Approximation Robinsonienne sur les PQ-arbres

Pascal Préa, François Brucker

Approximation Robinsonienne sur les PQ-arbres

Pascal Préa[*,**], François Brucker[*,**]

[*]LIS, Aix-Marseille Université, CNRS & Université de Toulon,
Marseille, France
[**]École Centrale Marseille,
Marseille, France
{pascal.prea, francois.brucker}@lis-lab.fr

Résumé. Une dissimilarité D sur un ensemble S est *Robinson* si il existe un ordre total sur S tel que $\forall x, y, z \in S, x < y < z \implies D(x, z) \geq \max\{D(x, y), D(y, z)\}$. Un tel ordre est dit *Robinson* ou *compatible* (avec D). Un *PQ-arbre* sur S est un arbre qui représente un ensemble de permutations de S. Étant donnée une dissimilarité de Robinson D, l'ensemble des ordres compatibles avec D peut être représenté par un PQ-arbre.

Dans ce papier, nous considérons le problème suivant : étant donnés une dissimilarité D et un PQ-arbre $\mathcal{T}$ sur un ensemble S, approximer D en une dissimilarité de Robinson R telle que tous les ordres représentés par $\mathcal{T}$ soient compatibles avec R.

Nous montrons que, dans la plupart ces cas, ce problème peut se ramener à une régression isotonique et nous donnons des algorithmes efficaces (les complexités varient entre $O(n^2)$ et $O(n^3 \log^3 n)$) pour résoudre les différentes versions (approxier selon la norme L_1, L_∞, ...). En plus, il est possible d'améliorer incrémentalement la solution obtenue. On obtient ainsi une dissimilarité de Robinson R, plus proche de D, mais tous les ordres représentés par $\mathcal{T}$ ne sont pas compatibles avec R.

1 Introduction

1.1 Dissimilarités de Robinson

Étant donné un ensemble (fini) X, une *dissimilarité* sur X est une fonction sur $X \times X$ à valeurs positives ou nulles, symétrique et telle que $\forall x \in X, d(x, x) = 0$. On appellera *espace* le couple (X, d). La *sériation (linéaire)* consiste, étant donné un espace (X, d), à déterminer si les points de X peuvent être disposés sur une droite de manière compatible avec d. La sériation a de nombreuses applications (archéologie, musicologie,...). Ce problème a été formalisé par Robinson (1951) de la manière suivante. Un ordre total est *compatible* si :

$$\forall x < y < z,\ d(x, z) \geq \max\{d(x, y), d(y, z)\} \tag{1}$$

Une dissimilarité (ou un espace) est (de) *Robinson* si elle admet un ordre compatible. De la même manière que les ultramétriques sont équivalentes aux hiérarchies, les dissimilarités de

Robinson (qui généralisent les ultramétriques) sont équivalentes aux pyramides (Diday, 1986; Durand et Fichet, 1988).

Il est possible de reconnaître les dissimilarités de Robinson (*ie.* déterminer si une dissimilarité est Robinson ou pas) de manière polynomiale (Chepoi et Fichet, 1997; Atkins et al., 1998; Laurent et Seminaroti, 2017), ou même optimale (Préa et Fortin, 2014) en $O(n^2)$, où $n = |S|$.

1.2 Problèmes d'approximation pour les dissimilarités de Robinson

L'*approximation Robinsonienne* consiste à déterminer la dissimilarité de Robinson R la plus proche d'une dissimilarité D donnée, où *plus proche* veut dire plus proche selon la norme L_1, L_2,..., ou la plus grande possible parmi les dissimilarités plus petites que D... L'approximation est dite *parfaite* si R est LA plus proche possible de D, et *imparfaite* sinon.

Approximer parfaitement une dissimilarité en une dissimilarité de Robinson est NP-dur pour les normes L_1 (Barthélemy et Brucker, 1998) et L_∞ (Chepoi et al., 2009). Aucun algorithme efficace n'est connu pour les autres normes. Il faut remarquer que cette NP-difficulté est pour le cas où il faut aussi trouver un ordre compatible. Quand un ordre compatible est imposé, le problème est plus simple ; par exemple :

— Ghandehari et Janssen (2019) présentent deux algorithmes (un en $O(n^4)$ et un en $O(n^6)$) qui donnent une approximation imparfaite pour la norme L_1.

