

多変量解析の学習における 3 次元可視化ツールの 活用の効果についての一考察

岸本 憲幸†

株式会社オープンストリーム†

1. はじめに

昨今の AI の進展に伴い、その礎を担う多変量解析等の統計学を学ぶ機会が増えている。数学あるいは統計学を学ぶ上で理論の厳密な理解は勿論重要であるが、まずは直感的に事象を捉えることが根底理解の大きな助けになることは間違いない。多変量解析は多次元空間を扱うが、人間の直感的理解が及ぶのはせいぜい 3 次元までであろう。したがって 4 次元以上の理解は 2 次元、3 次元でのイメージに倣うといった方法で学習を行うことになる。多くの場合 2 次元平面上での直感的理解で全容を把握するには事足りるが、中には 2 次元では表現が難しく、より高次元での直感的理解が全容把握の助けとなる箇所もあると思われる。本稿では 3 次元可視化ツールを用いてそのような例を 2 例挙げ、多変量解析の学習の際により深い理解が得られることを考察する。

2. 3 次元可視化例

本稿ではリニアングラフ 3D というフリーの 3 次元グラフソフトを使用した。以下に 3 次元可視化の例を 2 例紹介する。

1 例目は 2 変数を主成分分析により 1 つの合成変数にまとめる様子を 3 次元にて可視化する。2 つの変数を x_1, x_2 、これらを標準化したものを u_1, u_2 、合成変数（第一主成分）を z とし、

$$z = a_1 u_1 + a_2 u_2 \quad (1)$$

とおき、 z がもとのデータの情報をできるだけ多く有するように a_1, a_2 を決定する。以下の式を満たすことによりそれが実現できる。

$$\begin{bmatrix} 1 & r_{x_1 x_2} \\ r_{x_1 x_2} & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \lambda \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad (2)$$

ここで $r_{x_1 x_2}$ は、 x_1 と x_2 の相関係数を表す。

(2) 式より導いた a_1, a_2 を用いて u_1, u_2, z を表したグラフが図 1 である。赤色の平面が z を表している。青色の平面は u_1, u_2 の集合を表している。

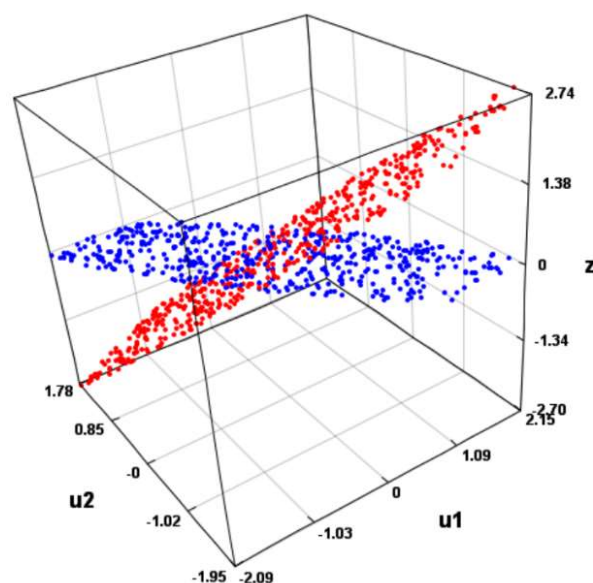
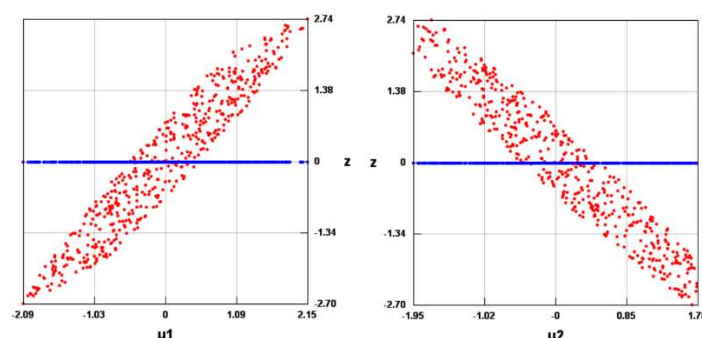


図 1 主成分分析

図 2 $u_1 - z$ (左), $u_2 - z$ (右) 投影図

A study of the effect of utilization of 3D visualization tools in learning multivariate analysis

† Noriyuki Kishimoto
Open Stream, Inc.

