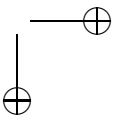
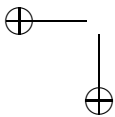


2004/8/17  
page i

---

# The Structural Representation of Proximity Matrices With MATLAB



# Contents

Preface	xi
<b>I (Multi- and Uni-dimensional) City-Block Scaling</b>	<b>1</b>
<b>1 Linear Unidimensional Scaling</b>	<b>3</b>
1.1 LUS in the $L_2$ -Norm . . . . .	4
1.1.1 A Data Set for Illustrative Purposes . . . . .	5
1.2 $L_2$ Optimization Methods . . . . .	6
1.2.1 Iterative Quadratic Assignment . . . . .	6
1.3 Confirmatory and Nonmetric LUS . . . . .	9
1.3.1 The confirmatory fitting of a given order using linfit.m	10
1.3.2 The monotonic transformation of a proximity matrix using proxmon.m . . . . .	11
1.4 The Dykstra-Kaczmarz Method . . . . .	15
<b>2 Linear Multidimensional Scaling</b>	<b>17</b>
2.1 The Incorporation of Additive Constants in LUS . . . . .	19
2.1.1 The $L_2$ Fitting of a Single Unidimensional Scale (with an Additive Constant) . . . . .	19
2.2 Finding and Fitting Multiple Unidimensional Scales . . . . .	22
2.3 Incorporating Monotonic Transformation of a Proximity Matrix	25
2.4 Confirmatory Extensions to City-Block Individual Differences Scaling . . . . .	27
<b>3 Circular Scaling</b>	<b>29</b>
3.1 The Mechanics of CUS . . . . .	31
3.1.1 The Estimation of $c$ and $\min\{ x_j - x_i , x_0 -  x_j - x_i \}$ for a Fixed Permutation and Set of Inflection Points	31
3.1.2 Obtaining Object Orderings and Inflection Points Around a Closed Continuum . . . . .	32
3.1.3 The Circular Unidimensional Scaling Utilities, cir- fit.m and cirfitac.m . . . . .	33
3.2 Circular Multidimensional Scaling . . . . .	39

---

<b>4</b>	<b>LUS for Two-Mode Proximity Data</b>	<b>45</b>
4.1	Reordering Two-Mode Proximity Matrices . . . . .	46
4.2	Fitting a Two-Mode Unidimensional Scale . . . . .	47
4.3	Multiple LUS Rearrangings and Fittings . . . . .	52
4.4	Some Useful Two-Mode Utilities . . . . .	56
4.5	Two-mode Nonmetric Bidimensional Scaling . . . . .	57
<b>II</b>	<b>The Representation of Proximity Matrices by Tree Structures</b>	<b>63</b>
<b>5</b>	<b>Ultrametrics for Symmetric Proximity Data</b>	<b>69</b>
5.1	Fitting a Given Ultrametric in the $L_2$ Norm . . . . .	71
5.2	Finding an Ultrametric in the $L_2$ Norm . . . . .	72
5.3	Graphically Representing an Ultrametric . . . . .	74
5.3.1	$\LaTeX$ Code for the Dendrogram of Figure 5.1 . . . . .	77
5.3.2	Plotting the Dendrogram with <code>ultraplot.m</code> . . . . .	80
<b>6</b>	<b>Additive Trees for Symmetric Proximity Data</b>	<b>83</b>
6.1	Fitting a Given Additive Tree in the $L_2$ -Norm . . . . .	84
6.2	Finding an Additive Tree in the $L_2$ -Norm . . . . .	85
6.3	Decomposing an Additive Tree . . . . .	87
6.4	Graphically Representing an Additive Tree . . . . .	89
6.5	An Alternative for Finding an Additive Tree in the $L_2$ -Norm . . . . .	90
<b>7</b>	<b>Fitting Multiple Tree Structures for a Symmetric Matrix</b>	<b>95</b>
7.1	Multiple Ultrametrics . . . . .	95
7.2	Multiple Additive Trees . . . . .	97
<b>8</b>	<b>Ultrametrics and Additive Trees for Two-Mode Data</b>	<b>101</b>
8.1	Fitting and Finding Two-Mode Ultrametrics . . . . .	102
8.2	Finding Two-Mode Additive Trees . . . . .	104
8.3	Completing a Two-Mode Ultrametric to one Defined on $S_A \cup S_B$ . . . . .	107
8.3.1	The goldfish_receptor data . . . . .	111
<b>III</b>	<b>The Representation of Proximity Matrices by Structures</b>	
	<b>Dependent on Order (Only)</b>	<b>113</b>
<b>9</b>	<b>Anti-Robinson (AR) Matrices for Symmetric Proximity Data</b>	<b>117</b>
9.0.2	Incorporating Transformations . . . . .	118
9.0.3	Interpreting the Structure of an AR matrix . . . . .	119
9.1	Fitting a Given AR Matrix in the $L_2$ -Norm . . . . .	121
9.1.1	Fitting the (In)-equality Constraints Implied by a Given Matrix in the $L_2$ -Norm . . . . .	122
9.2	Finding an AR Matrix in the $L_2$ -Norm . . . . .	124
9.3	Fitting and Finding a Strongly Anti-Robinson (SAR) Matrix in the $L_2$ -Norm . . . . .	126

