社会システム分析のための統合化プログラム８

― 正準相関分析・因子分析・ユーティリティ ―

福井正康・細川光浩

福山平成大学経営学部経営情報学科

概要

我々は教育での利用を主な目的に、社会システム分析に用いられる様々な手法を統合化したプログラムを作成してきたが、今回は多変量解析のうち、正準相関分析、因子分析をシステムに組み込んだ。また、授業の資料作成やサンプルデータの作成で利用されるユーティリティ機能についても説明する。

キーワード

社会システム分析，OR，統計，多変量解析，正準相関分析，因子分析，ソフトウェア，統合化プログラム

URL: http://www.heisei-u.ac.jp/mi/fukui/

１．はじめに

我々はこれまで主に教育を目的に、様々な分析手法をプログラム化してきたが[1-7]、今回は多変量解析の中で正準相関分析と因子分析を取り上げる。また、講義資料やレポートの作成に利用される数学関数の描画機能や授業に必要なサンプルデータを作成する機能等をユーティリティとしてまとめて紹介する。

正準相関分析は複数の変数を含む２群間の相関係数を決める手法である。それぞれの群の中で変数の線形結合による新しい変数を考え、その相関係数が最大になるように、線形結合のパラメータを決定する。線形結合のパラメータを調べることによって、２群間の関係の深さとそれに影響する変数の重要性等が読み取れる。

因子分析は主成分分析に類似している。主成分分析が変数の線形結合によって意味付け可能な新たな変数を作り出すのに比べ、因子分析は各変数に内在すると考えられる共通の因子の線形結合によってそれぞれの変数が作られるものと考え、その係数の値と因子についての意味付けを考える。主成分分析の解法が比較的容易で一意性があるのに比べて、因子分析の解法には様々な方法があり、あまり単純とは言えない。ここでは、古くから有名なセントロイド法と広く知られている主因子法の２つを取り上げている。その他にも部分的な計算法で細かい設定があるが、それについては章を改めて説明する。

ユーティリティとして取り上げるものは、メニュー［分析－数学・OR］の中の「関数グラフ」、［分析－基本統計］の中の「密度関数グラフ」、［ツール］の中の「データ発生」と「文字列結合」である。関数グラフはプリント等に簡単な関数グラフを挿入する場合利用するもので、密度関数グラフはこれを統計学の確率密度関数の描画に応用したものである。これらのグラフの特徴は、複数の設定のグラフを同時に描画できることである。例えばχ^2分布で自由度の異なる確率密度関数のグラフを比較表示したい場合などには便利である。

データ発生は列毎にセルをデータで埋める機能であるが、データの発生方法としては、同一データ、単調増加・減少、多項分布・正規分布・対数正規分布・一様分布の乱数発生が利用できる。これは統計学の演習用データ作りには欠かせない。最後に文字列結合は複数行のデータを文字列として結合し、新しいデータを作る機能である。これは３次元以上のデータの分割や複数の変数によって分割されたデータの統計量等を求める際に利用される。データの結合の際には、任意の文字列も加えることができる。

各分析には理論的にさらに深いレベルがあるが、それらを強化すれば機能が増え、不慣れな利用者には分かりにくくなる。機能と分かり易さのバランスを考えることは、利用者のレベルも考慮しなければならない課題である。

２．正準相関分析

正準相関分析は変数 $x_1, x_2, \cdots, x_r$ と変数 $y_1, y_2, \cdots, y_s$ を含む２群間の相関係数を、これらの変数を用いた 1 次関数間の相関係数と定義し、この相関係数が最大となるように係数を決める手法である。

まず、以下のような線形結合により、新しい変数 u，v を考える。

$$u = {}^t\mathbf{a}\mathbf{x},\quad v = {}^t\mathbf{b}\mathbf{y},$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{pmatrix},\quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix},\qquad ,$$

ここに、　，　は係数ベクトルである。

変数　　と変数　　の分散共分散行列をそれぞれ　，　とし、2 組の変数間の分散共分散行列を　（　　）とすると、　と　の相関係数　は以下となる。

但し係数ベクトルは　，　の分散が 1 になるように　，　と規格化している。

制約条件　　，　　を入れ、Lagrange の未定定数法を用いて　が最大となるように係数を求めると、以下の固有値問題に帰着する。

，　　，

ここに　は未定定数であるが、　に等しいことが上の計算過程から分かっており、最大の相関係数の 2 乗は最大の固有値に等しい。この固有値に対応する固有ベクトル　，　で決まる変量　，

