



Mathématiques et sciences humaines

Mathematics and social sciences

190 | Été 2010

Mathématiques discrètes : théories et usages.
Numéro en hommage à Bruno Leclerc

Un panorama des approximations en norme du supremum pour la classification

An overview of approximations with the supremum norm for classification

Bernard Fichet et Morgan Seston



Édition électronique

URL : <http://journals.openedition.org/msh/11746>

DOI : 10.4000/msh.11746

ISSN : 1950-6821

Éditeur

Centre d'analyse et de mathématique sociales de l'EHESS

Édition imprimée

Date de publication : 10 mars 2010

Pagination : 89-113

ISSN : 0987-6936

Référence électronique

Bernard Fichet et Morgan Seston, « Un panorama des approximations en norme du supremum pour la classification », *Mathématiques et sciences humaines* [En ligne], 190 | Été 2010, mis en ligne le 16 octobre 2010, consulté le 19 avril 2019. URL : <http://journals.openedition.org/msh/11746> ; DOI : 10.4000/msh.11746

UN PANORAMA DES APPROXIMATIONS EN NORME DU SUPREMUM POUR LA CLASSIFICATION

Bernard FICHET¹ et Morgan SESTON^{1,2}

RÉSUMÉ – *Dans un cadre général où les concepts de sous-dominante/sur-dominée jouent un rôle fondamental, nous dressons un vaste panorama d'approximations en norme du supremum pour nombre de structures de la classification : ultramétriques (partielles ou non), k-ultramétriques, régressions convexes et isotones. Pour les semi-distances/dissimilarités d'arbre et les dissimilarités de Robinson, nous montrons comment l'approche générale peut conduire à des algorithmes avec un facteur constant.*

MOTS CLÉS – Arbre, Dissimilarité, Norme du supremum, Régression, Ultramétrie

SUMMARY – *An overview of approximations with the supremum norm for classification. In a general framework where the concepts of subdominant/updominated play a crucial role, we give a vast overview of approximations with the supremum norm for many structures of classification : (partial or nonpartial) ultrametrics, k-ultrametrics, convex and isotonic regressions. For tree semi-distances/dissimilarities and Robinsonian dissimilarities, we show how the general approach leads us to algorithms with a constant factor.*

KEYWORDS – Dissimilarity, Supremum norm, Regression, Tree, Ultrametric

1. INTRODUCTION

C'est un lieu commun de dire que les mathématiques appliquées reposent sur des approximations. Et que dire alors de certaines branches de celles-ci, comme l'analyse des données, la classification, voire la statistique en général, où tout n'est qu'optimisation et approximation. Avec Sir R.A. Fisher, nous avons appris qu'un estimateur est n'importe quelle statistique, c'est-à-dire n'importe quelle transformation (on ajoute mesurable) des données. Mais c'est pour mieux dire que dans ce vaste univers, il convient de rechercher la meilleure estimation, au sens d'un certain critère, et donc d'optimiser. Le modèle statistique est parfois (souvent) absent en classification. Mais l'idée même du modèle reste présente (J-R. Barra a parlé de « modèle virtuel » en analyse des données), même en classification dite « non supervisée ». Comment peut-on faire de la phylogénie en oubliant la structure d'arbre, enraciné ou non, et donc sans rechercher l'arbre le plus proche, en un certain sens, des données ? De même,

¹Laboratoire d'Informatique Fondamentale, 163, avenue de Luminy - Case 901, F-13288 Marseille Cedex 9, bernard.fichet,morgan.seston@lif.univ-mrs.fr

²Unité des Virus Emergents, 27 boulevard Jean Moulin, 13385 Marseille Cedex 5, morgan.seston@ird.fr

la sériation est par nature liée à la recherche d'un ordre. Et serions-nous privés de tout modèle, qu'il nous plairait encore de disposer d'une structure de représentation visuelle simple. Nous serions ainsi confrontés au meilleur ajustement des données à cette dernière. Qui ne préfère regarder une carte d'Atlas, fausse bien sûr, puisque la sphère n'est pas développable, plutôt qu'un tableau (matrice) de distances ? Le « *multidimensional scaling* » (MDS), et les méthodes dites factorielles reposent sur ce principe.

Ainsi tout n'est qu'approximation. Les fervents de la norme L_1 aiment à rappeler que le premier critère utilisé pour la régression linéaire fut de ce type, avec les travaux de R.J. Boscovich au milieu du XVIII^e siècle, prolongés par P.S. de Laplace. Avec cette norme, les statisticiens y voient en outre un gage de robustesse. Mais nous ne pouvons oublier A.M. Legendre et K.F. Gauss pour la méthode des moindres carrés fondée sur la norme L_2 , avec la géométrie Euclidienne, la simplicité des démonstrations et des solutions analytiques. La loi normale, dite de Laplace-Gauss, fait merveille en probabilités et statistique, puisqu'elle porte en germe le carré de la distance euclidienne dans le coefficient exponentiel de sa densité. Le débat entre norme L_1 et norme L_2 a alimenté bien des critiques comparatives qu'illustre par exemple l'article de [Portnoy, Koenker, 1997].

La norme Euclidienne conduit en outre à une approximation moyenne, que A.-L. Cauchy a intensivement généralisée. Selon le bon mot que nous a suggéré M. Barbut, nous pourrions parler de « moyennes à la Cauchy ». Pour ces dernières, nous citons [Barbut, 1988] pour le vaste éventail et [Robertson, Wright, 1974] pour leurs implications statistiques.

Malheureusement, la norme L_2 , et même les normes L_p , p fini, ne sont d'aucun recours en classification, aussi bien pour le partitionnement avec la méthode des K -means ($p = 2$), la classification hiérarchique et les approximations ultramétriques, ses extensions aux arbres et à la classification pseudo-hiérarchique (pyramidale) avec ses classes empiétantes, en liaison avec les dissimilarités de Robinson. Toutes les approximations en norme L_p , p fini, liées à ces structures s'avèrent comptées comme NP-difficiles. Nous citons [Křivánek, Morávek, 1986] et [Křivánek, 1988] pour les ultramétriques, [Day, 1987] pour les distances d'arbre et [Barthélemy, Brucker, 2001] pour les dissimilarités dites fortement de Robinson.

En revanche, il en est tout autre pour la norme du supremum ($p = \infty$). Nous dressons dans cet exposé, un vaste panorama de ces approximations pour nombre de structures classiques de la classification et de l'analyse des données. Seront en particulier concernées, les structures ultramétriques et hiérarchiques, leurs extensions Robinsonniennes et arborées, que ces dernières soient métriques ou de dissimilarité, les métriques linéaires, ainsi que les régressions, soit convexes, soit monotones sur un ensemble partiellement ordonné. Tous ces résultats seront situés dans un cadre très général, tel qu'il fut décrit dans [Chepoi, Fichet, 2000], et où les concepts de sous-dominante et de sur-dominée jouent un rôle primordial. Le problème du consensus sera également traité dans ce cadre. Naturellement, l'exhaustivité ne peut être atteinte, et d'autres structures métriques sont absentes de cette approche, comme celles, actuellement très en vogue, reposant sur le concept d'ordres circulaires, [Hubert, Arabie, Meulman, 1998], [Brucker, Barthélemy, 2007], ou les métriques de type

L_1 circulairement décomposables, avérées comme équivalentes aux métriques de Kalmanson, [Chepoi, Fichet, 1998]. À notre connaissance, toutes souffrent d'un manque de résultats au niveau des approximations.

Les deux paragraphes suivants exposent respectivement l'environnement et les principaux résultats. Ils seront suivis d'une série d'exemples, conséquences directes des résultats généraux. Ceux-ci concernent de fait, presque toutes les structures sus-évoquées, avec un calcul analytique ou algorithmique aisé d'une sous-dominante et/ou d'une sur-dominée, à l'exception des structures arborées, Robinsonniennes et linéaires, pour lesquelles la NP-difficulté a été prouvée. La NP-difficulté de l'approximation par une distance arborée, peut être levée si l'on se restreint aux dissimilarités préservant les dissimilarités à un point pivot donné. Ce sera l'objet de l'avant dernier paragraphe, où il sera en outre montré, que l'approximation à pivot donné, conduit à une approximation générale avec un facteur 3. On rappelle qu'une approximation avec un facteur k est une approximation avec une erreur inférieure à k fois l'erreur globale optimale. Enfin le dernier paragraphe sera consacré aux distances linéaires et aux dissimilarités de Robinson, où seront encore décrits des algorithmes avec un facteur constant donné.

2. PRÉLIMINAIRE

Nous dressons ici le cadre général des approximations en norme du supremum. Toutes les structures considérées reposent dans un espace vectoriel réel \mathbb{E} de dimension finie p . Relativement à une base fixée une fois pour toutes, un vecteur u de \mathbb{E} a pour coordonnées u_1, \dots, u_p . L'espace est muni de la norme du supremum, dite aussi de Chebychev ou de la convergence uniforme, et notée simplement $\|.\|$, de sorte que $\|u\| = \max_j |u_j|$. On note également $\mathbf{1}$ le vecteur de coordonnées unité, générant la diagonale principale de \mathbb{E} , soit \mathcal{L} , c'est-à-dire l'ensemble des vecteurs $c.\mathbf{1}$, $c \in \mathbb{R}$.

Relativement à la base choisie, un ordre partiel usuel est défini sur \mathbb{E} par comparaison des coordonnées : $u \preceq v$ si et seulement si $u_j \leq v_j$, pour tout $j = 1, \dots, p$. Alors tout sous-ensemble non vide et borné A de \mathbb{E} possède une borne supérieure $\sup(A)$ et une borne inférieure $\inf(A)$, de jème coordonnées respectives $\sup\{u_j, u \in A\}$ et $\inf\{u_j, u \in A\}$. En d'autres termes, (\mathbb{E}, \preceq) est un treillis conditionnellement complet.

Par la suite, un sous-ensemble non vide \mathcal{K} de \mathbb{E} jouera le rôle de référence (de définition) d'une structure donnée que l'on cherchera à approcher à partir d'un ou plusieurs vecteurs de \mathbb{E} . Très souvent, et ce sera toujours le cas ici, \mathcal{K} est un cône (pointé) de \mathbb{E} , c'est-à-dire un ensemble de vecteurs invariant par multiplication par une constante positive ou nulle. Le problème d'approximation est alors le suivant : donnés n vecteurs w_1, \dots, w_n de \mathbb{E} , trouver \hat{z} de \mathcal{K} , solution (si elle existe) de : $\inf_{z \in \mathcal{K}} \max_{i=1, \dots, n} \|z - w_i\|$. C'est ici qu'apparaît une conséquence vertueuse de la norme du supremum. Notant $u := \inf(w_1, \dots, w_n)$ et $v := \sup(w_1, \dots, w_n)$, le problème général précédent est équivalent au problème simple suivant.

Donnés u et v de \mathbb{E} avec $u \preceq v$, trouver \hat{z} de \mathcal{K} , solution (si elle existe) de :

$$(P) : \inf_{z \in \mathcal{K}} \max[\|z - u\|, \|z - v\|] .$$

Cet infimum existe toujours. Il sera qualifié d'*erreur optimale*, et noté $\hat{\epsilon}$. Bien sûr, cette erreur n'est pas nécessairement atteinte. Remarquons une propriété immédiate de l'ensemble des solutions, due au choix de la norme du supremum : si z, z' , avec $z \preceq z'$, sont solutions de (P) , alors tout élément de l'intervalle $[z, z']$ de \mathcal{K} , i.e. l'ensemble des t de \mathcal{K} vérifiant $z \preceq t \preceq z'$, est solution.

Deux autres approximations, liées entre elles et à cette dernière, seront fort utiles. Elles consistent à approcher u et v par un élément de \mathcal{K} plus petit que u ou plus grand que v . Pour cela, nous adoptons les notations suivantes : pour tout t de \mathbb{E} , $\mathcal{K}_{\preceq}(t) := \{x \in \mathcal{K} : x \preceq t\}$ et $\mathcal{K}_{\succeq}(t) := \{x \in \mathcal{K} : x \succeq t\}$. On définit alors les deux problèmes suivants.

