

単調欠測データが一様共分散構造を持つ場合の 平行性仮説検定と水準差の信頼区間

東京理科大学大学院 理学研究科	佐伯 悠一郎
東京理科大学 理学部	八木 文香
東京理科大学 理学部	瀬尾 隆
防衛大学校 数学教育室	百武 弘登

1 はじめに

本研究ではプロフィール分析における平行性仮説検定とその水準差について考える。プロフィール分析とは例えば繰り返し測定データなどにおいて複数の平均成分の推移の比較を行うものである (Srivastava (2002) など参照)。ここでは単調欠測データの下で $\mu^{(1)}, \mu^{(2)}, \dots, \mu^{(g)}$ を g 個の母集団の平均ベクトルとして、次の平行性仮説検定問題を考える。

$$H: \mu^{(\ell)} - \mu^{(g)} = \gamma^{(\ell)} \mathbf{1}_p, \ell = 1, 2, \dots, g-1 \text{ vs. } A: \text{not } H,$$

ただし、 $\gamma^{(\ell)}$ は水準差と呼ばれる未知パラメータであり、 $\mathbf{1}_p = (1, 1, \dots, 1)'$ とする。平行性仮説検定は Srivastava (1987, 2002) や Rencher and Christensen (2012) など完全データの下での議論が行われている。2 群の場合についてはホテリングの T^2 型検定統計量が与えられ、正確な分布が得られている。多群の場合については尤度比検定統計量と漸近分布を用いた近似パーセント点が与えられている。また欠測データの下での議論については Onozawa, Takahashi and Seo (2013) で 2-step 単調欠測データを扱っている。その中では 2 群におけるホテリングの T^2 型検定統計量を用いた検定と、多群に対する尤度比検定統計量を与えており、どちらも検定統計量の漸近分布と近似上側パーセント点を提案している。一方、Sawatsuhashi and Hyakutake (2016) では観測ベクトルにランダム効果モデルと一様共分散構造を仮定した場合の 2 群における、一般の単調欠測データの下でのプロフィール分析が議論され、検定統計量とその正確な分布が得られている。ここで一様共分散構造とは繰り返し測定データに対して仮定される分散共分散行列の構造の一つであり、Srivastava (2002) 等で紹介されている。本報告では Sawatsuhashi and Hyakutake (2016) を基に、分散共分散行列に一様構造を持つ一般の単調欠測データに対して、多群の場合の平行性仮説検定における尤度比検定統計量を与え、その正確な分布を導出する。さらに、平行性が認められた場合の水準差の信頼区間について、簡単なケースである 2 群かつ 2-step 単調欠測データに対し近似信頼区間の提案を行う。最後に、水準差の信頼区間についてモンテカルロ・シミュレーションにより、完全データ部分のみを使用した信頼区間との被覆確率の比較を行う。

2 平行性仮説検定

k -step 単調欠測データとして、 $\mathbf{x}_{ij}^{(\ell)}$ ($i = 1, 2, \dots, k, j = 1, 2, \dots, n_i^{(\ell)}$) を第 ℓ 群の第 i ステップ内 j 番目にある観測ベクトルとし、互いに独立に $N_{p_i}(\mu_i^{(\ell)}, \Sigma_i)$ に従うとする。ただし、 $p = p_1 > p_2 > \dots > p_k$ である。また $\mu_i^{(\ell)}$ は $\mu^{(\ell)}$ の第 1 から第 p_i 成分で構成される $(p_i \times 1)$ ベクトルであり、 Σ_i は Σ の p_i 次主座小行列とする。 k -step 単調欠測データの観測行列は次の通りである。

となる．このことから，一般の k -step の場合の尤度関数を与えることができるので

$$H : \boldsymbol{\xi}^{(1)} = \boldsymbol{\xi}^{(2)} = \dots = \boldsymbol{\xi}^{(g)} \text{ vs. } A : \text{not } H$$

の下での尤度比は

$$\lambda_k = \left(1 + \frac{\sum_{m=1}^k \sum_{\ell=1}^g M_m^{(\ell)} (\bar{\mathbf{z}}_m - \bar{\mathbf{z}}_m^{(\ell)})' (\bar{\mathbf{z}}_m - \bar{\mathbf{z}}_m^{(\ell)})}{\sum_{m=1}^k \sum_{\ell=1}^g \sum_{s=1}^{M_m^{(\ell)}} (\mathbf{z}_{ms}^{(\ell)} - \bar{\mathbf{z}}_m^{(\ell)})' (\mathbf{z}_{ms}^{(\ell)} - \bar{\mathbf{z}}_m^{(\ell)})} \right)^{-\frac{1}{2}\nu_k}$$

