HAL open science

Correspondences between Pre-pyramids, Pyramids and Robinsonian Dissimilarities

Patrice Bertrand, Jean Diatta

To cite this version:

Patrice Bertrand, Jean Diatta. Correspondences between Pre-pyramids, Pyramids and Robinsonian Dissimilarities. 2012. hal-00703391

HAL Id: hal-00703391

https://hal.science/hal-00703391v1

Preprint submitted on 1 Jun 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Correspondences between Pre-pyramids, Pyramids and Robinsonian Dissimilarities

Patrice Bertrand[1] and Jean Diatta[2]

[1]CEREMADE, Université Paris Dauphine
[2]LIM-EA2525, Université de la Réunion

Abstract

We consider cluster structures in a general setting where they do not necessarily contain all singletons of the ground set. Then we provide a direct proof of the bijection between semi-proper robinsonian dissimilarities and indexed pre-pyramids. This result generalizes its analogue proven by Batbedat in the particular case of definite cluster structures. Moreover, the proposed proof shows that the clusters of the indexed pre-pyramid corresponding to a semi-proper robinsonian dissimilarity are particular 2-balls of the considered dissimilarity.

Keywords: Cluster structure, Compatible order, Robinsonian dissimilarity.

Structural bijections between cluster structures and dissimilarity types play an important role in cluster analysis. They not only validate the dissimilarity-based clustering approach, but make it possible to bring a cluster system specification problem back to a dissimilarity matrix determination (13; 9; 8; 10). The best known of these bijections is certainly the one between indexed hierarchies and ultrametrics (20; 5). Several extensions of this bijection have been obtained, e.g., between indexed quasi-hierarchies and quasi-ultrametrics (1; 15), between indexed pyramids and strongly robinsonian dissimilarities (17), and between weakly indexed pyramids and robinsonian dissimilarities (16). Generalizations and/or unifications of some of these bijections can be found in (3; 11; 6; 18; 7; 2; 12).

The present paper focuses on a bijection between semi-proper robinsonian dissimilarities and indexed pre-pyramids, in a general setting where cluster structures are not required to contain all singletons of the ground set, *i.e.*, are not necessarily definite. We give a direct proof of this bijection, showing that the clusters of the indexed pre-pyramid corresponding to a semi-proper robinsonian dissimilarity are particular 2-balls of the considered dissimilarity. This result generalizes its analogue proven by Batbedat (4) in the particular case of definite cluster structures. Moreover, it can be thought of as the third side of a triangular correspondence between indexed pre-pyramids, weakly indexed pyramids and semi-proper robinsonian dissimilarities. Indeed, on the one hand, a bijection between indexed pre-pyramids and weakly indexed pyramids derives as an instance of the general one between indexed cluster structures and weakly indexed closed cluster structures (14). On the other hand, the Diday's bijection (16) extends between weakly indexed (not necessarily definite) pyramids and semi-proper robinsonian dissimilarities.

1 Pyramids and Pre-pyramids

1.1 Pre-pyramids

Let E be a finite nonempty set. A *cluster structure* on E is a collection $\mathcal{C}$ of subsets of E, satisfying conditions (CS1), (CS2) and (CS2') below:

(CS1) the empty set is not a member of $\mathcal{C}$ whereas the ground set E is, i.e., $\varnothing \notin \mathcal{C}$ and $E \in \mathcal{C}$;

(CS2) the set of minimal members of $\mathcal{C}$ (w.r.t. set inclusion) partitions E; in other words, these minimal members are non-empty, pairwise disjoint, and they cover E (i.e. their union equals E);

(CS2') every non-minimal member of $\mathcal{C}$ is the union of members of $\mathcal{C}$ it properly contains, i.e., for all $X \in \mathcal{C}$: $\cup\{Y \in \mathcal{C} : Y \subset X\} \in \{\varnothing, X\}$;

(CS3) there is a linear order, say θ, on E, of which each member C of $\mathcal{C}$ is an interval, i.e., for all $x, y, z \in E$: $x, y \in C$ and $x\theta z\theta y$ imply $z \in C$.

