博 士（情報科学） 青 木 健 児

学 位 論 文 題 名

Study on Significance Verification of the Dimensions in Multivariate Analysis for Categorical Data

（質的データの多変量解析における次元の有意性検証に関する研究）

学位論文内容の要旨

実験や調査などによって得られるデータは2種類に大別される。一つは量的データ (quantitative data) であり、もう一つは質的データ (qualitative data) である。例えば、長さ、重さ、回数、個数などのように、数値として観測されたデータは量的データと呼ばれる。一方、性別、職業、学歴などのように、数値としてではなく、属性として観測されたデータは質的データと呼ばれる。また、観測の対象となる特徴量も同様に量的変数 (quantitative variable) と質的変数 (qualitative variable) の2種類に分けられる。上の例では、長さ、重さ、回数、個数を表わす変数が量的変数であり、性別、職業、学歴を表わす変数が質的変数である。人文、社会科学の分野で行なわれるアンケート調査や多肢選択式のテストなどでは、量的データよりも質的データが得られる機会が圧倒的に多い。したがって、これらの分野では、量的データよりも質的データを分析するための統計手法が必要とされる。

多変量解析とは、実験や調査などにおいて複数の特徴量が観測されたとき、それらの値をもとに変数間の従属関係や相関関係を調べたり、多数の変数を少数の変数に要約したり、観測された個体の判別規則を構築したりするための統計手法の総称である。多変量解析の手法の一つである正準相関分析は2組の変数セット間の相関構造を少数個の正準変数によって表わすための手法である。また、正準相関分析は多変量解析の中で数学的に最も一般的な手法であり、他のほとんどの手法は正準相関分析の特別な場合として定式化される。したがって、正準相関分析は理論的側面において重要な手法であるといえる。

次元縮小を目的とする正準相関分析において、分析対象である2組の変数セット間の相関構造がどの程度複雑であるのか、すなわち有意な正準相関係数の個数 (正準変数の個数) がいくつなのかを知ることは特に重要な問題である。この問題は、正準相関分析だけではなく、正準判別分析などの次元縮小を目的とする他の手法においても重要である。

ゼロでない正準相関係数の個数は2組の変数セット間の母集団における共分散行列のランクに等しく、正準相関分析における次元数と呼ばれる。したがって、有意な正準相関係数をゼロでない正準相関係数と定義したとき、有意な正準相関係数の個数を推定する問題は正準相関分析における次元数を推定する問題に帰着される。次元数を推定するための代表的な手段は共分散行列のランクの順次検定を行なうこと、すなわち次元検定問題 (dimensionality testing problem) を考えることである。量的変数に対する正準相関分析において、母集団が正規性を持つという仮定のもとで、この問

題のための検定量がいくつか導出されている。質的変数に対する正準相関分析においても、正規性を仮定した場合と同じ検定量を用いることが提案されている。しかし、質的変数に対しては正規性の仮定は満たされない。したがって、質的変数に対する正準相関分析において、これらの検定量を用いることの数学的な正当性はない。実際、数値シミュレーションにより、対象とする変数が質的変数である場合、これらの検定量の精度が悪いことが確かめられている。

正準相関係数や正準変数の分布に関しては、対象とする変数が量的、質的であるに関わらず、これまでに多くの研究がなされている。しかし、次元検定問題のための検定量の分布に関する研究は、対象とする変数が質的変数である場合については、量的変数の場合と比べ、十分に行なわれているとはいえない。

以上の背景のもと、本研究では、質的変数に対する正準相関分析における次元検定問題のための検定量を提案する。また、数学的な検討と数値シミュレーションを行なうことにより、提案した検定量の性質を明らかにするとともに、この検定量が、従来用いられてきた検定量と比べ、理論的にも実験的にも優れていることを示す。

質的変数に対する正準相関分析は対応分析や2次元分割表に対する数量化理論第3類を数学的に一般化したものであることから、本研究で提案する検定量はこれらの手法における次元検定問題においても用いることができる。また、この検定量の導出法は、他の仮説検定問題にも応用することが可能である。例えば、質的変数に対する正準判別分析として定式化される数量化理論第2類における次元検定問題のための検定量も同様の考え方で導出することができる。

質的データの多変量解析では、アイテムやカテゴリーが観測変数となる。本研究では、質的変数に対する正準相関分析におけるアイテムやカテゴリーの冗長性の定義を与え、アイテムやカテゴリーの冗長性問題を仮説検定問題として定式化し、この仮説検定問題のための検定量を導出することも行なう。この場合の検定量も次元検定問題における検定量と同様の考え方で導出される。

次元検定問題では、正準相関係数や相関比の値、すなわち固有値の値の大きさによってその有意性を評価している。しかし、固有値の値が大きいからといって必ずしもそれに対応する次元が有意であるとは限らない。このような状況の例の一つとして、対応分析における馬蹄形問題があげられる。馬蹄形問題は、質的変数の中でも特に順序尺度を持つ変数に対する対応分析において、しばしば生じる。馬蹄形問題が生じる原因は、分析の対象とする2次元分割表の背後に多変量正規分布が存在することを仮定することにより、理論的に説明されている。本研究では、この問題についても考察し、馬蹄形問題を解決するための方法を提案し、提案方法が有用であることを示す。

