

A Study on Principal Points for a Multivariate Binary Distribution

January 2015

Haruka Yamashita

A Thesis for the Degree of Ph.D. in Engineering

A Study on Principal Points
for a Multivariate Binary Distribution

January 2015

Graduate School of Science and Technology
Keio University

Haruka Yamashita

主 論 文 要 旨

報告番号	甲 乙 第	号	氏 名	山下 遥
主 論 文 題 目： A Study on Principal Points for a Multivariate Binary Distribution (多変量 2 値分布における Principal Points に関する研究)				
(内容の要旨) 近年、統計の分野において Principal Points に関する研究が数多く展開されている。Principal Points は“与えられた分布の代表点” として定義され、その理論的性質や応用について幅広く議論されてきたが、多変量 2 値分布に関する Principal Points に関する議論は今までされてこなかった。 本研究では、多変量 2 値分布における Principal Points を定義し、それを 2 値型 Principal Points と呼ぶことにする。この 2 値型 Principal Points は、 k 個の 2 値ベクトルとして与えられ、多変量 2 値分布をその実現値で要約することを可能とする。この方法は、従来までの一連の Principal Points 研究に対して、適用範囲の拡張を可能とするものである。しかしながら、実際のデータ解析に 2 値型 Principal Points を適用する場合、データから真の分布を推定して 2 値型 Principal Points を求める必要がある。また、この問題は NP 困難な問題であり、特に多くの変数を含む分布において最適解を求めることは難しい。 そこで、まず与えられた多変量 2 値データから 2 値型 Principal Points を推定するための 2 つの方法を提案する。1 つ目は、与えられたデータから得られる確率分布をそのまま真の分布として 2 値型 Principal Points を求める方法、もう 1 つは、与えられたデータから得られる確率分布は、対数線形モデルで表されるものと仮定し、最尤法によって推定されたパラメータから得られる真の確率分布を基に 2 値型 Principal Points を求める方法である。前者の 2 値型 Principal Points の推定方法をノンパラメトリック法、後者の推定方法をパラメトリック法と呼ぶことにする。 また、2 値型 Principal Points が NP 困難な問題であることに対して、本研究は 3 つの近似解法を提案する。1 つ目は、貪欲法による近似解法、2 つ目は k -means 法を基礎にした 2 値型 k -means 法、そして、3 つ目は、 k 個の 2 値型 Principal Points に対して $2k$ 個の候補を選び、近似解を求める方法である。1 つ目の方法は、2 値型 Principal Points の劣モジュラ性に基づいた近似解法で、その近似解は $(1-1/e)$ 倍以上の精度を保証することを理論的に示すことができる。2 つ目の方法は、Principal Points は少数個の代表点であることに注目して、問題のサイズを小さくしている。3 つ目の方法は、 k -means アルゴリズムを多変量 2 値分布の解析のために改良した手法であり、2 値型 Principal Points の問題設定に最もよくフィットするが、初期値の問題を有している。これら 3 つの方法を比較するためにシミュレーションデータおよびアンケートデータに 2 値型 Principal Points の推定および 3 つの近似解法を適用し、それらの結果を比較する。				

SUMMARY OF Ph.D. DISSERTATION

School Science and Technology	Student Identification Number	SURNAME, First name Yamashita Haruka
Title A Study on Principal Points for a Multivariate Binary Distribution		
Abstract In statistics, there are many studies on principal points. The concept of principal points which is proposed by Flury allows us to carry out an analysis in a variety of applications and properties of principal points also have been studied. Although principal points of a multivariate distribution have widely studied, there is no discussion of principal points for a multivariate binary distribution. In this thesis, we define the principal points for a multivariate binary distribution firstly. Since principal points for a multivariate binary distribution are selected from multivariate binary region, summarization of a multivariate binary distribution with binary vectors is available. However there are several problems of this definition. The first problem is data analysis with principal points for a multivariate binary distribution. Since principal points are the representative points for a given distribution, we have to estimate a multivariate binary distribution of the given data for the estimation of principal points for a multivariate binary distribution. Therefore, we propose two estimation methods of principal points for a multivariate binary distribution; a nonparametric estimation method and a parametric estimation method. The second problem is that the problem of principal points for a multivariate binary distribution is an NP-hard problem. We propose an approximation method based on the greedy algorithm. Since the property of principal points for a multivariate binary distribution, the approximated optimal value is at least $(1 - 1/e)$ times the optimal solution. We also propose two approximation methods. First one reduces the size of optimization by restricting the candidates from two view points, and solves the reduced optimization problem (points-oriented method). Another is based on k -means clustering method. This method finds the binary center point of each cluster in the each iteration of k -means algorithm. Finally we apply the estimation method and approximation method for principal points for a multivariate binary distribution to simulation study data and questionnaire survey data and compare the results.		