The pair of conditions (CS2) and (CS2') is often replaced by a stronger condition requiring each singleton to be a member of $\mathcal{C}$. Actually, a cluster structure satisfying this strong requirement is said to be total or definite.

A *pre-pyramid* on E is a cluster structure $\mathcal{C}$ onf E, satisfying the condition (CS3) below:

(CS3) there is a linear order, say θ, on E, of which each member C of $\mathcal{C}$ is an interval, i.e., for all $x, y, z \in E$: $x, y \in C$ and $x\theta z\theta y$ imply $z \in C$.

The order θ and the collection $\mathcal{C}$ are said to be *compatible* with each other, and subset collections satisfying condition (CS3) are said to admit a compatible order. The condition (CS3) is the specific condition of interval-type subset collections.

1.2 Pyramids

To every subset collection can be associated its closure consisting of arbitrary intersections of its members. As we are concerned with collections of nonempty subsets of finite sets, we will consider only finite nonempty intersections. The closure of a subset collection $\mathcal{C}$ under (finite) nonempty intersections will be denoted by $\overline{\mathcal{C}}$, and $\mathcal{C}$ will be said to be *closed* when it satisfies the condition (CS4) below:

(CS4) the intersection of two members of $\mathcal{C}$ is either empty or a member of $\mathcal{C}$, i.e., $X, Y \in \mathcal{C}$ implies $X \cap Y \in \mathcal{C} \cup \{\varnothing\}$.

A *pyramid* is a closed pre-pyramid. It can be checked that conditions (CS2) and (CS2') are equivalent under (CS1) and (CS4). Pyramidal classification has been introduced by Diday (16) and considered or investigated in several works among which we can mention (21; 19).

To show that $\overline{\mathcal{C}}$ is a pyramid when $\mathcal{C}$ is a pre-pyramid, it suffices to check that $\overline{\mathcal{C}}$ verifies (CS2). Now this derives from the fact that, by Lemma 1 and Lemma 2 below, every minimal member of $\mathcal{C}$ is also a minimal member of $\overline{\mathcal{C}}$.

Lemma 1 *(14) The conjunction of conditions (CS2) and (CS2') is equivalent to the conjunction of conditions (a) and (e), where:*

(a_1) $C, C' \in \mathcal{C}$ and C minimal in $\mathcal{C}$ imply $C \cap C' \in \{\varnothing, C\}$;

(a_2) minimal members of $\mathcal{C}$ cover E.

Lemma 2 *(14) The following conditions are equivalent for a collection $\mathcal{C}$ of nonempty subsets of E.*

(a_1) $C, C' \in \mathcal{C}$ and C minimal in $\mathcal{C}$ imply $C \cap C' \in \{\varnothing, C\}$.

(a_3) Every minimal member of $\mathcal{C}$ is minimal in $\overline{\mathcal{C}}$.

1.3 Indexed pre-pyramids and weakly indexed pyramids

Let $\mathcal{C}$ be a cluster structure on E. A *pre-index* on $\mathcal{C}$ is an order preserving map $f : (\mathcal{C}, \subseteq) \to (\mathbb{R}_+, \leq)$ taking the zero value on minimal members of $\mathcal{C}$, i.e.,

(i) C minimal implies $f(C) = 0$;

(ii) $C \subseteq C'$ implies $f(C) \leq f(C')$.

In the sequel, we will assume that a pre-index f takes the value zero only on minimal members, hence, $f(C) = 0$ if and only if C is minimal. A canonical pre-index f_c can be obtained by letting $f_c(C)$ be the number of elements of the union of members of $\mathcal{C}$ properly contained in C. An *index* on $\mathcal{C}$ is a *strict pre-index*, i.e., a pre-index f such that $C \subset C'$ implies $f(C) < f(C')$.
A *weak index* (6) on $\mathcal{C}$ is a pre-index f such that

$$[C \subset C' \text{ and } f(C) = f(C')] \text{ implies } [C = \cap\{C'' \in \mathcal{C} : C \subset C''\}].$$

