A modified principal component test for high-dimensional data

Siegfried Kropf[1], Guo-Chun Ding[2], Holger Heuer[2] and Kornelia Smalla[2]

[1] *University of Magdeburg, Germany, email: siegfried.kropf@med.ovgu.de*
[2] *Julius Kühn-Institut Braunschweig, Germany, email: guo-chun.ding@jki.bund.de, holger.heuer@jki.bund.de, kornelia.smalla@jki.bund.de*

Keywords: high-dimensional data, principal component test, rotation test.

Modern biochemical analysis techniques often deliver high-dimensional observation vectors, while only small sample sizes are feasible. As an example we consider a microarray (PhyloChip) data set for comparing the bacterial community structures in the rhizosphere of three potato cultivars grown at two sites (cf. to [4] for details).

In [3], Läuter and colleagues proposed a PC test that calculates the principal components from the total sums and squares and cross products matrix $\mathbf{W}$ of the data and carries out a test on the basis of the low-dimensional principal components. For multivariate normal data this yields left-spherically distributed components and hence an exact multivariate test for the usual multivariate test statistics. Another proposal for an exact multivariate test in this situation is the 50-50-MANOVA test by Langsrud in [1].

For extreme relations of sample size n and number of variables p (n=18 and p=2432 in the example), however, there arises a problem regarding the power. The sample size restricts the number q of principal components enclosed in the test. But omitting essential components may yield a loss of power.

Therefore, we use a modified test statistic

$$\tilde{F} = \frac{\left(\sum_{i=1}^{q} \lambda_i h_{ii}\right)/\nu_h}{\left(\sum_{i=1}^{q} \lambda_i g_{ii}\right)/\nu_g} \ ,$$

where the λ_i are the eigenvalues of $\mathbf{W}$, h_{ii} and g_{ii} are the hypothesis and residual related sums of squares, and ν_h and ν_g are the corresponding degrees of freedom as one would use in a univariate test. This test statistic does no longer follow an F-distribution under the null hypothesis, but one can find a Satterthwaite approximation and one can still use properties of left-spherically distributed data to derive an exact test on the basis of rotation tests as introduced by Langsrud in [2].

The power of the resulting tests is compared in the example and in simulation studies, demonstrating the good performance of the new test in this high-dimensional setting.

References

[1] Langsrud, Ø. (2005). 50-50-Multivariate analysis of variance for collinear responses, *The Statistician* **51**, 305–317.

[2] Langsrud, Ø. (2005). Rotation tests, *Statistics and Computing* **15**, 53–60.

[3] Läuter, J., Glimm, E., Kropf, S. (1996). New multivariate tests for data with an inherent structure, *Biometrical Journal* **38**, 5–23.

[4] Weinert, N., Piceno, Y., Ding, G.-C., Meinecke, R., Heuer, H., Berg, G., Schloter, M., Andersen, G., Smalla, K. (2011). PhyloChip hybridization uncovered an enormous bacterial diversity in the rhizosphere of different potato cultivars: many common and few cultivar-dependent taxa, *FEMS Microbiology Ecology* **75**, 497–506.