendantes du processus aléatoire de densité de Lebesgue f(x), défini sur R^d. Steele [20] a démontré le théorème suivant, en généralisant un résultat établi par Beardwood, Halton et Hammersley [5] :

$$\lim_{n \to \infty} \frac{L(X_n)}{n^{\frac{d-\gamma}{d}}} = \beta(\gamma, d) \int_{R^d} f(x)^{\frac{d-\gamma}{d}} dx \quad \text{(p.s.)}$$

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

$$\hat{H}_\nu(X_{n,k}) = \frac{1}{1-\nu} \ln \left(n^{-\nu} L(X_{n,k})^\gamma \right) + \beta(\nu, d) \tag{3}$$

la statistique construite à partir de la longueur $L^\gamma(X_{n,k}) = \sum_{e_{i,j} \in T_{n,k}} |e_{i,j}|^{(1-\nu)d}$, associée au k-MST $T_{n,k}$.

Nous avons établi dans [13] la propriété suivante :

Soit $\hat{L}(X_{n,k})$ la valeur approchée de $L(X_{n,k})$, obtenue par l'algorithme d'estimation des k-MST [13]. Si $k = \alpha n$, $\alpha \in [0,1]$, en substituant $\hat{L}(X_{n,k})$ à $L(X_{n,k})$ dans (3), on obtient un estimateur consistant et robuste de l'entropie de Rényi de la distribution de densité f :

$$\hat{H}_\nu(X_{n,k}) \to \min_{A : P(A) \geq \alpha} \frac{1}{1-\nu} \ln \int_A f^\nu(x) dx \quad \text{(p.s.)}$$

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 $P(A) = \int_A f(x)dx \geq \alpha$.

De cette expression, on déduit un certain nombre de propriétés remarquables.

- La valeur de β dans l'expression 3, est égale à l'entropie de Rényi d'une distribution de densité uniforme sur $[0,1]^d$. 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 [12, 13].
- 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.

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

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

Nous avons appliqué ce résultat dans [12] 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

Débruitage - séparation de mélange

Nous avons évoqué, dans les paragraphes précédents, la construction d'estimateurs consistant d'entropies à l'aide des MSTs. Dans cette application, nous mettons en évidence la sensibilité des MSTs 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 $f_1(x)$, étudié. On considère donc la densité de mélange suivante :

$$f(x) = (1-\varepsilon)f_1(x) + \varepsilon f_0(x)$$

Sur l'exemple considéré ici, f_1 est une densité définie sur 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, est représenté sur le graphique inférieur droit de la figure 3, contaminé par la présence de 50 points de distribution uniforme. Le MST construit sur ce mélange de distributions est représenté sur ce même graphique. 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 robuste les estimateurs précédents.

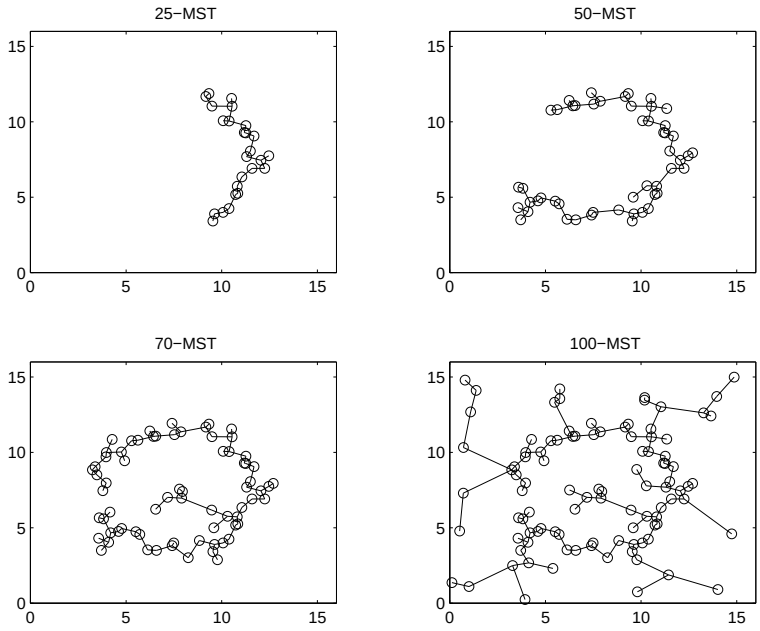


Figure 3: k-MST estimés pour différentes valeurs de k, sur le mélange de densité annulaire - uniforme.