When f is a pre-index (resp. an index, a weak index) on a cluster structure $\mathcal{C}$, the pair $(\mathcal{C}, f)$ is said to be *pre-indexed* (resp. indexed, *weakly indexed*). Let $(\mathcal{C}, f)$ be a pre-indexed cluster structure on E. Let $\mathrm{Inter}(\mathcal{C}, f)$ denote the pair $(\overline{\mathcal{C}}, \overline{f})$, where $\overline{f}$ is defined on $\overline{\mathcal{C}}$ by

$$\overline{f}(C) = \min\{f(C') : C' \in \mathcal{C} \text{ and } C \subseteq C'\}.$$

On the other hand, define an *f-maximal* member of $\mathcal{C}$ to be a non-minimal member $C \in \mathcal{C}$ such that there is no member $C' \in \mathcal{C}$ such that $C \subset C'$ and $f(C) = f(C')$. Let $\mathrm{Strict}(\mathcal{C}, f)$ denote the pair $(\underline{\mathcal{C}}, \underline{f})$, where $\underline{\mathcal{C}}$ is composed of minimal and f-maximal members of $\mathcal{C}$, and $\underline{f}$ the restriction of f on $\underline{\mathcal{C}}$. Then $\mathrm{Strict}(\mathcal{C}, f)$ is clearly an indexed cluster structure. Moreover, the following result shows that indexed pre-pyramids correspond to weakly indexed pyramids in a one-to-one way. A proof can be found in (14).

Theorem 1

(i) If $(\mathcal{C}, f)$ is an indexed cluster structure on E, then $\mathrm{Inter}(\mathcal{C}, f)$ *is a weakly indexed closed cluster structure on E. Moreover,* $\mathrm{Strict}(\mathrm{Inter}(\mathcal{C}, f)) = (\mathcal{C}, f)$.

(ii) Conversely, if $(\mathcal{C}, f)$ is a weakly indexed closed cluster structure on E, then $\mathrm{Strict}(\mathcal{C}, f)$ *is an indexed cluster structure on E. Moreover,* $\mathrm{Inter}(\mathrm{Strict}(\mathcal{C}, f)) = (\mathcal{C}, f)$.

2 Correspondences with robinsonian dissimilarities

2.1 Robinsonian dissimilarities

Let us recall that a dissimilarity on E is a map $d : E \times E \to \mathbb{R}_+$ satisfying $d(x, x) = 0$ and $d(x, y) = d(y, x)$. It is said to be *semi-proper* if $d(x, y) = 0$ implies $d(x, z) \leq d(y, z)$ for all z. A *robinsonian* dissimilarity is a dissimilarity admitting a compatible order, i.e., such that there exists a linear order θ on E such that

$$x\theta y\theta z \Longrightarrow \max\{d(x, y), d(y, z)\} \leq d(x, z).$$

Robinsonian dissimilarities play an important role in unidimensional scaling problems in archeology (23) and in the analysis of DNA sequences (22). See also (24) for more references.

2.2 Semi-proper robinsonian dissimilarities and indexed pre-pyramids

In this section, we show that semi-proper robinsonian dissimilarities correspond to indexed pre-pyramids via their associated 2-balls. Let d be a dissimilarity on E. Given $x \in E$ and $r \geq 0$, the d-ball of center x and radius r is the set $B^d(x,r)$ defined as $B^d(x,r) = \{z \in E : d(x,z) \leq r\}$.

Let now x, y be two (not necessarily distinct) elements of E. Then, the $(d,2)$-ball (or simply 2-ball) generated by the pair x, y is the set B^d_{xy} defined as $B^d_{xy} = B^d(x, d(x,y)) \cap B^d(y, d(x,y))$. It may be noted that $B^d_{xy} = B^d(x, d(x,x))$ when $x = y$.

Let diam_d be the function defined on nonempty subsets X of E by $\mathrm{diam}_d(X) = \max\{d(x,y) : x, y \in X\}$ and let $\mathcal{B}_d$ be the set defined by

$$\mathcal{B}_d = \{B^d_{xy} : x, y \in E \text{ and } \mathrm{diam}_d(B^d_{xy}) = d(x,y)\}.$$

We will use the two following lemmas to prove the bijection between indexed pre-pyramids and semi-proper robinsonian dissimilarities.

