Observations sur les distributions latentes aux matrices laplaciennes de graphes

Pierrick Bruneau
Benoît Otjacques

Luxembourg Institute of Science and Technology
5, avenue des Hauts-Fourneaux
L-4362 Esch-sur-Alzette
prénom.nom@list.lu

Résumé. L'algorithme de clustering spectral permet en principe d'extraire des clusters de formes arbitraires à partir de données numériques. Cette propriété a contribué à sa popularité, et même si ses bases théoriques sont établies depuis plus d'une décennie, des variantes en ont été proposées jusqu'à récemment. Son fonctionnement repose sur une transformation vers un espace latent dans lequel des formes de clusters arbitraires sont converties en structures faciles à traiter par un algorithme tel que k-means. Toutefois, les distributions dans cet espace latent n'ont été que peu discutées, beaucoup d'auteurs supposant que les propriétés prédites par la théorie sont vérifiées. Cet article propose alternativement une approche qualitative pour vérifier si cette structure idéale est effectivement obtenue en pratique. Le travail consiste également à identifier les paramètres de variabilité commandant à la transformation vers l'espace latent, via un état de l'art synthétique de la théorie sous-jacente au clustering spectral. Les observations tirées de nos expériences permettent d'identifier les combinaisons de paramètres efficaces, et les conditions de cette efficacité.

1 Introduction

L'algorithme de clustering spectral (Shi et Malik, 2000; Ng et al., 2002) se distingue des algorithmes à base de centroides (k-means, EM (Bishop, 2006)) principalement par sa capacité à extraire des clusters non convexes et sphériques.

Tous les algorithmes de clustering spectral de la litérature se basent sur l'analyse d'une matrice de similarité entre éléments d'un jeu de données, pouvant être vue comme un graphe pondéré. Les variantes diffèrent ensuite selon qu'elles étudient l'évolution de cette matrice selon un processus stochastique alternant expansion et inflation (van Dongen, 2008; Liu et al., 2013), ou qu'elles analysent les vecteurs propres du laplacien du graphe (Ng et al., 2002; Lin et Cohen, 2010). Dans ce dernier cas, sur lequel nous nous concentrons dans cet article, le reste de l'algorithme se résume essentiellement à appliquer l'algorithme k-means à une représentation extraite des vecteurs propres. La distribution des données dans cette représentation est donc essentielle au bon fonctionnement de l'algorithme de clustering spectral.

À notre connaissance, la distribution dans cet espace induit a été rarement discutée. Le cas échéant, seuls des cas très simples permettant de mettre en valeur les propriétés idéales de cet espace ont été discutés (e.g., matrices creuses avec similarités binaires, ou données tirées depuis un mélange de loi normales uni-dimensionnelles (Fowlkes et al., 2004; von Luxburg, 2007). Les études et variantes consécutives parues dans la littérature ne font soit pas du tout référence à la distribution de ces vecteurs, soit supposent que la structure idéale est vérifiée (Yan et al., 2009).

Alternativement, nous proposons une approche qualitative ayant pour objectif la vérification effective de cette structure idéale. Dans la section 2, une synthèse de l'état de l'art nous permet d'introduire le formalisme nécéssaire à l'étude, et d'identifier les paramètres commandant à la construction des vecteurs propres. À cette occasion, nous définissons un quantile au paramètre de largeur de bande, comblant ainsi une lacune de cet état de l'art formel.

Dans la section 3, des jeux de données illustrant une variété des difficultés pouvant être rencontrées en pratique sont introduits, et l'influence des paramètres identifiés en section 2 est évaluée de manière qualitative pour chacun d'eux. La section 4 synthétise nos observations.

2 Clustering spectral et graphes

2.1 Matrice de similarité

Soit un jeu de données numérique $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. En général, chacun des N éléments comporte un nombre identique de coordonnées, i.e., $\mathbf{x}_n = \{x_1, \ldots x_d\}$. d est alors appelé *dimension* du jeu de données dans son *espace initial*. Les éléments sont classiquement stockés en lignes d'une matrice, faisant de $\mathbf{X}$ une matrice à N lignes et d colonnes.

Alternativement on peut considérer que les N éléments sont les noeuds d'un graphe, dont les arêtes sont pondérées par les similarités entre paires d'éléments. Le graphe est matérialisé par la matrice $\mathbf{S}$, avec $\mathbf{S}_{nn'} \in [0, 1]$ la similarité entre les éléments n et n'.