Donnés u et v de \mathbb{E} avec $u \preceq v$, trouver \hat{z} , solution (si elle existe) de :

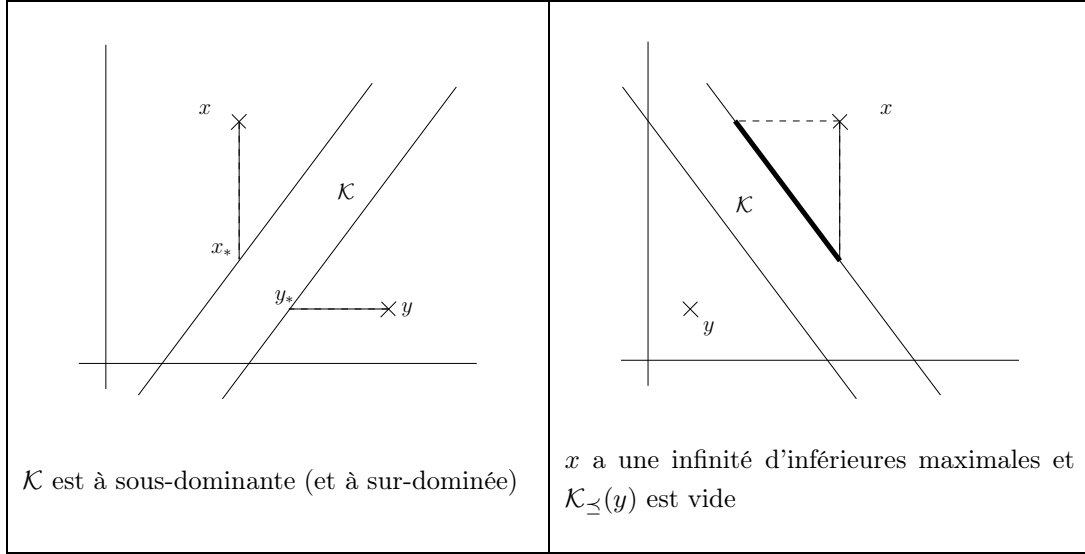
$$\begin{aligned} (P_*) &: \inf_{z \in \mathcal{K}_{\preceq}(u)} \|z - v\|. \\ (P^*) &: \inf_{z \in \mathcal{K}_{\succeq}(v)} \|z - u\|. \end{aligned}$$

Comme pour le problème (P) , les deux infima de (P_*) et (P^*) existent, dès que les ensembles $\mathcal{K}_{\preceq}(u)$ et $\mathcal{K}_{\succeq}(v)$ sont non vides. Ces infima seront encore appelés erreurs optimales, et notés respectivement ϵ_* et ϵ^* .

Notons déjà que si u (resp. v) appartient à \mathcal{K} , alors u (resp. v) est la plus grande (plus petite) solution de (P_*) (resp. (P^*)). On a alors : $\epsilon_* = \|v - u\|$ ($\epsilon^* = \|v - u\|$).

Il semble difficile d'aller plus avant dans l'existence ou le calcul d'une solution, si l'on ne dispose pas d'un minimum d'hypothèses topologiques, géométriques ou algébriques sur \mathcal{K} . Par exemple, si \mathcal{K} est fermé topologiquement, et il en est alors ainsi des ensembles $\mathcal{K}_{\preceq}(u)$ et $\mathcal{K}_{\succeq}(v)$, alors il y a existence d'une solution (généralement non unique). De même, si \mathcal{K} est convexe, ce qui entraîne également que $\mathcal{K}_{\preceq}(u)$ et $\mathcal{K}_{\succeq}(v)$ le sont aussi, alors l'ensemble des solutions est convexe fermé de \mathcal{K} , et donc de \mathbb{E} si \mathcal{K} est fermé. Notons que les deux propriétés élémentaires précédentes, l'une topologique, l'autre géométrique, ne dépendent nullement de la norme choisie. Voici une autre hypothèse géométrique, souvent vérifiée, et assurant le lien entre les trois problèmes. Nous dirons que \mathcal{K} vérifie la *condition du cylindre* s'il est invariant par translation le long de la diagonale \mathcal{L} , c'est-à-dire si pour tout x de \mathcal{K} , $x + c \cdot \mathbf{1}$ est dans \mathcal{K} pour tout réel c . L'ensemble \mathcal{K} est effectivement un cylindre de direction $\mathbf{1}$. Notons que si \mathcal{K} vérifie la condition du cylindre, alors pour tout u de \mathbb{E} , $\mathcal{K}_{\preceq}(u)$ et $\mathcal{K}_{\succeq}(u)$ sont non vides, et \mathcal{L} est incluse dans \mathcal{K} dès que l'origine o est dans \mathcal{K} .

Voici une hypothèse algébrique, qui associée à la condition du cylindre, conduira à des résultats forts et simples. Nous avons noté en définissant (P_*) (resp. (P^*)), l'existence évidente d'une plus grande solution si u (resp. v) est dans \mathcal{K} . Qu'en est-il si cette condition n'est pas remplie, mais que $\mathcal{K}_{\preceq}(u)$ (resp. $\mathcal{K}_{\succeq}(v)$) possède, lorsqu'il n'est pas vide, un plus grand (resp. plus petit) élément noté u_* (resp. v^*), appelé *sous-dominante* (resp. *sur-dominée*) de u (resp. v) dans \mathcal{K} ? Il est évident que u_* (resp. v^*) est encore la plus grande (resp. plus petite) solution de (P_*) (resp. (P^*)), donnant ainsi l'erreur optimale ϵ_* (resp. ϵ^*). Comme observé dans [Chepoi, Fichet, 2000], (cf. aussi [Jardine, Sibson, 1971] dans une définition plus lâche des sous-dominantes), affirmer l'existence d'une sous-dominante (resp. sur-dominée) pour tout élément de \mathbb{E} , est équivalent à dire que \mathcal{K} est *fermé supérieurement* (resp. *inférieurement*).

FIGURE 1. Deux exemples d'ensembles \mathcal{K} dans le plan.

Dans la terminologie classique des treillis ([Birkhoff, 1967]), \mathcal{K} est un sup (resp. inf)-demi-treillis conditionnellement complet, c'est-à-dire que pour tout sous-ensemble non vide et borné A de \mathcal{K} , $\sup(A)$ (resp. $\inf(A)$) est dans \mathcal{K} .

Que pour tout t de \mathbb{E} existe une sous-dominante et/ou une sur-dominée dans \mathcal{K} ([Brucker, Barthélemy, 2007] disent que \mathcal{K} est à sous-dominante et/ou sur-dominée), n'est malheureusement pas toujours le cas au sein des principales structures de classification. Il est toutefois possible, certes au prix de complications, de tirer avantage d'une propriété plus faible, à savoir l'existence dans $\mathcal{K}_{\preceq}(t)$ ($\mathcal{K}_{\succeq}(t)$) d'éléments maximaux (minimaux). Ceux-ci sont alors appelés *inférieures maximales* (*supérieures minimales*) de t dans \mathcal{K} . Pour peu que ces inférieures maximales (supérieures minimales) soient assorties d'autres propriétés, comme d'être en nombre fini, ou surtout que tout élément de $\mathcal{K}_{\preceq}(t)$ ($\mathcal{K}_{\succeq}(t)$) soit majoré (minoré) par l'une d'elles, il sera encore possible d'avancer vers des approximations en norme du supremum. Nous aborderons peu ce sujet ici. Pour plus de résultats dans cette voie, par exemple pour les ultramétriques, on peut consulter [Fichet, 2001], qui dut en particulier généraliser un résultat de [Leclerc, 1986] pour le calcul des supérieures minimales.

Notons enfin que notre terminologie est proche de celle de [Benzécri, 1973], qui nomme également la sous-dominante (sur-dominée) inférieure maxima (supérieure minima), alors que [Jardine, Sibson, 1971], emploient le terme de sous-dominante (sur-dominée) pour toutes les inférieures maximales (supérieures minimales). Enfin, [Brucker, Barthélemy, 2007] dénomment sous-dominante faible une inférieure maximale unique (mais non maximum!).

Tous ces concepts seront analysés par la suite pour chaque structure étudiée. Pour une étude approfondie de ces dernières, leurs implications et leurs imbrications, le lecteur peut consulter [Deza, Laurent, 1997] ou [Critchley, Fichet, 1997]. Le petit exemple de la Figure 1 est destiné à éclairer les précédentes notions.

3. RÉSULTATS GÉNÉRAUX

Nous commençons par deux propriétés très simples concernant le problème dit du sandwich et notre problème général (P) . La brièveté de la démonstration est à comparer à l'approche longue et algorithmique de [Farach, Kannan, Warnow, 1995], dans le cas particulier des ultramétriques et de l'approximation d'une seule dissimilarité. Elle dévoile la puissance de l'outil « sous-dominante ». Faisons état de l'originalité dans cet article de l'écriture de la résolution de (P) à partir du *problème du sandwich*, formulé ainsi : donnés u et v de \mathbb{E} avec $u \preceq v$, existe-t-il z de \mathcal{K} vérifiant $u \preceq z \preceq v$?

PROPRIÉTÉ 1. *Si \mathcal{K} est fermé supérieurement (inférieurement), le problème du sandwich a une solution si et seulement si $v_* \succ u$ ($u^* \preceq v$). Dans ce cas, v_* (u^*) est la plus grande (petite) solution.*

La preuve est immédiate.

PROPRIÉTÉ 2. *Si \mathcal{K} est fermé supérieurement (inférieurement) et vérifie la condition du cylindre, alors pour les deux problèmes $(P), (P_*)$, $((P), (P^*))$:*

- i) $\epsilon_* = 2\hat{\epsilon} = \|v - u_*\|$ ($\epsilon^* = 2\hat{\epsilon} = \|v^* - u\|$)
- ii) (P) a une plus grande (petite) solution, soit $u_* + \hat{\epsilon}\mathbf{1}$ ($v^* - \hat{\epsilon}\mathbf{1}$).

Preuve. Tout d'abord notons que u_* (v^*) est trivialement la plus grande (petite) solution de (P_*) ((P^*)). D'où la caractérisation de ϵ_* (ϵ^*) en terme de norme.

Nous donnons maintenant le complément de démonstration lorsque \mathcal{K} est fermé supérieurement, la totale symétrie de l'énoncé assurant le résultat dual. Pour tout $\epsilon \geq 0$, et pour z de \mathcal{K} , l'inégalité $\max[\|z - u\|, \|z - v\|] \leq \epsilon$ est équivalente à $u - \epsilon\mathbf{1} \preceq z \preceq u + \epsilon\mathbf{1}$ et $v - \epsilon\mathbf{1} \preceq z \preceq v + \epsilon\mathbf{1}$, i.e. $v - \epsilon\mathbf{1} \preceq z \preceq u + \epsilon\mathbf{1}$. (1)

Ainsi, pour qu'existe z à norme inférieure ou égale à ϵ de u et v , il est nécessaire que $v - u \preceq 2\epsilon\mathbf{1}$, soit $\epsilon \geq \frac{1}{2}\|v - u\|$ (ce qui est trivial *a priori*), et dans ce cas z doit répondre au problème du sandwich défini par (1). Il y a une solution si et seulement si $(u + \epsilon\mathbf{1})_* \succ v - \epsilon\mathbf{1}$. Mais par la condition du cylindre : $(u + \epsilon\mathbf{1})_* = u_* + \epsilon\mathbf{1}$, de sorte qu'il y a solution si et seulement si $v - u_* \preceq 2\epsilon\mathbf{1}$. Le plus petit ϵ possible est $\frac{1}{2}\|v - u_*\|$, soit $\frac{1}{2}\epsilon_*$, et satisfait bien la condition nécessaire. C'est donc l'erreur optimale $\hat{\epsilon}$, complétant ainsi i), et la plus grande solution est $u_* + \hat{\epsilon}\mathbf{1}$, prouvant ainsi ii).

Remarquons que si aux hypothèses de la propriété précédente, on ajoute la condition que v (u) a une sur-dominée (sous-dominante), par exemple si v (u) est dans \mathcal{K} , ou si \mathcal{K} est tout à la fois fermé supérieurement et inférieurement (sous-treillis de \mathbb{E}), alors (P) a une plus grande et une plus petite solution. L'ensemble des solutions est l'intervalle de \mathcal{K} : $[v^* - \hat{\epsilon}\mathbf{1}, u_* + \hat{\epsilon}\mathbf{1}]$.

De fait, la propriété précédente est un corollaire direct d'une propriété générale établie par [Chepoi, Fichet, 2000]. Nous ne faisons que la citer.