— Durand (1989) propose un algorithme en $O(n^2)$ qui rend la plus grande (resp. plus petite) dissimilarité de Robinson plus petite (resp. plus grande) d'une dissimilarité donnée.

Dans cette communication, nous nous intéressons à une généralisation de l'approximation à ordre fixé : l'approximation à PQ-arbre fixé. Ce problème semble être le plus général qui soit polynomial.

1.3 PQ-arbres

Un *PQ-arbre* sur un ensemble fini S représente un ensemble de permutations sur S. C'est un arbre planté dont les feuilles sont indexées par S et dont les sommets intérieurs sont des *P-nœuds* ou des *Q-nœuds*. On peut permuter le fils d'un P-nœud, alors qu'on ne peut qu'inverser ceux d'un Q-nœud. Par exemple, l'arbre de la figure 1 représente les permutations (1,2,3,4,5,6,7), (1,3,2,4,5,6,7), (2,1,3,4,5,6,7), (2,3,1,4,5,6,7), (3,1,2,4,5,6,7), (3,2,1,4,5,6,7), (7,6,5,4,1,2,3), (7,6,5,4,1,3,2), (7,6,5,4,2,1,3), (7,6,5,4,2,3,1), (7,6,5,4,3,1,2), (7,6,5,4,3,2,1).

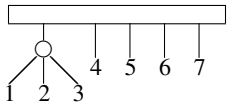

FIG. 1 – *Un PQ-arbre*

Cette structure de données a été introduite par Booth et Lueker (1976) pour résoudre le problème de la propriété des 1 consécutifs. Un autre intérêt des PQ-arbres est :

Proposition 1. *Étant donné un ensemble S et une dissimilarité de Robinson S sur S, l'ensemble des permutations compatibles de D peut être représenté par un PQ-arbre.*

Un PQ-arbre avec un seul nœud intérieur qui un P-nœud représente toutes les permutations sur S, il est alors dit *universel*; si cet unique nœud intérieur est un Q-nœud, le PQ-arbre ne représente qu'une permutation (et son opposée), il est alors dit *plat*

2 Notations et résultats préliminaires

Dans tout ce papier, S sera un ensemble à n éléments, $\mathcal{T}$ un PQ-arbre sur S, D une dissimilarité sur S et R une dissimilarité de Robinson sur S. Si α est un sommet de $\mathcal{T}$, S_α désignera l'ensemble des feuilles sous α. On note :
$D^+(\alpha,\beta) := \max\{D(x,y) : x \in S_\alpha, y \in S_\beta\}$,
$D^-(\alpha,\beta) := \min\{D(x,y) : x \in S_\alpha, y \in S_\beta, x \neq y\}$,
$D^+(\alpha) := \max\{D(x,y) : x,y \in S_\alpha\}$,
Si α est un P-nœud avec les fils $\nu_1,\ldots,\nu_p$, $D^-(\alpha) := \min\{D^-(\nu_i,\nu_j) : i \neq j\}$,
Si α est un Q-nœud avec les fils, dans cet ordre, $\nu_1,\ldots,\nu_p$, $D^-(\alpha) := D^-(\nu_1,\nu_p)$,
Si $D^+(\alpha) = D^-(\alpha)$, alors $D(\alpha) := D^-(\alpha)$,
Si $D^+(\alpha,\beta) = D^-(\alpha,\beta)$ alors $D(\alpha,\beta) := D^-(\alpha,\beta)$.

Proposition 2. *Pour tout sommet α de $\mathcal{T}$, si S est trié selon un ordre représenté par $\mathcal{T}$, alors S_α est un intervalle.*

Proposition 3. *Si toutes les permutations représentées par $\mathcal{T}$ sont compatibles avec R, alors, pour tout sommet α et tout $z \in S \setminus S_\alpha$, $R(z,\alpha)$ est bien défini et $R^+(\alpha) \leq R(z,\alpha)$, ie. $\forall x,y \in S_\alpha$, $R(x,y) \leq R(x,z) = R(y,z)$.*