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-paramètrique. 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 4, 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 [14].

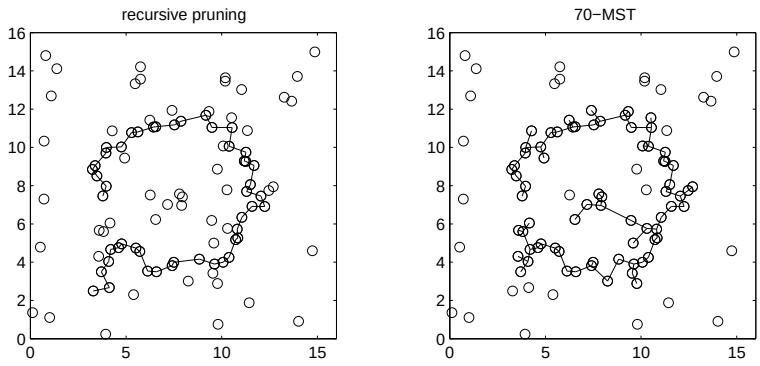


Figure 4: Comparaison des résultats de réjection de bruit (densité uniforme) pour -à gauche : l'algorithme de type "Banks", -à droite : l'algorithme basé sur les k-MSTs.

Sur la figure 3 sont présentés les k-MSTs calculés pour di□érentes valeurs du paramètre k. Il est évident que lorsque k augmente, de plus en plus de points "extérieurs" à la distribution f_1 sont rejetés. 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 di□cile 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 $\hat{L}(X_{n,k})$ 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, $\hat{L}(X_{n,k})$ 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. 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; $\hat{L}(X_{n,k})$ augmente alors beaucoup plus rapidement en fonction de k. Nous utilisons un critère de sélection de k construit sur la détection de cette rupture de pente. Sur la figure 5, nous présentons quelques exemples d'erreurs de prédiction ou de régression en fonction du paramètre k, dans le cadre de l'approximation linéaire de la fonction $\hat{L}(X_{n,k□}) = h(k□), k \leq k□,$. Nous postulons que la valeur k permettant la meilleure réjection de bruit, est la valeur maximale de k telle que pour $k□ \leq k$, l'hypothèse de linéarité de h(k) est vérifiée. 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 3.

[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

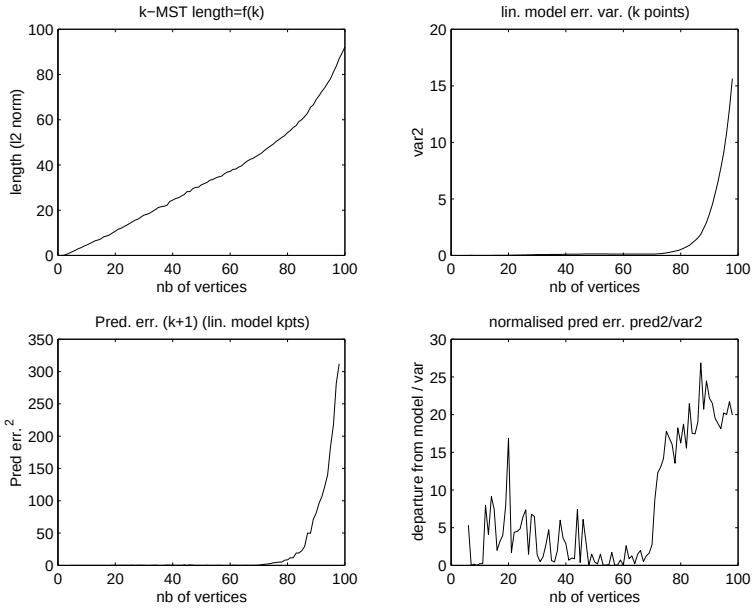


Figure 5: En haut -à gauche : $L(X_{n,k})$ en fonction de k; -à droite : erreur de régression linéaire sur $L(X_{n,k'})$ en fonction de k', pour $k' \leq k$. En bas-à gauche : variance d'erreur de prédiction de $L(X_{n,k'})$, pour $k' = k+1$, sous l'hypothèse de linéarité de $L(X_{n,k'})$, $k' \leq k$; -à droite : critère de détection construit sur l'erreur de prédiction, à partir du rapport défini par l'inégalité de Bienaymé-Tchébiche