PROPRIÉTÉ 3. *Si \mathcal{K} vérifie la condition du cylindre, alors pour les trois problèmes $(P), (P_*), (P^*)$:*

- i) $\epsilon_* = \epsilon^* = 2\hat{\epsilon}$,

ii) $\hat{z}, \hat{z} - \hat{\epsilon}\mathbf{1}, \hat{z} + \hat{\epsilon}\mathbf{1}$, sont solutions respectives des trois problèmes, dès que l'une d'entre elles est solution de son problème associé.

Les auteurs pré-cités en déduisent le résultat presque immédiat suivant.

COROLLAIRE 1. *Si \mathcal{K} est convexe, fermé supérieurement (inférieurement) et vérifie la condition du cylindre, et si de plus v (u) a une sur-dominée (sous-dominante), alors $\frac{1}{2}(u_* + v^*)$ est solution de (P) .*

Le *consensus*, appelé aussi *compromis* en analyse des données, joue un rôle fondamental en classification. C'est un domaine extrêmement riche qui a de nombreuses applications en sciences sociales via ce que l'on nomme la *théorie du choix social*. Il existe une vaste bibliographie sur le sujet, et le lecteur peut consulter par exemple les articles de [Barthélemy, Leclerc, Monjardet, 1986], [Leclerc, Monjardet, 1995] ou celui de [Hudry, Monjardet] dans ce volume.

Lorsqu'il peut être placé dans le cadre précis qui est le nôtre, le problème du consensus relève alors directement de nos résultats précédents. Il est défini comme suit : données n vecteurs w_1, \dots, w_n de \mathcal{K} , trouver \hat{z} de \mathcal{K} , solution (si elle existe) de $\inf_{z \in \mathcal{K}} \max_{i=1, \dots, n} \|z - w_i\|$. Comme déjà observé, on est ramené au problème (P) , en définissant $u := \inf(w_1, \dots, w_n)$ et $v := \sup(w_1, \dots, w_n)$. Si \mathcal{K} est fermé supérieurement (inférieurement) et vérifie la condition du cylindre, alors (P) a une plus grande (petite) solution. Mais dans ce cas, v (u) est dans \mathcal{K} , de sorte (P) a aussi une plus petite (grande) solution. C'est un cas déjà évoqué et l'on a ainsi :

COROLLAIRE 2. *Si \mathcal{K} est fermé supérieurement (inférieurement) et vérifie la condition du cylindre, alors le problème du consensus a pour solutions l'intervalle de \mathcal{K} : $[v - \hat{\epsilon}\mathbf{1}, u_* + \hat{\epsilon}\mathbf{1}]$ ($[v^* - \hat{\epsilon}\mathbf{1}, u + \hat{\epsilon}\mathbf{1}]$), où $\hat{\epsilon}$ est égal à $\frac{1}{2}\|v - u_*\|$ ($\frac{1}{2}\|v^* - u\|$).*

Comme noté par [Chepoi, Fichet, 2000], il est dans l'esprit même du consensus de proposer un compromis unique, en dépit du fait que la norme du supremum ne donne que rarement une solution unique. Sous les hypothèses du corollaire précédent, c'est un intervalle, que nous noterons $[u_1, v_1]$. Pour pallier cette profusion de solutions (supérieures aux données initiales!), les auteurs appliquent une procédure dite de *consensus universel* à celles-ci. Elle consiste à réduire récursivement un intervalle de \mathcal{K} à un sous-intervalle. Supposant que \mathcal{K} est fermé supérieurement, nous la décrivons comme suit (il existe bien sûr un algorithme similaire si \mathcal{K} est fermé inférieurement).

Partant de $[u_1, v_1]$ et nous trouvant à l'étape m avec $[u_m, v_m]$, on intersecte ce dernier avec l'intervalle, soit $[u'_m, v'_m]$, consensus de lui-même. Il est prouvé que c'est un intervalle de \mathcal{K} , soit $[u_{m+1}, v_{m+1}]$, où $u_{m+1} = \sup(u_m, u'_m)$ et $v_{m+1} = (\inf(v_m, v'_m))^*$. Les auteurs cités démontrent alors la propriété suivante.

PROPRIÉTÉ 4. *Si \mathcal{K} est fermé supérieurement et vérifie la condition du cylindre, la suite $\{[u_m, v_m]\}_{m \in \mathbb{N}}$ du consensus universel converge en au plus p étapes vers un singleton, appelé consensus universel de $[u_1, v_1]$.*

4. PREMIÈRES APPLICATIONS

Les résultats généraux précédents offrent des approximations en norme du supremum d'un calcul aisé, pour peu qu'existent ou soient établies des procédures numériques

simples pour l'obtention d'une sous-dominante et/ou d'une sur-dominée. Nous exhibons dans ce paragraphe un grand nombre de structures répondant à ce critère. Elles sont de deux types, les régressions et les dissimilarités.

4.1. LES RÉGRESSIONS

En première ébauche, le modèle classique de *régression* est celui d'une boîte noire, avec une variable d'entrée x qui peut être soit numérique uni ou multidimensionnelle, soit discrète, et une variable de sortie (généralement) numérique Y , qui en statistique est aléatoire. La variable d'entrée x , dite explicative, est appelée *prédicteur*. Elle peut être déterministe (modèle à effet fixe) ou aléatoire (modèle à effet aléatoire). La variable Y est dite expliquée, et nommée *réponse*. La fonction de régression, soit $y = f(x)$, n'est autre que l'espérance de Y à x fixé.

En pratique, on dispose d'un n -échantillon observé, que par commodité on peut ranger selon les entrées distinctes (choisies ou observées) $\{x_i, i = 1, \dots, p\}$, le i -ème niveau d'entrée donnant n_i sorties observées $\{y_{ij}, j = 1, \dots, n_i\}$, avec $\sum_i n_i = n$. Pour des régressions d'un type donné, précisé par le modèle, et dans un cadre non-paramétrique, une démarche consiste à estimer les valeurs de la régression aux points de l'ensemble $X = \{x_i, i = 1, \dots, p\}$, puis à étendre celle-ci au domaine général (si possible). Le critère d'estimation le plus usuel est celui des moindres carrés, voire des moindres écarts absolus. Nous donnons deux exemples avec la norme du supremum, pour lesquels existe une solution simple. Posant $z_i = f(x_i)$, on doit minimiser $\max_i \max_{j=1, \dots, n_i} |z_i - y_{ij}|$, ou de manière équivalente, comme nous l'avons déjà vu, minimiser $\max_i \max[|z_i - u_i|, |z_i - v_i|]$, où pour tout i , $u_i = \min_{j=1, \dots, n_i} y_{ij}$ et $v_i = \max_{j=1, \dots, n_i} y_{ij}$.

Exemple 1. Les régressions convexes.

Dans le modèle de régression *convexe* h -dimensionnelle, on considère l'espace vectoriel \mathbb{F} des fonctions numériques définies sur \mathbb{R}^h , et dans cet espace le cône convexe \mathcal{C} des fonctions convexes. Dans \mathbb{F} muni de la relation d'ordre partielle classique définie par $f \preceq g$ si et seulement si $f(x) \leq g(x)$ pour tout x de \mathbb{R}^h , il est bien connu et immédiat de constater, que le supremum d'une famille de fonctions convexes admettant un majorant, est convexe. Nous dirons encore que \mathcal{C} est fermé supérieurement. Bien que \mathcal{C} soit également invariant par translation le long de la droite engendrée par une fonction constante, remarquons que, dû à la dimension infinie, il peut exister une fonction de \mathbb{F} non minorée par un élément de \mathcal{C} (pour un contre-exemple, prendre $h = 1$ et $f(x) = -|x|$). En revanche, pour une fonction f minorée par un élément de \mathcal{C} , il existe un plus grand élément de \mathcal{C} minorant f , que l'on appellera encore sous-dominante (notée f_*) de f .

Revenant aux données observées et à la stratégie d'estimation proposée, l'espace vectoriel de base est l'ensemble des restrictions à X des éléments de \mathbb{F} , que l'on peut identifier à \mathbb{R}^X . L'ensemble de référence, noté \mathcal{K}_c , est celui des restrictions à X des éléments de \mathcal{C} . Ces deux ensembles héritent des propriétés de \mathbb{F} et \mathcal{C} , de sorte que \mathcal{K}_c est fermé supérieurement et vérifie la condition du cylindre. En outre, il existe ici une sous-dominante pour tout vecteur de \mathbb{R}^X . Les résultats généraux s'appliquent, tant pour la résolution du problème d'estimation, que pour une réponse au problème

du sandwich ou l'obtention d'un consensus universel. En particulier, l'estimation obtenue est toute extension convexe de $u_* + \frac{1}{2}\|v - u_*\|\mathbf{1}$.

Sur le plan algorithmique, tout repose donc sur le calcul d'une sous-dominante, soit t_* , d'une fonction t de \mathbb{R}^X , à extension convexe. Pour tout $i = 1, \dots, n$, notons M^i le point de coordonnées $(x_i, t(x_i))$ dans $\mathbb{R}^h \times \mathbb{R}$. Dans le cas uni-dimensionnel ($h = 1$), il existe un algorithme simple, intuitif, et aisément justifiable pour le calcul de t_* . Parmi les demi-droites partant de M^1 et passant par les différents $M^i, i > 1$, on cherche celle de plus petite pente, et soit j le plus grand indice tel que M_j est sur celle-ci. Puis, parmi les demi-droites partant de M^j et passant par les différents $M^i, i > j$, on cherche celle de plus petite pente, et soit k le plus grand indice tel que M^k est sur cette dernière. On poursuit ainsi jusqu'à rejoindre M^p . Les segments M^1M^j, M^jM^k, \dots ainsi construits, définissent le graphe d'une fonction linéaire par morceaux sur $[x_1, x_p]$, soit g , que l'on peut étendre à \mathbb{R} par extension des segments extrêmes. Cette fonction est convexe. Pour s'en convaincre, il suffit d'appliquer un théorème célèbre de la convexité, stipulant qu'une fonction numérique f sur \mathbb{R}^h est convexe, si et seulement si l'ensemble $\{(x, y) \in \mathbb{R}^h \times \mathbb{R} : x \in \mathbb{R}^h, y \geq f(x)\}$ est convexe, voir par exemple [Berge, 1966] ou [Rockafellar, 1970]. La fonction g a une restriction à X , soit $g|_X$, inférieure ou égale à t , et par l'inégalité de la convexité, toute fonction convexe f vérifiant $f|_X \preceq t$, est telle que $f \preceq g$ sur $[x_1, x_p]$. Ceci montre que $g|_X = t_*$, et même que g est la plus grande extension convexe sur $[x_1, x_p]$. Les valeurs $t_*(x_i)$ sont les ordonnées de la projection M_*^i de M^i sur g , parallèlement à l'axe des ordonnées. La Figure 2 illustre le calcul d'une sous-dominante convexe.

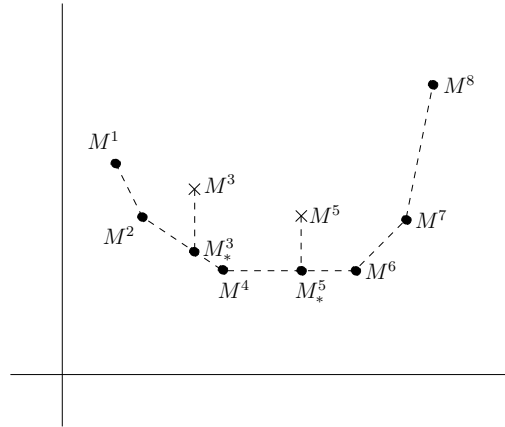


FIGURE 2. Un exemple d'approximation par en dessous pour les régressions convexes.

