High-dimensional data analysis: Selection of variables, data compression and graphics – Application to gene expression

Jürgen Läuter*[1,2], **Friedemann Horn**[1,3,4], **Maciej Rosołowski**[4,5], and **Ekkehard Glimm**[6]

[1] Interdisciplinary Centre for Bioinformatics (IZBI), University of Leipzig, Härtelstr. 16-18, 04107 Leipzig, Germany
[2] Otto von Guericke University Magdeburg, Mittelstr. 2/151, 39114 Magdeburg, Germany
[3] Institute of Clinical Immunology and Transfusion Medicine, Medical Faculty, University of Leipzig, Johannisallee 30, 04103 Leipzig, Germany
[4] Interdisciplinary Centre of Clinical Research (IZKF), Medical Faculty, University of Leipzig, Germany
[5] Institute of Medical Informatics, Statistics and Epidemiology (IMISE), Medical Faculty, University of Leipzig, Härtelstr. 16-18, 04107 Leipzig, Germany
[6] Novartis Pharma AG, Lichtstr. 35, 4056 Basel, Switzerland

Received 9 May 2008, revised 3 October 2008, accepted 21 December 2008

The paper presents effective and mathematically exact procedures for selection of variables which are applicable in cases with a very high dimension as, for example, in gene expression analysis. Choosing sets of variables is an important method to increase the power of the statistical conclusions and to facilitate the biological interpretation. For the construction of sets, each single variable is considered as the centre of potential sets of variables. Testing for significance is carried out by means of the Westfall-Young principle based on resampling or by the parametric method of spherical tests. The particular requirements for statistical stability are taken into account; each kind of overfitting is avoided. Thus, high power is attained and the familywise type I error can be kept in spite of the large dimension. To obtain graphical representations by heat maps and curves, a specific data compression technique is applied. Gene expression data from B-cell lymphoma patients serve for the demonstration of the procedures.

Key words: Data compression; Gene expression analysis; High-dimensional tests; Multivariate analysis; Selection of variables.

Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1002/bimj.200800207.

1 Selection of variables, a challenge for statisticians

Selection of variables and model choice play an important role in classical regression analysis, in multivariate analysis of variance and in discriminant analysis, as well as in logistic regression and other generalized linear models, and also in the Cox model of survival analysis. In all these cases, the response variables are related to the explanatory variables by a linear setup with unknown coefficients. Strategies for the search of essential variables are similar to the well-known methods from least-squares theory. Decisions on the choice of variables are based, if available, on

* Correspondence to: e-mail: juergen.laeuter@med.ovgu.de, Phone: +49391616427, Fax:+493915313937

multivariate overall tests, but mostly on conditional likelihood ratio tests for single variables. Most software packages provide an exhaustive evaluation of all possible sets of variables, or alternatively several variants of stepwise forward and backward selection procedures.

These approaches, however are fraught with difficulties. On a technical level, the computational burden can be prohibitive: For p variables, there are 2^p different sets of variables. Obviously, these cannot all be checked step by step if p is large. Thus, for example, if $p = 20\,000$ genes are investigated in gene expression analysis, then $2^{20\,000}$ sets of variables exist, and this number is greater than 10^{6000}.

Even if we do not consider this huge dimension, difficult problems arise. Many selection procedures, in which "best subsets of variables" are searched on the basis of a given sample, are disposed to evaluate the variables too optimistically. Merely minimizing the deviations between the model and the sample as, for example, in the "branch and bound algorithm" by Furnival and Wilson (1974) does not guarantee the relevance of a subset. If, directly or indirectly, a structural or causal analysis of the variables is carried out according to the method of least squares, then often "overfitted" conclusions are obtained. The models fit the given samples very well, but they do not sufficiently describe the true multivariate distributions at hand. The stepwise selection procedures mentioned above are plausible at first sight but, in many cases, they do not attain the sets of variables that are essential for practical applications. The important single variables and sets of variables are possibly hidden by intercorrelations, "masked", and therefore, they cannot be recognized reliably.

The theory of the multivariate linear model developed elegantly in the last half century could not yet overcome the great problems of statistical instability. Statisticians knew that a high dimension and a high correlation between the variables render difficulties in the selection procedures. Thus, many attempts of a computational stabilization have been made with "ridge", "shrinkage", "penalizing", "lasso", "lars" etc. methods (Hoerl and Kennard, 1970; Anderson and Blair, 1982; Tibshirani, 1996; Efron et al., 2004; Hastie, Tibshirani and Friedman, 2001). However, such interventions suffered always from some subjectivity, and their precise mathematical consequences remained unclear. The main deficiency was that the unified theoretical structure of the procedures was lost and, in particular, the type I error rate could not be strictly kept in such a selection.

This situation left statisticians with a dilemma for a long time: They wanted to get data with many variables, with much information. On the other hand, they could not analyze the high-dimensional data in a satisfactory way. If the algorithms implemented in the most popular statistical software tools were applied, for example, with a stepwise selection procedure or with subjective ridge-correction terms, then the derived statistical conclusions were uncontrollably biased. In the past, until the nineties, the only actually reliable methods of selection of variables were the classical parametric and non-parametric tests in connection with multiple testing procedures, like the closed test principle, ultimately based on the rules by Bonferroni and Bonferroni/Holm. However, if decisions on thousands or millions of unknown subsets must be made, then these methods were not sufficient. Later, the works by Westfall and Young (1993), Läuter (1996) and Läuter, Glimm and Kropf (1996, 1998) yielded an essential progress.

Goeman and Mansmann (2008) proposed a method to evaluate sets of variables arranged in a given graph structure by means of logical bottom-up and top-down procedures. Meinshausen and Bühlmann (2006) and Wasserman and Roeder (2007) investigated the selection of variables by means of Tibshirani's lasso method in asymptotical considerations, with the sample size n going to infinity. The asymptotical proofs are ingenious but rather complicated. The authors need some special assumptions about the relation of dimension p to size n, about the number of the "non-null" variables and about the covariance structure. The conclusions obtained are an important step forward in overcoming the uncertainty of selection. However, they are still not generally applicable to problems with a very large dimension p or a small sample size n in which no additional conditions for the means and covariances are fulfilled.

In this paper, we will report non-asymptotical methods that are exact for all dimensions p, all sample sizes n and arbitrary structures of the parameters. We search one or several subsets of variables that contain differential variables. Multiple testing based on the criterion of the familywise type I error rate (FWER) is used.

We believe that inferential investigations in multivariate statistical analysis should be directed toward methods which can substantially help to overcome the deficiencies of the last decades with respect to overfitting, multicollinearity and, more general, statistical instability. Our paper is intended to contribute to this aim.

2 An algorithm of selection based on the permutation method by Westfall and Young

In the Interdisciplinary Centre for Bioinformatics of the Leipzig University, we are working on procedures for the selection of variables which can be applied in gene expression analysis. The dimension p is usually very large, for example, $p \approx 20\,000$. The sample sizes are allowed to be much smaller, for example, $n \approx 100$. We would like to take into account the correlation structure of the given gene data. Generally, we assume that sets of genes with correlated patterns of expression are more informative for the understanding of the biological gene functions than single genes listed without giving their mutual relations. The sets of genes that originate from our data-dependent construction reveal possible relationship between the genes within the sets, but also allow to assess similarities and differences of all sets. Thus, they also support the biological interpretation, because they have a broader characterization.