を（第１）正準変量、その時の相関係数を（第１）正準相関係数という。これに倣ってα番めに大きい固有値に対応する固有ベクトルから同様に求まるものをそれぞれ第α正準変量、第α正準相関係数という。

個体（レコード）　について、変数　のデータを　，変数　のデータを　とするとこの個体の正準変量　，　は以下のように与えられる。

，

ここでは元のデータから分散共分散行列を用いて求める方法を示したが、変数の大きさ（ばらつき）に極端な差があるときは、各変数を標準化して相関行列から同様の計算を進める。

正準相関分析の実行画面を図 2.1 に示す。

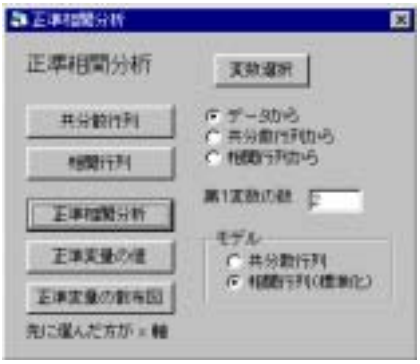

図 2.1 正準相関分析画面

図 2.2 正準相関分析出力画面

分析は、主成分分析等と同様、元データ、分散共分散行列、相関行列から実行できるが、正準変量の値と正準変量の散布図については、当然元データがないと求められない。計算のモデルは、データをそのまま利用する場合と、標準化して相関行列を用いて計算する場合のどちらかを選ぶようになっている。直感的に分り易いのはそのままの値を利用するものであるが、変数の大きさが相当違う場合や係数から重要性を読み取ろうとする場合には標準化した方がよい。図 2.2 は５つの変数を、２つと３つに分け、「正準相関分析」ボタンをクリックした実行結果である。この場合正準変量　に含まれる変数の数として 2 を指定する。結果は第１正準変量の値と第１正準相関係数の値を表示する。第２以降の正準変量や正準相関係数については今後どのように取り扱うか検討中である。

次に図 2.3 に「正準変量の値」ボタンをクリックした場合の実行結果を示す。各個体毎に正準変量の値を計算して表示している。ここでは標準化されたデータから計算を進めたので、結果は標準化された値となる。これらのデータから散布図を作ったものが、図 2.4 である。

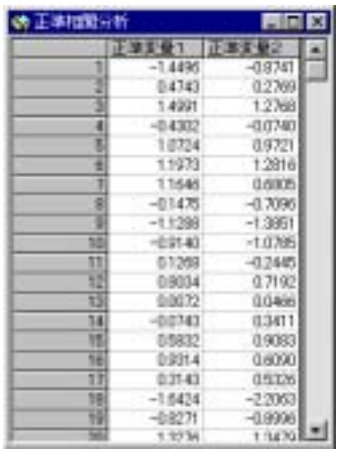

図 2.3 正準変量の値画面

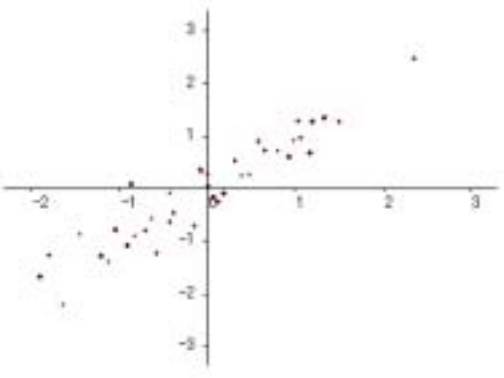

図 2.4 正準変量の散布図

第１正準変量を横軸に、第２正準変量を縦軸にとっているが、相当高い正準相関係数になることが見て取れる。

３．因子分析

因子分析が取り扱うデータは主成分分析等と同様に　変数、　個体（レコード）の変量（　　）である。これらのデータから各変数　に内在すると思われる因子を抽出することが因子分析のねらいである。

因子分析では変数　を標準化した変数　　を用いることが多いので、今後はこの変数　を用いて議論を進める。ここで　は変数　の標本平均、　は不偏分散から求めた標準偏差である。