La procédure numérique est plus compliquée dans le cas h -dimensionnel, mais repose sur les mêmes principes. Notons que dans le cas $h = 1$, la procédure proposée pour construire la fonction g linéaire par morceaux, n'est que le début du calcul de la fermeture convexe des M^i , le graphe de g n'étant autre que la partie inférieure de celle-ci. Dans le cas général, la frontière du convexe $C := \text{conv}\{M^i, i = 1, \dots, p\}$ de $\mathbb{R}^h \times \mathbb{R}$ se décompose encore en une partie supérieure C^* et une partie inférieure C_* : la droite D_x parallèle à l'axe des ordonnées et d'abscisse x de $C_0 := \text{conv}\{X\}$, intersecte C selon un intervalle fermé, la valeur supérieure (inférieure) de l'intervalle,

définissant l'élément d'abscisse x de C^* (C_*). Ainsi, C_* définit encore le graphe d'une fonction g définie sur C_0 , qui s'avère être encore convexe et linéaire par morceaux. Sa restriction $g|_X$ est trivialement inférieure à t et majore encore toute fonction convexe f vérifiant $f|_X \preceq t$. Ceci montre encore que $g|_X = t_*$. Pour tous les concepts sus-évoqués, tant théoriques que numériques, le lecteur peut consulter [Eldsbrunner, 1987]. Mais notons qu'ici, nous n'avons nul besoin de rechercher C , ni même C_* . Seules suffisent les ordonnées des M_*^i , qui peuvent être obtenues point par point par programmation linéaire.

Exemple 2. Les régressions isotones.

Une fonction de régression isotone est un élément de l'espace des fonctions numériques définies sur un espace partiellement ordonné (χ, \preceq) . Ce dernier peut être discret, mais aussi continu comme \mathbb{R}^h muni de l'ordre partiel usuel défini coordonnée par coordonnée, ou simplement \mathbb{R} et dans ce cas l'ordre est total. Une régression f , élément de \mathbb{R}^χ , est *isotone* si et seulement si : $x \preceq x'$ implique $f(x) \leq f(x')$. Pour des inégalités inverses sur f , la régression est dite *antitone*, et un simple changement de signe de f transforme tout résultat d'isotonie en un résultat similaire d'antitonie.

Depuis les travaux pionniers de [Ayer *et al.*, 1955] et de [van Eeden, 1958] pour l'estimation au sens des moindres carrés d'une régression isotone sur \mathbb{R} (le problème, dit maintenant du sandwich, étant même présent chez le dernier auteur), travail rendu populaire par [Kruskal, 1964] via l'algorithme des blocs, de nombreux résultats sont venus étoffer le sujet. Nous citons l'article de [Thompson, 1962] et l'ouvrage de [Barlow *et al.*, 1972], pour une extension aux arbres enracinés de la chaîne liée à l'ordre total, avec le critère des moindres carrés. Pour une étude complète avec un critère de type L_p ($1 \leq p < \infty$), mais restreinte aux ordres partiels ayant un arbre comme graphe de couverture, on peut consulter [Chepoi, Cogneau, Fichet, 1997]. Les travaux plus récents de [Stout, 2008] s'inscrivent également dans cette démarche. Transcendant les normes L_p pour le critère, [Robertson, Wright, 1974, 1980] étendent leur étude jusqu'aux « Cauchy mean value functions ». Néanmoins, et bien que le cône des régressions isotones soit polyédrique, se prêtant ainsi à des algorithmes spécifiques de programmation quadratique pour la norme L_2 , comme ceux présents dans [Lawson, Hanson, 1974] ou celui de [Dykstra, 1983], ou à des algorithmes de programmation linéaire pour la norme L_1 , il apparaît que nombre d'algorithmes proposés pour une norme L_p sont non polynomiaux, voir, par exemple, pour $p = 2$ et pour une régression bi-variée [Dykstra, Robertson, 1982].

Avec la norme du supremum ($p = \infty$), l'estimation peut être calculée aisément, [Chepoi, Fichet, 2000], quel que soit le graphe de couverture, voir également [Stout, 2009] pour des résultats de complexité. Notons d'abord, qu'à l'opposé des régressions convexes qui nécessitent directement ou implicitement une extension, pour une définition sur le domaine discret X des données, les régressions isotones sur échantillon, héritent de facto de la structure d'ordre restreinte à X . Elles sont donc ainsi définies, pour être ensuite étendues à l'espace entier χ . Ainsi notre espace de référence est encore \mathbb{R}^X et l'ensemble de référence, noté \mathcal{K}_{is} , est l'ensemble des fonctions isotones sur (X, \preceq) . On constate immédiatement que \mathcal{K}_{is} jouit de toutes les propriétés souhaitées : c'est un cône convexe, fermé supérieurement et inférieurement, qui vérifie la condition du cylindre. En outre, une

sous-dominante (sur-dominée) d'une fonction t de \mathbb{R}^X se calcule aisément via la formule : $t_*(x_i) = \min\{t(x_j) : x_j \succcurlyeq x_i\}$ ($t^*(x_i) = \max\{t(x_j) : x_j \preccurlyeq x_i\}$), admettant une plus grande (petite) extension sur le « support » de X dans χ , soit $t_*(x) = \min\{t(x_j) : x_j \succcurlyeq x\}$ ($t^*(x) = \max\{t(x_j) : x_j \preccurlyeq x\}$). Ici, nous avons appelé "support" de X , l'ensemble des x de χ tels qu'existent x_i et x_j de X , vérifiant : $x_i \preccurlyeq x \preccurlyeq x_j$. Ainsi notre problème général d'estimation possède une plus grande et une plus petite solution : $u_* + \frac{1}{2}||v - u_*||\mathbf{1}$ et $v^* - \frac{1}{2}||v^* - u||\mathbf{1}$ respectivement, qui toutes deux conduisent au compromis universel : $\frac{1}{2}(u_* + v^*)$, retrouvant ici un résultat de [Ubhaya, 1974] pour les régressions sur \mathbb{R}^h . La Figure 3 illustre la sous-dominante isotone dans le cas d'une régression uni-dimensionnelle.

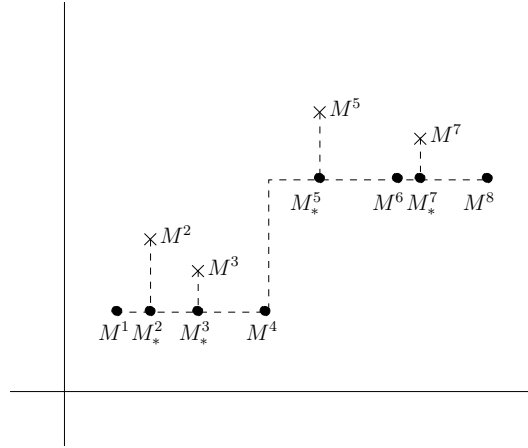


FIGURE 3. Un exemple d'approximation par en dessous pour les régressions isotones.

Signalons ici une propriété spécifique à la norme du supremum, qui sera extrêmement bénéfique dans certains exemples. Elle dérive des formules simples établissant sous-dominante et sur-dominée. L'erreur optimale $\hat{\epsilon}$ appartient à l'ensemble fini (polynomial en les données) : $\Delta := \{\frac{1}{2}(v_i - u_j) : x_j \succcurlyeq x_i, v_i \geq u_j\}$. Ainsi, fort de cette propriété : le problème général (P) se trouve être équivalent à un problème *a priori* plus simple :

(P') : pour $\epsilon \geq 0$ fixé, existe-t-il un vecteur z de \mathcal{K} tel que $\max[||z - u||, ||z - v||] \leq \epsilon$. Un tel vecteur s'il existe, sera dit ϵ - \mathcal{K} , en précisant selon la nature de \mathcal{K} . Ici, nous parlerons d'une régression z de \mathbb{R}^X ϵ -isotone. Pour résoudre (P) à partir de (P') , il suffit de parcourir les erreurs potentielles de Δ , pour trouver la plus petite possible, ce qui peut être fait par une recherche binaire en rangeant les éléments de Δ par ordre croissant, et en soumettant alors la médiane à (P') .

Naturellement, il est inutile dans le cadre des régressions isotones, de chercher à résoudre (P) par cette voie, puisque l'on a déjà l'erreur optimale et les solutions. Cependant, dans ce même cadre, il existe différentes applications de la propriété citée, que le lecteur trouvera dans [Chepoi, Cogneau, Fichet, 1997]. En voici une, que nous décrivons brièvement, et qui est à rapprocher de la méthode des « K-means », exprimée en termes de régressions isotones et non constantes, et avec la norme du supremum et non la norme euclidienne. Pour simplifier, nous supposons qu'il y a

une seule réponse y_i par entrée x_i . Ainsi $y = u = v$. On cherche une partition de X en K groupes X_1, \dots, X_K et une régression isotone optimale dans chaque groupe, telles que soit minimum la plus grande des erreurs optimales (résiduelles) dans les groupes. Il est clair que ce minimum appartient encore à l'ensemble $\Delta := \{\frac{1}{2}(y_i - y_j) : x_j \succ x_i, y_i \geq y_j\}$. On peut donc travailler à ϵ fixé dans Δ , et chercher une partition telle que dans chaque groupe l'erreur soit inférieure à ϵ . Pour ce faire, les auteurs introduisent un graphe Γ_ϵ , ayant pour sommets les éléments de X . Deux sommets x_i et x_j sont non-adjacents si et seulement si ils sont comparables, avec $(y_i - y_j) > 2 \cdot \epsilon$ lorsque $x_j \succ x_i$. Ils démontrent alors que l'erreur globale optimale est inférieure ou égale à ϵ si et seulement si Γ_ϵ peut être couvert par K cliques. Un tel problème de couverture par des cliques est en général NP-complet. Il est aisé en revanche pour K égal à deux, puisque cela revient à prouver que le graphe complémentaire $\overline{\Gamma}_\epsilon$ est biparti, les deux composantes donnant une bi-partition souhaitée de X .

4.2. LES DISSIMILARITÉS

Les dissimilarités, mesures de dissemblance entre objets, jouent un rôle capital en analyse des données et en classification. La plupart des méthodes utilisées dans ces disciplines reposent sur de telles mesures. Présentes parfois de manière implicite dans certaines méthodes classiques, comme l'analyse en composantes principales, elles sont souvent aussi calculées directement à partir des données premières observées.

Mathématiquement, une *dissimilarité* sur un ensemble X à n éléments, est une application d de X^2 dans \mathbb{R}_+ vérifiant *i)* $d(x, y) = d(y, x)$ et *ii)* $d(x, x) = 0$, pour tout x et tout y de X . Elle est dite *propre*, si $d(x, y) = 0$ implique $x = y$. Sur le plan théorique, il est commode d'accepter des valeurs négatives, on parle alors de *pré-dissimilarités*, avec des fonctions d à valeurs dans \mathbb{R} , tout en préservant symétrie et diagonale nulle. L'ensemble des pré-dissimilarités forme alors un espace vectoriel de dimension finie $p := n(n-1)/2$. Ce sera notre espace de base noté \mathcal{D} . Il sera muni de sa base canonique, et de la norme du supremum relativement à cette base. Ainsi les dissimilarités ne forment autre que l'orthant positif \mathcal{D}_+ de \mathcal{D} . Nous noterons d_0 et d_1 respectivement, la dissimilarité nulle de \mathcal{D} et la dissimilarité égale à 1 sur tout couple d'éléments distincts, engendrant la diagonale de \mathcal{D} .

Définies ainsi, les dissimilarités sont pauvres d'un point de vue axiomatique, et ne sauraient conduire à des résultats probants. Aussi fait-on appel à des dissimilarités particulières, en liaison souvent avec des modes de représentation graphique souhaitée ou imposée par un modèle sous-jacent, plans factoriels euclidiens ou rectilinéaires, arbres enracinés ou non, etc... En pratique, il convient d'approcher une dissimilarité issue d'un tableau de données par une dissimilarité d'un type fixé, *i.e.* appartenant à un ensemble de référence inclus dans \mathcal{D}_+ . Les exemples qui vont suivre caractérisent de telles dissimilarités, offrant en outre des approximations en norme du supremum d'un calcul aisé. Mais auparavant, faisons quelques remarques relativement à une structure, non nécessairement présente en classification et souvent non suffisamment riche lorsqu'elle est présente, quoique fondamentale dans le vaste champ des mathématiques : les espaces métriques.