Ce graphe peut être construit de diverses manières : par exemple, (von Luxburg, 2007) construit le *graphe k-NN* en fixant $\mathbf{S}_{nn'} = 1$ ssi les éléments n et n' appartiennent à leurs k plus proches voisins mutuels, 0 sinon. Dans cet article nous utilisons la fonction RBF (*Radial Basis Function* (Nabney, 2002)) :

$$\mathbf{S}_{nn'} = \exp\left(-\frac{||\mathbf{x}_n - \mathbf{x}_{n'}||^2}{\sigma^2}\right). \tag{1}$$

2.2 Largeur de bande de la fonction radiale

La fonction RBF comporte un paramètre libre, σ dans l'expression (1). Dans la littérature d'estimation empirique des densités ce paramètre est couramment nommé *largeur de bande* (Sheather et Jones, 1991), dans la mesure où il commande la largeur d'une gaussienne non-normalisée. Nous proposons d'affecter un quantile de la distribution empirique des distances euclidiennes à σ :

$$\mathbb{S} = \{||\mathbf{x}_n - \mathbf{x}_{n'}|| : n \in 1, \ldots, N-1 \wedge n' \in n+1, \ldots, N\} \tag{2}$$

$$\sigma_\alpha = \inf\{x \in \mathbb{S} : \alpha < F_{\mathbb{S}}(x)\}, \tag{3}$$

avec $F_{\mathbb{S}}$ la fonction de distribution cumulée empirique des normes, et σ_α le quantile à α (e.g., 10%) de cette distribution. En pratique, plus ce quantile est faible, plus la fonction RBF associée ressemblera à un pic, avec la fonction de Dirac bornée à 1 comme limite. La figure 1 montre les distributions de similarités obtenues avec la fonction RBF appliquée au jeu de données *synth1* (cf. figure 2) pour deux valeurs de quantiles. Avec un quantile faible la majorité des similarités devient nulle, rendant le graphe associé plus épars, comme l'est par construction un graphe k-NN. La seule différence est alors que les similarités non-nulles le sont sur le continuum $]0, 1]$. En augmentant le quantile, on tend vers un graphe complet pondéré.

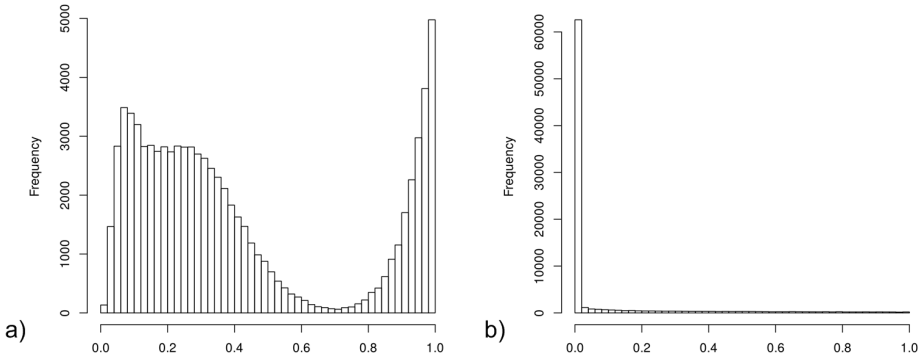

FIG. 1 – *Distribution de similarités obtenue* a) *avec* $s_{40\%}$, b) *avec* $s_{10\%}$.

(Zelnik-Manor et Perona, 2004) ont proposé un paramètre de largeur de bande local, en prenant le quantile pour chaque élément du jeu de données séparément, i.e., selon la distribution de $\mathbb{S}_n$ (cf. équation (2)) spécifique à chaque élément. Ces quantiles sont combinés dans une fonction RBF modifiée :

$$\mathbf{S}_{nn'} = \exp\left(-\frac{||\mathbf{x}_n - \mathbf{x}_{n'}||^2}{\sigma_{n,\alpha}\sigma_{n',\alpha}}\right), \tag{4}$$

avec $\sigma_{n,\alpha}$ le quantile à α spécifique à l'élément n. Par opposition, le paramètre obtenu selon l'équation (3) est dit de *largeur de bande global*. En fait, (Zelnik-Manor et Perona, 2004) (respectivement (Karatzoglou et al., 2004)) prennent déjà en quelque sorte un quantile en affectant à σ_n la distance à son $7^{\text{ème}}$ (respectivement $4^{\text{ème}}$) plus proche voisin. Toutefois cette définition dépend a priori de la taille du jeu de données considéré - ainsi tous les jeux de données considérés dans (Zelnik-Manor et Perona, 2004) ont $N \simeq 300$. On notera que dans ce cas, la distance au $7^{\text{ème}}$ élément est approximativement équivalente à $\sigma_{n,2\%}$.

2.3 Matrice laplacienne et espace latent

Soit $\mathbf{D}$ la matrice diagonale telle que $\mathbf{D}_{nn} = \sum_{n'=1}^{N} \mathbf{S}_{nn'}$. La matrice $\mathbf{D}$ comporte donc les *degrés* des éléments sur sa diagonale. Dans la littérature, la matrice $\mathbf{L} = \mathbf{D} - \mathbf{S}$ est alors appelée *laplacien* du graphe sous-jacent à $\mathbf{S}$. Le laplacien normalisé est formé par $\mathbf{L}_{\text{norm}} =$

$\mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}$. Contrairement au laplacien non-normalisé, son comportement a été établi comme consistant selon la théorie de la relaxation (Guattery et Miller, 1998).