因子分析では各データに内在すると考えられる共通因子　（　　）の線形結合によって、変数　のレコード　の変量　が以下のように表わされるものと考える。

係数　は　因子の因子負荷量と呼ばれている。ここで　は誤差であり、共通因子　との相関や互いの相関はないものと考える。

,　　（　　）

また共通因子　についても互いの相関はなく、平均 0、分散 1 に標準化されているものとする。

,

これらを利用すると変数　と　との相関係数　は以下のように表わせる。

（　　）,

ここで、　　と置くと、上式は以下のように表わされる。

,

,　　(1)

この中で特に　は共通性と呼ばれ、　　の関係を満たす。

共通性の和を取ると、

となるが、この関係式を利用し、誤差　　が 0 に近づけば左辺は　に近づくことを考えて、因子　の寄与率を以下のように定義する。

我々は (1) 式を解いて因子負荷量　　を求めようとするが、共通性　が定まらない限り一般には不可能である。そこで最初に適当な推定値　を用いて、因子負荷量　　を計算し、その値を使って再度　　で共通性　を計算し、それをまた推定値として再び因子負荷量を計算する。これを共通性　が収束するまで（このプログラムでは前回との差が 0.001 以下になるまで）繰り返すという方法で近似値を求める。その際最初の共通性　の推定値には変数　と他の変数の重相関係数や他との相関係数の中で最大のものなどが利用される。

さらに、共通性を仮定した後の因子負荷量　　の求め方にもセントロイド法、主因子法、最尤法、最小 2 乗法等種々の方法があるが [9]、ここでは歴史的に有名なセントロイド法と広く知られている主因子法を取り上げた。

セントロイド法は第 1 因子から逐次因子負荷量を求めていく手法で、

（　　）

の形で第 1 因子の因子負荷量を与える。次に　　として新たな相関行列を定義するが、その際対角要素は各行の非対角要素の絶対値の最大値を用い、負の相関係数をできるだけ少なくするために、参考文献 8) のアルゴリズムに従い座標反転を行なう。この相関行列を利用して新たに第 2 因子の因子負荷量を同様の方法で計算する。

さらに　　を用いて新たな相関行列を作り、上に述べた方法で対角要素と負の相関についての処理を行ない、次の因子の因子負荷量を計算して行く。

次に主因子法は寄与率を最大にするように因子負荷量　　を求める手法で、対角成分を共通性　で置き換えた相関行列　の固有値と固有ベクトルによって因子負荷量　　が計算される。即ち、第　因子の因子負荷量　　は、行列　の固有値　と規格化された固有ベクトル　　を使って、

のように与えられる。各因子負荷量ベクトル　　は直交し、寄与率は以下であたえられる。

次に各因子、各個体毎の因子得点　　の値について考える。前にも述べたとおり、誤差項が特定できない限り、一般に観測値　　から因子得点　　を決定することはできない。そこで我々は分散で重み付けされた誤差の 2 乗項

が最小になるように仮定して、因子得点　　を推定する。この解は成分が

　　,　　,　　,　　,

のように与えられる行列　,　,　,　を用いて以下のように求められる。

この推定法は Bartlett の重みつき最小 2 乗推定法と呼ばれる。この他にも回帰推定法と呼ばれるものがあるが [9]、ここでは省略する。

因子分析の実際の実行画面を図 3.1 に示す。データとしては主成分分析と同じように個体毎の元データ、共分散行列、相関行列が選択できる。因子負荷量を求める方法ではセントロイド法と主因子法が利用できる。いずれも共通性の推定の不完全さを補うために、共通性の値が一定値に近づくまで、近似計算を繰り返す。

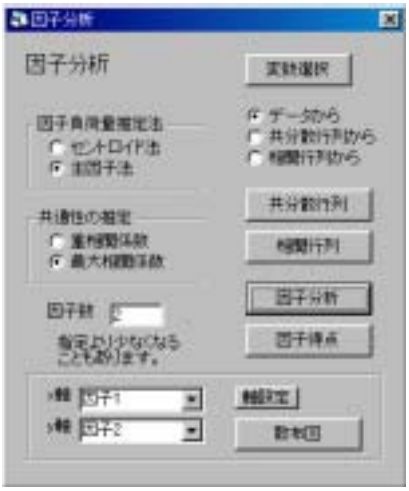

図 3.1 因子分析画面

図 3.2 因子分析出力画面

図 3.2 に「因子分析」のボタンをクリックした場合の出力画面を示す。因子数で指定した数だけ因子負荷量と寄与率、累積寄与率が表示されている。但し、因子数を指定しない場合は、セントロイド法で累積寄与率が 0.9 を超えたところで、主因子法では固有ベクトルの値が 0.5 未満になったところで因子の出力を停止する。また、因子数を指定した場合でも、主因子法で固有値が 0 に近い負の値を取ることも見つかっており、指定した個数より少なく表示される場合もある。こ