Une dissimilarité d sur X est une *semi-distance* (ou métrique) si elle vérifie l'*inégalité triangulaire* : $d(x, y) \leq d(x, z) + d(y, z)$, pour tout x, y, z de X . C'est une

distance si d est propre. Dans ce cas, (X, d) est appelé *espace métrique*. Géométriquement, l'ensemble noté \mathcal{D}_m des semi-distances forme un cône convexe polyédrique fermé de X , comme intersection de demi-espaces définis par les inégalités triangulaires (il est facile de voir que pour $n > 2$, celles-ci entraînent la positivité). En outre, \mathcal{D}_m est clairement fermé supérieurement, de sorte que toute dissimilarité (minorée comme telle par la semi-distance d_0) possède une sous-dominante métrique d_* . De plus, il existe un algorithme simple pour la calculer, par réductions successives de la plus grande valeur sur les triplets violant l'inégalité triangulaire. Cependant, \mathcal{D}_m ne vérifie pas la condition du cylindre. Elle ne reste invariante que par translation le long du demi-axe engendré par d_1 . En conséquence, nos résultats généraux ne s'appliquent pas. Tout juste pouvons-nous affirmer que l'erreur optimale pour une approximation en norme du supremum est au plus $\frac{1}{2}\|d - d_*\|$. Notons toutefois, que dûe à la structure polyédrique, une solution numérique peut être obtenue par programmation linéaire. Mais cette absence de caractérisation s'avèrera pénalisante lorsque nous étudierons les structures arborées.

Exemple 3. Les ultramétriques.

On rappelle qu'une dissimilarité d est dite (être une) *ultramétrique* si elle vérifie l'*inégalité ultramétrique* : $d(x, y) \leq \max[d(x, z), d(y, z)]$, pour tout x, y, z de X . On notera \mathcal{D}_u l'ensemble des ultramétriques et on a donc : $\mathcal{D}_u \subset \mathcal{D}_m$ (pour $n > 2$).

Les ultramétriques jouent un rôle capital en classification, pour être connectées aux hiérarchies indicées. Rappelons qu'une *hiérarchie* (totale) \mathcal{H} sur X est une classe de parties non vides de X vérifiant *i)* $X \in \mathcal{H}$, *ii)* $\{x\} \in \mathcal{H}$ pour tout x de X , *iii)* $H \cap H' \in \{H, H', \emptyset\}$, pour tout $H, H', H \neq H'$, de \mathcal{H} . Une hiérarchie (totale) *indicée* est un couple (\mathcal{H}, f) , où \mathcal{H} est une hiérarchie (totale) et où f est un *indice de niveau*, *i.e.* une fonction de \mathcal{H} dans \mathbb{R}_+ vérifiant *i)* $f(H) < f(H')$ dès que $H \subset H'$, *ii)* les éléments minimaux (singletons) ont pour niveau 0. Graphiquement, une hiérarchie est représentée sous forme d'un *dendrogramme*, *i.e.* un arbre enraciné qui n'est autre que le graphe de couverture de \mathcal{H} pour l'ordre donné par l'inclusion. Une bijection bien connue met alors en correspondance l'ensemble des hiérarchies (totales) indicées et l'ensemble des ultramétriques (propres), voir [Johnson, 1967], ou les ouvrages de [Jardine, Sibson, 1971], [Benzécri, 1973].

Par commodité, il sera utile d'étendre \mathcal{D}_u aux pré-dissimilarités qui vérifient l'inégalité ultramétrique pour tout triplet $\{x, y, z\}$ d'éléments distincts de X (la restriction aux dissimilarités, étant équivalente à notre définition première). On notera \mathcal{D}'_u cet ensemble. Alors \mathcal{D}'_u vérifie la condition du cylindre. On voit facilement que, comme \mathcal{D}_u , il est fermé supérieurement, de sorte que pour d de \mathcal{D} (\mathcal{D}_+) il existe une sous-dominante d_* dans \mathcal{D}'_u (\mathcal{D}_u , puisque $d_0 \in \mathcal{D}_u$). En outre, on dispose de nombreux algorithmes performants pour calculer cette sous-dominante. Les deux plus usuels sont l'algorithme dit du « lien simple », algorithme de construction ascendante d'une hiérarchie indicée directement à partir de d , qui s'avère être celle associée à d_* [Johnson, 1967], ainsi que celui construisant l'ultramétrique associée à un arbre couvrant minimum issu du graphe complet pondéré donné par d [Gower, Ross, 1969]. Pour d'autres algorithmes, signalons [Roux, 1968], voir aussi [Jardine, Sibson, 1971], par réductions successives des triangles, [Lerman, 1970] par la même

technique, mais après avoir rangé les données par ordre croissant, et l'algorithme « SLINK » de [Sibson, 1972].

Ainsi, les résultats généraux s'appliquent ici. En particulier, la plus grande approximation de d , est donnée par : $\hat{d} = d_* + \frac{1}{2}||d - d_*|| \cdot d_1$. Notons que \hat{d} est toujours propre si d n'est pas ultramétrique, ce qui peut paraître choquant. Nous laissons au lecteur à titre d'exercice, le soin de construire une solution optimale s'annulant sur les couples où s'annule d . Le consensus, avec sa solution universelle, est obtenu aussi selon la voie tracée.

Terminons par une remarque. Il peut être intéressant d'approcher d à hiérarchie fixée \mathcal{H} , c'est-à-dire trouver l'indigage f sur \mathcal{H} telle que l'ultramétrique en correspondance avec (\mathcal{H}, f) soit une plus proche de d . Par exemple, dans un algorithme de construction ascendante, on peut s'interroger non pas sur la hiérarchie obtenue, mais sur l'indigage proposé par l'algorithme. Pourquoi ne pas adapter d directement à la hiérarchie ? En théorie du consensus, différentes approches ne proposent qu'une hiérarchie, comme par exemple la hiérarchie de la règle majoritaire. On pourrait alors indiquer celle-ci en approchant au mieux les ultramétriques originelles (si elles existent). Ceci peut être fait aisément au sens de toute norme L_p . En particulier, pour $p = \infty$, nous sommes face à un problème de régression isotone comme traité précédemment.

Exemple 4. Ordres compatibles et dissimilarités de Robinson.

Les mesures de dissemblance/ressemblance qui concernent cet exemple ont été introduites par [Robinson, 1951] en termes de similarité, pour des problèmes de sériation chronologique en archéologie, et portent aujourd'hui son nom. Elle reposent sur le concept de compatibilité. Une dissimilarité d et un ordre total \preceq sur X sont dits *compatibles*, si pour tout x, y, z de X , avec $x \preceq y \preceq z$, on a : $d(x, z) \geq \max[d(x, y), d(y, z)]$. En termes de matrice, lorsque l'on range lignes et colonnes selon l'ordre de X , la matrice de dissimilarité est croissante en lignes et colonnes au dessus de la diagonale principale. Une dissimilarité d est dite *de Robinson* (ou Robinsonienne), s'il existe un ordre compatible avec d . Reconnaître si une dissimilarité est de Robinson, c'est-à-dire répondre par oui ou non à l'existence d'ordres compatibles et décrire ceux-ci s'ils existent, est un vrai problème algorithmique. Longtemps d'ailleurs il fut fait usage d'algorithmes non polynomiaux, bien qu'efficaces dans leurs contextes. Pour des algorithmes polynomiaux citons ceux de [Mirkin, Rodin, 1984] fondés sur la reconnaissance des graphes d'intervalles, ceux de [Chepoi, Fichet, 1997] ou de [Seston, 2008] fondés sur une stratégie de « diviser pour régner », ou celui de [Préa, Fortin, 2009] de complexité optimale.

Par nature, les dissimilarités de Robinson, jouent un rôle en sériation. Mais par delà cet aspect, elles sont à la base d'une généralisation de la classification hiérarchique, en offrant un dendrogramme pyramidal pourvu de classes empiétantes. On définit une *pseudo-hiérarchie* (totale) \mathcal{H} sur X , comme une classe de parties non vides de X , satisfaisant les axiomes *i*), *ii*) d'une hiérarchie, l'axiome *iii*) d'emboîtement étant remplacé par *iii*)' : $H \cap H' \in \mathcal{H} \cup \{\emptyset\}$, pour tout H, H' de \mathcal{H} , et où s'ajoute *iv*) : il existe un ordre total sur X tel que tout H de \mathcal{H} est un intervalle selon cet ordre. L'indigage se définit de la même manière, en distinguant

cette fois, l'indigage (strict par définition) d'un indigage faible où deux classes emboîtées peuvent être au même niveau. Deux bijections ont alors été établies, entre d'une part pseudo-hiérarchies totales faiblement indicées et dissimilarités propres Robinsoniennes [Bertrand, Diday, 1985], et d'autre part pseudo-hiérarchies (totales) indicées et dissimilarités (propres) Robinsoniennes particulières, appelées fortement de Robinson [Durand, Fichet, 1988].

Nous notons \mathcal{D}_{\preceq} l'ensemble des dissimilarités compatibles avec l'ordre \preceq . Comme pour les ultramétriques, nous étendons cet ensemble aux pré-dissimilarités obéissant sur les couples d'éléments distincts aux mêmes inégalités de définition. On note \mathcal{D}'_{\preceq} cette extension. Alors, clairement \mathcal{D}'_{\preceq} est un cône convexe polyédrique, fermé supérieurement et inférieurement, et vérifie la condition du cylindre. Toute dissimilarité d possède donc une sous-dominante d_* dans \mathcal{D}'_{\preceq} (de fait, dans \mathcal{D}_{\preceq}) et une surdominée d^* . En outre, des formules très simples et immédiates les caractérisent. Pour tout x, y de X , avec $x \prec y$, on a : $d_*(x, y) = \min[d(z, t) : z \preceq x \prec y \preceq t]$ et $d^*(x, y) = \max[d(z, t) : x \preceq z \prec t \preceq y]$. Ainsi, il existe une plus grande et une plus petite approximation de d dans \mathcal{D}'_{\preceq} , conduisant toutes deux à une solution consensus universel, de fait dans \mathcal{D}_{\preceq} , $\hat{d} = \frac{1}{2}(d_* + d^*)$, l'erreur optimale étant $\hat{\epsilon} = \frac{1}{2}\|d - d_*\| = \frac{1}{2}\|d^* - d\|$.

Remarquons, que comme pour les régressions isotones, l'erreur optimale appartient à un ensemble fini $\Delta = \{\frac{1}{2}|d(x, y) - d(z, t)| : x, y, z, t \in X\}$. Ce n'est d'ailleurs pas une coïncidence. La solution trouvée, est celle d'un problème de régression isotone sur l'ensemble des paires de X , muni de l'ordre partiel défini par $\{x, y\} \preceq \{z, t\}$ si et seulement si $z \preceq x \prec y \preceq t$ sur X . Pour une approximation globale dans l'ensemble, noté \mathcal{D}_r , des dissimilarités de Robinson, ce qui sera étudié ultérieurement, il est clair que l'erreur optimale est égale au minimum, sur tous les ordres totaux sur X , des erreurs optimales à ordre fixé. Ainsi elle appartient encore à Δ , et il y a équivalence entre le problème général et la recherche d'un élément de \mathcal{D}_r à distance de d inférieure à ϵ fixé dans Δ . Comme pour les régressions isotones, on parlera alors de dissimilarités ϵ -Robinsoniennes ou de dissimilarités et d'ordres ϵ -compatibles.

Notons enfin, que comme pour les ultramétriques, on peut travailler encore à pseudo-hiérarchie fixée, et donc à ordre fixé, pour trouver un indigage (au sens large), et donc une dissimilarité Robinsonienne, approchant au mieux une dissimilarité d . Le consensus selon la règle majoritaire de pseudo-hiérarchies possédant un ordre compatible commun, en est encore un exemple d'application. Là encore, nous sommes face à un problème de régression isotone.

La Figure 4 présente une dissimilarité d , son ultramétrique sous-dominante d'' et la sous-dominante d' compatible avec un ordre choisi compatible avec d'' . Ainsi d'' est la sous-dominante ultramétrique de d' .

Exemple 5. Les ultramétriques partielles.

On appelle *vecteur partiel* de \mathbb{E} un vecteur qui n'est défini que par ses coordonnées le long d'un sous-ensemble précis de vecteurs de base de \mathbb{E} . Il appartient donc de fait, non pas à \mathbb{E} , mais à un espace $\overline{\mathbb{E}}$, que l'on peut identifier à un sous-espace de \mathbb{E} . La structure de référence \mathcal{K} a une trace sur $\overline{\mathbb{E}}$, soit $\overline{\mathcal{K}}$, identifiable à une projection. Ainsi, dans l'esprit de ce qui a été fait pour les régressions, nous dirons qu'un vecteur

d	x_2	x_3	x_4	x_5	x_6
x_1	1	5	2	3	3
x_2	0	1	2	2	3
x_3		0	7	2	9
x_4			0	1	1
x_5				0	4

d'	x_2	x_3	x_4	x_5	x_6
x_1	1	2	2	3	3
x_2	0	1	2	2	3
x_3		0	2	2	3
x_4			0	1	1
x_5				0	1

d''	x_2	x_3	x_4	x_5	x_6
x_1	1	1	2	2	2
x_2	0	1	2	2	2
x_3		0	2	2	2
x_4			0	1	1
x_5				0	1

FIGURE 4. Une dissimilarité d , sa sous-dominante compatible avec l'ordre $x_1 \preceq x_2 \preceq x_3 \preceq x_4 \preceq x_5 \preceq x_6$ (d') et sa sous-dominante ultramétrique (d'').