Proposition 4. *Soient α et β deux sommets tels que $S_\alpha \cap S_\beta = \emptyset$. Si toutes les permutations représentées par $\mathcal{T}$ sont compatibles avec R, alors $R(\alpha,\beta)$ est bien défini et on a $R(\alpha,\beta) \geq R^+(\alpha), R^+(\beta)$.*

Proposition 5. *Soit α un P-nœud de fils $\nu_1,\ldots,\nu_p$. Si toutes les permutations représentées par $\mathcal{T}$ sont compatibles avec R, alors, pour tout i,j,i',j' avec $1 \leq i \neq j \leq p$ et $1 \leq i' \neq j' \leq p$, $r(\nu_i,\nu_j) = R(\nu_{i'},\nu_{j'})$.*

Proposition 6. *Soit α un Q-nœud de fils $\nu_1,\ldots,\nu_p$ dans cet ordre. Si toutes les permutations représentées par $\mathcal{T}$ sont compatibles avec R, alors, pour tout $i,j \in \{1,\ldots,p\}$ avec $i < j-1$, $R(\nu_i,\nu_j) \geq R(\nu_i,\nu_{j-1}), R(\nu_{i+1},\nu_j)$.*

Étant donné un PQ-arbre $\mathcal{T}$, on construit le graphe orienté acyclique $G_\mathcal{T} = (V_\mathcal{T}, E_\mathcal{T})$ suivant :
$V_\mathcal{T} = \{v_\alpha, \alpha \text{ sommet de } \mathcal{T}\} \cup \{v^\alpha_{\nu_i,\nu_j}, \alpha \text{ Q-nœud de } \mathcal{T}, \nu_i,\nu_j \text{ fils distincts de } \alpha\}$
Si α est un Q-nœud de fils, dans cet ordre, $\nu_1,\ldots,\nu_p$, on identifie v_α et $v^\alpha_{\nu_1\nu_p}$
Si α est un P-nœud et β un fils de α, alors $v_\alpha v_\beta \in E_\mathcal{T}$
Si α est un Q-nœud avec comme fils $\nu_1,\ldots,\nu_p$ dans cet ordre, alors :
$\forall i,j \in \{1,\ldots,p\} : i < j-1$, $v^\alpha_{\nu_i,\nu_j} v^\alpha_{\nu_i,\nu_{j-1}}, v^\alpha_{\nu_i,\nu_j} v^\alpha_{\nu_{i+1},\nu_j} \in E_\mathcal{T}$
$\forall i \in \{1,\ldots,p-1\}$, $v^\alpha_{\nu_i,\nu_{i+1}} v_{\nu_i}, v^\alpha_{\nu_i,\nu_{i+1}} v_{\nu_{i+1}} \in E_\mathcal{T}$
Le graphe $G_\mathcal{T}$ pour le graphe de la figure 1 est représenté en Figure 2

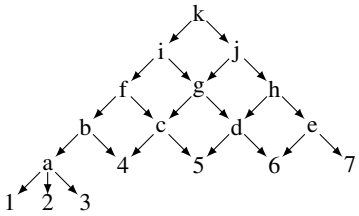

FIG. 2 – *Le graphe $G_\mathcal{T}$ correspondant au PQ-arbre de la figure 1. Les nœuds labellés '1', '2',..., '7' correspondent aux feuilles ; le sommet labellé 'a' au P-nœud, 'k' avec la racine (ou à la paire $\{a, 7\}$), 'c' à la paire $\{4, 5\}$ 'g' à la paire $\{4, 6\}$, ...*

À une dissimilarité D et un PQ-arbre $\mathcal{T}$, on peut associer un vecteur $\mathfrak{b} \in \mathbb{R}^{V_\mathcal{T}}$ de la manière suivante :

Si v est une feuille, $\mathfrak{b}_D(v) := 0$

Si v correspond à un P-nœud α, $\mathfrak{b}_D(v) := \max\{D(x, y) : x, y \in S_\alpha\}$

Si v est un nœud $v^\alpha_{\nu_i,\nu_j}$, $\mathfrak{b}_D(v) := \max\{D(x, y) : x \in S_{\nu_i}, y \in S_{\nu_j}\}$

Inversement, si $\mathfrak{b} \in \mathbb{R}^{V_\mathcal{T}}$ est à valeur positive avec $\mathfrak{b}(v) = 0$ pour toute feuille, on construit à partir de $\mathfrak{b}$ la dissimilarité $D_\mathfrak{b}$ avec :