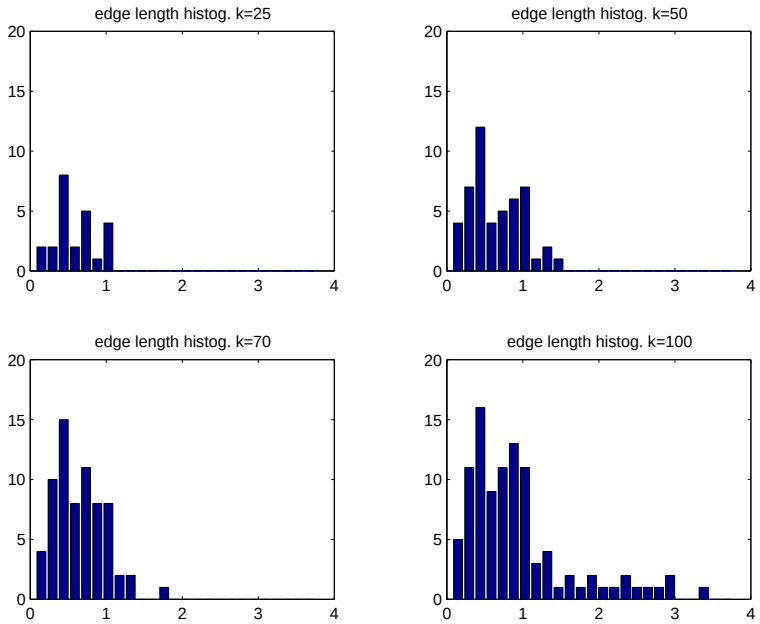


Figure 6: Histogrammes des longueurs de segments des k-MST, pour di□érentes valeurs de k; les k-MST sont estimés sur le mélange de distributions annulaire - uniforme.

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. L'ensemble des maxima relatifs de la distribution est identifié dans une première étape. Chacun de ces maxima relatifs peut être considéré comme une réalisation d'un processus aléatoire tridimensionnel; les variables considérées sont du type $x = [t, \nu, E(t,\nu)]$, où $E(t,\nu) \in R$ est la valeur prise par la distribution temps-fréquence à la date $t \in \Delta T$ et à la fréquence $\nu \in \Delta F$. Le problème de détection des composantes est alors reformulé comme un problème de séparation de mélange $f = (1-\varepsilon)f_1 + \varepsilon f_2$, 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 $\Delta T \times \Delta F \times R$. 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 déchantillonnage F_e de la série temporelle , ni du nombre de bins fréquentiels N_b utilisés dans l'estimation de la distribution TF. Cette propriété peut être obtenu en introduisant deux constantes K et $K' \in$ (dimensionellement homgène à un temps) et la définition de distance suivante entre 2 points $P_i = (t_i, \nu_i)$, $(i \in \{1,2\})$ du plan TF :

$$D_{12} = \sqrt{\left(\frac{t_1 - t_2}{K F_e}\right)^2 + \left(\frac{K' F_e}{2N_b}(\nu_1 - \nu_2)\right)^2}$$

Dans la suite, nous ne considérerons que le cas $F_e = 1$ et $K = K' = 1$ et $N_b = \frac{N}{2}$. 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 discutées dans cette étude; la première est une transposition directe des méthodes proposées dans des travaux antérieurs, en dimension deux [14]. 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. L'algorithme d'élagage utilisé a été proposé par Banks [2] dans un contexte de régression non paramètrique.

L'ensemble des maxima les plus forts est déterminé par seuillage sur l'energie. Le seuil de réjection est fixé par un critère de détection de rupture de la dérivée seconde de la fonction de distribution cumulative hauteurs des maxima (figure (7)).

La seconde approche présentée est appliquée directement en trois dimension. L'energie est normalisée de sorte que les dynamiques sur chacun des axes temps, fréquence et énergie sont numériquement identiques. Soit T_n^* le MST construit sur la distribution des maxima relatifs de la distribution TF et $\{e_{i,j}\}$ l'ensemble