A basic principle of our selection procedures is that we start by considering each variable $i_1 = 1, \ldots, p$ as the source or the centre of potential subsets of variables. A variable i may be added to the source variable i_1 if the correlation between both variables exceeds a fixed minimum value. As an application of our general method, we will treat in this paper the comparison of the mean vectors $\boldsymbol{\mu}^{(1)\prime}$ and $\boldsymbol{\mu}^{(2)\prime}$ of two p-dimensional samples

$$\mathbf{x}_{(j)}^{(1)}\prime = (x_{j1}^{(1)}\, x_{j2}^{(1)} \ldots x_{jp}^{(1)}),\, j = 1, \ldots, n^{(1)},$$

$$\mathbf{x}_{(j)}^{(2)}\prime = (x_{j1}^{(2)}\, x_{j2}^{(2)} \ldots x_{jp}^{(2)}),\, j = 1, \ldots, n^{(2)}.$$

Then, we are interested in finding subsets m of variables, in which some of the mean values are different: $\mu_i^{(1)} \neq \mu_i^{(2)}$ for at least one index $i \in m$. The corresponding data matrix with $n = n^{(1)} + n^{(2)}$ rows is

$$\underset{(n^{(1)}+n^{(2)})\times p}{\mathbf{X}} = \begin{pmatrix} \underset{n^{(1)}\times p}{\mathbf{X}^{(1)}} \\ \underset{n^{(2)}\times p}{\mathbf{X}^{(2)}} \end{pmatrix}.$$

The usual "total correlation coefficient" of the columns $\mathbf{x}_{i_1}$ and $\mathbf{x}_i$ of $\mathbf{X}$

$$r_{i_1 i}^2 = \frac{((\mathbf{x}_{i_1} - \bar{\mathbf{x}}_{i_1})'(\mathbf{x}_i - \bar{\mathbf{x}}_i))^2}{(\mathbf{x}_{i_1} - \bar{\mathbf{x}}_{i_1})'(\mathbf{x}_{i_1} - \bar{\mathbf{x}}_{i_1}) \cdot (\mathbf{x}_i - \bar{\mathbf{x}}_i)'(\mathbf{x}_i - \bar{\mathbf{x}}_i)}$$

serves as the measure of similarity of the variables i_1 und i. The total mean vectors $\bar{\mathbf{x}}_{i_1} = \mathbf{1}_n \bar{x}_{i_1}$ and $\bar{\mathbf{x}}_i = \mathbf{1}_n \bar{x}_i$, without separating the groups (1) and (2), are applied. Here, $\mathbf{1}_n$ denotes the n-dimensional vector consisting only of ones. The variable i is considered as a potential partner of the source variable i_1 if it fulfills the necessary correlation condition $r_{i_1 i}^2 \geq c$, with c being a fixed positive constant.

Of course, there are many different subsets starting all from the same source i_1. The smallest subset is $\{i_1\}$, consisting only of the source variable, the largest subset is the so-called "maximum set" of source i_1, M_{i_1}, that contains all variables i satisfying the above correlation condition. Let $bas(i_1,\mathbf{X},c)$ denote the set of all possible subsets generated by variable i_1. The concept leads to a "basis totality" of all candidate sets of our method, which is the union of the p partitions corresponding to the sources $i_1 = 1,\ldots,p$:

$$BAS(\mathbf{X},c) = \bigcup_{i_1=1,\,\ldots\,,p} bas(i_1,\mathbf{X},c).$$

This strategy has already been described in Läuter, Glimm and Eszlinger (2005) and Läuter (2007). The idea of forming "correlation neighbourhoods" of the single variables was also applied by Tibshirani and Wasserman (2006).

We will assume that the n row vectors $(n \geq 3)$

$$\mathbf{y}_{(j)}^{(1)}{}' = (\mathbf{x}_{(j)}^{(1)} - \boldsymbol{\mu}^{(1)})', j = 1,\ \ldots\ ,n^{(1)},\ \mathbf{y}_{(j)}^{(2)}{}' = (\mathbf{x}_{(j)}^{(2)} - \ \boldsymbol{\mu}^{(2)})', j = 1,\ \ldots\ ,n^{(2)}$$

have a joint np-dimensional distribution which does not change with permuting the vectors $\mathbf{y}_{(1)}^{(1)}{}',\ \ldots, \mathbf{y}_{(n^{(1)})}^{(1)}{}', \mathbf{y}_{(1)}^{(2)}{}',\ \ldots, \mathbf{y}_{(n^{(2)})}^{(2)}{}'$. As a special case, the n vectors $\mathbf{y}_{(j)}^{(1)}{}'$ and $\mathbf{y}_{(j)}^{(2)}{}'$ can have independent p-dimensional normal distributions $\mathrm{N}_p(\mathbf{0}',\Sigma)$ with the same covariance matrix Σ. Then, the widespread resampling procedure by Westfall and Young (1993, Section 2.3) can be applied to all subsets $m \in BAS(\mathbf{X},c)$. For a set m consisting of the variables $i_1, i_2, \ldots, i_s$, we use the test statistic $F_m = F_m(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_s})$. A necessary condition for our procedure is that F_m must be monotone increasing with respect to the sets m, i.e., $F_{m_1} \leq F_{m_2}$ for $m_1 \subseteq m_2$. Mostly, we employ the sum of the univariate beta statistics belonging to m

$$F_m = \sum_{i=i_1,\,\ldots\,,i_s} \mathrm{B}_i = \sum_{i=i_1,\,\ldots\,,i_s} \frac{n^{(1)}n^{(2)}}{n^{(1)}+n^{(2)}}\,\frac{(\bar{x}_i^{(1)} - \bar{x}_i^{(2)})^2}{(\mathbf{x}_i - \bar{\mathbf{x}}_i)'(\mathbf{x}_i - \bar{\mathbf{x}}_i)}.$$

Each value B_i lies between 0 and 1. The Westfall-Young procedure is essentially based on the empirical distribution of the maximal F_m values of all candidate subsets m. In the literature, the denotation "max$T$ procedure" is used.

In this strategy, sets with many variables tend to give higher values of F_m than sets with few variables. Sets with much information get priority over sets with little information. This corresponds to our principle that repeatedly observable biological findings should be emphasized in the statistical analysis rather than isolated ones. In particular, the maximum set M_{i_1} produces a higher value of F_m than any smaller subset $m \in bas(i_1,\mathbf{X},c)$. The definitions of the subsets m of variables and of the statistics F_m do not contain any unstable elements. The only metric used substantially in the procedure is the Euclidean metric of the n-dimensional space formed by the n independent measurements. However, the space of the p variables does not need a special structure. Our sets of variables are adaptively created at places where many variables are concentrated. Redundance in the data is consciously utilized for stabilization.

Based on this construction of the subsets m and the statistics F_m, a very effective Westfall-Young algorithm can be applied. The maximum set M_{i_1} "majorizes" all subsets $m \in bas(i_1,\mathbf{X},c)$ with respect to the statistic F_m. Therefore, only the p maximum sets $M_1,\ldots,M_p$ must be taken into consideration when determining the empirical maximum distribution according to the Westfall-Young technique:

$$F^* = \max_{i_1=1,\,\ldots\,,p} F^*_{M_{i_1}}.$$

The asterisk in this formula refers to the resampling permutations of the rows of $\mathbf{X}$, corresponding to the two-group comparison. We can take all possible row permutations or, if this number is to large, we can resort to a randomly generated choice of them. In the special case of very small sample sizes $n^{(1)}$ and $n^{(2)}$, data rotations of the rows can be used instead of permutations (Läuter *et al.*, 2005).