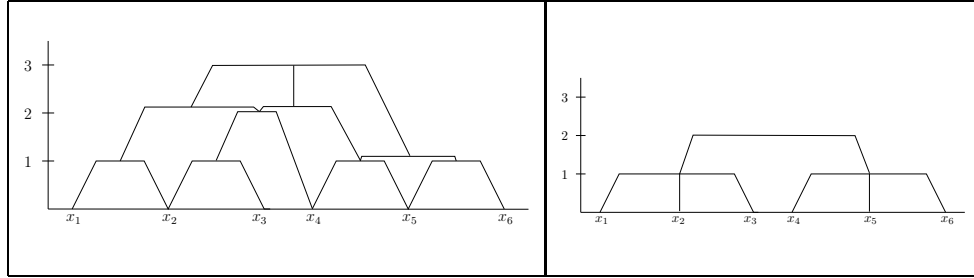


FIGURE 5. Les représentations pseudo-hiérarchiques et hiérarchiques des dissimilarités d' et d'' de la figure 4.

est partiel- \mathcal{K} , en précisant selon la nature de \mathcal{K} , s'il est un élément de $\overline{\mathcal{K}}$, *i.e.* s'il admet une extension dans \mathcal{K} .

Dans cet exemple, nous portons notre intérêt sur les *ultramétriques partielles* qui ne sont définies que sur un ensemble A de paires de X . Elles appartiennent à $\overline{\mathcal{D}}'_u$ de $\overline{\mathcal{D}}$, et une dissimilarité partielle d de \mathcal{D} est ultramétrique si elle admet une extension dans \mathcal{D}'_u , de fait dans \mathcal{D}_u . La structure partielle $\overline{\mathcal{D}}'_u$ hérite des propriétés de \mathcal{D}'_u : elle est fermée supérieurement et vérifie la propriété du cylindre dans $\overline{\mathcal{D}}$. Toute (pré)-dissimilarité partielle d admet donc une sous-dominante ultramétrique partielle d_* .

Vérifier si une dissimilarité partielle d est ultramétrique est aisé. En effet, si \underline{d} est une extension adéquate de d , alors $d' := \sup(\underline{d}, d_0)$ est encore une extension ultramétrique, puisque \mathcal{D}'_u est fermé supérieurement, et $d'' := \inf(\underline{d}, M \cdot d_1)$ où M est le maximum des valeurs de d , est une autre extension ultramétrique, comme cela se constate facilement. On est donc ramené au problème du sandwich : d est ultramétrique partielle si et seulement si $d''_* \succcurlyeq d'$, *i.e.* si d''_* coïncide avec d sur les paires de A . On note que pour le calcul de d''_* , rechercher un arbre couvrant minimum du graphe pondéré complet associé à d'' , revient à rechercher une forêt couvrante minimum du graphe pondéré (X, A) associé à d (appliquer l'algorithme de Kruskal ou celui de Prim pour s'en convaincre). Ainsi, lorsque d est une ultramétrique partielle, d''_* est la plus grande extension ultramétrique de d si (X, A) est connexe. Si d n'est pas ultramétrique, la restriction de d''_* aux paires de A est clairement la sous-dominante d_* de d . Alors $d_* + \frac{1}{2}||d - d_*||_{d_1|_A}$ est la plus grande ultramétrique partielle optimale.

Exemple 6. Les k -ultramétriques.

Au chapitre 8 de leur ouvrage, [Jardine, Sibson, 1971] introduisent les k -ultra-

métriques faibles et fortes pour tout entier $k \leq n - 2$. Une dissimilarité d sur X est dite faiblement (fortement) k -ultramétrique, si pour tout sous-ensemble Y de X de cardinalité $(k + 2)$ et si pour tout x et tout y distincts de Y on a : $d(x, y) \leq \max\{d(z, t) : z, t \in Y, z \neq t, \{z, t\} \neq \{x, y\}\}$ (inégalité k -ultramétrique faible), ou $d(x, y) \leq \max\{\max[d(x, z), d(y, z)] : z \in Y \setminus \{x, y\}\}$ (inégalité k -ultramétrique forte). En d'autres termes, il y a au moins deux paires distinctes dans Y réalisant la valeur maximum de d sur Y , quelconques pour l'inégalité faible, et avec un sommet commun pour l'inégalité forte. Clairement, ces inégalités généralisent l'inégalité ultramétrique. Les deux ensembles structurels correspondant, sont fermés supérieurement, et pour peu qu'ils soient étendus aux pré-dissimilarités, ils vérifient la condition du cylindre. Comme noté par les auteurs, il y a donc existence dans les deux cas d'une sous-dominante, dont le calcul peut être fait par réductions successives d'une plus grande valeur sur certains $(k + 2)$ -uplets de X . Nous pouvons encore en déduire un calcul d'une approximation optimale. Notons toutefois que ceci est avant tout un exemple illustratif, car même si ces structures sont reliées à des classifications fondées sur la recherche de cliques maximales (« ML-sets »), de l'aveu même des auteurs, ces dernières sont rudimentaires.

5. SEMI-DISTANCES ET DISSIMILARITÉS DE TYPE ARBORÉ

Une (semi)-distance d sur X est dite *(semi)-distance de type arboré* si (X, d) est isométriquement plongeable dans un espace métrique arboré (V, ρ) , associé à un arbre fini $T = (V, E)$ pondéré sur les arêtes par des nombres strictement positifs, *i.e.* s'il existe une application ϕ de X dans V telle que $\rho(\phi(x), \phi(y)) = d(x, y)$, pour tout x, y de X . On rappelle que la distance $\rho(u, v)$ entre deux sommets u et v de T , est la somme des poids des arêtes le long de l'unique chemin reliant ces sommets. On note que dans un tel plongement, qui est démontré unique à un isomorphisme et des réductions près, il peut exister des sommets de T non étiquetés par X , ou en d'autres termes, que ϕ n'est pas nécessairement surjective. En revanche elle est trivialement injective si seulement si d est une distance. Nous noterons \mathcal{D}_{tm} le sous-ensemble de \mathcal{D} des semi-distances de type arboré. C'est un cône (évident) fermé (presque immédiat). La caractérisation des semi-distances de type arboré repose sur ce que l'on nomme la *condition des quatre points* suivante : pour tout x, y, z, t de X :

$$d(x, y) + d(z, t) \leq \max[d(x, z) + d(y, t), d(x, t) + d(y, z)] \quad (1)$$

Il a alors été démontré qu'une semi-distance d est de type arboré si et seulement si l'on a (1). Il existe de nombreuses démonstrations de ce résultat, les plus connues étant celles de [Zaretskii, 1965] et [Buneman, 1974], voir aussi l'ouvrage de [Barthélemy, Guénoche, 1991]. Il est facile de constater que la condition (1) est équivalente à une double condition : *métricité* et *condition faible des quatre points*, à savoir l'inégalité de (1) pour tout x, y, z, t distincts de X . Mais la métricité ne découle plus de cette dernière. Forts de ce fait, [Bandelt, Steel, 1995] et [Leclerc, 1995] introduisent des *dissimilarités de type arboré*. Ce sont celles qui vérifient la condition faible. Elles conduisent à un plongement « isométrique » dans un arbre pondéré, mais avec des poids qui peuvent être négatifs sur les arêtes terminales. Pour s'en convaincre, à partir d'une telle dissimilarité d , il suffit de considérer $d + c \cdot d_1$, c suffisamment grand,

pour avoir une semi-distance de type arboré, donc un plongement dans un arbre pondéré. Une réduction des arêtes à partir des sommets représentatifs de X , qu'on introduit de nouvelles feuilles, donne le plongement souhaité pour (X, d) . Nous noterons \mathcal{D}_{td} l'ensemble des dissimilarités de type arboré, et \mathcal{D}'_{td} son extension aux pré-dissimilarités du même type.

De fait, il existe un lien étroit entre dissimilarités (métriques) de type arboré et ultramétriques. Pour le montrer, nous faisons choix d'un pivot a de X , et on note $X^a := X \setminus \{a\}$ et \mathcal{D}^a l'espace des pré-dissimilarités sur X^a . Pour une famille de réels $\{r_x, x \in X\}$, on note encore $\mathcal{D}^{a,r}$ le sous-ensemble de \mathcal{D} dont les éléments d ont une dissimilarité fixe par rapport à a : $d(a, x) = r_x$ pour tout x de X . C'est une variété linéaire de \mathcal{D} , que l'on peut identifier à \mathcal{D}^a . On considère alors la transformation τ^a (endomorphisme de \mathcal{D}), qui à d fait correspondre $d^a := \tau^a(d)$ de \mathcal{D} , défini comme suit sur les paires de X :

$$d^a(x, y) = d(x, y) - d(a, x) - d(a, y), \text{ pour tout } x, y, x \neq y, \text{ de } X. \quad (2).$$

On observe que $d^a \in \mathcal{D}^{a,r_0}$, r_0 étant la famille de réels nuls, et que τ^a est idempotent. Une analyse un peu plus fine, montre que τ^a est le projecteur sur le sous-espace \mathcal{D}^{a,r_0} , opérant parallèlement au sous-espace des pré-dissimilarités étoilées de centre a . En particulier, la variété $\mathcal{D}^{a,r}$ est projetée globalement sur son espace de direction \mathcal{D}^{a,r_0} .

La quantité $-\frac{1}{2}d^a(x, y)$ est appelée *produit de Gromov* de x, y dans X^a , voir [Ghys, de la Harpe, 1990], tandis que d^a , à une constante positive près, est appelée *transformée de Farris*. Le résultat le plus général pour la caractérisation de \mathcal{D}'_{td} a été donné par [Leclerc, 1995] : d est une (pré)-dissimilarité de type arboré si et seulement si $d|_{X_a} \in \mathcal{D}^{a'}_u$. Notons la portée de ce résultat, qui montre que la condition faible des quatre points, valide pour tout quadruplet d'éléments distincts, est équivalente à cette même condition pour a fixé et pour tout triplet d'éléments distincts de X^a . Une condition semblable d'ultramétrie existe pour les semi-distances de type arboré, lorsque par hypothèse d est métrique. On peut consulter les travaux de [Farris, Kluge, Eckart, 1970], redécouverts par [Klotz, Blanken, 1981], [Brossier, 1985] ou [Bandelt, 1990]. Avec des hypothèses minimales de métricité a priori pour d , [Chepoi, Fichet, 2000] donnent la caractérisation suivante : $d \in \mathcal{D}_{tm}$ si et seulement si $d^a \in \mathcal{D}'_u$ et pour tout $x, y \in X$, $|d(a, x) - d(a, y)| \leq d(x, y)$. (3)

[Agarwala *et al.*, 1996] ont montré que l'approximation d'un élément d de \mathcal{D}_m (et donc de \mathcal{D}_+) par un élément de \mathcal{D}_{tm} est NP-difficile. En revanche, par un va-et-vient astucieux via la transformée de Farris, ils ont obtenu une solution simple pour approcher un élément d de $\mathcal{D}^{a,r}_m := \mathcal{D}_m \cap \mathcal{D}^{a,r}$ par un élément de $\mathcal{D}^{a,r}_{tm} := \mathcal{D}_{tm} \cap \mathcal{D}^{a,r}$. La procédure est la suivante. Partant d'une semi-distance d , ils calculent sa transformée de Farris d^a (à une constante additive près), puis par l'algorithme de [Farach *et al.*, 1995], une meilleure approximation ultramétrique \hat{d}^a , enfin par la transformée inverse de Farris un élément \hat{d} . Pour s'assurer que \hat{d} est dans $\mathcal{D}^{a,r}$, il faut que \hat{d}^a soit dans \mathcal{D}^a , ce que l'on obtient facilement en le remplaçant par $\inf(\hat{d}^a, d_0)$ sans en altérer l'optimalité. Clairement, \hat{d} est dans \mathcal{D}_{td} . Il reste à montrer la métricité, et les auteurs reconnaissent avoir dû modifier légèrement à cette fin l'algorithme pour l'obtention de l'ultramétrie. Forts de ce résultat, par un jeu de retouches sur la longueur

des arêtes, ils montrent que \hat{d} est une approximation de facteur 3 pour le problème global d'approximation dans \mathcal{D}_{tm} , *i.e.* que l'erreur trouvée $\hat{\epsilon} := \|d - \hat{d}\| = \|d^a - \hat{d}^a\|$ est inférieure à $3 \cdot \epsilon_0$, où ϵ_0 est l'erreur optimale globale.