となる．ただし， $M_m = \sum_{\ell=1}^g M_m^{(\ell)}$ であり

$$\bar{\mathbf{z}}_m^{(\ell)} = \frac{1}{M_m^{(\ell)}} \sum_{s=1}^{M_m^{(\ell)}} \mathbf{z}_{ms}^{(\ell)}, \quad \bar{\mathbf{z}}_m = \frac{1}{M_m} \sum_{\ell=1}^g \sum_{s=1}^{M_m^{(\ell)}} \mathbf{z}_{ms}^{(\ell)}$$

である．また， $\nu_k = M_1(q_1 - 1) + M_2q_2 + \dots + M_kq_k$ である．さらにこの尤度比を用いた検定統計量 T_k の分布について

$$T_k = \frac{\nu_k - g(q-1)}{(g-1)(q-1)} (\lambda_k^{-\frac{2}{\nu_k}} - 1) \sim F_{(g-1)(q-1), \nu_k - g(q-1)}$$

であることが示される．ただし， $q = \sum_{m=1}^k q_m$ である．補足として，いくつかのパラメータ設定でのモンテカルロ・シミュレーションを行ったところ，数値的にも F 分布に一致することを確認している．よって， $\{\nu_k - g(q-1)\}\{(g-1)(q-1)\}^{-1}(\lambda_k^{-2/\nu_k} - 1) \geq F_{(g-1)(q-1), \nu_k - g(q-1)}(\alpha)$ ならば平行性仮説 H を棄却する．ここで， $F_{(g-1)(q-1), \nu_k - g(q-1)}(\alpha)$ は自由度 $(g-1)(q-1), \nu_k - g(q-1)$ の F 分布の上側 $100\alpha\%$ 点である．

3 水準差の信頼区間

平行性が認められた場合，次に関心があるのは水準差 γ の信頼区間である．本報告では，扱う水準差のパラメータが一つである 2 標本問題に限定し，2-step 単調欠測データを対象とする．観測ベクトル $\mathbf{x}_{ij}^{(\ell)}$ ($i = 1, 2, j = 1, 2, \dots, n_i^{(\ell)}, \ell = 1, 2$) は $N_{p_i}(\boldsymbol{\mu}_i^{(\ell)}, \boldsymbol{\Sigma}_i)$ に従うとし，平行性 $\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)} = \gamma \mathbf{1}_p$ が認められているとする．また $\boldsymbol{\Sigma}_i$ は一様構造であるとする．2-step 単調欠測データの場合の観測行列は以下の通りである．

$$\left(\begin{array}{c|ccc} \hline \overbrace{\begin{array}{c} \mathbf{x}_{11}^{(\ell)'} \\ \vdots \\ \mathbf{x}_{1n_1^{(\ell)}}^{(\ell)'} \end{array}}^{p_1} & & & \\ \hline \mathbf{x}_{21}^{(\ell)'} & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ \underbrace{\mathbf{x}_{2n_2^{(\ell)}}^{(\ell)'}}_{p_2} & * & \cdots & * \\ \hline \end{array} \right) \left. \begin{array}{l} \vphantom{\begin{array}{c} \mathbf{x}_{11}^{(\ell)'} \\ \vdots \\ \mathbf{x}_{1n_1^{(\ell)}}^{(\ell)'} \end{array}} \right\} n_1^{(\ell)} \\ \left. \vphantom{\begin{array}{c} * \\ \vdots \\ * \end{array}} \right\} n_2^{(\ell)} \end{array}$$

欠測データにおけるパラメータ ρ , σ^2 の MLE(最尤推定量) は容易に求められないので、これらを既知パラメータとみなし最後に推定量を代入する。

$$n_i^* = \left(\frac{1}{n_i^{(1)}} + \frac{1}{n_i^{(2)}} \right)^{-1}, \quad \bar{\mathbf{x}}_i^{(\ell)} = \frac{1}{n_i^{(\ell)}} \sum_{j=1}^{n_i^{(\ell)}} \mathbf{x}_{ij}^{(\ell)}, \quad \mathbf{u}_i = \bar{\mathbf{x}}_i^{(1)} - \bar{\mathbf{x}}_i^{(2)}, \quad i = 1, 2$$