Lemma 3 *(17) If d is a robinsonian dissimilarity, then for all compatible order θ and all $x, y \in E$, B^d_{xy} is an interval of θ.*

Lemma 4 *(17) If d is a robinsonian dissimilarity, then for any compatible order θ, each member B of $\mathcal{B}_d$ is of the form B^d_{xy}, where x and y are the bounds of B relatively to θ.*

Theorem 2 *If d is a semi-proper robinsonian dissimilarity on E, then $(\mathcal{B}_d, \mathrm{diam}_d)$ is an indexed pre-pyramid. Conversely, if $(\mathcal{C}, f)$ is an indexed pre-pyramid on E, then there exists a unique semi-proper robinsonian dissimilarity d such that $(\mathcal{B}_d, \mathrm{diam}_d) = (\mathcal{C}, f)$.*

Proof. Let d be a a semi-proper robinsonian dissimilarity on E. Clearly, $E \in \mathcal{B}_d$ since $E = B^d_{xy}$ with x, y such that $d(x,y) = \mathrm{diam}_d(E)$. Moreover, as d is semi-proper, the 2-balls of the form B^d_{xx} are the minimal members of $\mathcal{B}_d$. Hence they cover E, and, for all x, y, z, $B^d_{xx} \cap B^d_{yz} \in \{\varnothing, B^d_{xx}\}$. Finally, by Lemma 3, the elements of $\mathcal{B}_d$ are intervals of any order compatible with d, proving that $\mathcal{B}_d$ is a pre-pyramid on E. On the other hand, it is a clear fact that, for $B \in \mathcal{B}_d$, $\mathrm{diam}_d(B) = 0$ if and only if $B = B^d_{xx}$ for some $x \in E$, as d is semi-proper, i.e., if and only if B is minimal. Moreover, it follows obviously from Lemma 4 that diam_d is an index on $\mathcal{B}_d$, proving the direct assertion.

Conversely, let $(\mathcal{C}, f)$ be an indexed pre-pyramid on E. For all $x, y \in E$, denote by $\mathcal{C}(x,y)$ the collection of members of $\mathcal{C}$ that contain x and y, *i.e.* $\mathcal{C}(x,y) = \{C \in \mathcal{C} : x, y \in C\}$. Let d be the dissimilarity on E defined by $d(x,y) = \min\{f(C) : C \in \mathcal{C}(x,y)\}$.
The dissimilarity d is robinsonian. Let $x\theta y\theta z$, where θ is an order compatible with $\mathcal{C}$. Then every member of $\mathcal{C}$ containing x and z contains also y. Therefore,

$$\max\{d(x,y), d(y,z)\} = \max\{\min_{C \in \mathcal{C}(x,y)} f(C), \min_{C \in \mathcal{C}(y,z)} f(C)\} \leq \min_{C \in \mathcal{C}(x,z)} f(C) = d(x,z),$$

as required.
The dissimilarity d is semi-proper. Indeed, let $x, y \in E$ such that $d(x,y) = 0$. Then there is a member of $\mathcal{C}$ containing x, y, say C_{xy}, such that $f(C_{xy}) = 0$. On the other hand, by condition (CS2), there is a minimal member of $\mathcal{C}$ containing x, say C_x. Thus $C_{xy} = C_x$ since $f(C_x) = 0 = f(C_{xy})$ and f is an index. Let $z \in E$. Now C_{xy} being minimal, for all $C \in \mathcal{C}(x,z)$ we have, by Lemma 1-(a_1), $C_{xy} \subseteq C$. Therefore, every member of $\mathcal{C}$ containing x, z contains also y, z, so that

$$d(y,z) = \min_{C \in \mathcal{C}(y,z)} f(C) \leq \min_{C \in \mathcal{C}(x,z)} f(C) = d(x,z).$$