A set $m \in BAS(\mathbf{X}, c)$ shows a significant difference between groups (1) and (2) at significance level α if the corresponding beta sum $F_m = \sum_{i \in m} \mathrm{B}_i$ fulfills the condition $F_m > F^*_{1-\alpha}$, where $F^*_{1-\alpha}$ is the $(1-\alpha)$ quantile from the empirical F^* distribution. If $F^{(1)}, F^{(2)}, \ldots, F^{(r)}$ are the increasingly ordered resampling values of F^*, then $F^*_{1-\alpha} = F^{(k)}$ with k arising from $r(1-\alpha)$ by upward rounding to the next integer number. In practice, the p maximum sets are checked for significance first. Subsequently, if significance of a maximum set M_{i_1} has been attained, the smaller subsets of the source i_1 can also be evaluated. This procedure keeps the familywise type I error rate α in the strong sense: In the series of all significant subsets m, some falsely significant subsets–subsets consisting only of non-differential variables–may appear with probability α, at most. Our procedure is suitable for arbitrary patterns of differential and non-differential variables.

The multiple level α is strictly kept in this Westfall-Young procedure, because the "total sums of products matrix" $\mathbf{W} = (\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}})$ is used for the construction of the totality $BAS(\mathbf{X},c)$ of subsets. Here, $\bar{\mathbf{X}} = \mathbf{1}_n \bar{\mathbf{x}}'$ is the total mean matrix calculated without utilizing the groups (1) and (2). This matrix $\mathbf{W}$, the derived correlations $r_{i_1 i}$ and the generated subsets do not change with permutations of the rows of $\mathbf{X}$. These facts are essential for the non-differential variables but, additionally, they bring about that the unknown differences between the groups (1) and (2) have a promoting effect on the subset generation and, thus, the power of the selection procedure can be improved. Above all, it is important for the exactness of the selection procedure that the empirical F^* distribution dominates all null distributions, which can arise in $BAS(\mathbf{X},c)$. For the justification of the procedure in more detail, we refer to Läuter *et al.* (2005) and Läuter (2007, http://www.izbi.uni-leipzig.de/izbi/Working%20Paper/2007/WP_15_Statistik.pdf).

Furthermore, one should note for reasons of mathematical strictness that different maximum sets can randomly occur in this procedure, even if the means $\boldsymbol{\mu}^{(1)\prime}$ and $\boldsymbol{\mu}^{(2)\prime}$ are fixed and only n fixed values $(\mathbf{x}^{(1)}_{(j)} - \boldsymbol{\mu}^{(1)})'$ and $(\mathbf{x}^{(2)}_{(j)} - \boldsymbol{\mu}^{(2)})'$ in different random permutations of "practice" are considered as a special discrete distribution. In such a case, the distribution symmetry of practice does not completely correspond to the suppositions on the null-related symmetry in our algorithm, if variables with $\mu^{(1)}_i \neq \mu^{(2)}_i$ are included in the subsets. Obviously, the subsets consisting only of variables with $\mu^{(1)}_i = \mu^{(2)}_i$ remain unchanged in all permutations of "practice". However, in spite of this random variability in the subset generation, our procedure fulfills always the requirements on the familywise error rate because of the monotonicity property of the test statistic F_m.

In the literature (see, for example, Dudoit, Shaffer und Boldrick (2002)), the Westfall-Young procedure is usually presented in a sequential, stepdown version. After some significances have been obtained, the corresponding variables are removed from the data matrices. Then, the procedure is started anew with a reduced number of variables in order to recognize further significances. Thus, the power of the procedure can be increased. However, this strategy must not be used in our selection procedure, because two tasks have to be fulfilled correctly in close connection: the generation of the subsets as well as the tests for significance. Careless application of the stepdown technique to our random maximum sets could lead to the violation of the multiple level of significance (Läuter, 2007, Chapter 7).

3 Application to the genomic exploration of B-cell lymphomas

Based on the above considerations, we will analyze gene expression data obtained from tumor tissue in order to retrieve information on the oncogenic pathways involved in their pathogenesis and to distinguish tumor subtypes exhibiting different clinical parameters or therapeutic outcomes. For these studies, we utilize previously published gene expression data derived from B-cell lymphomas (Hummel *et al.*, 2006). The cases investigated there comprise patients with Burkitt's lymphomas and diffuse large B-cell lymphomas (DLBCL) from the Molecular Mechanisms in Malignant

Lymphomas Network Project. For the analyses considered here, we use 108 lymphoma cases defined as training data set in that study. The Affymetrix microarray employed provides 22 277 different gene probesets. Based on transcriptional and genomic profiling, Hummel *et al.* have proposed a molecular definition of Burkitt's lymphoma (molecular Burkitt lymphoma, mBL).

Furthermore, to gain insight into the functional implications of different oncogenic pathways of these lymphoma entities, we make use of a set of expression data generated by Bild *et al.* (2006). This group of authors manipulated mammary epithelial cell cultures by overexpressing various oncogenes and determined the expression patterns by genome-wide expression analysis. For our considerations, we apply the data obtained from mammary cells overexpressing the c-myc oncogene that is known to be highly expressed in many B-cell lymphomas and other tumors and to be causally involved in their pathogenesis (Boxer and Dang, 2001). The data from 10 biological replicates of c-myc overexpressing cells and 10 control cultures are compared.

We apply the Westfall-Young selection procedure of Section 2 to these lymphoma data. However, we will use a special "two-matrices modification" in which the generation of subsets of variables and the tests for significance are performed on two separate data sets. The generation of the subsets is based on the clinical data matrix $\mathbf{X}_0$ with 108 rows corresponding to the lymphoma patients and with 22 277 columns belonging to the genes. Thus, we obtain subsets of genes with correlated expression. Each single gene is taken as a source for potential gene sets. The subsequent tests for significance are based on the cell culture expression matrix $\mathbf{X}$ of the format $20 \times 22\,277$ from cells overexpressing oncogenes versus control cells. This way, specific biological pathways are considered with respect to their clinical implications.

The selection procedure of Section 2 is correct also in this two-matrices design. The restrictions for the application of this procedure are even weakened, because the clinical and the experimental data are stochastically independent. In this particular case, the stepdown version of the Westfall-Young selection procedure as mentioned above may be employed as well (but we have not done this).

In this application, the following parameters are set:

1. Only the genes i whose clinical sums of squares fulfill the condition $(\mathbf{x}_i - \bar{\mathbf{x}}_i)'(\mathbf{x}_i - \bar{\mathbf{x}}_i) \geq 15$ are included in the proper algorithm of the procedure. This is the case for 6374 of the original 22 277 genes. The expression values of these genes are replaced by the corresponding ranks $1, \ldots, 108$.
2. The expression values of the cell lines are used without any changes.
3. The resampling is carried out with 1000 random permutations of the rows of $\mathbf{X}$.
4. The multiple level of significance is $\alpha = 0.05$.
5. The generation of the clinical gene subsets is performed under the condition $r_{i_1 i}^2 \geq 0.65$. Additionally, the correlations $r_{i_1 i}$ are required to be positive.
6. The beta-sums statistic F_m of a subset m is calculated only from the 13 highest beta values (if more than 13 genes are present).

All these modifications do not impair the exact control of the multiple significance level. Further details of the parameters of the selection procedure are given by Läuter (2007).

