Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 50 (2006) 21–39
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Principal component analysis of binary data by iterated singular value decomposition

Jan de Leeuw*

Department of Statistics, University of California, Los Angeles, 8130 Math Sciences Blvd., Los Angeles, CA 90095-1554, USA

Available online 21 August 2004

Abstract

The maximum-likelihood estimates of a principal component analysis on the logit or probit scale are computed using majorization algorithms that iterate a sequence of weighted or unweighted singular value decompositions. The relation with similar methods in item response theory, roll call analysis, and binary choice analysis is discussed. The technique is applied to 2001 US House roll call data.

Keywords: Multivariate analysis; Factor analysis; Binary data; Item response models; Applications to social sciences

1. Introduction

Suppose $P = \{p_{ij}\}$ is an $n \times m$ binary data matrix, i.e. a matrix with elements equal to zero or one (representing yes/no, true/false, present/absent, agree/disagree). For the moment we suppose that P is complete, the case in which some elements are missing is discussed in a later section.

There are many examples of such binary data in the sciences. We give a small selection in Table 1, many more could be added.

Many different statistical techniques have been developed to analyze data of this kind. One important class is latent structure analysis (LSA), which includes latent class analysis, latent trait analysis and various forms of factor analysis for binary data. Alternatively,

* Tel.: +1-310-825-9550; fax: +1-310-206-5658.
E-mail address: deleeuw@stat.ucla.edu (Jan de Leeuw).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.07.010

Table 1
Binary data

Discipline	Rows	Columns
Political science	Legislators	Roll calls
Education	Students	Test items
Systematic zoology	Species	Characteristics
Ecology	Plants	Transects
Archeology	Artefacts	Graves
Sociology	Interviewees	Questions

by recoding the data as a 2^m table, log-linear decompositions and other approximations of the multivariate binary distribution become available. There are also various forms of cluster analysis which can be applied to binary data, usually by first computing some sort of similarity measure between rows and/or columns. And finally there are variations of principal component analysis (PCA) specifically designed for binary data, such as multiple correspondence analysis (MCA).

In this paper, we combine ideas of LSA, more particularly item response theory and factor analysis of binary data, with PCA and MCA. This combination produces techniques with results that can be interpreted both in probabilistic and in geometric terms. Moreover, we propose algorithms that scale well, in the sense that they can be fitted efficiently to large matrices.

Our algorithm is closely related to the logistic majorization algorithm proposed by Groenen et al. (2003). We improve on their somewhat heuristic derivation, propose an alternative uniform logistic majorization, and a uniform probit majorization.

2. Problem

The basic problem we solve in this paper is geometric. We want to represent the rows of the data matrix as points and the columns as hyperplanes in low-dimensional Euclidean space $\mathbb{R}^r$, i.e. we want to make a drawing of our binary matrix. Rows i are represented as points a_i and the hyperplanes corresponding with columns j are parametricized as vectors of slopes b_j and as scalar intercepts c_j. The parameter r is the dimensionality of the solution. It is usually chosen to be equal to two, but drawings in different dimensionalities are also possible.

The drawing should be constructed in such a way that points a_i for which $p_{ij}=1$ should be on one side of hyperplane (b_j, c_j) and the points for which $p_{ij}=0$ should be on the other side. Or, equivalently, if we define the point sets $\mathscr{A}_{j1} = \{a_i \mid p_{ij} = 1\}$ and $\mathscr{A}_{j0} = \{a_i \mid p_{ij} = 0\}$, the convex hulls of $\mathscr{A}_{j1}$ and $\mathscr{A}_{j0}$ should be disjoint. Of course we want these disjoint convex hulls for all columns j simultaneously, and this is what makes the representation restrictive. Depending on the context, such a representation, if possible, is known as an inner product representation, a vector representation, or a compensatory representation. In the multidimensional scaling literature the algebraic version of the compensatory or vector model is usually attributed to Tucker (1960), although Coombs (1964) reviews some earlier work by his students and co-workers. The vector representation is most often applied to preference rank orders, but also quite often to binary choices and paired comparisons.