Si $x \in S_\nu, y \in S_\mu$, avec ν, μ fils d'un P-nœud α, alors $D_\mathfrak{b}(x, y) := \mathfrak{b}(v_\alpha)$

Si $x \in S_\nu, y \in S_\mu$, avec ν, μ fils d'un Q-nœud α, alors $D_\mathfrak{b}(x, y) := \mathfrak{b}(v^\alpha_{\nu\mu})$

et on a $\mathfrak{b}_{D_\mathfrak{b}} = \mathfrak{b}$. Passer de D à $\mathfrak{b}_D$ ou de $\mathfrak{b}$ à $D_\mathfrak{b}$ se fait en $O(n^2)$.

Soit $G = (V, E)$ un graphe orienté acyclique. Un vecteur $\mathfrak{b} \in \mathbb{R}^V$ est *isotonique* si $uv \in E \implies \mathfrak{b}_u \geq \mathfrak{b}_v$. La *régression isotonique* consiste à déterminer le vecteur isotonique le plus proche d'un vecteur donné. Il existe de nombreux algorithmes pour ce problème dont les complexités varient, en fonction de la norme et du type de graphe (arbre, planaire,...), de $O(|V|)$ à $O(|V|^4)$ (Stout, 2013; Kyng et al., 2015).

Proposition 7. *Un vecteur $\mathfrak{b} \in \mathbb{R}^{+V_\mathcal{T}}$ à valeur nulle sur les feuilles est isotonique si et seulement si $D_\mathfrak{b}$ est Robinson. De plus, toutes les permutations représentées par $\mathcal{T}$ sont compatibles avec $D_\mathfrak{b}$.*

3 Algorithmes

Nous allons maintenant présenter, pour chaque définition de "proche", un algorithme qui, étant donnés une dissimilarité D et un PQ-arbre $\mathcal{T}$, donne la dissimilarité de Robinson R la plus proche de D telle que toutes les permutations représentées par $\mathcal{T}$ soient compatibles avec R.

1. La première étape de ces algorithmes consiste à construire le graphe $G_\mathcal{T}$ et le vecteur $\mathfrak{b}_D$. Ceci se fait en $O(n^2)$.
2. Pour déterminer la plus grande des dissimilarités de Robinson plus petites que D, on fait un parcours top-down de $G_\mathcal{T}$, ce qui prend $O(n^2)$.

3. Pour déterminer la plus petite des dissimilarités de Robinson plus grandes que D, on fait un parcours bottom-up de $G_{\mathcal{T}}$, ce qui prend $O(n^2)$.
4. Pour déterminer la dissimilarité de Robison la plus proche de D selon la norme L_p, on calcule la régression isotonique de $\mathfrak{b}_D$ selon la norme L_p. Ceci peut se faire en temps $O(n^2 \log n)$ pour la norme L_∞, $O(n^2 \log^2 n)$ pour la norme L_1 et $O(n^3 \log^3 n)$ pour la norme L_2.
5. Après les étapes 2, 3 ou 4, on obtient un vecteur isotonique $\mathfrak{r}$. On calcule la dissimilarité $D_{\mathfrak{r}}$, ce qui se fait en $O(n^2)$. Cette dissimilarité est la meilleure approximation de D sur $\mathcal{T}$.

Il est possible d'utiliser cet algorithme pour obtenir une approximation imparfaite (on obtient un optimum local) dans le cas général :

1. On part d'une permutation σ_0 et on construit le PQ-arbre plat $\mathcal{T}_{\sigma_0}$ représentant σ_0 et son opposée.
2. On approxime D sur $\mathcal{T}_\sigma$. On obtient une dissimilarité de Robinson R_0.
3. Si le PQ-arbre de R_0 représente d'autres permutations que σ_0 et son opposée, on choisit une autre permutation σ_1 ; on applique 1 et 2 avec σ_1 à la place de σ_0. On obtient une dissimilarité de Robinson R_1, qui est au moins aussi proche de D que R_0.
4. Si R_1 est plus proche que R_0, on recommence.