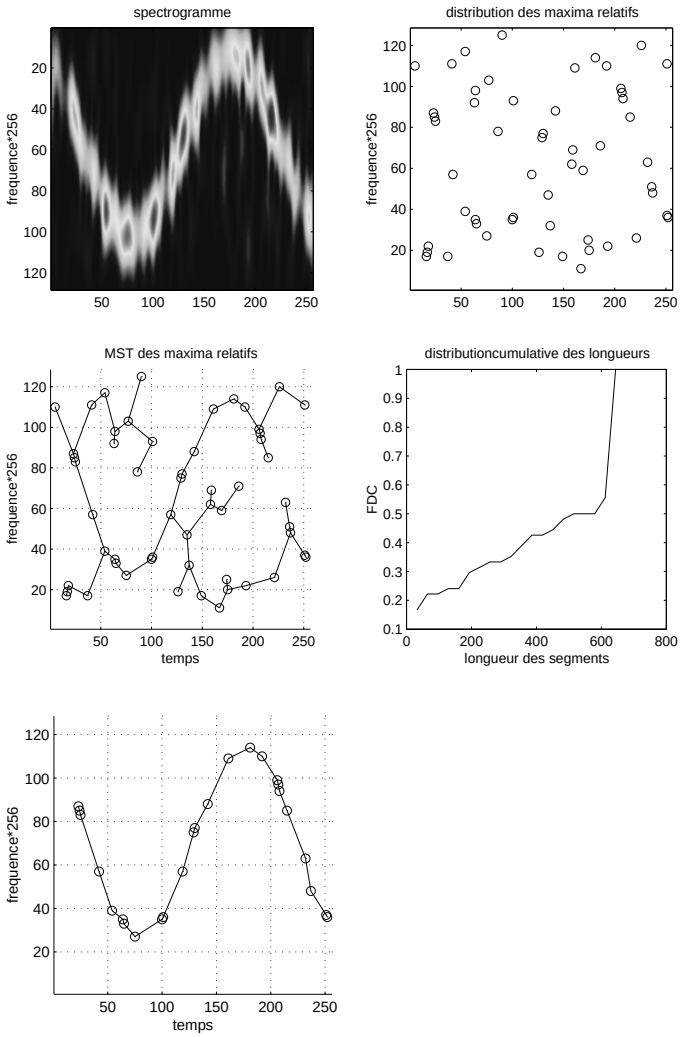


Figure 7: Extraction automatique d'une modulation sinusoïdale de fréquence (RSB=5dB); En haut, à gauche: spectrogramme. En haut, à droite : carte des maxima locaux. Centre, à gauche : Projection du MST 3-D de la carte des maxima locaux. Centre, à droite: Fonction de distribution cumulative des longueurs (fdc) de segments du MST 3-D. En bas, à gauche : Structure extraite par seuillage de la fdc, et élagage.

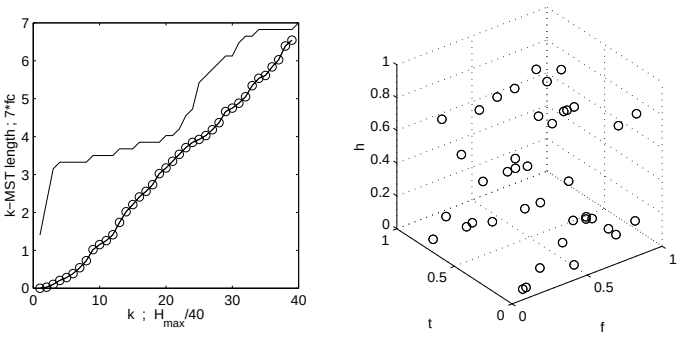


Figure 8: Sur le graphique droit sont superposées deux courbes : la fonction cumulative de la distribution des hauteurs des maxima relatifs (l'échelle est modifiée à des fins de représentation), et la longueurs des k-MST en fonction de k, pour la distribution des maxima relatifs considérés comme des variables définies dans R^3, distribution représentée sur le graphique de gauche. Aucune de ces courbes ne permet de définir aisemment un seuil.

de ses segments. Soit alors c une coupure sur ce MST, définissant deux sous-ensemble de points S_1,S_2. On cherche c tel que

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

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 choisi 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 'clusering' sur l'ensemble des maxima relatifs. Les MST bidimensionnels sont alors appliqués sur chacun de ces sous ensembles (voir figure (9)). On peut noter que dans le cas présenté, l'utilisation de la norme euclidienne usuelle ($\gamma = 1$, cf sections précédentes) dasn 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.

La figure (8), montre que dans le cas étudié, ni les longueurs des k-MST [14], ni la fonction de distribution cumulative des hauteurs des maxima ne permettent de séparer les maxima de signal des maxima de bruit.