青色平面の集合が(1)式により赤色平面上の z の値に合成されたことが分かる. また, 図2は図1をそれぞれ $u_1 - z$ 方向, $u_2 - z$ 方向に投影したグラフである. それぞれ右肩上がり, 右肩下がりのグラフで, u_1 と z , u_2 と z の相関が高いことが見てとれる. つまり, z が元の u_1 , u_2 の情報をできるだけ多く有していることがわかる.

2例目は重回帰分析における共線性が存在する様子を説明変数2個で表現する. 説明変数2個なので, 推定式は以下のようにおく.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \quad (3)$$

このときの回帰係数予測値は以下の式で導き出せる.

$$\hat{\beta}_1 S_{11} + \hat{\beta}_2 S_{12} = S_{1y} \quad (4)$$

$$\hat{\beta}_1 S_{12} + \hat{\beta}_2 S_{22} = S_{2y} \quad (5)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 \quad (6)$$

ここで

$$S_{11} = \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 \quad S_{22} = \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2$$

$$S_{12} = \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

$$S_{1y} = \sum_{i=1}^n (x_{i1} - \bar{x}_1)(y_i - \bar{y})$$

$$S_{2y} = \sum_{i=1}^n (x_{i2} - \bar{x}_2)(y_i - \bar{y})$$

である.

ここで2つの説明変数 x_{i1} , x_{i2} に $x_{i2} = ax_{i1} + b$ の関係を持たせる. つまり x_{i1} と x_{i2} の相関係数=1である. これでこのモデルには共線性が存在する. さらに $y_i = cx_{i1} + dx_{i2} + e$ の関係を持たせ, 回帰係数が解不定となる状態とする. この3次元グラフを図3に示す.

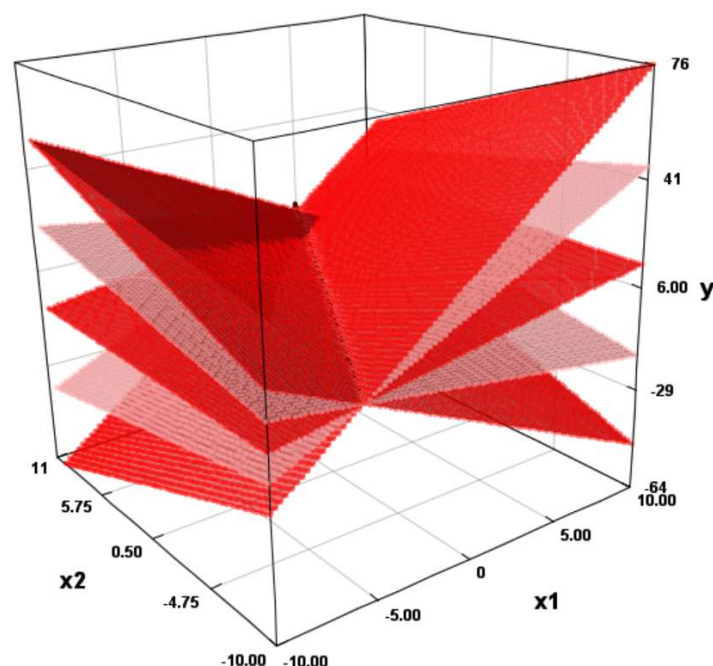


図3 共線性 (解不定)

複数の平面が1つの直線上で交わっている. この直線が x_{i1} , x_{i2} に対応する y_i の実測値の集合である. この直線にフィットする平面は図3のように無数に存在するので, 回帰係数の予測値 $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ を一意に定めることができず解不定となることが見てとれる.

3. おわりに

冒頭でも述べたが, 多変量解析を学ぶ上で理論を緻密に追って理解することは重要であるが, まず取り掛かりとして直感的な絵的なイメージを持つことができれば, その後の理解に大いに助けになる. 本稿では多変量解析のなかでも比較的初歩の内容にて事例を挙げたが, もう少し高度な内容でこのような事例を挙げることができれば, 自身の学習あるいは他者への教育においても, 成果を上げられるのではないかと考える.

参考文献

永田 靖・棟近雅彦「多変量解析法入門」