とおくと ρ が既知の場合の γ の MLE とその従う分布は

$$\hat{\gamma} = \frac{n_1^* \tau_{11}^{-1} \mathbf{1}_{p_1}' \mathbf{u}_1 + n_2^* \tau_{12}^{-1} \mathbf{1}_{p_2}' \mathbf{u}_2}{n_1^* \tau_{11}^{-1} p_1 + n_2^* \tau_{12}^{-1} p_2} \sim N \left(\gamma, \frac{\sigma^2}{n_1^* \tau_{11}^{-1} p_1 + n_2^* \tau_{12}^{-1} p_2} \right)$$

である。ただし、 $\tau_{1i} = 1 + (p_i - 1)\rho$ である。ここで

$$\mathbf{V}_i^{(\ell)} = \sum_{j=1}^{n_i^{(\ell)}} (\mathbf{x}_{ij}^{(\ell)} - \bar{\mathbf{x}}_i^{(\ell)}) (\mathbf{x}_{ij}^{(\ell)} - \bar{\mathbf{x}}_i^{(\ell)})', \quad \mathbf{V}_i = \sum_{\ell=1}^2 \mathbf{V}_i^{(\ell)}, \quad \mathbf{S}_i = \frac{1}{n_i - 2} \mathbf{V}_i, \quad n_i = \sum_{\ell=1}^2 n_i^{(\ell)}, \quad i = 1, 2$$

と定義する。Siotani, Hayakawa and Fujikoshi (1985) で与えられている完全データの下での MLE を基に、パラメータ ρ と σ^2 の推定量として

$$\tilde{\rho} = \frac{\hat{\tau}_{21} - \text{tr}(\mathbf{S}_1) + \hat{\tau}_{22} - \text{tr}(\mathbf{S}_2)}{\{p_1(p_1 - 1) + p_2(p_2 - 1)\} \tilde{\sigma}^2}, \quad \tilde{\sigma}^2 = \frac{1}{(n_1 - 2)p_1 + (n_2 - 2)p_2} \left\{ \text{tr}(\mathbf{V}_1) + \text{tr}(\mathbf{V}_2) \right\}$$

を提案する。ただし $\hat{\tau}_{2i} = \mathbf{1}_{p_i}' \mathbf{S}_i \mathbf{1}_{p_i}$ とする。 $\tilde{\rho}$, $\tilde{\sigma}^2$ は共に MLE ではないが、どちらも 2-step 単調欠測データの全ての観測値を使用した統計量である。Jensen の不等式とこれらの推定量を用いて γ の $100\alpha\%$ 近似信頼区間

$$\tilde{\gamma} - \left(\sum_{i=1}^2 \frac{n_i^* p_i^2}{\hat{\tau}_{2i} t_{n_i-2}^2(\frac{\alpha}{2})} \right)^{-\frac{1}{2}} < \gamma < \tilde{\gamma} + \left(\sum_{i=1}^2 \frac{n_i^* p_i^2}{\hat{\tau}_{2i} t_{n_i-2}^2(\frac{\alpha}{2})} \right)^{-\frac{1}{2}}$$

が得られる。ただし

$$\tilde{\gamma} = \frac{\tilde{\tau}_{11}^{-1} n_1^* \mathbf{1}_{p_1}' \mathbf{u}_1 + \tilde{\tau}_{12}^{-1} n_2^* \mathbf{1}_{p_2}' \mathbf{u}_2}{\tilde{\tau}_{11}^{-1} n_1^* p_1 + \tilde{\tau}_{12}^{-1} n_2^* p_2}, \quad \tilde{\tau}_{1i} = 1 + (p_i - 1)\tilde{\rho}$$

であり、 $t_{n_i-2}(\alpha/2)$ は自由度 $n_i - 2$ の t 分布の上側 $100(\alpha/2)\%$ 点である。

4 シミュレーション結果

この節では、前節で導出した近似信頼区間を C1 とし、完全データ部分のみを使用した信頼区間を C2 とし、被覆確率をシミュレーションによって比較する。 $n_2 = 0$ とすると γ の MLE は $\hat{\gamma}_1 = (1/p_1) \mathbf{1}_{p_1}' \mathbf{u}_1$ となり、 $100\alpha\%$ 信頼区間 C2 は

$$\hat{\gamma}_1 - t_{n_1-2} \left(\frac{\alpha}{2} \right) \left(\frac{n_1^* p_1^2}{\hat{\tau}_{21}} \right)^{-\frac{1}{2}} < \gamma < \hat{\gamma}_1 + t_{n_1-2} \left(\frac{\alpha}{2} \right) \left(\frac{n_1^* p_1^2}{\hat{\tau}_{21}} \right)^{-\frac{1}{2}}$$