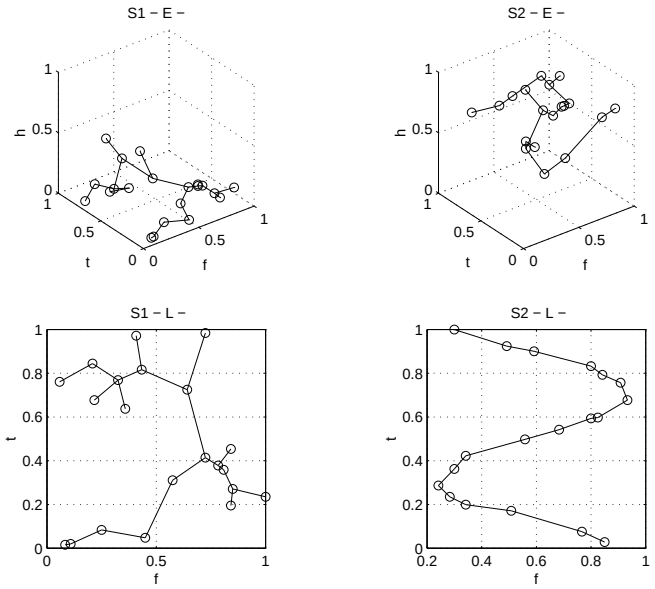


Figure 9: Graphiques supérieurs : représentation dans R^3 des composantes identifiées S1 et S2, par le critère de séparation entropique défini dans le texte. Graphiques inférieurs : MST calculés dans le seul plan temps-fréquence (la composante 'energie' est négligée), pour chacune des composantes identifiées S1 et S2.

5 Estimation de I-divergences

Il est montré dans cette section que la I-divergence entre deux densités f et g peut être obtenue par changement de variables, permettant de passer de la densité g à la densité f. Sous des hypothèses assez générales, cela permet d'utiliser des algorithmes reposant sur la mise en oeuvre des k-MST dans un cadre de classification de densités.

5.1 Analyse et résultat

Soit $g(x)$ une densité de référence définie sur $\mathbf{R}^d$, qui domine la densité $f(x)$, $x = [x^1, \ldots, x^d]^T$, au sens où pour tout x tel que $g(x) = 0$, on a $f(x) = 0$. Pour tout x vérifiant $g(x) > 0$, on factorise $g(x)$ sous la forme $g(x) = g(x^1)g(x^2|x^1)\ldots g(x^d|x^{d-1}, \ldots, x^1)$ où $g(x^k|x^{k-1}, \ldots, x^1)$ est la densité conditionnelle de la k-ième composante, associée à $g(x)$. Dans la suite, l'ensemble $\{x : g(x) = 0\}$ est négligé. Sur cet ensemble, $f(x) = 0$, et donc ce dernier est de probabilité nulle. On construit alors le vecteur $y = [y^1, \ldots, y^d]^T \in \mathbf{R}^d$ par la transformation suivante :

$$\begin{aligned} y^1 &= G(x^1) \\ y^2 &= G(x^2|x^1) \\ &\vdots \quad \vdots \end{aligned} \tag{4}$$

$$y^d = G(x^d|x^{d-1},\ldots,x^1)$$

où

$$G(x^k|x^{k-1},\ldots,x^1) = \int_{-\infty}^{x^k} g(\tilde{x}^k|x^{k-1},\ldots,x^1)d\tilde{x}^k$$

est la fonction de distribution cumulative de la k-ième composante. C'est une fonction monotone croissante, sauf sur les ensembles de probabilité nulle $\{x : g(x) = 0\}$, de probabilités nulles. Par conséquent, sauf sur ces derniers ensembles, la fonction de distribution conditionnelle admet une fonction inverse

$$x^k = G^{-1}(y^k|x^{k-1},\ldots,x^1) = G^{-1}(y^k|y^{k-1},\ldots,y^1)$$

On établit (à partir de la formule du Jacobien classique dans le cadre des changements de variable) que la fonction de densité conjointe du vecteur y ainsi calculée, h(y), s'exprime :

$$h(y) = \frac{f(G^{-1}(y^1),\ldots,G^{-1}(y^d|y^{d-1},\ldots,y^1))}{g(G^{-1}(y^1),\ldots,G^{-1}(y^d|y^{d-1},\ldots,y^1))} \quad (5)$$