Par la suite, nous considérons la décomposition en valeurs propres du laplacien, établie comme $\mathbf{L} = \mathbf{U}\Lambda\mathbf{U}^T$. Les colonnes de la matrice $\mathbf{U}$ sont les vecteurs propres de $\mathbf{L}$. (von Luxburg, 2007) a alors montré que le nombre de clusters idéal K est lié à la multiplicité de la valeur propre 0, et que cette dernière est associée aux vecteurs propres mineurs de $\mathbf{U}$ (i.e., ses dernières colonnes). Ce théorème est essentiellement justifié en termes de coupe minimale du graphe associé, permettant ainsi d'établir des clusters de noeuds.

Toutefois, comme la plupart des implémentations de clustering spectral basées sur l'analyse des vecteurs propres du laplacien, dans cet article nous décomposons plutôt $\mathbf{S}$ (ou $\mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}$ selon la normalisation voulue). En effet, le nombre de clusters idéal K est alors associé à la multiplicité de la valeur propre 1, elle-même associée aux K vecteurs propres majeurs, identiques au signe près aux K vecteurs propres mineurs de $\mathbf{L}$ (Fowlkes et al., 2004). L'extraction des vecteurs propres est alors plus stable numériquement. Il devient également possible d'utiliser des algorithmes de décomposition partielle efficaces (e.g., Lanczos), procédant en général dans l'ordre décroissant des valeurs propres (Baglama et Reichel, 2006).

Dans cet article, nous laissons sciemment de côté la détermination automatique de K, qui constitue un problème à part entière, traité par ailleurs dans la littérature (cf. e.g., (Zelnik-Manor et Perona, 2004; Bruneau et al., 2014)). Soit $\mathbb{1}_k$ le vecteur indicateur du $k^{\text{ème}}$ cluster, i.e., $\mathbb{1}_{kn} = 1$ ssi l'élément n appartient au $k^{\text{ème}}$ cluster, 0 sinon. Les propositions 2 et 4 dans (von Luxburg, 2007) peuvent alors être résumées :

Proposition 1

- *l'image des K premiers vecteurs propres du laplacien non-normalisé a pour base l'ensemble de vecteurs $\mathbb{1}_k, k \in \{1 \dots K\}$,*
- *l'image des K premiers vecteurs propres du laplacien normalisé a pour base l'ensemble de vecteurs $\mathbf{D}^{-1/2}\mathbb{1}_k$.*

En d'autres termes, les K premières colonnes de $\mathbf{U}$ sont nécessairement une combinaison linéaire de ces indicateurs. Nous appelons *vecteur des coordonnées latentes de l'élément n* les $n^{\text{èmes}}$ coordonnées des K premiers vecteurs propres (i.e., la $n^{\text{ème}}$ ligne de $\mathbf{U}$ restreinte à ses K premières colonnes). L'espace dans lequel ces vecteurs de coordonnées prennent leurs valeurs est alors appelé *espace latent* du laplacien. Cette dénomination est à contraster avec l'*espace initial* dans lequel sont représentées les données, sur lequel a été appliqué la fonction RBF entre paires d'éléments.

L'intuition du clustering spectral est de formaliser les clusters en termes de voisinages dans un graphe, permettant de transformer l'espace initial en une représentation simplifiée selon la proposition 1. Les éléments prennent alors en théorie une position distincte parmi K dans l'espace latent, facilitant grandement leur traitement par un algorithme tel que k-means. De plus, la dimensionalité des données d devient implicite aux similarités, et la complexité de k-means passe de $O(d)$ à $O(K)$, classiquement faible dans les applications du clustering.

La proposition 1 précise que dans le cas du laplacien normalisé les coordonnées non-nulles de la base de l'espace propre sont pondérées selon les degrés des éléments correspondants : en d'autres termes, les clusters sont alors concentrés sur des variétés linéaires plutôt que sur des positions discrètes. Ce phénomène peut être problématique si l'algorithme k-means est appliqué sur cette représentation latente.

Les algorithmes basés sur le laplacien normalisé proposent ainsi de normaliser les vecteurs de coordonnées (i.e., par $\sqrt{\sum_{k=1}^{K} \mathbf{U}_{nk}^2}$), permettant de concentrer les clusters en des positions discrètes dans l'espace latent (Ng et al., 2002; von Luxburg, 2007).