The selection algorithm yields 99 significant maximum sets of genes, 68 of which follow the direction of c-myc expression changes in the cell experiments while 31 respond reciprocally. For clarity of exposition, we restrict our interpretation to maximum sets that do not overlap with respect to the genes included. Under this additional condition, eight significant maximum sets remain. Fig. 1 displays a corresponding biplot according to Gabriel (1971). It shows both the 240 genes comprised in the eight chosen sets and the 108 patients.

Inspection of the genes included in the obtained sets reveals that they can be allocated rather clearly to cell biological processes. This is particularly remarkable as the statistical method generating these gene sets does not make use of any biological or clinical information. As an example,

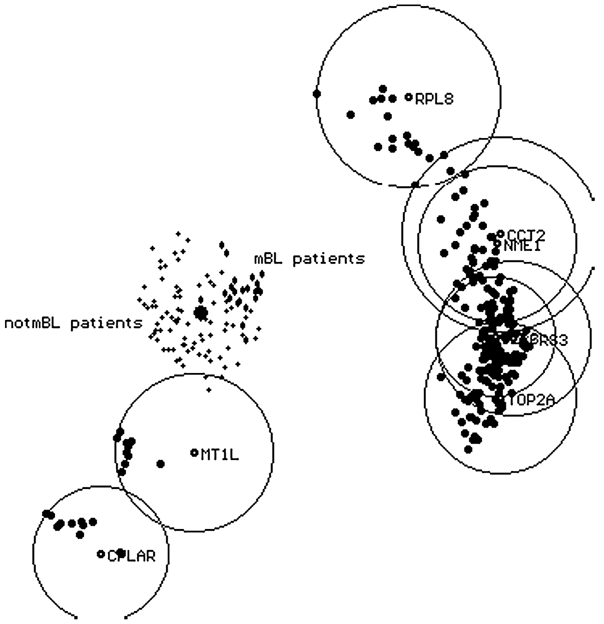

Figure 1 Biplot of the clinical gene expression data. Representation of 108 lymphoma patients and 8 disjoint gene sets that correspond to the c-myc overexpression. The circles are centered at the source genes and include all genes of the respective sets. The gene sets are denoted by their source genes:

200936_at	RPL8	ribosomal protein	19 genes
201577_at	NME1	non-metastatic cells 1, protein (NM23A) expr.	33 genes
201947_s_at	CCT2	chaperonin containing TCP1, subunit beta	19 genes
211375_s_at	ILF3	interleukin enhancer binding factor 3	12 genes
201292_at	TOP2A	topoisomerase (DNA) II alpha	63 genes
208673_s_at	SFRS3	splicing factor, arginine/serine-rich 3	71 genes
204326_x_at	MT1L	metallothionein 1L	11 genes
210564_x_at	CFLAR	CASP8 and FADD-like apoptosis regulator	12 genes.

the maximum set with the source gene for RPL8 (ribosomal protein L8) also contains, among others, 11 additional ribosomal protein genes (L3, L4, L10a, L14, L15, L18a, L36, S9, S16, S20, S21), two genes encoding eukaryotic translation initiation factors (EIF3S3, EIF3F6), and the eukaryotic translation elongation factor EEF1B2. Therefore, this gene set clearly represents the cellular protein synthesis machinery. Likewise, the maximum set characterized by the source gene TOP2A (topoisomerase (DNA) II alpha) includes a high number of genes coding for proteins implicated in cell cycle regulation, like the cyclins A2 and B2, the cyclin-dependent kinase CDC2, the ki-67 antigen, and the transcription factor E2F, to name only a few. Hence this gene set correlates closely with cell proliferation. Other significant gene sets are enriched in genes involved in mitochondrial functions, splicing, or apoptosis regulation.

As evident from Fig. 1, two gene subsets represented by the apoptosis (programmed cell death) inhibitor CFLAR (CASP8 and FADD-like apoptosis regulator/c-FLIP) and the metallothionein family of metal-binding proteins (MT1L), are clearly separated from the other ones. Interestingly, the mBL cases shift towards the six subsets to the right, indicating higher expression of those genes in mBL, whereas the other cases (notmBL) preferentially cluster close to the CFLAR and MT1L subsets. Hence, the selected gene sets are able to separate mBL and not-mBL patients on the basis of their expression levels. It is to be emphasized that our algorithm did not use the information on mBL diagnosis to produce and select the gene subsets.

4 Data compression and graphical representation

In this section, we will discuss how the relation between the individuals and the identified sets of variables can be visualized. For the application from Section 3, the clinical data matrix $\mathbf{X}_0$ with $n = 108$ patients and $p = 22\,277$ genes is used. We define an "individual coordinate"

$$k_{ji} = \sqrt{\frac{n}{n-1}} \frac{x_{ji} - \bar{x}_i}{\sqrt{(\mathbf{x}_i - \bar{\mathbf{x}}_i)'(\mathbf{x}_i - \bar{\mathbf{x}}_i)}}$$

for a patient j and a gene i. Moreover, we define an "individual set coordinate"

$$k_{jm} = \frac{\sum_{i \in m} k_{ji}}{\sqrt{\sum_{h \in m} \sum_{i \in m} r_{hi}}}$$

for a patient j and a gene set m, where r_{hi} in the denominator are the usual correlations of the genes

$$r_{hi} = \frac{(\mathbf{x}_h - \bar{\mathbf{x}}_h)'(\mathbf{x}_i - \bar{\mathbf{x}}_i)}{\sqrt{(\mathbf{x}_h - \bar{\mathbf{x}}_h)'(\mathbf{x}_h - \bar{\mathbf{x}}_h) \cdot (\mathbf{x}_i - \bar{\mathbf{x}}_i)'(\mathbf{x}_i - \bar{\mathbf{x}}_i)}}.$$

This construction is statistically justified because each k_{jm}^2 is beta distributed, $k_{jm}^2 \sim \tilde{\mathrm{B}}(\frac{1}{2}, \frac{n-2}{2})$, and the equation $\sum_{j=1}^{n} k_{jm}^2 = \frac{n}{n-1}$ is valid, if the rows of $\mathbf{X}_0$ are independently and identically distributed as $N_p(\boldsymbol{\mu}_0', \Sigma)$. These facts mean, that the individual coordinates k_{jm} of the sets m have no differences with respect to the scales and the covariance structures of the variables. The individual set coordinates k_{jm} form a matrix $\mathbf{K}$, so that each patient j receives a unique numerical characterization for each gene set m. In our example from Section 3, matrix $\mathbf{K}$ has the format 108×8.

To obtain graphical representations of the patients and the gene sets, matrix $\mathbf{K}$ is subjected to a two-dimensional projection with a smoothing effect. The n-dimensional eigenvalue problem

$$(\mathbf{K}\mathbf{K}')\mathbf{V} = \mathbf{V}\Lambda, \quad \mathbf{V} = (\mathbf{v}_1\ \mathbf{v}_2), \quad \Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \quad \mathbf{V}'\mathbf{V} = \mathbf{I}_2$$

is solved by the first two eigenvectors $\mathbf{v}_1, \mathbf{v}_2$ and the corresponding eigenvalues λ_1, λ_2. The projection is provided by $\mathbf{K}_{(2)} = \mathbf{V}\mathbf{V}'\mathbf{K}$. Thus, a compressed description of the individual gene expression of all sets m is obtained, which roughly shows the behaviour of the observed data. Here, $\mathbf{I}_2$ denotes the 2×2 identity matrix.

