

多変量解析におけるパーミュテーションテストについて

杉山 高一 (中央大学理工)

山田 智哉 (中央大学理工)

1 はじめに

パーミュテーションテストは、以前からよく知られている方法である。Fisher(1935)により導入され、Pitman(1937,1938)によって詳細に展開されている。検定を行う際に、統計量の分布が問題になるが、パーミュテーションテストではデータに対して並び替えを行い、並べ替えられたデータに対して逐次統計量を計算してヒストグラムを与えることにより分布を求めている。多変量解析に適用する際の問題点は、かなりの計算時間がかかることである。この点は計算機のスピードが飛躍的に改良されたことにより解決された。その他にも問題点はあるが、それは本論の中で触れることにする。

ここでは多変量解析におけるパーミュテーションテストの特性を把握するために、主成分分析の個々の固有ベクトルの検定および固有値の検定について、また正準相関係数に関する検定について述べることにする。

2 固有ベクトルの同等性におけるパーミュテーションテスト

$X^{(i)}$ ($i = 1, 2$) を平均ベクトル、分散共分散行列を持つ p 変量ベクトルとする。また $\Sigma^{(i)}$ の固有値を $\lambda_1^{(i)} \geq \dots \geq \lambda_p^{(i)}$ 、固有ベクトルを $\beta_1^{(i)}, \dots, \beta_p^{(i)}$ とする。この時、帰無仮説として

$$H_0 : \beta_i^{(1)} = \beta_i^{(2)}.$$

を考える。

$x_1^{(i)}, \dots, x_{N_i}^{(i)}$ を i 番目の母集団 Π_i から得られた標本とする。ここで平均ベクトル $\mu^{(i)}$ 、分散共分散行列 $\Sigma^{(i)}$ は未知とする。標本平均を $\bar{X}^{(i)}$ 、標本分散共分散行列を $S^{(i)}$ 、また $S^{(i)}$ の固有値を $\ell_1^{(i)} \geq \dots \geq \ell_p^{(i)}$ 、固有ベクトルを $b_1^{(i)}, \dots, b_p^{(i)}$ とする。特に帰無仮説の下では

$$\beta_i^{(1)'} \beta_i^{(2)} = 1$$

となることから検定統計量として次の統計量を考える：

$$T_1 = b_j^{(1)'} b_j^{(2)}.$$

この T_1 は帰無仮説の下では 1 に近い値をあることになる。また

$$T_2 = nk_1 k_2 (b_i^{(1)} - b_i^{(2)})' (k_2 \Psi_1 + k_1 \Psi_2)^{-1} (b_i^{(1)} - b_i^{(2)})$$

なども考えられる。ここで $n_i = N_i - 1, n = n_1 + n_2, k_i = n_i/n$ 、

$$\Psi_g = \sum_{j \neq i} \frac{\ell_i^{(g)} \ell_j^{(g)}}{(\ell_j^{(g)} - \ell_i^{(g)})^2} \mathbf{b}_j^{(g)} \mathbf{b}_j^{(g)'}$$

とする。この T_2 については塚田などにより提案されているものである。

3 固有値の同等性におけるパーミュテーションテスト

次に固有値に関する仮説検定を考える。帰無仮説としては

$$H_0 : \lambda_j^{(1)} = \lambda_j^{(2)}$$

を考える。帰無仮説の下では

$$\text{Var} [\boldsymbol{\beta}_j^{(1)'} \mathbf{X}_1] = \text{Var} [\boldsymbol{\beta}_j^{(2)'} \mathbf{X}_2]$$

が成り立つ。そこで $\boldsymbol{\beta}_j^{(1)'} \mathbf{X}_1$ と $\boldsymbol{\beta}_j^{(2)'} \mathbf{X}_2$ の分散の同等性について考える。

我々は $\boldsymbol{\beta}_j^{(i)}$ のかわりに $\mathbf{b}_j^{(i)}$ を用いてパーミュテーションテストを適用する。今

$$y_{jk}^{(i)} = \mathbf{b}_j^{(i)'} (\mathbf{x}_k^{(i)} - \bar{\mathbf{x}}^{(i)}) \quad k = 1, 2, \dots, N_i, \quad i = 1, 2,$$

とする。また j 番目の固有値に対応する第 j 主成分は $(y_{j1}^{(i)}, y_{j2}^{(i)}, \dots, y_{jN_i}^{(i)})$ と表せる。一般に $y_{jk}^{(i)}$ と $y_{jk'}^{(i)}$ は独立ではないということが知られている。しかし $y_{jk}^{(i)}$ と $y_{jk'}^{(i)}$ の相関は母集団分布が正規分布のときには $O(N_i^{-2})$ となることから。サンプルサイズが大きくなれば $\mathbf{T}(y_{j1}^{(1)}, y_{j2}^{(1)}, \dots, y_{jN_1}^{(1)})$ と $(y_{j1}^{(2)}, y_{j2}^{(2)}, \dots, y_{jN_2}^{(2)})$ との間で並び替えを行ってパーミュテーションテストを考えることができる。用いる検定統計量としては

$$T = \text{Var} [y_j^{(1)}] - \text{Var} [y_j^{(2)}].$$

を考える。

4 正準相関係数におけるパーミュテーションテスト

正準相関係数におけるパーミュテーションテストの適用を考える。 $X(p \times 1), Y(q \times 1)$ を母分散 Σ をもつ確率変数とする。ここで $p \leq q$ とし

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$$

とする。この時、正準相関係数は、次の固有方程式により求まる：

$$|\Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} - \rho^2 \Sigma_{xx}| = 0$$