の原因は現在考察中である。

「因子得点」ボタンをクリックすると図 3.3 のように個体毎の因子得点が表示される。ここでは因子得点の推定に、Bartlett の重みつき最小 2 乗推定法を用いている。「散布図」ボタンをクリックすると図 3.4 のように因子得点 1 を横軸に因子得点 2 を縦軸にした散布図を作成する。

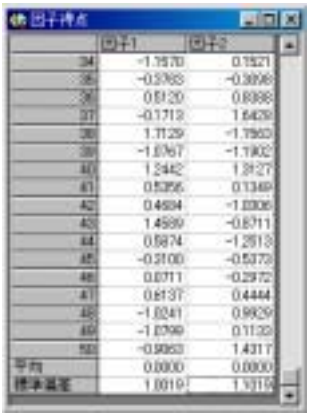

図 3.3 因子得点出力画面

図 3.4 因子得点散布図

4．ユーティリティ

ここでは、主要な分析として作られてはいないが、資料の作成や授業のためのデータ作成に役立つ機能をユーティリティとして紹介する。

最初は 1 変数関数のグラフを表示するユーティリティである。メニュー［分析－数学・OR－関数グラフ］を選択すると図 4.1 に示す実行画面が表示される。

図 4.1 関数グラフ画面

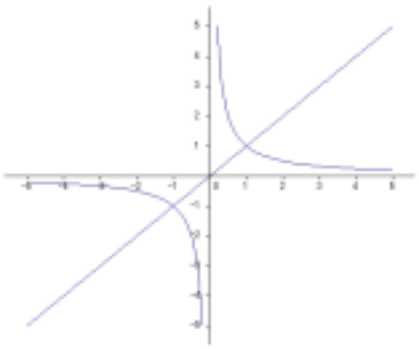

図 4.2 関数グラフ描画例

この中で数式を入力し、最低限ｘ軸の下限と上限、目盛間隔を入力して「グラフ描画」ボタンをクリックすると簡単にそのグラフが描かれる。通常はグラフのｘ軸の下限が目盛の下限であるが、

これを変えたいとき目盛下限のテキストボックスに値を入力する。ｙ軸の値を変えたいときは同様にしてｙ軸の下限と上限、目盛間隔に値を入力する。数式は基本的に Basic の書式で変数名を x として入力する。

グラフの滑らかさを決定するのが区間分割数である。これはグラフの変域をいくつに分割して計算するかを決めるもので、標準で 30 に設定しているが、グラフの変動が大きいような場合は大きくして調節する。この関数グラフ描画機能には複数の関数を同時に表示させる機能がある。最初にあるグラフを新規のオプションボタンを選んで表示し、その後別のグラフを追加のオプションボタンを選んで表示すると２つのグラフが同時に表示される。実際に図 4.2 ではこの機能を用いて　　と　　の２つのグラフを表示している。

次に関数グラフ描画機能を統計学の授業等で利用するために、正規分布に関係した基本的な分布の確率密度関数のグラフを描く機能を追加した。メニュー［分析－基本統計－密度関数グラフ］を選択すると図 4.3 の実行画面が表示される。利用法は関数グラフを描画する場合とほぼ同じであり、統計学のプリント等を作成する場合には手軽に利用できる。図 4.4 は自由度を１から５まで変えた χ^2 分布の確率密度関数を描いたものである。

図 4.3 密度関数グラフ画面

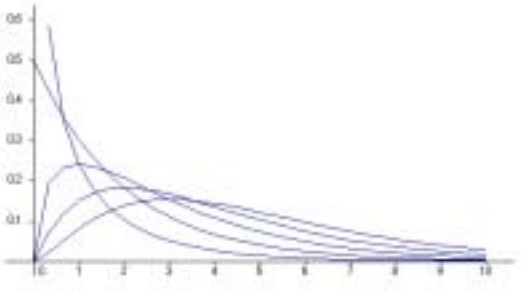

図 4.4 自由度を変えた χ^2 分布の確率密度関数

次にデータ作成によく利用するユーティリティを紹介する。統計の授業において演習は重要であり、そのためには簡単にデータ作りのできるツールは必要不可欠である。そこで我々は図 4.5 に示すようなデータ発生用のユーティリティを作成した。元々基本統計の中に乱数発生の機能を付けていたが、同一値のデータや、単調増加・単調減少データを付け加えて、ユーティリティとしてメニュー［ツール－データ発生］で使えるように訂正した。