The pairs $(\mathcal{B}_d, \mathrm{diam}_d)$ and $(\mathcal{C}, f)$ are equal. Let $B := B^d_{xy} \in \mathcal{B}_d$. By Lemma 4, we assume, w.l.g., its bounds be x and y. Let $A_{xy} \in \mathrm{Argmin}\{f(C) : C \in \mathcal{C}(x,y)\}$. Then $B \subseteq A_{xy}$ since A_{xy} is an interval of θ. Let us prove the reverse inclusion. Denoting u and v the bounds of A_{xy}, we have $d(u,v) \leq f(A_{xy}) = d(x,y)$.
Now, as $x, y \in A_{xy}$ we get $d(x,y) \leq d(u,v) = \mathrm{diam}(A_{xy})$.
Then, by compatibility of d with θ, we have

$$\max\{d(x,u), d(x,v), d(y,u), d(y,v)\} \leq d(u,v) = d(x,y),$$

i.e., $u, v \in B$, proving that $B = A_{xy}$, hence $\mathcal{B}_d \subseteq \mathcal{C}$. To prove the reverse inclusion, let $C \in \mathcal{C}$ with bounds x and y ($x\theta y$). Then, as B^d_{xy} is an interval of θ, $C \subseteq B^d_{xy}$. In order to prove the reverse inclusion, let $v \notin C = [x,y]$. Let us first consider the case where $v\theta x$. Then for all $X \in \mathcal{C}(v,y)$, $d(x,y) \leq f(C) < f(X)$, because $C \subset X$, so that

$$d(x,y) < \min_{X \in \mathcal{C}(v,y)} f(X) = d(v,y).$$

Assume now that $y\theta v$. By an argument of symetry, we can deduce from the previous case that

$$d(x,y) < \min_{X \in \mathcal{C}(x,v)} f(X) = d(x,v).$$

Thus $v \notin B^d_{xy}$, proving that $B^d_{xy} \subseteq C$. Therefore, $C = B^d_{xy}$ and, by the way, x and y are the bounds of B^d_{xy}. Hence $C \in \mathcal{B}_d$, proving that $\mathcal{B}_d = \mathcal{C}$. In addition every $C(x,y)$ contains C, so that $f(C) = d(x,y) = \mathrm{diam}_d(B^d_{xy})$, proving that $(\mathcal{B}_d, \mathrm{diam}_d) = (\mathcal{C}, f)$.
The dissimilarity d is unique. Let d' be a semi-proper robinsonian dissimilarity such that $(\mathcal{B}_{d'}, \mathrm{diam}_{d'}) = (\mathcal{C}, f)$. Let $x, y \in E$ with $x\theta y$. Let u be the smallest element of E (w.r.t. θ) such that $d'(u,y) = d'(x,y)$ and let v be the greatest element of E such that $d'(u,v) = d'(u,y)$ ($u\theta x$ and $y\theta v$). Then

$$d'(u,v) = d'(x,y) \text{ and } B^{d'}_{uv} = [u,v] \in \mathcal{B}_{d'} = \mathcal{C} = \mathcal{B}_d.$$

Thus, as $x, y \in B^{d'}_{uv}$ (since $u\theta x\theta y\theta v$ and $B^{d'}_{uv}$ is an interval containing u and v),

$$d(x,y) \leq f(B^{d'}_{uv}) = \mathrm{diam}_{d'}(B^{d'}_{uv}) = d'(u,v) = d'(x,y).$$

Assume now that there exists a member $X \in \mathcal{C}(x,y)$ such that $f(X) < d'(x,y)$. If w and z are the bounds of X with $w\theta z$, then $w\theta x\theta y\theta z$. But $X \in \mathcal{C} = \mathcal{B}_{d'}$ implies $X = B^{d'}_{wz}$, so that $f(X) = \mathrm{diam}_{d'}(B^{d'}_{wz}) = d'(w,z)$. It follows then that $d'(w,z) < d'(x,y)$, which contradicts the compatibility of d' with θ. Hence

$$d'(x,y) \leq \min_{C \in \mathcal{C}(x,y)} f(C) = d(x,y),$$

proving that $d'(x,y) = d(x,y)$. Therefore $d' = d$, as required. □

It should be noticed that Theorem 2 still holds and the Diday's bijection (16) can be derived form it, when "$\mathcal{B}_d$" and "indexed" are replaced by "$\overline{\mathcal{B}_d}$" and "weakly indexed", respectively (17).