Dans le cas limite où $K = 1$, la proposition 1 admet par ailleurs le corollaire suivant de manière triviale :

Corollaire 1 *Si $K = 1$:*
- *le premier vecteur propre du laplacien non-normalisé est proportionnel à $\mathbb{1}$,*
- *le premier vecteur propre du laplacien normalisé est proportionnel à $\mathbf{D}^{-1/2}\mathbb{1}$.*

3 Etude de cas

La section 2 a permis d'identifier les paramètres influançant la construction de la représentation latente des données, finalement traitée par l'algorithme k-means :
- normalisation du laplacien,
- normalisation des vecteurs de coordonnées latentes,
- largeur de bande locale ou globale,
- quantile de la largeur de bande.

Dans un esprit pragmatique, nous proposons d'évaluer qualitativement l'influence de ces paramètres. Dans la littérature la structure idéale est généralement obtenue pour des jeux de données présentant des clusters convexes, sphériques et sans bruit de fond. Un ensemble de jeux de données artificiels de tailles modestes sera utilisé ici pour incarner des motifs souvent rencontrés à des degrés divers dans des jeux de données réels (voir Figure 2) :
- *synth1* : le cas simple, avec 4 clusters sphériques de taille égale (100 éléments dans chacun). Chaque cluster est généré par une distribution normale 2D isotropique,
- *synth2* : comporte un cluster non-convexe (300 éléments),
- *synth3* : comporte deux clusters très compacts superposés à du bruit de fond (300 éléments),
- *synth4* : comporte des clusters imbriqués (300 éléments).

Pour chaque jeu de données, nous allons estimer l'influence des paramètres sur la distribution obtenue dans l'espace latent. En particulier, nous allons relever dans quelle mesure elle dévie de la structure idéale identifiée en section 2.3. Ces observations seront effectuées de manière qualitative, en utilisant des matrices de nuages de points comme support visuel. Bien qu'inconnues au calcul des représentations latentes, les classes réelles serviront à colorer leurs élément respectifs dans les nuages de points, ainsi qu'à faciliter la conjecture de propriétés dans l'espace latent.

3.1 Résultats sur les données *synth1*

L'essentiel des observations relatives à *synth1* dans cette section a déjà été partiellement fait dans (von Luxburg, 2007, Figure 1). Nous avons étendu l'analyse à un jeu de données 2D, ainsi qu'à la prise en compte explicite des paramètres identifiés dans la section 2.3.

En utilisant la largeur de bande globale et un quantile faible (i.e., jusqu'à 10%), la distribution dans l'espace lentent montrée sur la figure 3a est obtenue. En décomposant le laplacien

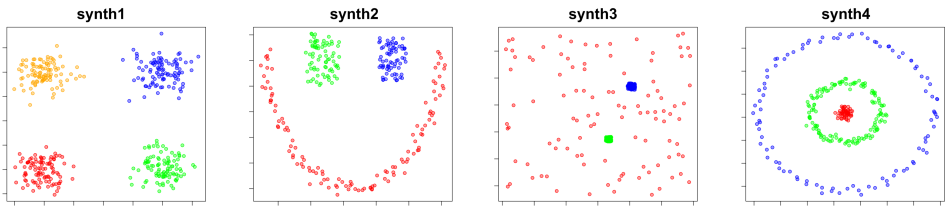

FIG. 2 – *Jeux de données synthétiques supportant l'évaluation qualitative. Les points sont colorés selon leur classe.*

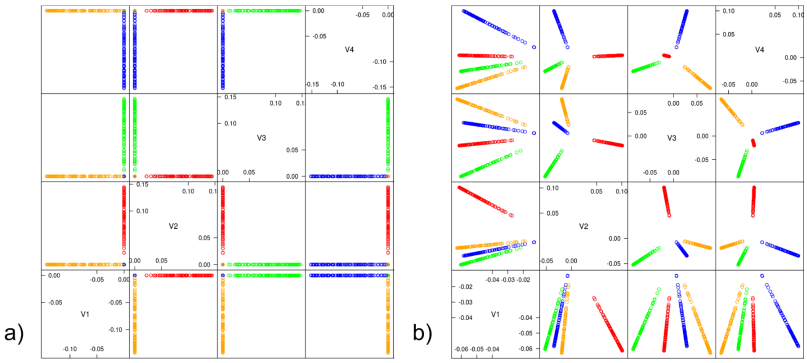

FIG. 3 – *Représentation latente au laplacien de* synth1 *avec le paramètre de largeur de bande global et le quantile à 10% :* a) *laplacien non-normalisé,* b) *laplacien normalisé.*

non-normalisé, toutes les dimensions de l'espace latent sauf une sont nulles pour chaque élément du jeu de données, comme attendu selon la proposition 1 - à ceci près que les valeurs non-nulles s'échellonnent sur un continuum dans la figure 3a. Ceci est vraisemblablement un effet du passage des similarités binaires (e.g., graphe k-NN) aux graphes creux pondérés dont un exemple d'histogramme est donné en figure 1b.