Contents	v
9.4 The Use of Optimal Transformations and the m-function prox-mon.m . . . . .	129
9.5 Graphically Representing SAR Structures . . . . .	134
9.6 Representation Through Multiple (Strongly) AR Matrices . . .	137
<b>10 Circular-Anti-Robinson (CAR) Matrices</b>	<b>145</b>
10.1 Fitting a Given CAR Matrix in the $L_2$ -Norm . . . . .	147
10.2 Finding a CAR Matrix in the $L_2$ -Norm . . . . .	149
10.3 Finding a Circular Strongly-Anti-Robinson (CSAR) Matrix in the $L_2$ -Norm . . . . .	150
10.4 Graphically Representing CSAR Structures . . . . .	154
10.5 Representation Through Multiple (Strongly) CAR Matrices . .	154
<b>11 Anti-Robinson (AR) Matrices for Two-Mode Proximity Data</b>	<b>163</b>
11.1 Fitting and Finding Two-Mode AR Matrices . . . . .	164
11.2 Multiple Two-Mode AR Reorderings and Fittings . . . . .	167
<b>Bibliography</b>	<b>173</b>
<b>A Header comments for the mentioned m-files</b>	<b>179</b>
<b>Indices</b>	<b>210</b>
Author Index . . . . .	210
Subject Index . . . . .	212

# List of Tables

- 1.1 The number.dat data file extracted from Shepard, Kilpatric, and  
Cunningham (1975) . . . . . 6
  
- 3.1 A proximity matrix, morse\_digits.dat, for the ten Morse code sym-  
bols representing the first ten digits (data from Rothkopf, 1957) . 31
  
- 4.1 The goldfish\_receptor.dat data file constructed from Schiffman and  
Falkenberg (1968) . . . . . 46
- 4.2 The two unidimensional scalings of the goldfish\_receptor data . . . 53
  
- 9.1 Order-constrained least-squares approximations to the digit prox-  
imity data of Shepard *et al.* (1975); the upper-triangular portion  
is anti-Robinson and the lower-triangular portion is strongly-anti-  
Robinson . . . . . 135
- 9.2 The 45 subsets listed according to increasing diameter values that  
are contiguous in the object ordering used to display the upper-  
triangular portion of Table 9.1. The 22 subsets given in italics  
are redundant in the sense that they are proper subsets of another  
listed subset with the same diameter. . . . . 136
- 9.3 The fourteen (nonredundant) subsets listed according to increasing  
diameter values are contiguous in the linear object ordering used  
to display the lower-triangular SAR portion of Table 9.1. . . . . 139
  
- 10.1 The fifteen (nonredundant) subsets listed according to increasing  
diameter values are contiguous in the circular object ordering used  
to display the CSAR entries in Table 10.2. . . . . 155
- 10.2 A circular strongly-anti-Robinson order-constrained least-squares  
approximations to the digit proximity data of Shepard *et al.* (1975). 155

# List of Figures

4.1	Two-dimensional joint biplot for the <code>goldfish_receptor</code> data obtained using <code>biplotm.m</code> . . . . .	57
4.2	Two-dimensional joint biplot for the <code>goldfish_receptor</code> data obtained using <code>bimonscaltmac.m</code> and <code>biplotm.m</code> . . . . .	61
5.1	A dendrogram (tree) representation for the ultrametric described in the text having VAF of .4941 . . . . .	78
5.2	Dendrogram plot for the number data obtained using <code>ultraplot.m</code> . . . . .	82
6.1	A dendrogram (tree) representation for the ultrametric component of the additive tree described in the text having VAF of .6359 . . . . .	91
6.2	A graph-theoretic representation for the additive tree described in the text having VAF of .6359 . . . . .	92
9.1	Two $4 \times 4$ submatrices and the object subsets they induce, taken from the anti-Robinson matrix in the upper-triangular portion of Table 9.1. For (a), a graphical representation of the fitted values is possible; for (b), the anomaly indicated by the dashed lines prevents a consistent graphical representation from being constructed.	138
9.2	A graphical representation for the fitted values given by the strongly-anti-Robinson matrix in the lower-triangular portion of Table 9.1.	139
10.1	A graphical representation for the fitted values given by the circular strongly-anti-Robinson matrix in the lower-triangular portion of Table 10.2 (VAF = 72.96%). Note that digit 3 is placed both in the first and the last positions in the ordering of the objects with the implication that the sequence continues in a circular manner. This circularity is indicated by the curved dashed line. . . . .	156