Soit $\hat{L}(Y_{n,k}^*)$ la longueur du k-MST obtenu par l'algorithme d'approximation proposé dans [13], appliqué aux variables aléatoires y, et $Y_{n,k}^*$ l'ensemble des k points connectés par ce k-MST approché. Alors, d'après les résultats du paragraphe précédent, on a la propriété suivante :

$$\hat{H}_\nu(Y_{n,k}^*) \to \frac{1}{1-\nu} \ln \int h^\nu(y)dy \quad \text{(a.s.)} \quad (6)$$

Si la transformation inverse $y \to x$ spécifiée par (4) est appliquée dans l'intégrale précédente, en remarquant que par la formule du Jacobien $dy = g(x)dx$, et en utilisant l'équation (5) pour h, l'intégrale dans le membre de droite de l'équation (6) s'exprime comme la I-divergence de Rényi entre f(x) et g(x) :

$$\frac{1}{1-\nu} \ln \int h^\nu(y)dy = \frac{1}{1-\nu} \ln \int \left(\frac{f(x)}{g(x)}\right)^\nu g(x)dx.$$

$\hat{H}_\nu(Y_{n,k}^*)$ est par conséquent un estimateur consistent de la I-divergence de Rényi.

Les résultats établis dans [13] se généralisent donc sans difficulté à l'estimation des I-divergences. Les calculs rapidement présentés dans cette section montrent que cette nouvelle approche utilisant la théorie des graphes peut être appliquée dans le contexte de problèmes de classification par rapport à une distribution de référence arbitraire f_0, et non uniquement par rapport à la distribution uniforme comme nous l'avons proposé dans [12].

5.2 Application

Soit la densité f définie par un mélange de densités bidimensionnelles (définies sur R^2)

$$f = (1-\varepsilon)f_1 + \varepsilon f_0, \quad (7)$$

où $f_1(x) = (\frac{1}{2} - |x^1 - \frac{1}{2}|)(\frac{1}{2} - |x^2 - \frac{1}{2}|)$ est une densité triangulaire séparable et $f_0 = 1$ représente la densité uniforme; toutes deux sont définies sur le support $x = (x^1, x^2) \in [0,1]^2$. Le paramètre de mélange α varie sur l'intervalle $[0,1]$.

Dans chacune des simulations présentées, 256 réalisations de la variable aléatoire x, de densité f, ont été considérées. Les I-divergences de Rényi $I(f, f_0)$ et $I(f, f_1)$ sont estimées respectivement par $\hat{H}_\nu(X_n)$ et $\hat{H}_\nu(Y_n)$, pour le cas $\nu = \frac{1}{2}$ ($\gamma = 1$) dans la construction du k-MST. L'ensemble Y_n est obtenu par application de la transformation $y = (y^1, y^2) = (F_1(x^1), F_1(x^2))$ de l'ensemble des réalisations échantillonnées X_n; $F_1(u)$ est la fonction de distribution cumulative marginale associée à la densité triangulaire.

Lors d'une première série de simulations, les valeurs estimées $\hat{H}_\nu(X_n)$ et $\hat{H}_\nu(Y_n)$ des I-divergences de Rényi $I(f, f_0)$ et $I(f, f_1)$, sont comparées à un seuil permettant de tester l'hypothèse $H_0 : \alpha = 0$ contre l'hypothèse $H_1 : \alpha \neq 0$, et l'hypothèse $H_0 : \alpha = 1$ contre l'hypothèse $H_1 : \alpha \neq 1$, respectivement. Les "Caractéristiques opérationnelles de réception" -les courbes COR- sont présentées sur les figures 10-a et 10-b. Comme on pouvait le prévoir, on observe bien que les performances de détection sont d'autant plus grandes que les hypothèses H_0 et H_1 sont bien séparées.

Dans une seconde série d'expériences, nous avons retenu deux réalisations différentes du mélange (7) de deux densités, l'une uniforme et l'autre triangulaire. Le premier mélange considéré correspond à $\alpha = 0.1$, le second correspond au mélange "symétrique", $\alpha = .9$. Lorsque $\alpha = 0.1$, la densité dominante est la densité triangulaire; au contraire, dans le cas $\alpha = .9$, la densité f résultant du mélange est dominée par la densité uniforme.