発生するデータは、同一の値を示すもの（数字または文字列）、単調に増加・減少するもの、多項分布するもの、正規分布または対数正規分布するもの、一様分布するものが選択できる。出

力列を選択して、左上の開始行、個数、Seed、小数点桁数を入力し、乱数を選択して出力ボタンをクリックする。但し、開始行と個数はデフォルトでそれぞれ 1 と個体（レコード）数に設定されている。乱数の種類は今後必要に応じて付け加えて行く。

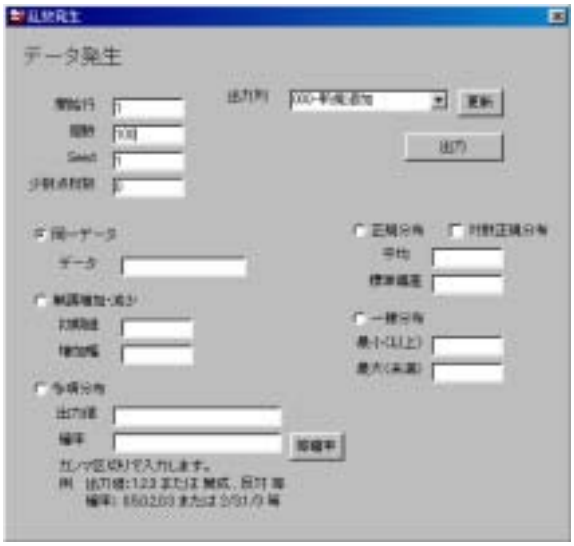

図 4.5 データ発生画面

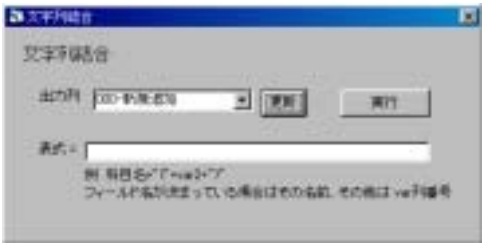

図 4.6 文字列結合画面

最後にいくつかの変数のデータを個体（レコード）毎に文字列として結合させて新しいデータを作成する文字列結合機能について説明する。メニュー［ツールー文字列結合］を選択すると、図 4.6 に示される文字列結合実行画面が表示される。これは例えば 2 次元分割表より次元の高い分割表を考えようとする場合や複数列の変数で群分けしたデータの基本統計量等を求める際に利用される。また自由に文字列を付け加えることができれば利用範囲も広がると考え、任意の文字列を加えて自由にレイアウトできるようにした。これを用いると「科目名 +”(“ +var2+”)”」は「科目名」変数のデータと 2 列目の変数（var2）のデータを結合し、例えば“基礎数学(金 3)” のような新たな文字列データを作成する。

5．おわりに

今回の論文では前回やり残した多変量解析の中の正準相関分析と因子分析を中心に、このソフトウェアを授業で利用する際に必要となった関数グラフ描画やデータ発生のユーティリティについて説明した。以下それぞれの分析について今後の改良点などをまとめておく。

正準相関分析のプログラムにはまだ課題が多い。この中で我々は単に第 1 正準相関係数と第 1 正準変量とを求めただけであり、第 2 正準相関係数以降については何も触れていない。また正準相関分析自体の妥当性の検定やいくつまで正準相関係数を求めることができるかという検定につ

いても同様である。今後授業で活用する中で必要に応じてこれらの機能を加え、より完全なものに近づけたい。

因子分析は当初主成分分析で代用できると考え、システムに組み込む予定ではなかった。しかし、様々な分野での利用を考えるとやはり避けて通れないものと考え、必要最低限な部分についてプログラムを作成した。そのためまだ不充分な点が多い。例えば因子負荷量の推定法について、このプログラムには歴史的なセントロイド法と比較的よく利用される主因子法が入っているが、最尤法や最小 2 乗法といった手法は含まれていない。また、因子の解釈を容易にするためのバリマックス回転等の機能もない。因子得点を求める手法でも Bartlett の重みつき最小 2 乗推定法と呼ばれる方法だけで回帰推定法と呼ばれる方法には触れていない。このように不備な点も多く、因子分析を専門に取り扱っておられる方には不満足なものであろう。