The two-dimensional biplot with respect to the gene sets is given by the matrix of two columns, $\begin{pmatrix} \mathbf{V} \\ \mathbf{K}'\mathbf{V} \end{pmatrix}$. The rows of $\mathbf{V}$ correspond to the patients, the rows of $\mathbf{K}'\mathbf{V} = \mathbf{K}'_{(2)}\mathbf{V}$ to the gene sets. We use this information to get an order of the patients and an order of the gene sets which correspond to the main tendencies of the gene expression: Each row of $\mathbf{V}$ and $\mathbf{K}'\mathbf{V}$ is standardized to norm 1, and it is put on the unit circle in the biplot plane. Then, a sorting of the patients and of the gene sets is obtained from the corresponding points on the unit circle, so that neighbouring patients and neighbouring gene sets are correlated as highly as possible. The largest gap on the unit circle between the patients and between the gene sets, respectively, marks the beginning and the end of the ordered patient or gene-set sequence. Fig. 2 shows the positions of all eight significant gene sets on the unit circle in our example.

This leads to the representation of the clinical expression data $\mathbf{X}_0$ in the heat map of Fig. 3. The ordered patients are drawn in the horizontal and the ordered gene sets in the vertical direction. Light fields indicate high gene expression, dark ones low expression. We can observe that the molecular Burkitt patients (dark) lie compactly on the left-hand side in the ordered sequence of patients. Also the order of the gene sets reflects well the similarities in their expression. Thus, large blocks of patients and gene sets are obtained, in which a uniformly high or uniformly low gene expression occurs.

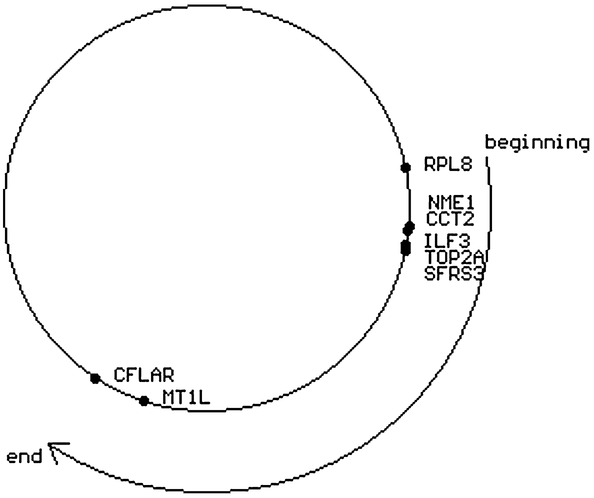

Figure 2 Ordered gene-set sequence on the unit circle in the biplot plane that corresponds to matrix **K**.

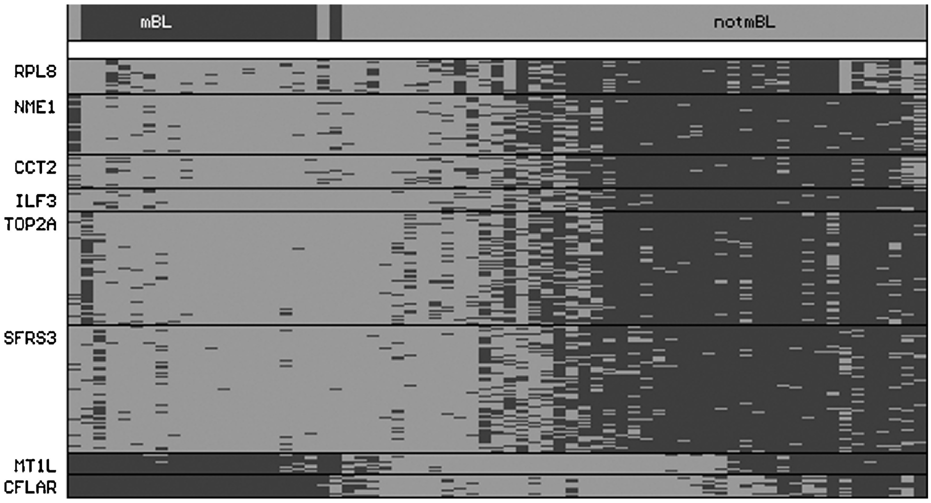

Figure 3 Heat map to show the expression of 240 genes belonging to 8 ordered gene sets in 69 ordered lymphoma patients. Light-grey = high expression, dark-grey = low expression.

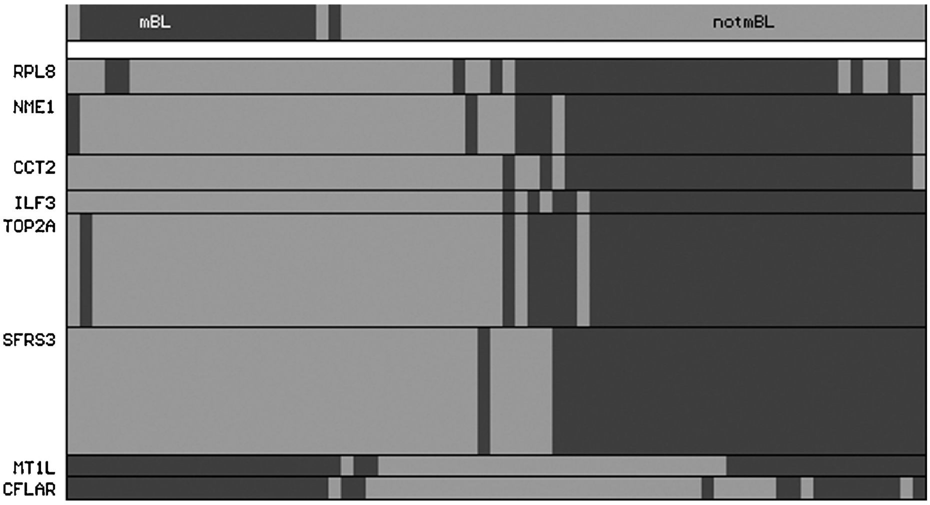

Figure 4 Expression as in Fig. 3, but in the representation by set coordinates (matrix **K**).

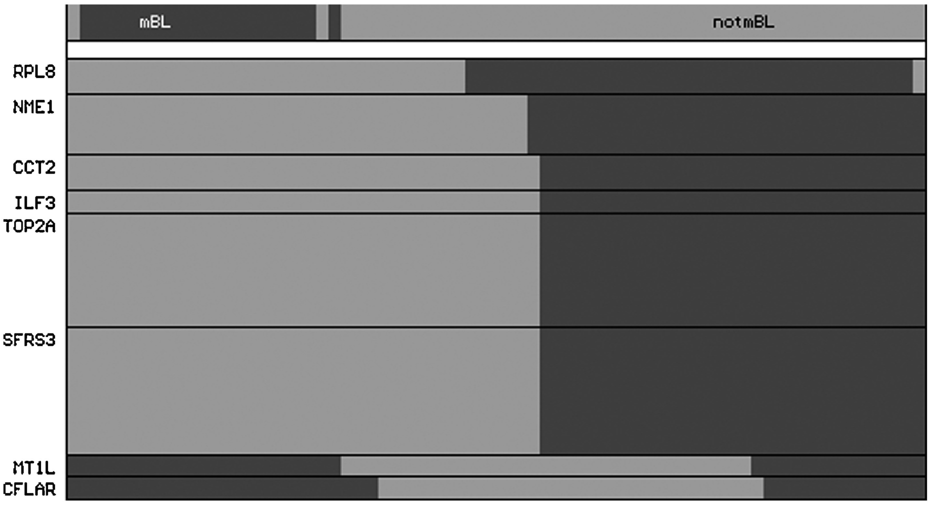

Figure 5 Expression as in Fig. 3, but in the representation by smoothed set coordinates (matrix $\mathbf{K}_{(2)}$).