Selon la proposition 1 un résultat assez similaire était attendu pour le laplacien normalisé. En particulier, les coordonnées latentes sont alors proportionnelles à la racine inverse du degré des noeuds correspondants ce qui explicite le continuum observé. Les variétés supportant les clusters dans la figure 3b suivent une rotation, ce qui est acceptable dans la mesure où les vecteurs propres sont définis à une rotation près.

Cependant, le premier vecteur propre ne semble alors pas du tout caractéristique des clusters, ce qui semble indiquer une combinaison uniforme de tous les vecteurs indicateurs (voir proposition 1). Cette observation devient valable quelque soit la normalisation du laplacien en prenant un quantile plus élevé (cf. figure 4a). Ce phénomène peut être vu comme une transition vers le corrolaire 1 : un tel premier vecteur propre est caractéristique d'un seul cluster, alors que les $K - 1$ vecteurs propres restants restent conformes à la proposition 1.

Certaines implémentations (e.g., (Fowlkes et al., 2004; Karatzoglou et al., 2004)) exploitent cette observation en excluant le premier vecteur propre de la représentation latente. Nous avons

vu que cela doit être évité avec le laplacien non-normalisé et un quantile faible, car la première coordonnée latente contient alors de l'information utile. Cette perte est d'autant plus manifeste en normalisant les vecteurs de coordonnées latentes. Alors que cette opération permet de concentrer les variétés observées en figure 3b, rendant la représentation latente plus facile à traiter pour l'algorithme k-means, la structure de clusters peut aussi être complètement perdue (cf. figure 4b).

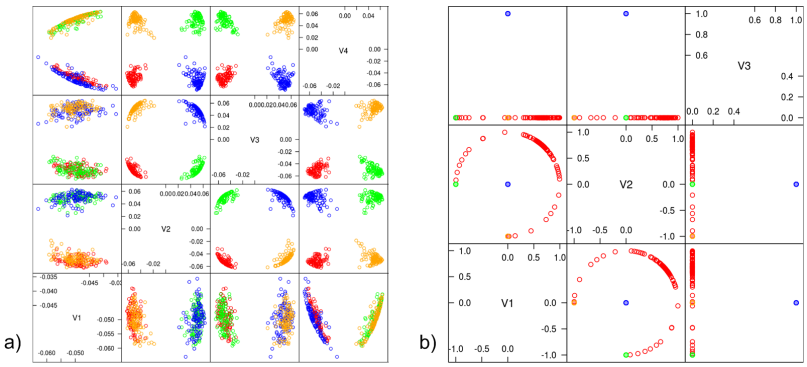

FIG. 4 – a) *: représentation latente au laplacien non-normalisé de* synth1 *avec la largeur de bande globale et le quantile à 40%.* b) *pour le quantile à 10%, effet de la normalisation des vecteurs de coordonnées après exclusion du premier vecteur propre.*

L'utilisation de la largeur de bande locale amène des conclusions similaires pour tous les quantiles considérés dans cette section. Seul le quantile à 2% se distingue : les coordonnées latentes sont alors considérablement dégradées avec la largeur de bande globale. En revanche un tel quantile est proche des recommandations de la littérature dans le contexte d'un paramètre local (cf. section 2.2), les coordonnées latentes sont donc satisfaisantes dans ce dernier cas.

3.2 Résultats sur les données *synth2*

Avec le paramètre global de largeur de bande et un quantile faible (i.e., ici 2 ou 10%), les coordonnées latentes au laplacien non-normalisé deviennent assez éloignées de la structure décrite par la proposition 1. Les points du cluster rouge (i.e., concave dans l'espace d'origine) se confondent ainsi en zéro sur toutes les coordonnées latentes avec le laplacien non-normalisé (cf. figure 5a). La structure de clusters est alors complètement perdue en cas de normalisation des vecteurs de coordonnées latentes.

Avec le quantile à 2%, les coordonnées latentes au laplacien normalisé sont conformes à la structure idéale de la proposition 1 (cf. figure 6a). À mesure de l'augmentation du quantile, les clusters restent bien délimités, mais le cluster rouge (concave dans l'espace initial, cf. figure 2) ne suit alors plus une variété linéaire (cf. figure 6b). La normalisation des vecteurs de coordonnées latentes ne le concentre donc plus en un seul point mais en une variété non-linéaire, ce qui est potentiellement dommageable pour l'algorithme k-means.

