一般化した分布の仮定の下での高次元 MANOVA 問題

滋賀大学データサイエンス学部　姫野哲人
鹿児島大学総合教育機構共通教育センター　山田隆行

近年、ビッグデータが様々な分野で注目されるようになり、高次元データの解析手法の重要性が増してきている。そこで、本研究では 2 元配置多変量分散分析モデル

$$\boldsymbol{y}_{ijk} = \boldsymbol{\eta}_{ij} + \boldsymbol{\varepsilon}_{ijk} \quad (i = 1, \ldots, a,\ j = 1, \ldots, b,\ k = 1, \ldots n_{ij})$$

を考える。ここで、$\boldsymbol{\varepsilon}_{ijk}$ は全て独立な p 次元の誤差ベクトルで、平均ベクトル $\mathbf{0}$、分散共分散行列を Σ とする。また、平均ベクトルに関し、

$$\boldsymbol{\eta}_{ijk} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij}$$

と表されるとする。ただし、主効果と交互作用に対しては次の制限を仮定している。

$$\sum_{i=1}^{a} n_{i\cdot}\boldsymbol{\alpha}_i = \mathbf{0},\ \sum_{j=1}^{b} n_{\cdot j}\boldsymbol{\beta}_j = \mathbf{0},\ \sum_{i=1}^{a} n_{i\cdot}\boldsymbol{\gamma}_{ij} = \mathbf{0},\ \sum_{j=1}^{b} n_{\cdot j}\boldsymbol{\gamma}_{ij} = \mathbf{0}.$$

ここで、$n_{i\cdot} = \sum_{j=1}^{b} n_{ij}, n_{\cdot j} = \sum_{i=1}^{a} n_{ij}$ である。$N(= \sum_{i,j} n_{ij}), p$ ともに十分大きいという状況の下で、次の 3 つの帰無仮説

$$\begin{aligned} H_\alpha &: \boldsymbol{\alpha}_i = \mathbf{0}\ (i = 1, \ldots, a) \\ H_\beta &: \boldsymbol{\beta}_i = \mathbf{0}\ (j = 1, \ldots, b) \\ H_\gamma &: \boldsymbol{\gamma}_{ij} = \mathbf{0}\ (i = 1, \ldots, a,\ , j = 1, \ldots, b) \end{aligned}$$

を考える。

上記のモデルは多変量線形モデルにより一般化される。この場合に、誤差ベクトルに正規分布を仮定するとき、これらの検定に対する検定統計量は、$n = N - ab > p$ では尤度比統計量、$p > n$ では Dempster 型検定統計量が代表的である。しかし、高次元データの誤差分布が多変量正規モデルに従うという仮定はかなり強い仮定であるうえ、その仮定が満たされるかどうかを診断する方法もほとんど提案されていない。

そこで、近年の高次元データに対する研究では正規性を仮定しない手法が数多く提案されている。それらの研究の多くは標準化された誤差 $\boldsymbol{z} = \Sigma^{-1/2}\boldsymbol{\varepsilon}(= (z_1, \ldots, z_p))$ に対し、

$$\mathrm{E}[\boldsymbol{z}] = \mathbf{0},\ \mathrm{Cov}(\boldsymbol{z}) = I_p,\ \mathrm{E}[z_i] < \infty\ (i = 1, \ldots, p)$$
$$\mathrm{E}(z_{\ell_1}^{v_1} \cdots z_{\ell_r}^{v_r}) = \mathrm{E}(z_{\ell_1}^{v_1}) \cdots \mathrm{E}(z_{\ell_r}^{v_r})\ (\textstyle\sum_{i=1}^{r} v_i \leq 8,\ \ell_1 < \cdots < \ell_r)$$

が仮定されている。この仮定の下、二標本問題における平均ベクトルの検定（Chen and Qin, 2010, Ann. Statist.）や 1 元配置多変量線形モデルに対する一般化線形仮説検定（Zhou et al., 2017, JSPI）等が提案されている。しかし、この誤差分布の仮定は $\boldsymbol{z}$ の全成分が独立であれば成り立つが、独立でない場合は、このような仮定を満たす分布はほとんど存在しない。

本発表では、大標本、高次元の下、誤差分布として楕円分布を含むようなクラスを仮定したうえで、2 元配置分散分析に対する検定統計量の提案及びその漸近正規性を示し、その近似精度について数値実験の結果を示す。

参考文献

[1] Chen, S. X. and Qin, Y. L. (2017), A two-sample test for high-dimensional data with applications to gene-set testing, *The Annals of Statistics*, 38, 808-835.

[2] Zhou, B., Guo, J., and Zhang, J. T. (2017), High-dimensional general linear hypothesis testing under heteroscedasticity, *Journal of Statistical Planning and Inference*, 188, 36-54.