の形で書くことができる。ここに C2 は正確な信頼区間である。以下の表 1 と表 2 は有意水準 5%, 繰り返し回数 100,000 回のシミュレーションで得られた C1 と C2 の被覆確率である。表 1 より、 $\rho = -0.1$ 以外においては C1 の方が被覆確率が高く、かつその確率は 0.95 を大きく上回っていない。サンプルサイズを増やした表 2 においても同様の結果となっている。詳細は当日報告する。

表 1. シミュレーション 1 ($n_1 = 20$, $n_2 = 10$, $\alpha = 0.05$)

CP		$p_1 = 5, p_2 = 3$				$p_1 = 10, p_2 = 6$			
σ^2	ρ	-0.1	0.0	0.5	0.9	-0.1	0.0	0.5	0.9
0.1	C1	0.9500	0.9495	0.9510	0.9510	0.9428	0.9505	0.9505	0.9509
	C2	0.9499	0.9495	0.9496	0.9498	0.9502	0.9503	0.9503	0.9504
0.5	C1	0.9491	0.9502	0.9505	0.9509	0.9426	0.9502	0.9512	0.9512
	C2	0.9502	0.9499	0.9489	0.9498	0.9500	0.9498	0.9496	0.9493
1.0	C1	0.9489	0.9512	0.9504	0.9511	0.9442	0.9507	0.9510	0.9515
	C2	0.9494	0.9509	0.9488	0.9492	0.9508	0.9504	0.9502	0.9504
5.0	C1	0.9507	0.9505	0.9514	0.9515	0.9443	0.9502	0.9504	0.9504
	C2	0.9501	0.9496	0.9504	0.9509	0.9511	0.9491	0.9497	0.9501
10.0	C1	0.9486	0.9509	0.9510	0.9513	0.9436	0.9503	0.9518	0.9514
	C2	0.9484	0.9507	0.9503	0.9504	0.9499	0.9497	0.9501	0.9502

表 2. シミュレーション 2 ($n_1 = 60$, $n_2 = 30$, $\alpha = 0.05$)

CP		$p_1 = 5, p_2 = 3$				$p_1 = 10, p_2 = 6$			
σ^2	ρ	-0.1	0.0	0.5	0.9	-0.1	0.0	0.5	0.9
0.1	C1	0.9495	0.9497	0.9510	0.9512	0.9487	0.9504	0.9514	0.9515
	C2	0.9492	0.9502	0.9495	0.9500	0.9502	0.9505	0.9505	0.9498
0.5	C1	0.9494	0.9503	0.9506	0.9505	0.9478	0.9498	0.9503	0.9520
	C2	0.9495	0.9501	0.9499	0.9504	0.9493	0.9487	0.9486	0.9514
1.0	C1	0.9498	0.9492	0.9512	0.9506	0.9487	0.9504	0.9509	0.9508
	C2	0.9497	0.9491	0.9508	0.9504	0.9511	0.9501	0.9491	0.9505
5.0	C1	0.9506	0.9509	0.9500	0.9507	0.9486	0.9511	0.9506	0.9509
	C2	0.9506	0.9499	0.9494	0.9498	0.9506	0.9504	0.9497	0.9502
10.0	C1	0.9485	0.9509	0.9500	0.9517	0.9482	0.9499	0.9513	0.9501
	C2	0.9507	0.9502	0.9493	0.9509	0.9493	0.9493	0.9496	0.9496

参考文献

- [1] Koizumi, K. and Seo T. (2009). Testing equality of two mean vectors and simultaneous confidence intervals in repeated measures with missing data. *Journal of The Japanese Society of Computational Statistics*, **22**, 33-41.
- [2] Onozawa, M., Takahashi, S. and Seo, T. (2013). Tests for profile analysis based on two-step monotone missing data. *Discussiones Mathematicae Probability and Statistics*, **33**, 171-190.
- [3] Rencher, A. C. and Christensen, W. F. (2012). *Methods of Multivariate Analysis, Third Edition*. New York, NY: Wiley.
- [4] Sawatsuhashi, K. and Hyakutake, H. (2016). Profile analysis for random effects model in two sample problem with monotone missing. *Far East Journal of Theoretical Statistics*, **52**, 235-251.
- [5] Siotani, M., Hayakawa, T. and Fujikoshi, Y. (1985). *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook*, American Sciences Press, Inc., Columbus, Ohio.
- [6] Srivastava, M. S. (1987). Profile analysis of several groups. *Communications in Statistics -Theory and Methods*, **16**, 909-926.
- [7] Srivastava, M. S. (2002). *Methods of Multivariate Statistics*. New York, NY: Wiley.