Toujours avec la largeur de bande globale et le quantile désormais à 40%, toutes les configurations de paramètres restants convergent vers une distribution proche de celle de la fi-

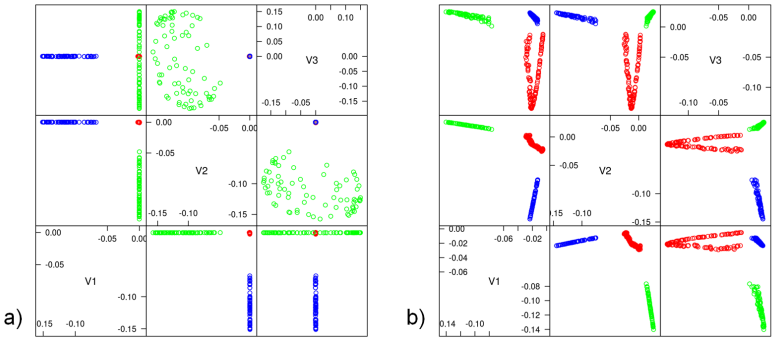

FIG. 5 – *Représentations latentes au laplacien non-normalisé de* synth2 *obtenues avec le quantile à 10% :* a) *avec le paramètre de largeur de bande global,* b) *avec le paramètre local.*

gure 6c. Alors qu'en section 3.2 augmenter le quantile tendait à rendre le premier vecteur propre inutile, cette coordonnée semble désormais caractériser partiellement les clusters. La variété non-linéaire du cluster rouge, problématique pour k-means, est toujours présente.

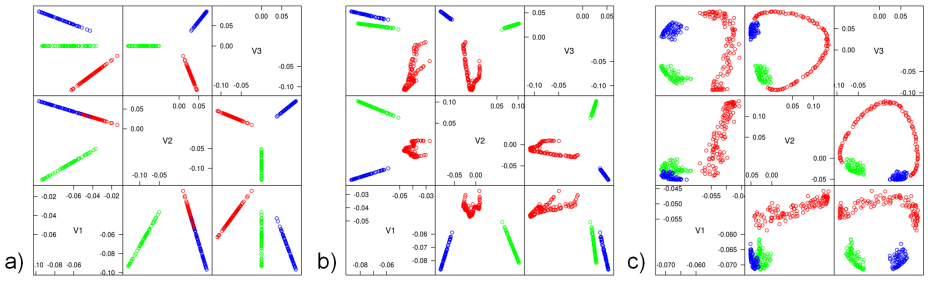

FIG. 6 – *Représentations latentes au laplacien normalisé de* synth2 *avec la largeur de bande globale et* a) *le quantile à 2%,* b) *le quantile à 10%,* c) *le quantile à 40%.*

L'utilisation de la largeur de bande locale amène des résultats sensiblement identiques, si ce n'est pour le laplacien non-normalisé avec le quantile à 10%, où les clusters apparaissent plutôt bien délimités malgré la présence récurrente de la variété non-linéaire (cf. figure 5b).

3.3 Résultats sur les données *synth3*

La figure 7 montre le résultat obtenu pour les laplaciens de *synth3* avec la largeur de bande globale et le quantile à 10%. La distribution latente au laplacien non-normalisé est sensiblement identique à celle observée pour *synth2* en figure 5a. En revanche un impact très négatif, de nouveau lié au cluster inorthodoxe (bruit de fond uniforme dans le cas de *synth3*) est observé

dans le cas du laplacien normalisé (cf. figure 7b), rendant peu probable le succès du traitement de cette distribution par k-means. Ces observations restent valables avec un quantile à 40%.

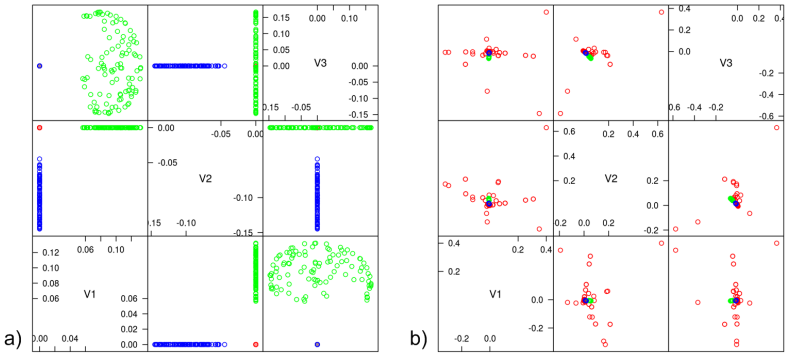

FIG. 7 – *Représentations latentes au laplacien de* synth3 *avec la largeur de bande globale et le quantile à 10% :* a) *laplacien non-normalisé,* b) *laplacien normalisé.*

L'utilisation de la largeur de bande locale (quantile à 10%) n'amène pas d'amélioration significative dans le cas du laplacien non-normalisé. En revanche la distribution idéale en variétés linéaires bien séparées est alors retrouvée dans le cas du laplacien normalisé. Cette observation reste valable pour un quantile plus faible, mais la représentation latente au laplacien normalisé dégradée présentée en figure 7b est obtenue avec un quantile significativement plus élevé (40%).