[Chepoi, Fichet, 2000] étendent ces résultats par usage des sous-dominantes calculées en termes d'ultramétries puis transférées par transformée de Farris inverse. Ils démontrent d'abord que tout d de $\mathcal{D}^{a,r}$ a une sous-dominante dans $\mathcal{D}_{td}^{a,r} := \mathcal{D}'_{td} \cap \mathcal{D}^{a,r}$. L'existence est évidente puisque $\mathcal{D}_{td}^{a,r}$ est trivialement fermé supérieurement et vérifie la condition du cylindre dans $\mathcal{D}^{a,r}$ identifié à \mathcal{D}^{a,r_0} (la diagonale est celle de \mathcal{D}^{a,r_0}). Son calcul est aisé grâce aux caractérisations ultramétriques via les transformées de Farris. Les auteurs montrent de plus que d de $\mathcal{D}^{a,r}$ a une sous-dominante dans $\mathcal{D}_{tm}^{a,r}$ si et seulement si la condition de métricité (3) précédente est vérifiée. En outre, si $d \in \mathcal{D}_m$ les deux sous-dominantes coïncident. Les auteurs retrouvent ainsi les résultats précédents d'Agarwala *et al.*, \hat{d}^a étant la plus grande solution ultramétrique, garantie d'une démonstration de la métricité de \hat{d} . Ils obtiennent également une meilleure approximation dans $\mathcal{D}_{td}^{a,r}$, qui est aussi une approximation globale de facteur 3. Enfin, sous la seule condition (3) de métricité locale pour une approximation dans $\mathcal{D}_{tm}^{a,r}$, malgré la condition du cylindre dans $\mathcal{D}_{td}^{a,r}$ et du demi-cylindre dans \mathcal{D}_m , ils ne peuvent faire usage d'une translation à partir de la sous-dominante pour l'obtention d'une solution, car la préservation des dissimilarités à a n'assurerait plus la métricité globale. Cependant, réalisant une translation le long de la diagonale globale de \mathcal{D} , ils obtiennent un élément de \mathcal{D}_{tm} qui n'est pas dans $\mathcal{D}^{a,r}$, mais qui s'avère encore être prouvé comme étant une approximation de facteur 3 globale. À notre connaissance, seul le problème général d'approximation d'une dissimilarité quelconque par un élément de \mathcal{D}_{tm} n'a pas encore de réponse en terme d'algorithme à facteur constant.

6. DISSIMILARITÉS DE ROBINSON ET SEMI-DISTANCES UNIDIMENSIONNELLES

Nous avons vu précédemment que toute dissimilarité d sur X possède une sous-dominante compatible avec un ordre donné. L'ensemble de ces sous-dominantes pour tous les ordres est fini, et possède des éléments maximaux. Ce sont exactement les inférieures maximales Robinsoniennes de d . Donc, dans le cas des dissimilarités de Robinson, on a une sous-dominante pour un ordre donné mais pas dans le cas général. Pour les semi-distances linéaires (aussi appelées unidimensionnelles ou de chaîne), cas particulier des semi-distances d'arbre, il est à noter qu'il n'existe pas de sous-dominante, même pour un ordre donné. Le triangle équilatéral est un exemple simple avec une infinité d'inférieures maximales. Par contre, le problème d'approximation d'une dissimilarité par une dissimilarité de Robinson ou par une semi-distance linéaire est polynomial si on fixe l'ordre : via les sous-dominantes pour les dissimilarités de Robinson comme il a été vu, et via la programmation linéaire pour les semi-distances linéaires [Håstad, Ivansson, Lagergren, 2003]. Malheureusement, dans le cas général, ces deux problèmes sont NP-complets.

Les preuves d'NP-complétude reposent sur une réduction à un problème de satisfaisabilité d'une formule booléenne (NOT-ALL-EQUAL-3-SAT). Elles montrent de plus, qu'il est NP-difficile d'approcher une solution par une semi-distance linéaire

avec un facteur inférieur à $7/5$, [Håstad, Ivansson, Lagergren, 2003], et avec un facteur inférieur à $3/2$ par une dissimilarité de Robinson, [Chepoi, Fichet, Seston, 2009]. Néanmoins, il existe des algorithmes d'approximation avec un facteur constant : facteur 2 pour les semi-distances linéaires, [Håstad, Ivansson, Lagergren, 2003], et facteur 16 pour les dissimilarités de Robinson, [Chepoi, Seston, 2009]. Nous allons maintenant présenter brièvement ces deux algorithmes.

Intéressons-nous d'abord aux semi-distances linéaires. La technique utilisée est similaire à celle utilisée pour le problème d'approximation d'une dissimilarité par une semi-distance de type arboré en conservant les distances à un point donné. En effet pour chaque point a , on construit une semi-distance linéaire telle que tout point x a pour coordonnée $d(a, x)$. On montre alors que parmi celles-ci, une plus proche de d donne une approximation avec un facteur 3. Il est possible d'obtenir un facteur 2 en décalant chaque point à gauche ou à droite d'une faible quantité.

La conception d'un algorithme avec un facteur constant pour l'approximation par une dissimilarité de Robinson se révèle beaucoup plus ardue que pour les problèmes précédemment cités. En effet, les algorithmes naïfs fondés sur la conservation des distances par rapport à un point pivot, ne permettent pas d'obtenir un facteur d'approximation. Ici, nous allons dresser les grandes lignes d'un algorithme d'approximation avec un facteur 16, utilisant le fait que l'erreur optimale $\hat{\epsilon}$ appartient à une liste compacte et bien définie $\Delta = \{\frac{1}{2}|d(x, y) - d(z, t)| : x, y, z, t \in X\}$. Pour une valeur donnée $\epsilon \in \Delta$, l'algorithme, soit trouve qu'il n'existe pas d'ordre ϵ -compatible, soit retourne un ordre 16ϵ -compatible. Si ϵ est la plus petite valeur telle que l'algorithme ne renvoie pas de réponse négative, alors $\hat{\epsilon} \geq \epsilon$. L'ordre \preceq retourné est $16\hat{\epsilon}$ -compatible et un optimum \hat{d}_{\preceq} pour cet ordre vérifie : $\|\hat{d}_{\preceq} - d\| \leq 16\hat{\epsilon}$. Nous allons maintenant voir comment construire l'ordre \preceq . Pour un ϵ donné, l'algorithme construit une relation binaire \preceq telle que tout ordre ϵ -compatible raffine \preceq ou son dual. Si \preceq n'est pas un ordre partiel, alors l'algorithme retourne une réponse négative. Supposons que \preceq soit un ordre partiel. S'il est total, alors c'est un ordre ϵ -compatible pour d . Sinon, on choisit une chaîne maximale par inclusion $P = (a_1, a_2, \dots, a_p)$ relativement à cet ordre. On dira que deux éléments consécutifs a_i, a_{i+1} forment un *trou* H_i et qu'un élément x tel que $a_i \preceq x \preceq a_{i+1}$ est *placé* dans le trou H_i . Pour obtenir un ordre total, il faut alors placer chaque élément de $X^\circ = X \setminus P$ dans un trou et fixer un ordre total entre les éléments de X° placés dans un même trou. Pour chaque élément x , on peut définir un ensemble $H(x)$ de trous admissibles. Un trou H_i est dit admissible pour x si pour tout y , il existe un trou H_j (admissible pour y) tel que l'ordre obtenu en ajoutant les relations $a_i \preceq x \preceq a_{i+1}$ et $a_j \preceq y \preceq a_{j+1}$ à l'ordre \preceq , est ϵ -compatible. Parmi les trous de $H(x)$, on distinguera le trou le plus à gauche et le trou le plus à droite que l'on appellera « trous frontières ». Chaque élément x de X° sera alors placé dans l'un de ses trous frontières. Le choix entre les deux se fait à l'aide d'une formule 2-SAT. Maintenant, pour fixer l'ordre entre deux éléments x et y placés dans un même trou H_i , on distingue deux cas. Si x et y n'ont qu'un seul trou frontière H_i en commun, soient H_j le trou frontière de x différent de H_i et H_k le trou frontière de y différent de H_i . On fixe alors $x \preceq y$ si $j < k$ et $y \preceq x$ sinon. Enfin, on appelle récursivement l'algorithme sur les éléments partageant les mêmes trous frontières et étant placés dans le même trou après avoir fixé certaines contraintes d'ordre entre ces éléments. Si aucune étape de

l'algorithme ne retourne la réponse « non », alors on montre que l'ordre obtenu est 16ϵ -compatible avec $\epsilon \leq \hat{\epsilon}$. Via la sous-dominante, on calcule la dissimilarité optimale compatible avec cet ordre, et on obtient une approximation avec un facteur 16 pour le problème.

Cette description, volontairement générale pour ne pas embrouiller le lecteur, ne précise pas toutes les facettes et les stratégies de l'algorithme, en particulier celles faisant appel au graphes. Pour l'éclairer un peu plus avant, reprenons la dissimilarité d de la Figure 4 pour illustrer par un exemple simple, le déroulement de l'algorithme. Certes un exemple plus complexe élargirait le champ du possible, mais interdirait une résolution à la main. La Figure 6 redonne la dissimilarité d ainsi que la dissimilarité d_R retournée par l'algorithme. Notons d'abord que l'ensemble Δ des erreurs potentielles ϵ , est le suivant : $\{0, 5; 1; 1, 5; 2; 2, 5; 3; 3, 5; 4\}$. À la lecture des données, on peut aisément constater qu'il n'existe pas d'ordre 0, 5-compatible et que l'ordre $x_3 \prec x_5 \prec x_2 \prec x_1 \prec x_4 \prec x_6$ compatible avec d_R , est 1-compatible avec d . C'est ce que l'algorithme va découvrir.

Commençons avec $\epsilon = 0, 5$. Afin de construire la relation binaire \preceq , l'algorithme fixe x_1 comme plus petit que tous les autres éléments (il sera fait de même avec tous les éléments de X). À partir des valeurs de d , de la valeur de ϵ et d'un ensemble de règles non décrites ici, l'algorithme enrichit \preceq des relations suivantes : $x_1 \prec x_2, x_4, x_5, x_6 \prec x_3$ et $x_6 \prec x_1$. Il détecte que la relation \preceq n'est pas un ordre partiel et de même pour les relations binaires obtenues en fixant x_2, x_3, x_4, x_5 ou x_6 comme élément inférieur à tous les autres. Ainsi d n'est pas 0, 5-Robinson, et l'algorithme teste alors $\epsilon = 1$. En utilisant la même démarche que précédemment, il détecte que les relations binaires construites avec x_1, x_2, x_4 ou x_5 comme point le plus petit, ne sont pas des ordres partiels. Voyons plus en détails la construction de \preceq lorsque x_3 est le point le plus petit. À partir des distances à x_3 et de ϵ , il obtient les relations : $x_1 \prec x_6; x_2 \prec x_1, x_4, x_6; x_5 \prec x_1, x_4, x_6$. De même, à partir des distances à x_5 et de ϵ , il obtient $x_4 \prec x_6$. On a alors l'ordre partiel $x_3 \prec x_2, x_5 \prec x_1 \prec x_4 \prec x_6$. La chaîne $x_3 \prec x_2 \prec x_1 \prec x_4 \prec x_6$ est choisie comme chaîne maximale de l'ordre \preceq . Le seul élément n'appartenant pas à cette chaîne est x_5 qui peut être placé soit dans le trou (x_3, x_2) soit dans le trou (x_2, x_1) . Comme x_5 est le seul n'appartenant pas à la chaîne, l'algorithme placera arbitrairement x_5 dans l'un de ces trous, par exemple entre x_3 et x_2 . Il calcule alors une dissimilarité optimale pour l'ordre 16-compatible obtenu : $x_3 \prec x_5 \prec x_2 \prec x_1 \prec x_4 \prec x_6$. De même, en fixant x_6 comme point le plus petit, l'algorithme obtient un ordre 16-compatible, par exemple $x_6 \prec x_4 \prec x_1 \prec x_5 \prec x_2 \prec x_3$. L'algorithme calcule alors l'erreur associée à chacun de ces ordres et choisira le meilleur. Dans notre cas les deux erreurs sont identiques, et l'algorithme choisit arbitrairement l'un d'eux, disons $x_3 \prec x_5 \prec x_2 \prec x_1 \prec x_4 \prec x_6$.

d	x_2	x_3	x_4	x_5	x_6
x_1	1	5	2	3	3
x_2	0	1	2	2	3
x_3		0	7	2	9
x_4			0	1	1
x_5				0	4

d_R	x_2	x_3	x_4	x_5	x_6
x_1	2	6	2	2	4
x_2	0	2	2	2	4
x_3		0	8	2	10
x_4			0	2	2
x_5				0	5

FIGURE 6. Une dissimilarité d et la dissimilarité de Robinson d_R retournée par l'algorithme d'approximation.

