Title	Interpretable Multivariate Analysis Procedures by Means of Their Matrix-Intensive Refinements
Author(s)	山下, 直人
Citation	
Issue Date	
Text Version	ETD
URL	https://doi.org/10.18910/76335
DOI	10.18910/76335
rights	
Note	

Abstract of Thesis

Name (Naoto Yamashita)

Title: Interpretable Multivariate Analysis Procedures by Means of Their Matrix-Intensive Refinements
(線形代数学的アプローチによる解釈容易な多変量解析法に関する研究)

Abstract of Thesis

In empirical sciences, data are 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. Seven 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.

In Chapter 2, 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, and the relationships between Permutimin and Procrustes rotation is discussed.

In the next chapter, rotation toward simple structure is extended to canonical correlation analysis (CANO). 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.

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 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.

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.

In Chapter 6, a novel framework of multivariate analysis procedure is proposed that is called 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 a 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. Further, layered PCA is also proposed, in which a loading matrix is constrained to have a layered simple structure.

In Chapter 7, 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.

The final section is devoted for summarization and general discussions throughout the foregoing chapters. The thesis finally concludes that the studies successfully developed procedures that are able to simplify solution matrices, by accomplishing sparsity and simplicity at the same time, which is the main contribution of the thesis to the studies on multivariate data analysis.

様式 7

論文審査の結果の要旨及び担当者

氏　名　（　山下　直人　）		
	（職）	氏　名
論文審査担当者	主査　教授 副査　教授 副査　教授	足立　浩平 八十島　安伸 入戸野　宏

論文審査の結果の要旨

人間科学で観測されるデータの多くは，多変量，つまり，各個体から観測される複数種の変数の数値集合であり，その身近な例は，各受験者の複数テストの得点である．変数どうしの関係を考慮しながら，上記の多変量データを分析する統計解析の諸方法は多変量解析と総称され，人間科学の重要なツールとなっているが，ユーザーの多様なニーズに応えるため，統計科学では新たな解析手法の開発が続いている．本論文の諸研究もそれに含まれるが，特に，分析法が出力する解の解釈可能性の向上(平易に言えば，解釈が容易になること)というユーザーのニーズに応えることを目指した研究開発である．本論文において解釈可能性の向上に関わるキーワードとして，クラスター化・単純構造・スパース性・可視化が掲げられる．クラスター化とは個体や変数の少数群への類別，単純構造とは類別に加えて群間相違が明確な解の構造を指す．スパース性は解を構成する数値の多くが零であることを指し，これは単純構造を強めた性質とも位置づけられ，可視化とは高次元のデータ空間を可視の3次元以下の解空間に縮約することを指す．これらの各キーワードまたはその複合を多変量解析の個別手法に実現させるアプローチが，本論文で考案および論考され，そのための数学的基礎は，特に，行列(マトリックス)の数理を扱う行列代数(線形代数)である．

冗長さを抑えた上で133頁に及ぶ英文の本論文では，1章(序論)と最終8章(総合考察)の間を占める2～7章で考案手法が記される．まず，2章と3章は単純構造に関わる．複数組の解が存在する場合に，それらの中で，単純構造を持つ解を数理的に探索する方法は，回転法と総称されるが，2章では，置換(Permutation)行列の最適化を伴う稀有でかつ有用な因子回転法(Permutimin回転)が考案される．3章では，回転が許容されないがために解の解釈が容易でなかった正準相関分析を，回転が許容されるように再定式化できることを示す新たな定理が証明され，正準相関分析のための回転法が考案される．4章では，可視化とクラスター化を組み合わせた新たなバイプロット法(個体と変数を同一空間内に布置させる方法)，すなわち，把握できる以上の多数の個体・変数からなるビッグデータを考慮して，個体と変数のクラスター化とバイプロットを同時に行う解析法が考案される．続いて，5章では，スパースなクラスター特徴行列と個体のクラスター化を同時に行う新たなクラスター分析法が考案される．また，6章では，多変量回帰分析や主成分分析などの各種の多変量解析法が出力する解の行列を，層と称するスパースな行列の和になるように再定式化した手法(層化多変量解析法)が考案され，この方法によって，ユーザーは，解の各層を参照することで，その構造を容易に把握できる．そして，7章では，機械学習でよく使われるペナルティ関数を発展させ，単純構造のターゲット行列に解をできるだけ近づけることを目指したProcrustesペナルティという関数が考案される．

上述の各章の論述からは，いずれの考案手法も優れた洞察に育まれることが読みとれ，各手法が解を求めるために要するアルゴリズム(解法)もエレガントに仕上がっている．また，理論的には，行列代数に基づく幾つかの新たな定理が提示されたことに見られるように，考案手法は，着実な数理的考究に基礎を置く．その一方で，シミュレーション(数値実験)によって，考案手法による解の真値の再現精度を評価する手続きもとられ，新たな解析手法の挙動の実証的な検証もなされている．さらに，考案手法の行動科学データへの適用例は，手法の有用性を示すのに説得力の高いものとなっている．また，序論の1章では，既存の諸方法が手際よく再構成的にレビューされ，最終8章では，考案された諸方法の立体的な位置づけと再考がなされる．本論文で考案された多変量解析の諸方法は，実証科学の諸分野で利用可能であり，さらに，解の解釈可能性の向上を目指した手法であることから，人間科学の殆どの分野に浸透可能であり，人間科学の方法論の発展に大きく寄与できるものと高く評価される．以上の論文審査の結果，本論文は博士(人間科学)の学位を授与するのにふさわしいものと判定した．