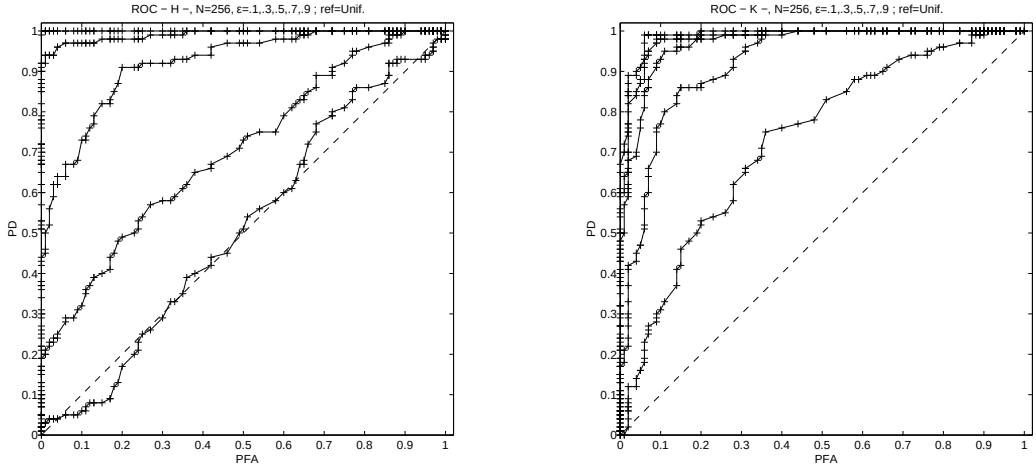
Pour chacune de ces situations, nous avons estimé le k-MST approché (avec $k/n = 0.9$), par l'algorithme de Ravi [16]. Le graphe issu de cette procédure d'estimation est utilisé pour réaliser une segmentation de l'ensemble des observations en deux sous ensembles : les points connectés par le graphe estimé d'une part, et d'autre part les points non connectés au graphe. L'objet de cette segmentation est de sélectionner les réalisations issues de la densité dominante.

D'un point de vue pratique, dans le cas $\alpha = 0.1$ pour lequel la densité dominante est la densité triangulaire, l'algorithme de segmentation est appliqué directement sur les données X_n, alors que dans le cas $\alpha = .9$, la segmentation par le k-MST est calculée sur les données transformées Y_n.

Cela peut être justifié rapidement par l'argument suivant : la longueur du k-MST est une fonction croissante de l'entropie de la distribution des sommets connectés qui forment le graphe. La recherche du graphe minimal (au sens de sa longueur) est donc équivalente à la recherche du sous ensemble d'entropie minimal, dans l'ensemble de départ. Si $\alpha = 0.1$ la distribution dominante -triangulaire- est la distribution de plus faible entropie dans le mélange. La méthode segmentation proposée est menée dans l'espace des observations

directes X_n. Dans le cas $\alpha = .9$, le k-MST tend à connecter entre eux les points de la distribution d'entropie minimale. Sur X_n, ce sont encore les points associés à la distribution triangulaire. Dans l'espace transformé Y_n, les observations qui sont issues de la densité triangulaire ont désormais une densité uniforme... d'entropie maximale. Peut importe qu'il soit ou non possible de déterminer analytiquement la nouvelle densité des observations, de densité uniforme de X_n, cette dernière est de toute façon d'entropie minimale dans l'espace transformé. Le k-MST cherche donc à connecter les observations qui, dans l'espace direct, ont une densité uniforme.

Les quantités déduites des k-MST, $\hat{H}_\nu(X_{n,k})$ et $\hat{H}_\nu(Y_{n,k})$ peuvent être interprétées comme des estimateurs robustes des divergences d'information de Rényi $I(f_1, f_0)$ et $I(f_0, f_1)$ respectivement. Ces résultats sont illustrés sur les figures 11, et 12; ces figures illustrent le potentiel de telles approches dans le cadre de la segmentation ou de la réjection de bruit (conduisant à la présence de points supplémentaires non désirables dans l'espace des réalisations).



(a) (b)

Figure 10: (a) : Courbes COR de l'algorithme détection du mélange de densités triangulaire-uniforme $f = (1 - \alpha)f_1 + \alpha f_0$ (H_1) contre l'hypothèse $f = f_0$ (H_0). Les différentes courbes correspondent à différentes valeurs de α, variant de $\alpha = .1$ (courbe supérieure), à $\alpha = .9$ (courbe inférieure). (b) : Courbes COR de l'algorithme détection du mélange de densités triangulaire-uniforme$f = (1 - \alpha)f_1 + \alpha f_0$ (H_1) contre l'hypothèse $f = f_1$ (H_0). Les différentes courbes correspondent à différentes valeurs de α, variant de $\alpha = .1$ (courbe inférieure), à $\alpha = .9$ (courbe supérieure).