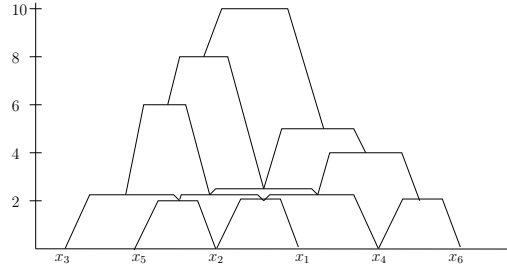


FIGURE 7. La représentation pseudo-hiérarchique de la dissimilarité d_R de la Figure 6. Notons que la sous-dominante ultramétrique de d_R n'est pas celle de d .

Remerciements. Les auteurs remercient deux rapporteurs anonymes, pour leur lecture minutieuse de l'article, leurs suggestions et leur apport bibliographique.

Les auteurs ont été en partie soutenus par l'ANR BLAN06-138894 (projet OPTICOMB).

BIBLIOGRAPHIE

- AGARWALA R., BAFNA V., FARACH M., NARAYANAN B., PATERSON M., THORUP M. (1996), "On the approximability of numerical taxonomy : Fitting distances by tree metrics", *Proceedings of the 7th Annual ACM-SIAM Symposium on Discrete Algorithms*.
- AYER M., BRUNK H.D., EWING G.M., REID W.T., SILVERMAN E. (1955), "An empirical distribution function for sampling with incomplete information", *Ann. Math. Statist.* 26, p. 641-647.
- BANDELT H.-J. (1990), "Recognition of tree metrics", *SIAM Journal of Discrete Mathematics* 3, p. 1-6.
- BANDELT H.-J., STEEL M.A. (1995), "Symmetric matrices representable by weighted trees over a cancellative abelian monoid", *SIAM Journal of Discrete Mathematics* 8, p. 517-525.
- BARBUT M. (1988), « Sur une classe de résumés statistiques : les valeurs centrales », *L'à peu près. Aspects anciens et modernes de l'approximation*, Centre d'Analyse et de Mathématiques Sociales, Éditions de l'École des Hautes Études en Sciences Sociales, Paris, p. 109-141.
- BARLOW R. E., BARTHOLOMEW D.J., BREMNER J.M., BRUNK H.D. (1972), *Statistical Inference under Order Restrictions*, New York, Wiley.
- BARTHÉLEMY J.-P., BRUCKER F. (2001), "NP-hard approximation problems overlapping clustering", *Journal of Classification* 18, p. 159-183.

- BARTHÉLEMY J.-P., GUÉNOCHE A. (1991), *Trees and proximity representations*, Chichester, Wiley.
- BARTHÉLEMY J.-P., LECLERC B., MONJARDET B. (1986), "On the use of ordered sets in problems of comparison and consensus of classifications", *Journal of Classification* 3, p. 187-224.
- BENZÉCRI J.-P. (1973), *L'analyse des données. 1. La Taxinomie*, Paris, Dunod.
- BERGE C. (1966), *Espaces topologiques. Fonctions multivoques*, [deuxième édition], Paris, Dunod.
- BERTRAND P., DIDAY E. (1985), "A visual representation of the compatibility between an order and a dissimilarity index : the pyramids", *Comput. Statist. Quart.* 2, p. 31-44.
- BIRKHOFF G. (1967), *Lattice theory*, [3rd edition], Providence (RI) : Amer. Math. Soc.
- BROSSIER G. (1985), « Approximation des dissimilarités par des arbres additifs », *Mathématiques et Sciences humaines* 91, p. 5-21.
- BRUCKER F., BARTHÉLEMY J.-P. (2007), *Eléments de classification. Aspects combinatoires et algorithmiques*, Paris, Hermes, Lavoisier.
- BUNEMAN P. (1974), "A note on the metric properties of trees", *Journal of Combinatorial Theory Ser. B* 17, p. 48-50.
- CHEPOI V., COGNEAU D., FICHET B. (1997), "Polynomial algorithms for isotonic regression problem", Y. Dodge (ed.), *L_1 -statistical procedures and related topics*, Institute of Mathematical Statistics, Lecture Notes Monograph Series, 31, p. 147-160.
- CHEPOI V., FICHET B. (1997), "Recognition of Robinsonian dissimilarities", *Journal of Classification* 14, p. 311-325.
- CHEPOI V., FICHET B. (1998), "A note on Circular Decomposable Metrics", *Geometriae Dedicata* 69, p. 237-240.
- CHEPOI V., FICHET B., SESTON M. (2009), "Seriation in the presence of errors : NP-hardness of l_∞ -fitting Robinson structures to dissimilarity matrices", *Journal of Classification* 26, p. 279-296.
- CHEPOI V., SESTON M. [in press], "Seriation in the presence of errors : a factor 16 approximation algorithm for l_∞ -fitting Robinson structures to distances", *Algorithmica*.
- CRITCHLEY F., FICHET B. (1997), "The partial order by inclusion of the principal classes of dissimilarity on a finite set, and some of their basic properties", van Cutsem B. (ed.), *Classification and Dissimilarity Analysis*, Lecture Notes In Statistics, Springer-Verlag, Berlin, p. 5-65.
- DAY W.H.E. (1987), "Computational complexity of inferring phylogenies from dissimilarity matrices", *Bulletin of Mathematical Biology* 49, p. 461-467.
- DEZA M., LAURENT M. (1997), *Geometry of Cuts and Metrics*, Berlin, Springer-Verlag.
- DURAND C., FICHET B. (1988), "One-to-one correspondences in pyramidal representation : a unified approach", Bock H.H. (ed.), *Classification and Related Methods of Data Analysis*, Amsterdam, North Holland, p. 85-90.
- DYKSTRA R.L. (1983), "An algorithm for restricted least squares regressions", *J. Am. Statist. Assoc.* 78, p. 837-842.
- DYKSTRA R.L., ROBERTSON T., "An algorithm for isotonic regression for two or more independent variables", *Ann. Statist.* 10, p. 708-716.
- EDELSBRUNNER H. (1987), *Algorithms in combinatorial geometry*, New-York, Springer-Verlag.

- EEDEN VAN C. (1956), *Testing and estimating ordered parameters of probability distributions*, Dissertation thesis, Amsterdam, University of Amsterdam.
- FARACH M., KANNAN S., WARNOW T. (1995), "A robust model for finding optimal evolutionary trees", *Algorithmica* 13, p. 155-179.
- FARRIS J.S., KLUGE A.G., ECKARDT M.J. (1970), "A numerical approach to phylogenetic systematics", *Systematic Zoology* 19, p. 172-189.
- FICHET B. (2001), "Ultramétriques supérieures minimales sous contraintes", *Proceedings of the 8th Annual meeting of SFC*, Guadeloupe, p. 147-150.
- GHYS E., de la HARPE P. (1990), « Les Groupes Hyberboliques d'après M. Gromov », *Progress in Mathematics*, (Vol. 83), Basel, Birkhauser.
- GOWER J.C., ROSS G.K.S. (1969), "Minimum spanning tree and single linkage cluster analysis", *Applied Statistics* 18, p. 54-64.
- HUBERT L., ARABIE P., MEULMAN J. (1998), "Graph-theoretic representation for proximity matrices through strongly-anti-Robinson or circular-anti-Robinson matrices", *Psychometrika* 63, p. 341-358.
- HÅSTAD J., IVANSSON L., LAGERGREN J. (2003), "Fitting points on the real line and its application to RH mapping", *Journal of Algorithms* 49, no. 1, p. 42-62.
- JARDINE N., SIBSON R. (1971), *Mathematical Taxonomy*, New York, Wiley.
- JOHNSON S.C. (1967), "Hierarchical clustering schemes", *Psychometrika* 32(3), p. 241-254.
- KLOTZ L.C., BLANKEN R.L. (1981), "A practical method for calculating evolutionary trees from sequence data", *Journal of Theoretical Biology* 91, p. 261-272.
- KŘIVÁNEK M. (1988), "The complexity of ultrametric partitions on graphs", *Information Processing Letters* 27(5), p. 265-270.
- KŘIVÁNEK M., MORÁVEK J. (1986), "NP-Hard problems in hierarchical-tree clustering", *acta informatica* 23, p. 311-323.
- KRUSKAL J.B (1964), "Nonmetric multidimensional scaling : a numerical method", *Psychometrika* 29, p. 115-129.
- LAWSON C., HANSON R.J. (1974), *Solving Least Squares Problems*, Prentice Hall.
- LECLERC B. (1986), « Caractérisation, construction et dénombrement des ultramétriques supérieures minimales », *Statistique et Analyse des données* 11(2), p. 26-50.
- LECLERC B. (1995(a)), "Minimum spanning trees for tree metrics : abridgment and adjustments", *Journal of Classification* 12, p. 207-241.
- LECLERC B., MONJARDET B. (1995(b)), "Latticial Theory of Consensus", Barnett W., Moulin H., Salles M., Schofield N. (eds.), *Social Choice, Welfare and Ethics*, Cambridge, Cambridge University Press, p. 145-160.
- LERMAN I.C. (1970), *Les bases de la classification automatique*, Paris, Gauthier-Villars.
- MIRKIN B., RODIN S. (1984), *Graphs and Genes*, Berlin, Springer-Verlag.
- PORTNOY S., KOENKER R. (1997), "The Gaussian hare and the Laplacian tortoise : computability of squared error versus absolute-error estimators", *Statistical Science* 12, p. 279-300.
- PRÉA P., FORTIN D. [soumis], "An optimal algorithm to recognize Robinsonian dissimilarities", *Algorithmica*.
- ROBERTSON T., WRIGHT F.T. (1974), "A norm reducing property for isotonized Cauchy mean value functions", *Ann. Statist.* 33(6), p. 1302-1307.
- ROCKAFELLAR R.T. (1970), *Convex analysis*, Princeton, Princeton Univ. Press.

- ROUX M. (1968), *Un algorithme pour construire une hiérarchie particulière*, Dissertation thesis, Paris, Université de Paris VI.
- SESTON M. (2008), "A simple algorithm to recognize Robinsonian dissimilarities", P. Brito (éd.), *Proceedings in Computational Statistics, Physica-Verlag* 2, p. 241-248.
- SIBSON R. (1972), "SLINK : an optimally efficient algorithm for the single link cluster method", *The Computer Journal* 16.
- STOUT Q.F. (2008), "Unimodal Regression via Prefix Isotonic Regression", *Statistics and Data Analysis* 53, p. 289-297.
- STOUT Q.F. (2009), "Algorithms for L_∞ Isotonic Regression", <http://www.eecs.umich.edu/~qstout/pap/LinfinityIso090731.pdf>
- THOMPSON W.A. (1962), "The problem of negative estimates of variance components", *Ann. Math. Statist.* 33, p. 273-289.
- UBHAYA V.A. (1974), "Isotone optimization, I, II", *Journal of Approximation Theory* 12, p. 146-159.
- ZARETSKII K. (1965), "Constructing trees from the set of distances between pendant vertices", *Uspehi Matematicheskikh Nauk* 20, p. 90-92.