4 Conclusion

L'approximation à PQ-arbre fixé est une extension naturelle de l'approximation à ordre fixé, et nous obtenons des algorithmes au moins aussi bons que ceux proposés pour ce problème, efficaces, souvent optimaux ou proches de l'optimal. On peut ainsi traiter des matrices de relativement grande taille : sur un portable de 2014 (processeur Intel Core i5 à 1,4 GHz), en norme $L1$, une matrice 100×100 se traite en 0.8s, et une matrice 200×200 en 6.6s.

De plus, cela prouve que la NP-difficulté de l'approximation Robinsonienne est entièrement due à la détermination des ordres compatibles.

Références

Atkins, J. E., E. G. Boman, et B. Hendrickson (1998). A spectral algorithm for seriation and the consecutive ones problem. *SIAM J. Computing 28*, 297–310.

Barthélemy, J.-P. et F. Brucker (1998). NP-hard approximation problems in overlapping clustering. *Journal of Classification 18*, 159–183.

Booth, K. S. et G. S. Lueker (1976). Testing for the consecutive ones property, interval graphs and graph planarity using pq-tree algorithm. *Journal of Computer and System Sciences 13*, 335–379.

Chepoi, V. et B. Fichet (1997). Recognition of Robinsonnian dissimilarities. *Journal of Classification 14*, 311–325.

Chepoi, V., B. Fichet, et M. Seston (2009). Seriation in the presence of errors: NP-hardness of l_∞-fitting Robinson structures to dissimilarity matrices. *Journal of Classification 26*, 379–296.

Diday, E. (1986). Orders and overlapping clusters by pyramids. In J. de Leeuw, W. Heiser, J. Meulman, et F. Critchley (Eds.), *Multidimensionnal Data Analysis*, pp. 201–234. DSWO.

Durand, C. (1989). *Ordres et graphes pseudo-hiérarchiques*. Thèse de doctorat, Université Aix-Marseille I.

Durand, C. et B. Fichet (1988). One-to-one correspondences in pyramidal representation: an unified approach. In H. Bock (Ed.), *Classification and Related Methods of Data Analysis*, pp. 85–90. North Holland.

Ghandehari, M. et J. Janssen (2019). An optimization parameter for seriation of noisy data. *SIMA J. on Discr. Math. 32*, 712–730.

Kyng, R., A. Rao, et S. Sachdeva (2015). Fast, provable algorithms for isotonic regression in all l_p-norms. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, et R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pp. 2719–2727. North Holland.

Laurent, M. et M. Seminaroti (2017). A lex-bfs-based recognition algorithm for Robinsonian matrices. *Discr. Appl. Math. 222*, 151–165.

Préa, P. et D. Fortin (2014). An optimal algorithm to recocognize robinsonnian dissimilarities. *Journal of Classification 31*, 351–385.

Robinson, W. S. (1951). A method for chronologically ordering archeological deposits. *American Antiquity 16*, 293–301.

Stout, Q. S. (2013). Isotonic regression via partitioning. *Algorithmica 66*, 93–112.

Summary

A dissimilarity D on a n-set S is *Robinson* if there exists a linear order on S such that $\forall x, y, z \in S,\ x < y < z \implies D(x,z) \geq \max\{D(x,y), D(y,z)\}$. Such an order is said to be *Robinson* or *compatible* with D. A *PQ-tree* on S is a tree which represents a set of permutations of S. Given a Robinson dissimilarity D, the set of the orders compatible with D can be represented by a PQ-tree.

In this paper, we consider the following problem: given a dissimilarity D and a PQ-tree $\mathcal{T}$ on a n-set S, approximate D into a Robinson dissimilarity R such that all orders represented by $\mathcal{T}$ are compatible with R. This problem generalizes the classical problem of approximating a given dissimilarity into a Robinson one with a given compatible order.

We show that, in most cases, this problem can be settled as isotonic regression and give efficient algorithms (complexities in the worst case range from $O(n^2)$ to $O(n^3 \log^3 n)$ where $n = |S|$) to solve its different versions (approximate along the L_1 norm, the L_∞ norm,...). In addition, it may be possible to incrementally improve the solution of these problems, we then get a Robinson dissimilarity R which is closer from D, but all orders represented by $\mathcal{T}$ are not compatible with R.

Ce travail a bénéficié d'un financement de l'ANR via le projet DISTANCIA (ANR-17-CE40-0015).