References

[1] Bandelt, H.-J. Four point characterization of the dissimilarity functions obtained from indexed closed weak hierarchies. Mathematisches Seminar, Universität Hamburg, Germany, 1992.

[2] Bandelt, H.-J., and Dress, A. W. M. An order theoretic framework for overlapping clustering. *Discrete Mathematics 136* (1994), 21–37.

[3] Batbedat, A. Les isomorphismes HTS et HTE (après la bijection de Benzécri/Johnson) (première partie). *Metron 46* (1988), 47–59.

[4] Batbedat, A. Application des isomorphismes HTS et HTE (vers la classification arborée) (seconde partie). *Metron 47* (1989), 35–51.

[5] Benzécri, J.-P. *L'Analyse des données : la Taxinomie.* Dunod, Paris, 1973.

[6] Bertrand, P. Set systems and dissimilarities. *Europ. J. Combinatorics 21* (2000), 727–743.

[7] Bertrand, P., and Janowitz, M. F. The k-weak hierarchical representations: an extension of the indexed closed weak hierarchies. *Discrete Applied Mathematics 127* (2003), 199–220.

[8] Brucker, F. Sub-dominant theory in numerical taxonomy. *Discrete Applied Mathematics 154* (2006), 1085–1099.

[9] Chepoi, V., and Fichet, B. l_∞-approximation via subdominants. *J. Math. Psychology 44* (2000), 600–616.

[10] Chepoi, V., and Seston, M. Seriation in the presence of errors: an approximation algorithm for fitting robinson structures to dissimilarity matrices. *Algorithmica 59* (2011), 521–568.

[11] Critchley, F., and Cutsem, B. V. An order theoretic unification and generalisation of certain fondamental bijections in mathematical classification. In *Classification and Dissimilarity Analysis*, B. V. Cutsem, Ed., Lecture Notes in Statistics. Spinger, 1994, pp. 87–141.

[12] Diatta, J. Dissimilarités multivoies et généralisations d'hypergraphes sans triangles. *Math. Inf. Sci. hum. 138* (1997), 57–73.

[13] Diatta, J. Approximating dissimilarities by quasi-ultrametrics. *Discrete Mathematics 192* (1998), 81–86.

[14] Diatta, J. One-to-one correspondence between cluster structures and weakly indexed closed cluster structures. In *Selected contributions in Data Analysis and Classification* (2007), P. Brito, P. Bertrand, G. Cucumel, and F. de Carvalho, Eds., Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, pp. 477–482.

[15] Diatta, J., and Fichet, B. From Apresjan hierarchies and Bandelt-Dress weak hierarchies to quasi-hierarchies. In *New Approaches in Classification and Data Analysis* (1994), E. Diday, Y. Lechevalier, M. Schader, P. Bertrand, and B. Burtschy, Eds., Springer-Verlag, pp. 111–118.

[16] Diday, E. Une représentation visuelle des classes empiétantes : les pyramides. Tech. Rep. 291, INRIA, France, 1984.

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

[18] Fichet, B. Data analysis: geometric and algebraic structures. In *Proceedings of the First World Congress of the BERNOULLI SOCIETY* (Tachkent, 1986), Y. A. Prohorov and V. V. Sazonov, Eds., V.N.U. Science Press, 1987, Vol. 2: 123–132.

[19] Gaul, W., and Schader, M. Pyramidal classification based on incomplete dissimilarity data. *Journal of Classification 10* (1993), 219–240.

[20] Johnson, S. C. Hierarchical clustering schemes. *Psychometrika 32* (1967), 241–254.

[21] Mirkin, B. *Mathematical Classification and Clustering.* Kluwer Academic Press, Dordrecht, 1996.

[22] Mirkin, B., and Rodin, S. N. *Graphs and Genes.* Sringer-Verlag, 1984.

[23] Robinson, W. S. A Method for Chronologically Ordering Archeological Deposits. *American Antiquity 16* (1951), 295–301.

[24] Warrens, M. J. On robinsonian dissimilarities, the consecutive ones property and latent variable models. *Advances in Data Analysis and Classification 3* (2009), 169–184.