正準相関分析は 2 変量間の関連性を与えることや、正準変量同士が無相関になるようとられていることなどから、重回帰分析、主成分分析の応用として位置付けられている。また標本空間の場合には Σ の代わりに標本分散共分散行列 S を用いて

$$\left| S_{xy} S_{yy}^{-1} S_{yx} - \hat{\rho}^2 S_{xx} \right| = 0$$

を計算すればよい。さてここで興味のある仮説検定として第 1 正準相関係数 ρ_1 がある値であるという検定、つまり

$$H_{10}; \quad \rho_1 = \rho_{10} \quad (\rho_{10} \neq 0 : \text{given})$$

を考える。この仮説検定に対してノンパラメトリック検定の一つであるパーミュテーションテストを試みる。

正準相関係数に対する検定も帰無仮説が正準相関スコア同士が無相関であるという検定つまり正準相関係数が 0 であるという仮説検定に対しては、並べ替えた正準相関スコアも無相関になることから有用である。しかし我々の考えている仮説は、正準相関係数がある値 $\rho_{10} \neq 0$ であるという問題であるから、このままパーミュテーションテストを行うことはできない。今回我々が提案する手法は得られたデータに対して無相関になるよう変数変換を行い、その後にパーミュテーションテストを行うという手法である。正準相関分析により得られた X の第 1 正準相関スコアを u_1 、 Y の第 1 正準相関スコアを u_2 とすると、変数変換として

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = A^{-1} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \quad \text{where} \quad AA' = \begin{pmatrix} 1 & \rho_{10} \\ \rho_{10} & 1 \end{pmatrix}$$

を考える。このとき新しいスコア v_1, v_2 は帰無仮説の下では無相関となることから、これらに対して並べ替えを行い、統計量を用いて検定を行えばよい。ここでは、統計量として Fisher の z 変換

$$z = \frac{\sqrt{N-3}}{2} \log \frac{1+r}{1-r}$$

を用いて検定を行う。以上の事柄をまとめて、 H_{10} についての仮説検定を考える場合には、以下のアルゴリズムを用いて検定を行う。

1. 正準相関分析を行い、正準相関係数、重みベクトルを与え、それを用いて X, Y の第一正準相関スコア u_1, u_2 を計算する。
2. 変数変換として $v = A^{-1}u$ を行い、新しいスコアを v_1, v_2 とする。
3. v_1, v_2 を用いて相関係数 r を計算し、Fisher の z 変換を行う。
4. v_1, v_2 を並び替えることにより得られたデータ v_1^*, v_2^* を用いて相関係数 r^* を計算し、Fisher の z 変換を行う。
5. 4 を繰り返し計算することにより、 z に関する経験分布関数を求める。
6. 3 で得られた z が 4 で与えられた経験分布関数に従っているとして検定を行う。

本発表では、第 1 正準相関係数に関する検定論を中心に行ったが、正準相関係数に関するいくつかの仮説検定は同様の手法で行うことが可能である。また同様の検定論は相関係数の検定にも有効である。

5 数値結果

多変量正規分布におけるシミュレーションによりこの方法の信頼性について吟味する。上の結果は 10000 回の繰り返し数で 2000 回の並び替えを行ったものである。非正規分布における数値解析は現在行っている最中である。

5.1 固有値に関するパーミュテーションテストの数値解析

T_1 における数値結果は以下ようになる。

Table 1. The simulation results for p-values.

$$\Sigma_1 = \Sigma_2 = \text{diag}(7, 2, 1)$$

	$N_1 = N_2 = 50$					$N_1 = N_2 = 100$				
H_0	.050	.100	.500	.900	.950	.050	.100	.500	.900	.950
$\beta_1^{(1)} = \beta_1^{(2)}$	.0510	.0965	.5090	.9025	.9500	.0495	.0990	.5065	.9000	.9500
$\beta_2^{(1)} = \beta_2^{(2)}$	.0500	.1040	.5030	.8985	.9490	.0515	.1005	.4975	.8970	.9465
$\beta_3^{(1)} = \beta_3^{(2)}$	.0500	.0990	.4980	.9010	.9520	.0505	.1000	.5085	.9035	.9512

	$N_1 = N_2 = 200$				
H_0	.050	.100	.500	.900	.950
$\beta_1^{(1)} = \beta_1^{(2)}$	.0510	.0965	.5090	.9025	.9500
$\beta_2^{(1)} = \beta_2^{(2)}$	.0500	.1040	.5030	.8985	.9490
$\beta_3^{(1)} = \beta_3^{(2)}$	.0500	.0990	.4980	.9010	.9520

5.2 固有値に関するパーミュテーションテストの数値解析

多変量正規分布におけるシミュレーションによりこの方法の信頼性について吟味する。ここでは 5 次元正規分布を考える。一様性については次の様になる。

Table 1. The simulation results for p-values.

$$\Sigma_1 = \Sigma_2 = \text{diag}(40, 10, 5, 2, 1)$$

	$N_1 = N_2 = 100$					$N_1 = N_2 = 200$				
H_0	.050	.100	.500	.900	.950	.050	.100	.500	.900	.950
$\lambda_1^{(1)} = \lambda_2^{(1)}$	.0558	.1078	.5092	.9026	.9518	.0512	.1022	.5004	.9010	.9490
$\lambda_1^{(2)} = \lambda_2^{(2)}$	.0506	.1000	.4896	.9000	.9473	.0534	.1034	.5004	.9010	.9490
$\lambda_1^{(3)} = \lambda_2^{(3)}$	.0534	.0996	.5032	.9060	.9523	.0514	.1018	.5078	.9014	.9500