High-dimensional supervised search for relevant lumps of variables – general algorithms keeping the familywise error exactly

Jürgen Läuter
Otto von Guericke University Magdeburg, Germany
and
Interdisciplinary Centre for Bioinformatics, University of Leipzig, Germany

E-Mail: juergen.laeuter@med.ovgu.de

In many high-dimensional biological applications, for example, in gene expression analysis, investigating sets of correlated variables offers a number of advantages over methods which focus on identifying single variables. First of all, "lumps" of correlated variables provide more insight into the biological mechanisms than single variables. Additionally, sets of highly correlated explanatory variables are statistically more stable, they can be detected earlier and more reliably than single variables, and they are less affected by spurious observations. We will present algorithms for the selection of variables which, for the first time, combine the two essential tasks, namely the generating of suitable subsets of variables and their testing for significance, in a way strictly observing the familywise type I error α. In the literature, there are many proposals to determine the sets of variables in a heuristical manner, so that the level of significance cannot be strictly kept. Sometimes, prespecified sets of variables are used, for example, those from "Gene Ontology" (Ashburner et al., 2000).

Goeman and Mansmann (2008) and Meinshausen (2008) proposed methods, in which given hierarchies of subsets of variables are analyzed by logical bottom-up and top-down considerations. However, these procedures are not applicable if the subsets are not prespecified.

We will investigate the correlations between the p columns of the $n \times p$ explanatory matrix $\mathbf{X}$ and the $n \times 1$ response vector $\mathbf{y}$. The dimension p can be very large in comparison to the sample size n. In the context of the parametric test methods, we assume that $(\mathbf{y} \quad \mathbf{X})$ consists of n independent rows that have the same normal distribution $N_{1+p}\left(\begin{pmatrix} \mu_y & \boldsymbol{\mu}' \end{pmatrix}, \begin{pmatrix} \sigma_{yy} & \boldsymbol{\sigma}' \\ \boldsymbol{\sigma} & \boldsymbol{\Sigma} \end{pmatrix} \right)$. We want to recognize which of the elements σ_i of $\boldsymbol{\sigma}' = (\sigma_1 \quad \ldots \quad \sigma_p)$ are unequal to zero. An exact test of the global null hypothesis $\boldsymbol{\sigma}' = \mathbf{0}'$ is given by

$$B = \frac{(\mathbf{g}'\mathbf{d})^2}{g_{yy} \cdot \mathbf{d}'\mathbf{G}\mathbf{d}} \quad \geq \quad B_{1-\alpha}\left(\frac{1}{2}, \frac{n-2}{2}\right),$$

where $g_{yy} = (\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})$, $\mathbf{g}' = (\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{X} - \bar{\mathbf{X}})$, $\mathbf{G} = (\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}})$, and $\mathbf{d}$ is an arbitrary function of $\mathbf{G}$ (Läuter, 1996; Läuter, Glimm and Kropf, 1996, 1998). For example, $\mathbf{d}$ can be the first eigenvector of the eigenvalue problem $\mathbf{G}\mathbf{d} = \mathbf{d}\lambda$. An extension to several eigenvectors is possible. In this representation, the regression setup for $\mathbf{y}$ depending on $\mathbf{X}$ is consciously avoided. Regression coefficients do not appear. The regression setup is too restrictive for our test and selection problems.

In a corresponding way, the comparison of two independent samples of the sizes $n^{(1)}$, $n^{(2)}$ with $n = n^{(1)} + n^{(2)}$ can be carried out:

$$B = \frac{n^{(1)}n^{(2)}}{n^{(1)}+n^{(2)}} \; \frac{((\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})'\mathbf{d})^2}{\mathbf{d}'\mathbf{G}\mathbf{d}} \quad \geq \quad B_{1-\alpha}(\frac{1}{2}, \frac{n-2}{2}).$$

The considerations can be generalized to non-normal data, in particular, to discrete data. Permutation tests are used to show whether the sets of **X** columns are depending on **y**.

Our first algorithm for selection of variables is based on the Westfall-Young resampling principles (Westfall and Young, 1993). Until now, the Westfall-Young procedure was only applied for searching single relevant variables. We extend it to sets of variables, the "lumps". To generate the sets, each variable is considered as a centre of potential subsets of highly correlated variables. This is a very flexible, adaptive method to analyze data with unknown covariance structure. The selection procedure controls strongly the familywise type I error in spite of the fact that the subsets can have random fluctuations, because

- the intercorrelations between the **X** variables are determined without utilizing the response variable **y** ("total covariances"),
- a simple correlation condition is used,
- the applied test statistics have certain monotonicity properties.

Another algorithm for generating and testing the subsets is based on the exact high-dimensional parametric tests by Läuter et. al. (1996, 1998) and the test procedure by Kropf and Läuter (2002). This algorithm utilizes a data-depending ordering of the subsets according to their sums of products. A special shortcut strategy serves for managing the huge number of subsets.

In the procedures, the particular requirements for statistical stability are taken into account. Any kind of overfitting is avoided. The natural interaction between sets of correlated variables and existing groups of individuals is systematically utilized. Once sets of variables have been found, our procedures offer new interesting possibilities for sorting and structuring the given *n* individuals. The applications refer to gene-expression data of lymphoma patients.

References

Ashburner, M. et al. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**(1), 25-29.

Goeman, J.J. and Mansmann, U. (2008). Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics Advance Access*, http://bioinformatics.oxfordjournals.org/cgi/reprint/btm628v1

Kropf, S. and Läuter, J. (2002). Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression data. *Biometrical Journal* **44**, 789-800.

Läuter, J. (1996). Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics* **52**, 964-970.

Läuter, J., Glimm, E. and Kropf, S. (1996). New multivariate tests for data with an inherent structure. *Biometrical Journal* **38**, 5-23, Erratum: *Biometrical Journal* **40**, 1015.

Läuter, J., Glimm, E. and Kropf, S. (1998). Multivariate tests based on left-spherically distributed linear scores. *The Annals of Statistics* **26**, 1972-1988, Correction: *The Annals of Statistics* **27**, 1441.

Meinshausen, N. (2008). Hierarchical testing for variable importance. *Biometrika* **95**, 2, 265-278.

Westfall, P.H. and Young, S.S. (1993). *Resampling-based multiple testing.* John Wiley & Sons, New York.