Interpretable Multivariate Analysis Procedures by Their Matrix-Intensive Refinements

Naoto Yamashita

1. Introduction

In empirical sciences, data is collected in various forms regardless of qualitative or quantitative, and structured or non-structured. The ultimate goal of empirical sciences is to understand the laws of nature deeply, and successively, to predict its future. In order to accomplish the goal, researchers often conduct their researches in hypothetico-deductive manner, where a hypothesis creation and its validation is repeated. For example, for the purpose of understanding a fundamental mechanism of psychological process of human, a psychologist establishes a research hypothesis based on the previous researches. Then, some psychological experiments would be planned and conducted to verify the hypothesis. Since the verification process is totally based on data collected by experiments or observations, data play a critical role in today's scientific research and its data-based hypothesis verification.

The thesis firstly focuses on multivariate data analysis procedures and their exploratory usages, which is called exploratory data analysis. They are often used for extracting clues for hypothesis creation. For example, Principal Component Analysis (PCA) can extract essential components of a data matrix, excluding redundant information by compressing multivariate data into lower dimensional space.

In exploratory data analysis, it is considered that how easily a resulting solution can be interpreted, that is, the *interpretability* of a solution, is a very important property. Interpretability is also important in the context of prediction, because it assures the accountability of prediction models. However, in general, interpretability is not always considered in the most of multivariate data analysis procedures, while they sorely consider how well their solution fit to the given data.

Therefore, in this thesis, the author proposes a series of procedures to improve the interpretability of various multivariate data analysis procedures, addressing that interpretability is an essential property in their exploratory use cases. Two sub-concepts of interpretability are proposed, sparseness (how many a solution matrix contains zero elements) and simple structure. The existing procedures are modified by matrix-intensive refinements so as to their solutions are highly interpretable. Six procedures proposed in the thesis are categorized into the following four families; rotation of solution matrices, combination with clustering, sparse estimation with cardinality constraint, and some emerging techniques.

2. Factor Rotation to Simple Structure with Permutation of Variables

A new rotation technique is proposed, to overcome the critical drawback of the existing procedures for target rotation; the correspondence of the variables in a target matrix to a loading matrix is unknown. In the proposed procedure, a loading matrix is rotated simultaneously with a permutation of the rows of the target matrix, so that the rotated loading matrix is optimally matched with the row-permuted target matrix. Its algorithm is presented, with Thurstone’s definition of simple structure modified so as to specify the target matrix uniquely. Permutimin is illustrated with real data examples, as presented in Table 1, and the relationships between Permutimin and Procrustes rotation is discussed.

Table 1: Rotated loading matrices for the box problem by Permutimin, Promax, and Geomin.