The heat map of Fig. 4 provides the representation of the individual set coordinates **K** instead of the original gene expression values. The single genes are not depicted in this graph. Furthermore, Fig. 5 summarizes the data via the smoothed individual set coordinates $\mathbf{K}_{(2)}$. Here, only the main tendencies of the gene expression are visible.

The heat maps of Figs. 3–5 have been additionally modified in such a way that patients with a weak gene expression are removed. We apply the beta test

$$\frac{n}{n-1}(v_{j1}^2 + v_{j2}^2) \geq \mathrm{B}_{0.50}\left(1, \frac{n-3}{2}\right)$$

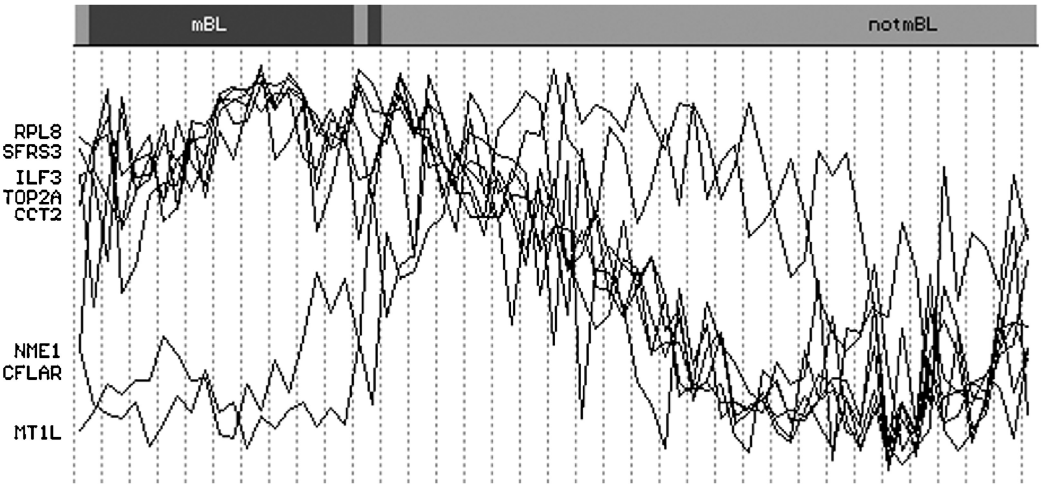

Figure 6 Expression curves of 8 gene sets in 69 ordered lymphoma patients (matrix **K**).

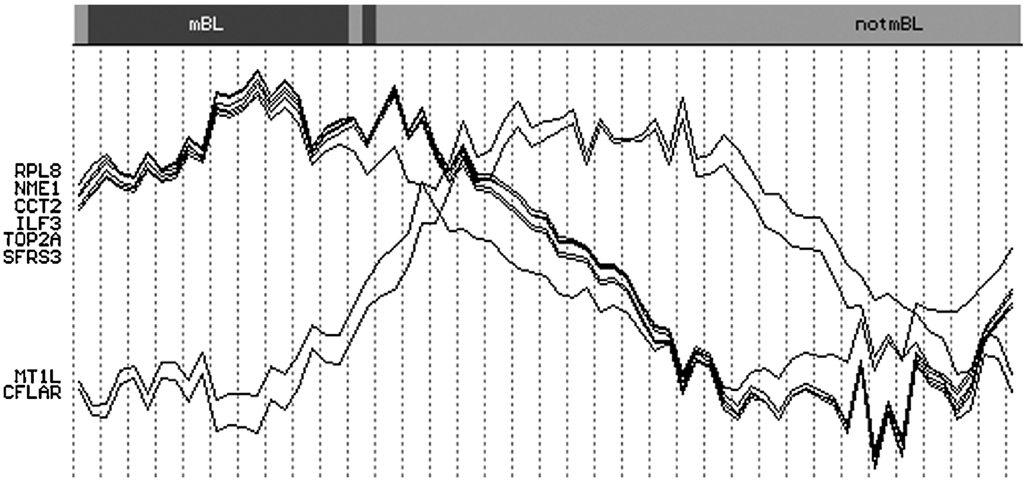

Figure 7 Smoothed expression curves of 8 gene sets in 69 ordered lymphoma patients (matrix $\mathbf{K}_{(2)}$).

to each of the patients $j = 1, ..., n$, where v_{j1} and v_{j2} are the elements of the jth row of **V**. Patients who are not significant are excluded. Thus, in the given example, the number of patients is reduced from 108 to 69.

A further way of illustrating the results of the analysis is to draw curves instead of the heat maps. Then the expression of the significant gene sets does no longer appear in the rough light-dark representation, but in a continuous curve representation. Fig. 6 shows the curves corresponding to **K** and Fig. 7 the smoothed curves corresponding to $\mathbf{K}_{(2)}$.

5 An algorithm of selection based on parametric tests

The preceding sections have shown that the non-parametric Westfall-Young procedure is an effective tool for high-dimensional selection of variables. This procedure has the advantage that an arbitrary test statistic F_m may be used which increases with increasing set m.

Since 1996 exact high-dimensional parametric tests are available (Läuter, 1996; Läuter *et al.*, 1996, 1998). This raises the question whether corresponding parametric selection procedures with high effectivity can also be developed from these methods. Sections 5 and 6 are devoted to this problem.

A basic algorithm for searching single significant variables in the comparison of $\boldsymbol{\mu}^{(1)\prime}$ and $\boldsymbol{\mu}^{(2)\prime}$ has been introduced by Kropf (2000). The p variables i are sorted according to decreasing values of the "total sums of squares" $(\mathbf{x}_i - \bar{\mathbf{x}}_i)'(\mathbf{x}_i - \bar{\mathbf{x}}_i)$. Then, univariate beta tests (or corresponding F tests) without an α adjustment are carried out in the obtained order, as long as significances result. This procedure of stepwise testing keeps the familywise type I error rate α (FWER) in the strong sense. However, this procedure is not scale invariant, because large total sums of squares are preferred. Nevertheless, this approach is sensible since large total sums of squares reflect large differences between the groups (1), (2) and since variables with very small fluctuations which could be artefacts are suppressed.

Some modifications of this procedure have been proposed: by Westfall, Kropf and Finos (2004), by Hommel and Kropf (2005), and by Läuter (2007). The modification by Hommel and Kropf uses tests on the stricter significance level α/k but, correspondingly, the sequence of the tests is stopped only when the kth non-significant result is obtained. Here, k is a fixed positive integer.

The application of this stepwise principle to sets of variables would demand that all candidate subsets are sorted according to an order criterion and are successively tested until non-significances appear. If the number of variables is very high and the subsets are not uniquely pre-specified, then this procedure becomes computationally expensive. In the past, we have tried to construct a suitable algorithm (Läuter *et al.*, 2005), but we could not yet find an acceptable method for high dimensions. Therefore, a further proposal is now presented.