以上で述べたように、本研究の目的は、質的データの次元縮小を目的とする多変量解析の手法における次元の有意性を、より正確に検証するための方法を提案するとともに、提案方法の性質と有用性を示すことである。

学位論文審査の要旨

主　査　　教　授　　佐　藤　義　治
副　査　　教　授　　宮　腰　政　明
副　査　　教　授　　工　藤　峰　一
副　査　　准教授　　今　井　英　幸

学位論文題名

Study on Significance Verification of the Dimensions in Multivariate Analysis for Categorical Data

(質的データの多変量解析における次元の有意性検証に関する研究)

質的データ解析は広くは離散データ解析に含まれるが,離散確率変数の統計理論は組み合わせの数 (離散確率変数の取り得るすべての組み合わせ) が問題となるため,実際には極めて困難である.分割表 (contingency table) の分析においては,フィシャー (R.A.Fisher) の精密確率を計算する方法が知られているが,現実に適用可能な分割数は高々2または3程度であり,一般的な分割表の場合には実用的な範囲を遥かに超えている.

現実問題においては,多くの場合連続分布による近似,特に中心極限定理による正規近似に基づく種々の統計量によって推定あるいは検定の問題が議論されている.

本論文で扱われている対応分析 (Correspondence Analysis) は種々の定式化が提案されているが,いずれも同値であることが示されている.従って,議論の対象に最も妥当な定式化を用いることになる.本論文の主題である次元検定の問題には正準相関分析の枠組みを用いている.ここで,次元の問題とは,データとして得られる分割表の行または列の類似性を矛盾なく表現するために必要なユークリッド空間の次元である.これを直感的に理解するためには,正準相関分析の枠組みよりも多次元尺度構成法の枠組みで考えた法が理解しやすい.すなわち,多次元尺度構成法の枠組みでは,分割表の相対度数 (確率) に関する行間のカイ二乗距離,列間のカイ二乗距離をその距離を何次元のユークリッド空間の距離関係で表現可能かという問題である.しかし,この直感的な次元を理論的に扱うことは困難であり,本論文では正準相関分析による定式化を用いている.

連続変量の正準相関分析において,二組の変量の線形結合間の正準相関係数はある種の分散共分散行列の固有値問題として定式化され,固有ベクトルとして最適な線形結合の係数が求まる.従って,ここでの次元問題は非零となる固有値の個数,すなわち共分散行列のランクを検定することと同値である.そのための統計量は正規性の仮定の下で漸近的にカイ二乗分布することが用いられている.

一方,離散データを用いた正準相関分析を行ったとき,連続変量と同様に共分散行列の固有値,固有ベクトルの問題として定式化ができ,次元問題を行列のランクとして捉えられることは同様であ

るが，問題はそのための検定統計量の分布を連続量として近似することに大きな乖離が生ずる．実際に著者は本論文において，シミュレーションを用いて従来の連続な場合の統計量をそのまま近似として利用することは検定精度の低下が実用的な範囲を大きく超えていることを示している．

著者はこの近似の悪さの原因を離散変量の分布が中心極限定理によって，多変量正規分布に近づいたとしても，分散共分散行列の標本分布がウィシャート (J.Wishart) 分布から大きく乖離することによって生ずることを実証した．

近年，情報技術の発展により，観測される標本数はますます膨大なものとなる傾向をもち，統計的データ解析においても，古典的な小標本理論は重要ではあるが，それに執着する必要はなく，十分に標本が得られる状況での議論も有用であるものと考えられる．本論分はその立場から，標本をいくつかのブロックに分割することによって，各ブロックに中心極限定理を適用し，各ブロックからの統計量を多変量正規分布からの標本と見なすことにより，分散共分散行列の固有値に関する検定統計量を導出し，その検定統計量が漸近的に従来の多変量正規分布により検定統計量に分布収束することを証明した．

対応分析を正準相関分析の立場から見た場合，分散共分散行列の固有値が正準相関係数の平方を表しており，その値が大きい程対応する固有ベクトルの布置がもつ情報が有効であるものと見なされる．しかし，対応分析における馬蹄形問題という現象が古くから知られており，固有値の大きさだけが必ずしも有効な次元を表現するものではないことが指摘されていた．馬蹄形問題の発生要因については様々な研究がなされているが，著者は分割表の背後に二変量正規分布が存在するものと仮定するならば馬蹄形問題生ずるという研究に注目し，馬蹄形問題の解消のため新しい方法を提案した．その基本的な考え方は，最大固有値に対応する固有ベクトルのエルミート多項式を 2 次,3 次，と以下順次必要なまで，データとして与えられる分割表から差し引いた残差分割表に対応分析を適用することによって，最大固有値の固有ベクトルの多項式成分,(2 次成分が馬蹄形となる) 除去できることを示した．

これを要するに，著者は離散データ解析における正準相関分析ともいうべき対応分析においてその有効次元数の検定統計量を提案するとともにその有用性を示し，されに対応分析における馬蹄形問題として古くから知られていた問題にたいしてその解決法を与えたものとして，統計科学におけるデータ解析学やコンピュータサイエンスにおけるデータマイニングの分野に貢献するところ大なるものがある．

よって著者は，北海道大学博士 (情報科学) の学位を授与される資格あるものと認める．