3.4 Résultats sur les données *synth4*

La figure 8 montre les résultats obtenus sur *synth4* avec la largeur de bande globale et le quantile à 10%. Sans normaliser les vecteurs de coordonnées (cf. figure 8a et b), les clusters semblent être le mieux délimités par le premier vecteur propre pour les deux types de laplacien. Cette observation est surprenante au regard de la proposition 1 et du corollaire 1, car le premier vecteur propre devrait soit être caractéristique d'un seul cluster, soit d'aucun cluster en particulier (i.e., égal à $\mathbb{1}$). Pour le laplacien normalisé, on peut supposer que les degrés des éléments sont caractéristiques des clusters, expliquant ainsi le phénomène (i.e., première coordonnée égale à $\mathbf{D}^{-1/2}\mathbb{1}$) : mais cette hypothèse ne tient pas dans le cas du laplacien non-normalisé.

Les autres coordonnées latentes font figurer des structures circulaires (cf. figure 8a et c), alors que le but de l'approche spectrale au clustering est justement de transformer les données comportant de telles structures vers un espace où elles deviennent convexes et denses. Ce problème, amplifié par l'utilisation d'un quantile plus grand (e.g., 40%) existe dans une moindre mesure avec le laplacien normalisé (cf. figure 8b), car seule une structure circulaire clairement séparée des autres clusters subsiste alors. La normalisation des vecteurs de coordonnées semble alors avoir des effets négatifs sur la délimitation des clusters dans tous les cas (cf. figure 8c et d).

En utilisant la largeur de bande locale et un quantile suffisamment faible (2% ici), les variétés linéaires de la figure 3 sont retrouvées pour le laplacien normalisé de *synth4*. Dans le

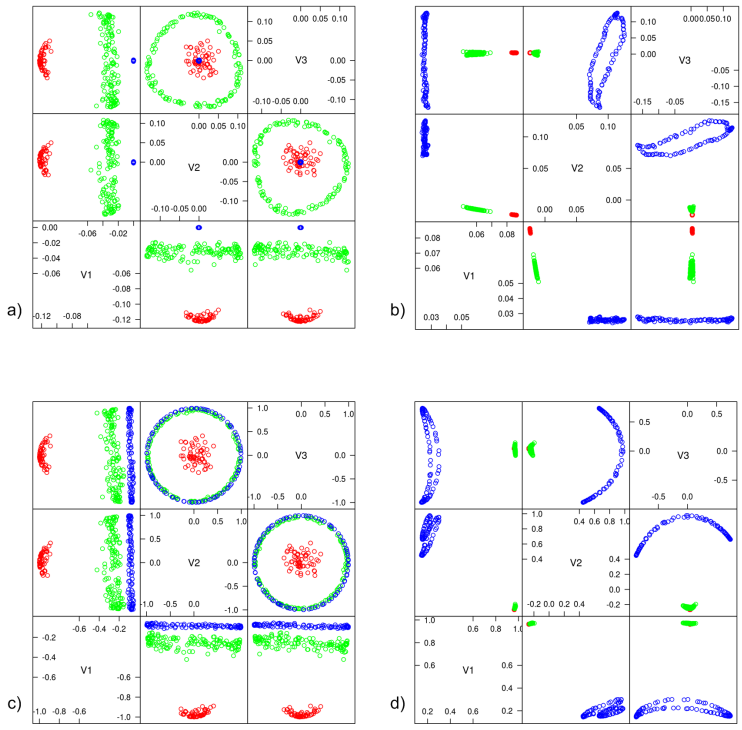

FIG. 8 – *Distributions latentes au laplacien de* synth4 *avec la largeur de bande globale et le quantile à 10% :* a et c) *laplacien non-normalisé,* b et d) *laplacien normalisé,* a et b) *sans normalisation des vecteurs de coordonnées,* c et d) *avec normalisation des vecteurs de coordonnées.*

cas non-normalisé, la distribution latente est proche de celle obtenue avec *synth2* en figure 5a, limitant son intérêt pour l'application de k-means. En revanche, l'utilisation d'un quantile plus grand (i.e., 10% ici) engendre des distributions similaires à celles de la figure 8, rendant alors indistinguables les deux types de largeur de bande.

4 Discussion

Dans (Guattery et Miller, 1998), un contre-exemple est utilisé pour justifier l'inconsistance du laplacien non-normalisé. Toutefois cet exemple est assez irréaliste dans le cas de données numériques continues converties en similarités par une fonction RBF comme dans cet article. Notre objectif est plutôt d'estimer empiriquement l'influence de paramètres à partir d'études de cas. Nous pouvons ainsi établir que seule l'utilisation du laplacien normalisé avec la largeur de bande locale et un quantile faible (e.g., 2%) permet d'obtenir des résultats satisfaisants dans tous les cas de figure présentés dans cet article - nous confirmons ainsi les recommandations de

(Zelnik-Manor et Perona, 2004) et (Karatzoglou et al., 2004), en ayant au préalable généralisé leur définition en termes de quantiles. La largeur de bande locale s'est par ailleurs révélée indispensable à l'extraction du bruit de fond (*synth3*) et des clusters concentriques ou imbriqués (*synth4*).