We start once more from our basis totality of subsets $BAS(\mathbf{X}, c) = \bigcup_{i_1=1,\ldots,p} bas(i_1, \mathbf{X}, c)$, i.e. from the union over all partitions $bas(i_1, \mathbf{X}, c)$ defined by the source variables $i_1 = 1, \ldots, p$. To test a subset $m \in BAS(\mathbf{X}, c)$, the univariate beta statistic of the corresponding source variable i_1 is used:

$$\mathrm{B}_{i_1} = \frac{n^{(1)}n^{(2)}}{n^{(1)} + n^{(2)}} \frac{(\bar{x}_{i_1}^{(1)} - \bar{x}_{i_1}^{(2)})^2}{(\mathbf{x}_{i_1} - \bar{\mathbf{x}}_{i_1})'(\mathbf{x}_{i_1} - \bar{\mathbf{x}}_{i_1})}, \quad \mathrm{B}_{i_1} \sim \mathrm{B}\left(\frac{1}{2}, \frac{n^{(1)} + n^{(2)} - 2}{2}\right) \text{ for } \mu_{i_1}^{(1)} = \mu_{i_1}^{(2)}.$$

Thus, all subsets having the same source variable i_1 are assessed by the same test statistic. This is the crucial point in the procedure proposed here. For the exactness of the beta test, we assume that the rows of $\mathbf{X}$ have independent multivariate normal distributions:

$$\begin{aligned} \mathbf{x}_{(j)}^{(1)\prime} &\sim \mathrm{N}_p(\boldsymbol{\mu}^{(1)\prime}, \boldsymbol{\Sigma}), \quad j = 1, \ldots, n^{(1)}, \\ \mathbf{x}_{(j)}^{(2)\prime} &\sim \mathrm{N}_p(\boldsymbol{\mu}^{(2)\prime}, \boldsymbol{\Sigma}), \quad j = 1, \ldots, n^{(2)}. \end{aligned}$$

The order criterion is given by

$$O(m) = (\mathbf{x}_{i_1} - \bar{\mathbf{x}}_{i_1})'(\mathbf{x}_{i_1} - \bar{\mathbf{x}}_{i_1}) \sum_{i \in m} \mathrm{abs}(r_{i_1 i}).$$

It depends on the total sum of squares of the source variable i_1 analogously to Kropf's procedure. Additionally, the value of $O(m)$ increases with increasing size of the subset m, namely according to the total correlations $r_{i_1 i}$ between the source variable i_1 and the partner variables i (see Section 2). By this strategy, a source variable i_1 and the pertaining univariate beta test get a high weight if many highly-correlated partners i exist for i_1. This corresponds to our principles of multivariate stabilization. Thus, the seemingly univariate testing method becomes a properly multivariate method.

Still, the aim is to identify subsets m containing at least one differential variable. The application of our general searching method (Kropf and Läuter, 2002) requires that all subsets

$m \in BAS(\mathbf{X}, c)$ are evaluated in decreasing order of $O(m)$ as long as significances occur. An α adjustment is not necessary. All subsets recognized as significant in this stepwise process satisfy the conditions of a multiple testing procedure with the familywise type I error rate α. The testing procedure is based on the general rules of the spherical tests. If m_0 is the set of all variables i with $\mu_i^{(1)} = \mu_i^{(2)}$, then the first subset m with $m \subseteq m_0$ in the stepwise process is uniquely determined (with probability 1) by the total sums of products matrix belonging to m_0, $(\mathbf{X}_{m_0} - \bar{\mathbf{X}}_{m_0})'(\mathbf{X}_{m_0} - \bar{\mathbf{X}}_{m_0})$. Therefore, the decisive B_{i_1} test is also uniquely determined and, thus, the multiple level α of the procedure is exactly kept, according to the theorems in Läuter *et al.* (1996, 1998).

Under the conditions of the given special selection procedure, where the subsets m are assessed only by the univariate statistic B_{i_1}, a shortcut strategy can be applied. In this case, the first non-significant testing result will always appear at a maximum set M_{i_1} (see Section 2), because $O(m) \leq O(M_{i_1})$ for $m \in bas(i_1, \mathbf{X}, c)$.

Therefore, in practice, the p maximum sets should be tested separately until the first non-significant maximum set, $M_{i_1^{non}}$, occurs. Thus, the threshold value for the order criterion, $O(M_{i_1^{non}})$, is found. All preceding maximum sets, $M_{i_1^{(1)}}, M_{i_1^{(2)}}, M_{i_1^{(3)}}, \ldots$, prove to be significant. Furthermore, some particularly interesting subsets m contained in $bas(i_1, \mathbf{X}, c)$, $i_1 = i_1^{(1)}, i_1^{(2)}, i_1^{(3)}, \ldots$, can be investigated for significance besides the maximum sets $M_{i_1^{(1)}}, M_{i_1^{(2)}}, M_{i_1^{(3)}}, \ldots$. One has only to apply the condition $O(m) > O(M_{i_1^{non}})$. Additional requirements, for example, with respect to a small number of variables belonging to m can then also be fulfilled.

It should be noted that the above mentioned modification by Hommel and Kropf is valid in this procedure. However, the modifications by Westfall *et al.* (2004) and by Läuter (2007) can no longer be used.

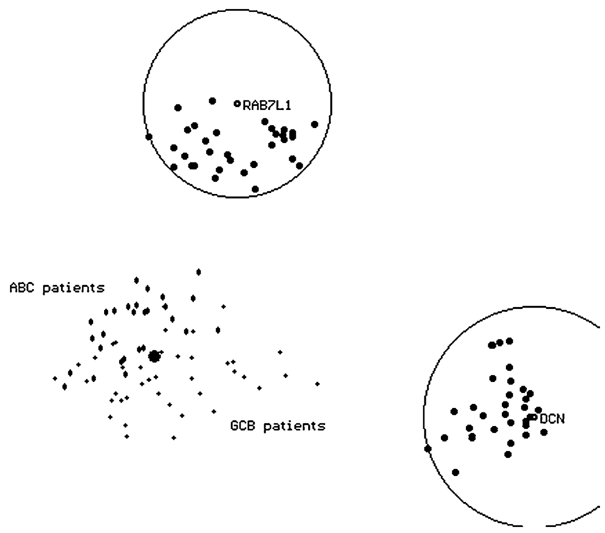

Figure 8 Biplot of the DLBCL patients. Representation of 68 patients and 2 disjoint gene sets with their circles of the included genes. The gene sets are denoted by the source genes:

218700_s_at	RAB7L1	RAB7, member RAS oncogene family-like 1	33 genes
211896_s_at	DCN	decorin	34 genes.

6 Application of the parametric selection procedure

Previously, DLBCL patients have been divided into two subgroups based on a molecular signature developed by Alizadeh *et al.* (2000) and Rosenwald *et al.* (2002). These two entities, the activated B-cell-like lymphomas (ABC) and the germinal center B-cell-like lymphomas (GCB) differ significantly with respect to prognosis. We will employ our selection procedure to find gene subsets that separate the patients of these lymphoma types. Expression data from 28 ABC and 40 GCB cases (Hummel *et al.*, 2006) are applied for this analysis.

The following special values of the parameters are used in the selection procedure:

1. Only the genes i whose sums of squares fulfill the conditions $(\mathbf{x}_i - \bar{\mathbf{x}}_i)'(\mathbf{x}_i - \bar{\mathbf{x}}_i) \geq 25$ are included in the proper algorithm of the procedure. Then 1433 genes remain from the given 22 283 genes.
2. The multiple level of significance is $\alpha = 0.05$.
3. The generation of the gene subsets is performed under the condition $r^2_{i_1 i} \geq 0.40$. Additionally, the positivity of the correlations $r_{i_1 i}$ is required.
4. A gene i is assigned to the source gene i_1 only if gene i has a smaller sum of squares than gene i_1.
5. The correlation sum $\sum_{i \in m} \text{abs}(r_{i_1 i})$ of the order criterion $O(m)$ is calculated only from the 30 highest correlations (if more than 30 genes are present).