# Preface

As the title of this monograph implies, our main goal is to provide and illustrate the use of functions (by way of m-files) within a MATLAB<sup>1</sup> computational environment to effect a variety of structural representations for proximity information assumed available on a set of objects. The structural representations that will be of interest have been discussed and developed primarily in the applied (behavioral science) statistical literature (e.g., in psychometrics and classification), although interest in these topics has now extended much more widely (for example, to bioinformatics and chemometrics). We subdivide the monograph into three main sections depending on the general class of representations being discussed. Part I will develop linear and circular uni- and multi-dimensional scaling using the city-block metric as the major representational device; Part II is concerned with characterizations based on various graph-theoretic tree structures, and specifically with those usually referred to as ultrametrics and additive trees; Part III uses representations defined solely by order properties, and particularly to what are called (strongly) anti-Robinson forms. Irrespective of the part of the monograph being discussed, there generally will be two kinds of proximity information analyzed: one-mode and two-mode. One-mode proximity data are defined between the  $n$  objects from a *single* set, and usually given in the form of a square ( $n \times n$ ) symmetric matrix with a zero main diagonal; two-mode proximity data are defined between the objects from two distinct sets containing, say,  $n_a$  and  $n_b$  objects, respectively, and given in the form of a rectangular ( $n_a \times n_b$ ) matrix. Also, there will generally be the flexibility to allow the fitting (additively) of multiple structures to either the given one- or two-mode proximity information.

It is not the intent of the monograph to present formal demonstrations of the various assertions we might make along the way, such as for the convergence of a particular algorithm or approach. All of this is generally available in the literature (and much of it by the authors of the current monograph), and the references to this source material is given when appropriate. The primary interest here is to present and demonstrate how to actually find and fit these structures computationally with the help of some sixty-five functions (though m-files) we provide that are usable within a MATLAB computational environment. The usage header information for each of these functions is given in Appendix A (listed alphabetically). The m-files themselves can be downloaded individually from

---

<sup>1</sup>MATLAB is a registered trademark of The MathWorks, Inc.

[http://cda.psych.uiuc.edu/srpm\\_mfiles](http://cda.psych.uiuc.edu/srpm_mfiles)

Also, there is a “zipped” file called `srpm_mfiles.zip` at this site that includes them all, as well as the few small data sets used throughout the monograph to illustrate the results of invoking the various m-files (or we might say, invoking the various “m-functions”); thus, the reader should be able to reproduce all of the examples given in the monograph (assuming, obviously, access to a MATLAB environment).

The computational approach implemented in the provided m-files for obtaining the sundry representations, are by choice, invariably least-squares, and based on what is called the Dykstra-Kaczmarz (DK) method for solving linear inequality constrained least-squares tasks. The latter iterative strategy is reviewed in Chapter 1 (Section 1.4, in particular). All of the representations of concern (over all three monograph Parts) can be characterized by explicit linear inequalities; thus, once the latter constraints are known (by, for example, the identification of certain object permutations through secondary optimization problems such as quadratic assignment), the actual representing structure can be obtained by using the iterative DK strategy. Also, as we will see particularly in Part II dealing with graph-theoretic tree structures (ultrametrics and additive trees), the DK approach can even be adopted heuristically to first identify the inequality constraints that we might wish to impose in the first place. And once identified in this exploratory fashion, a second application of DK could then do a confirmatory fitting of the now fixed inequality constraints.

As noted above, our purpose in writing this monograph is to provide an applied documentation source for a collection of m-files that would be of interest to applied statisticians and data analysts but also accessible to a notationally sophisticated but otherwise substantively focused user. Such a person would typically be most interested in analyzing a specific data set by adopting one (or some) of the structural representations we discuss. The background we have tried to assume is at the same level required to follow the documentation for good, commercially available optimization subroutines, such as the Numerical Algorithms Group (NAG) Fortran subroutine library, or at the level of one of the standard texts in applied multivariate analysis usually used for a graduate second-year methodology course in the behavioral and social sciences. An excellent example of the latter would be the widely used text now in its fifth edition by Johnson and Wichern (2002). Draft versions of the current monograph have been used as supplementary material for a course relying on the latter text as the primary reference.

The research reported in this monograph has been partially supported by the National Science Foundation through Grant No. SES-981407 (to LH), and by the Netherlands Organization for Scientific Research (NWO) through Grant No. 575-67-053 for the ‘PIONEER’ project ‘Subject Oriented Multivariate Analysis’ (to JM).

Lawrence Hubert  
Phipps Arabie  
Jacqueline Meulman  
September, 2004