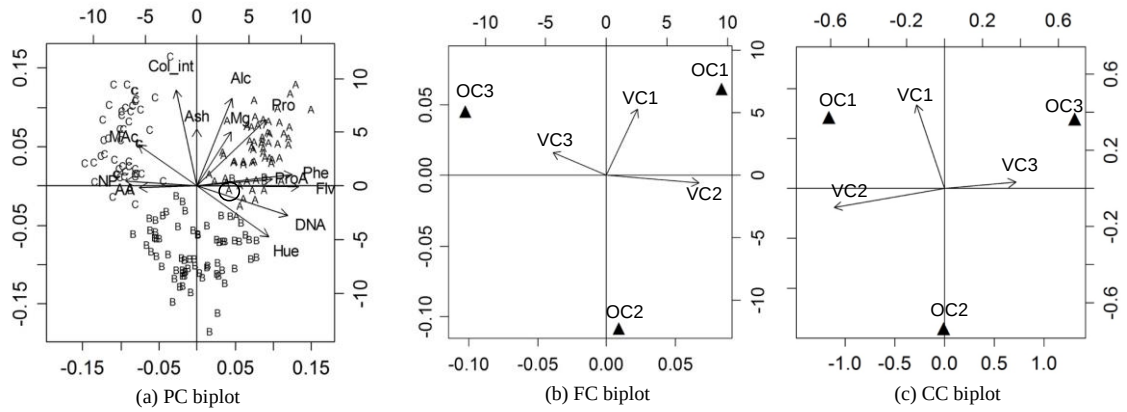
	Permutimin			Promax			Geomin		
	F1	F2	F3	F1	F2	F3	F1	F2	F3
x	**0.98**	0.03	0.03	**-0.52**	**0.48**	**0.88**	**1.49**	0.00	**0.88**
y	0.02	**0.96**	0.03	**0.46**	**0.82**	**-0.37**	0.03	**0.58**	**-0.75**
z	0.03	0.04	**0.96**	**0.84**	**-0.32**	**0.43**	0.07	**-0.50**	**-0.79**
xy	**0.61**	**0.67**	0.00	-0.04	**0.86**	0.26	**0.93**	**0.42**	0.06
xy^2	**0.35**	**0.84**	0.03	0.22	**0.87**	-0.03	**0.54**	**0.50**	**-0.35**
$2x+2y$	**0.54**	**0.75**	-0.04	0.01	**0.91**	0.15	**0.82**	**0.49**	-0.04
$(x^2+y^2)^{1/2}$	**0.51**	**0.74**	-0.02	0.03	**0.88**	0.14	**0.78**	**0.47**	-0.07
xz	**0.61**	0.01	**0.67**	0.24	0.04	**0.83**	**0.94**	**-0.36**	0.01
xz^2	**0.42**	-0.03	**0.82**	**0.46**	-0.13	**0.75**	**0.67**	**-0.47**	-0.25
$2x+2z$	**0.58**	-0.02	**0.71**	0.27	-0.01	**0.84**	**0.91**	**-0.40**	-0.02
$(x^2+z^2)^{1/2}$	**0.55**	-0.02	**0.71**	0.30	-0.02	**0.81**	**0.85**	**-0.40**	-0.05
yz	-0.03	**0.60**	**0.63**	**0.85**	0.26	0.00	-0.02	0.02	**-1.01**
yz^2	-0.01	**0.43**	**0.77**	**0.88**	0.07	0.14	0.00	-0.16	**-0.98**
$2y+2z$	-0.01	**0.61**	**0.63**	**0.84**	0.28	0.00	0.00	0.03	**-1.00**
$(y^2+z^2)^{1/2}$	0.01	**0.61**	**0.59**	**0.79**	**0.30**	0.01	0.04	0.05	**-0.95**

3. Rotation in Canonical Correlation Analysis as Maximizing Sum of Squared Correlations

Rotation toward simple structure is extended to canonical correlation analysis (CANO). There, a new formulation of CANO is firstly proposed, which is proved to be equivalent to the exiting one. Two canonical structure matrices have freedom with respect to orthogonal rotation under the new formulation. The study thus proposes an orthogonal rotation method for rotating two structure matrices individually for their simplicity and interpretability, while the existing formulation allows only simultaneous rotation of the matrices. It is illustrated that the proposed method facilitates the interpretation of solutions of CANO by a real data example.

4. Biplot Procedures with Joint Classification Objects and Variables

Chapter 4 deals with biplot, a technique for obtaining a low-dimensional configuration of a data matrix. Biplot with a large number of objects and variables is known to be difficult to interpret. Therefore, this study proposes a new biplot procedure where objects and variables are classified into a small number of clusters



(a) PC biplot (b) FC biplot (c) CC biplot

Figure 1: Principal component (PC) biplot and fuzzy (FC) and crisp (CC) biplots for wine data.

Table 2: Estimated centroid matrix by CCKM with row/column-cardinality constraint and four clusters for job impression data; exact-zero elements are shown as blank cells.