数学を得意としない学生に対して授業を行ない、多くの手法の中からどれかを選択させる場合、教え方に困る場合がある。違いを数式で表現するのは簡単であるが、言葉だけで表現するには不可能と思われる場合さえある。またいくつかの選択肢の説明に時間を掛け過ぎると全体が忘れ去られがちになる。因子分析は近似解法であるため様々な選択肢があり、特に難しい。共通性の推定方法でも選択肢としてではなく、例えばセントロイド法では最大相関係数、主因子法では重相関係数というように自動で固定化させようかとも考える。この当り読者の方はどう考えられるのであろうか、今後の大きな課題である。

さて、我々がこれまで追加してきた機能は実際に授業を行う際必要になったものが多い。例えば、表をコピーしたものを貼り付ける際、表の行数や列数が不足する場合がある。以前は表の行数や列数を増やしてから貼り付けるようにしていたが、学生の反応を見るとそこで思考の中断が起こる。これを防ぐためにはどうしても自動的に範囲を拡張する機能も必要になる。また、例えば 1 次元分割表についても、このプログラムでは見易さを重視して縦に表示するが、学生にレポートを作らせると縦の表はどうしても間延びしてしまうので、表示の縦横を交換するような機能も必要になる。さらにこの論文で説明したユーティリティ機能は教員側がこのソフトウェアを現場で利用する際、必要と感じたものであった。このようにソフトウェアの粗は使ってみて初めて認識できる。その意味ではまだ現場で試していない分析機能もあり、今後様々な場面での活用が必要である。

このソフトウェアの中で統計学については、部分的な変更は必要なものの、今回で一応の完結を見た。授業でも 1 年間利用してきて、学生が分かりにくいところなどに細かな訂正を加えてきた。今後は OR や意思決定の手法を充実させて行きたい。

参考文献

1) 福井正康・田口賢士, 社会システム分析のための統合化プログラム, 福山平成大学経営情報研究, 3 号, 109-127, 1998.

2) 福井正康・田口賢士，社会システム分析のための統合化プログラム２ －産業連関分析・KSIM・AHP－，福山平成大学経営情報研究, 3 号, 129-144, 1998.
3) 福井正康・増川純一，社会システム分析のための統合化プログラム３ －線形計画法・待ち行列シミュレーション－，福山平成大学経営情報研究, 4 号, 99-115, 1999.
4) 福井正康，社会システム分析のための統合化プログラム４ －基本統計－，福山平成大学経営情報研究, 5 号, 89-100, 2000.
5) 福井正康，社会システム分析のための統合化プログラム５ －システムの改良・ISM－，福山平成大学経営情報研究, 6 号, 91-104, 2001.
6) 福井正康，細川光浩，社会システム分析のための統合化プログラム６ －DEA・実験計画法・クラスター分析－，福山平成大学経営情報研究, 7 号, 65-83, 2002.
7) 福井正康，細川光浩，社会システム分析のための統合化プログラム７ －多変量解析－，福山平成大学経営情報学研究, 7 号, 85-106, 2002.
8) 河口至商，多変量解析入門Ⅰ，森北出版, 1973.
9) 田中豊・垂水共之編, Windows 版 統計解析ハンドブック 多変量解析，共立出版社, 1995.
10) 田中豊・脇本和昌，多変量統計解析法，現代数学社, 1983.

Multi-purpose Program for Social System Analysis 8

- Canonical Correlation Analysis, Factor Analysis, Utilities -

Masayasu FUKUI and Mitsuhiro HOSOKAWA

Department of Management Information, Faculty of Management,
Fukuyama Heisei University

Abstract

We have been constructing unified software on social system analysis for the purpose of education. Now we added the canonical correlation analysis and the factor analysis, which are in the group of the multivariate analysis, to our system. We also explain some utility programs which are used for writing reports and making sample data.

Keywords

social system analysis, statistics, multivariate analysis, canonical correlation analysis, factor analysis, software, unified program

URL: http://www.heisei-u.ac.jp/mi/fukui/