L'utilisation de quantiles élevés (i.e., supérieurs à 10%) est à proscrire quelque soit la configuration des autres paramètres, car elle ne semble être efficace que pour des données déjà faciles à traiter par k-means dans leur espace initial (*synth1*). Contrairement à certains avis (Fowlkes et al., 2004), nous conseillons de ne pas exclure le premier vecteur propre de l'analyse : au mieux selon le corollaire 1 il ne contient effectivement pas d'information, mais au pire de l'information utile peut être ignorée (cf., e.g., section 3.1). Au cas où seule la largeur de bande globale serait applicable, le laplacien normalisé semble rester le meilleur choix sauf pour les jeux de données comportant du bruit de fond (cf. section 3.3).

Nos expériences n'ont pas permis de clairement établir si il est préférable d'utiliser la normalisation des vecteurs de coordonnées. Son principal intérêt est en principe de permettre de concentrer des variétés linéaires (cf. figure 3) en positions quasi-discrètes, facilitant le traitement des représentations latentes par k-means. Toutefois nous avons pu observer que cette opération pouvait aussi avoir des effets négatifs (cf. figure 8). Seul un algorithme de clustering traitant des variétés linéaires permettrait alors de se passer de l'étape de normalisation des vecteurs de coordonnées.

5 Conclusion

Guidée par l'identification des paramètres influançant la construction de la représentation latente au clustering spectral, une approche qualitative a été suivie dans cet article. Elle a permis, d'une part, de souligner que la plupart des configurations de paramètres engendre des distributions latentes impropres à leur traitement par k-means, d'autre part, d'identifier les quelques configurations traitant avec succès les cas de figure évoqués dans l'article.

Les jeux de données utilisés dans cet article ont été sciemment choisis très simples, de manière à pouvoir associer visuellement les propriétés des distributions initiale et latente. Les conclusions que nous avons tiré de leur analyse devront être confirmées sur des données à plus haute dimension et/ou plus volumineux. Dans ce dernier cas, la complétion de Nyström permet de limiter la complexité du clustering spectral de $O(N^2)$ à $O(N)$ au prix d'une approximation (Fowlkes et al., 2004). L'interaction entre le paramétrage identifié dans notre article et cette technique pourrait également être étudié.

Références

Baglama, J. et L. Reichel (2006). Restarted block Lanczos bidiagonalization methods. *Numerical Algorithms 43*(3), 251–272.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bruneau, P., O. Parisot, et B. Otjacques (2014). A Heuristic for the Automatic Parametrization of the Spectral Clustering Algorithm. In *International Conference on Pattern Recognition*, pp. 1313–1318.

Fowlkes, C., S. Belongie, F. Chung, et J. Malik (2004). Spectral grouping using the Nystrom method. *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*(2), 214–225.

Guattery, S. et G. L. Miller (1998). On the quality of spectral separators. *SIAM Journal on Matrix Analysis and Applications 19*(3), 701–719.

Karatzoglou, A., A. Smola, K. Hornik, et A. Zeileis (2004). kernlab - an S4 package for kernel methods in R. *Journal of Statistical Software 11*(9), 1–20.

Lin, F. et W. W. Cohen (2010). Power iteration clustering. In *International Conference on Machine Learning*, pp. 655–662.

Liu, H., J. Latecki, et S. Yan (2013). Fast Detection of Dense Subgraphs with Iterative Shrinking and Expansion. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*(9), 2131–2142.

Nabney, I. (2002). *NETLAB : algorithms for pattern recognition*. Springer.

Ng, A. Y., M. I. Jordan, et Y. Weiss (2002). On spectral clustering : Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pp. 849–856.

Sheather, S. J. et M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society series B*(53), 683–690.

Shi, J. et J. Malik (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(8), 888–905.

van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications 30*(1), 121–141.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing 17*(4), 395–416.

Yan, D., L. Huang, et M. I. Jordan (2009). Fast approximate spectral clustering. In *ACM SIGKDD*, pp. 907–916.

Zelnik-Manor, L. et P. Perona (2004). Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, pp. 1601–1608.

Summary

Spectral clustering is motivated by the extraction of arbitrary cluster shapes in numerical data. This property enhanced its popularity, and despite its theoretical base having been established for more than a decade now, variants have still been introduced until recently. This algorithm basically transforms data to a latent space where arbitrary cluster shapes become easy to process by a clustering algorithm such as k-means. However distributions actually observed in this latent space have received little attention, and many authors assume theory predictions are verified. Alternatively, this paper follows a qualitative approach to check if the ideal, theoretical structure is indeed obtained in practice. A theoretical state-of-the-art summary serves the identification of parameters commanding at the transformation. Observations drawn from our experiments lead to the identification of effective parameter combinations with associated conditions.