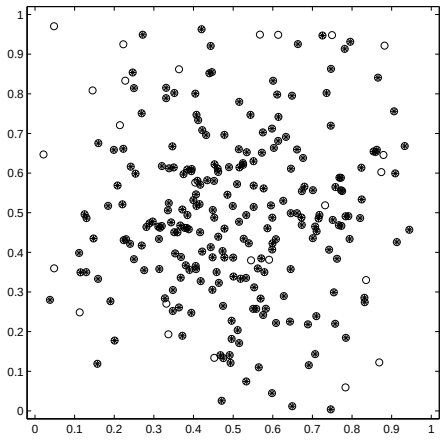
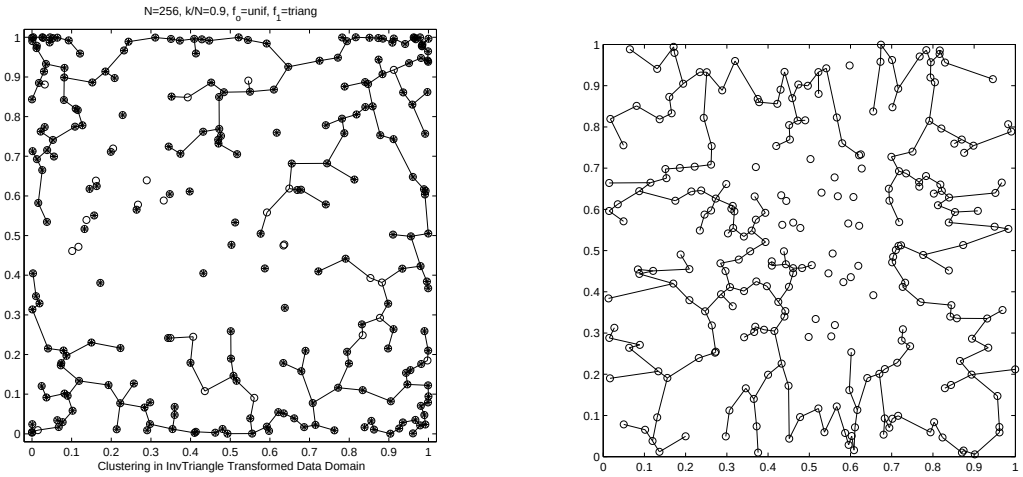


Figure 11: Échantillon de 256 réalisations de la variable $x \in R^2$, de densité $f = (1 - \varepsilon)f_1 + \varepsilon f_0$, pour $\varepsilon = .1$. Les réalisations issues de la densité triangulaire sont marquées du symbole '*', celles issues de la densité uniforme sont identifiées par 'o'. Le k-MST obtenu pour $k = 230 \approx .9 * 256$ est superposé au plan des observations.



(a) (b)

Figure 12: (a) : Échantillon de 256 réalisations de la variable transformée $y = G(x)$, $x \in R^2$; G est ici l'inverse de la fonction de distribution cumulative de la densité triangulaire. x a pour de densité $f = (1 - \varepsilon)f_1 + \varepsilon f_0$, pour $\varepsilon = .9$. L'espace représenté est l'espace transformé. Les réalisations issues de la densité triangulaire sont marquées du symbole 'o', celles issues de la densité uniforme sont identifiées par '*'. Le k-MST obtenu, pour $k = 230 \approx .9 * 256$ est superposé au plan des observations. (b) : Même figure que la figure (a), cette fois l'espace de représentation est l'espace direct. (le k-MST a été calculé dans l'espace transformé, c'est le même que sur la figure refUnifTrig.

6 Conclusion

Les estimateurs d'entropie et de divergence informationnelle construits sur la base d'outils issus de la théorie des graphes permettent de proposer des méthodes robustes débruitage, de segmentation ou de séparation de mélange. 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. Enfin, l'exploitation des relations fortes qui existent entre entropie et graphe minimaux a conduit à proposer un estimateur non paramètrique de divergence informationelle, et de construire un test statistique de détection dans un mélange de processus aléatoires multi-variés.

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 et al., Fiches descriptives d'algorithmes de segmentation de signaux, Traitement du Signal, Vol.9, No.1, 1992.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

[13] A.O.Hero, O.Michel : "Asymptotic theory of greedy approximations to minimal K-point random graphs.", accepté pour publication dans IEEE Trans. on Information Theory, 1999.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.