	C1	C2	C3	C4
admirable	0.34			
useful	0.56	-0.39		
good				
large		-0.42		
powerful		-0.80	1.03	
strong		-0.71	1.01	
fast			0.95	-1.10
noisy			0.87	-1.21
young		0.69		-1.55
honest	0.44			
stubborn	0.57	-1.24		
busy	0.59			-1.24

by means of *K*-means clustering. The resulting clusters are simultaneously biplotted in lower-dimensional space. The resulting biplot is thus composed of fewer points and vectors and therefore easily captured. An extension of the proposed method to fuzzy *K*-means clustering is also proposed. A simulation study and real data example are also provided to demonstrate the effectiveness of the proposed procedures. Figure 1 shows illustrative biplots of the existing and the two proposed procedures.

5. A Modified *K*-means clustering for obtaining a simple centroid matrix

Chapter 5 considers the interpretability in *K*-means clustering since it does not allow any post-hoc transformation of solutions. The study proposes a new procedure for obtaining a centroid matrix, so that it has a number of exactly zero elements by cardinality constraint. This allows easier interpretation of the matrix, as we may focus on only the nonzero centroids. The development of an iterative algorithm for the constrained minimization is described. A special case of the proposed procedure is also proposed, in which some restrictions are imposed on the positions of nonzero elements. Behaviors of the proposed procedure were evaluated in simulation studies and are illustrated with three real data examples, which demonstrate that the performance of the procedure is promising. One of these examples is presented in Table 2, which exhibits highly simple structure.

6. Layered Multivariate Regression with Its Applications

A novel framework of multivariate analysis procedure is proposed that is called as Layered Multivariate Analysis (LMA), which includes Layered Multivariate Regression (LMR) as a special case. In LMR, a regression coefficient matrix is assumed to be the sum of several sparse matrices, which is called layer. Therefore, the sparseness of the resulting coefficient matrix is controlled by how many layers are used. It is theoretically guaranteed that an LMR solution converges to the unconstrained solution as the number of layers increases. LMR is assessed in a simulation study and illustrated with a real data example. Table 3 shows the resulting coefficient matrices obtained by LMR and the one by unconstrained estimation. Further, Layered PCA is also proposed, in which a loading matrix is constrained to have a layered simple structure.

Table 3: Estimated coefficient matrices by LMR with $L = 1, \ldots, 4$ and unconstrained solution with proportion of variance explained (Vexp.); element equaling to zero shows a blanc cell.

	LMR (L = 1)			LMR (L = 2)			LMR (L = 3)			LMR (L = 4)			unconstrained		
	Brate	Sug	Nic	Brate	Sug	Nic	Brate	Sug	Nic	Brate	Sug	Nic	Brate	Sug	Nic
N		-0.69			-0.65	0.51		-0.52	0.39		-0.58	0.30	0.10	-0.58	0.29
Cl	-0.61			-0.60	0.38		-0.60	0.41	-0.31	-0.60	0.39	-0.32	-0.58	0.39	-0.32
K	0.62			0.52	0.26		0.52	0.14		0.52	0.20	0.11	0.45	0.20	0.11
P		0.17		-0.14	0.20		-0.14	0.22		-0.14	0.22		-0.13	0.22	-0.05
Ca	0.33			0.47			0.47		0.18	0.47	0.11	0.25	0.41	0.11	0.24
Mg			0.73	-0.26		0.43	-0.26	-0.21	0.40	-0.26	-0.21	0.47	-0.32	-0.22	0.48
Vexp.	0.569			0.701			0.740			0.742			0.744		

7. Procrustes Penalty Function for Matching Matrices to Targets

A new penalty function that can be used for penalized estimation in various multivariate analysis procedures is proposed. The new penalty function shrinks a solution matrix towards a target matrix with simple structure. The proposed function is a generalization of the existing ones, in that it includes LASSO and ridge penalties as special cases.

8. Discussion

In order to conclude the thesis, the procedures proposed in the foregoing chapters are reviewed, and they are categorized into some classes according to their strategies for enhancing the interpretability of solutions. Their limitations and directions for the future studies are also discussed. Finally, it is discussed that the proposed procedures are superior to the existing ones, in that they are able to achieve sparseness and simple structure simultaneously, which is essential for interpretability. (Behavioral Statistics)