These modifications do not impair the maintenance of the multiple significance level.

The parametric selection algorithm yields 13 significant gene maximum sets that are able to discriminate ABC from GCB cases. One maximum set shows higher expression in ABC than in GCB, 12 maximum sets have a lower expression in ABC than in GCB. The testing modification by Hommel and Kropf identifies 15 significant maximum sets if the number $k = 2$ is used. When the additional condition of disjointed sets is applied, only two gene maximum sets remain.

The first gene set characterized by the source gene RAB7L1 (RAB7, member RAS oncogene family-like 1) includes, among others, the trancription factor STAT3 and several of its target genes, all related to apoptosis control. The other gene set represents a remarkable collection of genes encoding proteins of the extracellular matrix, including DCN (decorin), collagens and fibronectin.

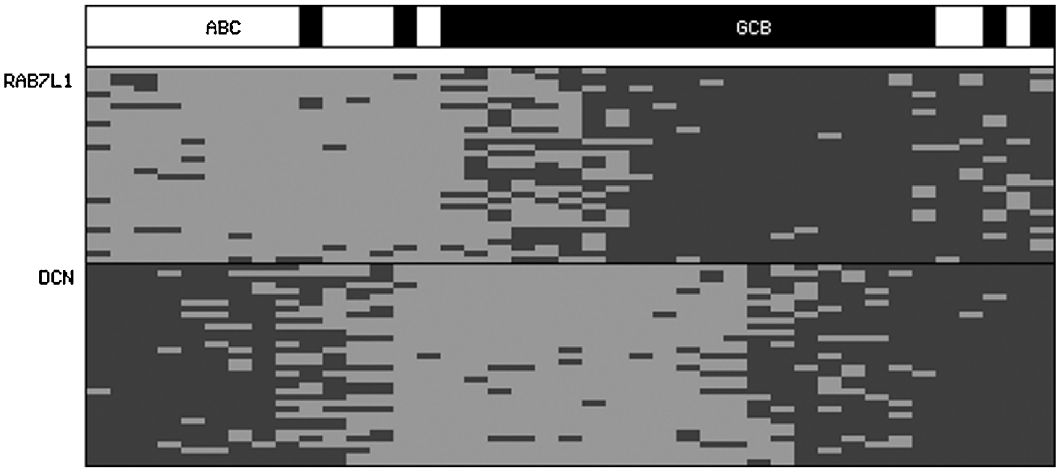

Figure 9 Heat map to show the expression of 67 genes belonging to two disjoint gene sets in 41 ordered DLBCL patients. Light-grey = high expression, dark-grey = low expression.

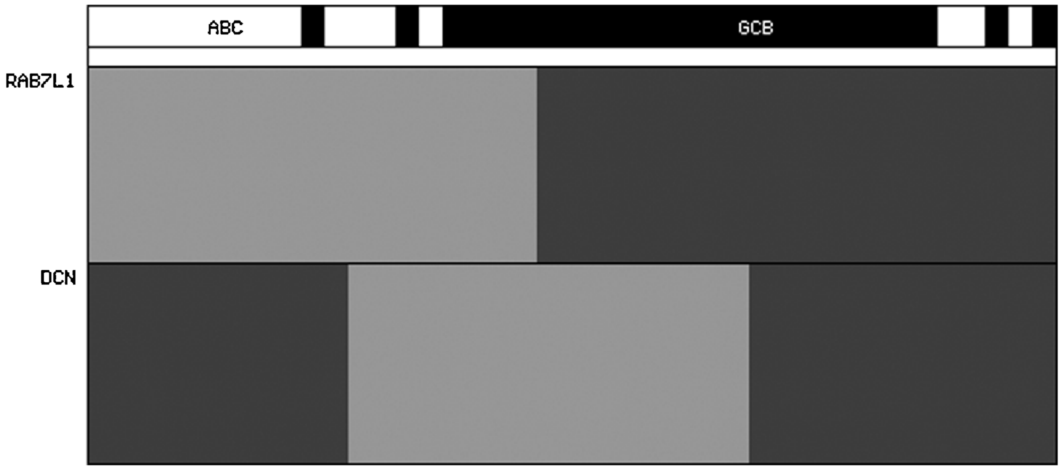

Figure 10 Expression as in Figure 9, but in the representation by set coordinates (matrix $\mathbf{K} = \mathbf{K}_{(2)}$).

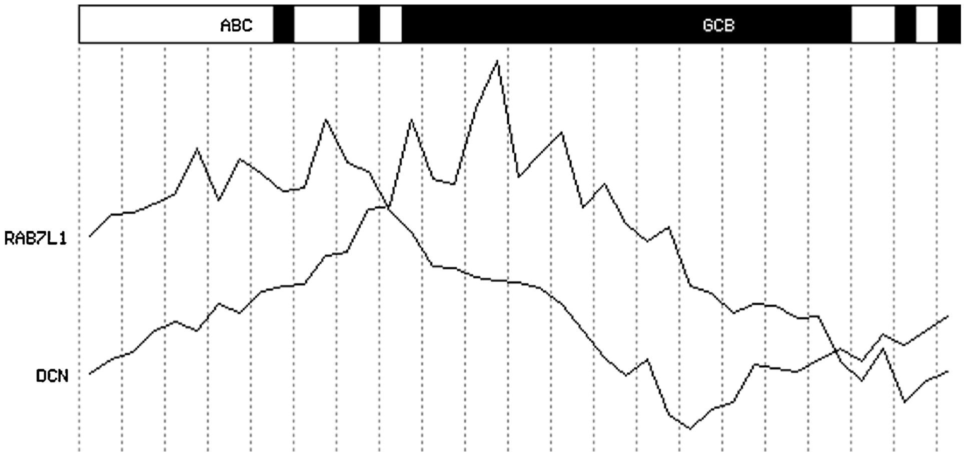

Figure 11 Expression curves of two gene sets in 41 ordered DLBCL patients (matrix $\mathbf{K} = \mathbf{K}_{(2)}$).

Fig. 8 reveals the biplot of both gene sets and the 68 patients. In Figs. 9, 10 and 11, the corresponding heat maps and the expression curves are presented. In the same way as in Section 4, the ordered sequence of patients, with reduction to the essential part (41 instead of 68), is used.

As is apparent from Figs. 9–11, the RAB7L1 subset discriminates quite precisely between ABC and GCB cases (ABC > GCB), whereas the DCN subset is low in most ABC cases but subdivides the GCBs in high and low expressors. It is intriguing that according to our recent analyses, this latter subset of extracellular matrix genes is also able to discriminate between patients of different survival prognosis (manuscript in preparation).

Taken together, these studies demonstrate that our algorithms allow to identify biologically meaningful sets of correlated genes in tumor gene expression data bases and to use such gene sets to discriminate between different disease entities.

7 Conclusion

This paper presents effective and mathematically exact algorithms of selection of variables which are also applicable in cases with a very large dimension. Both non-parametric and parametric strategies are used. The particular requirements of the statistical stability are taken into account, so that high power is attained and the familywise type I error can be kept in spite of the large dimension. The methodology associates biological hypothesis generation and biological interpretation with mathematical and computational conciseness. The `R` package `SETGEN` implementing our Westfall-Young procedure is available at http://www.people.imise.uni-leipzig.de/maciej.rosolowski/software.html

Acknowledgements The authors are grateful to the referees for their valuable comments. These have much helped to improve and to complete the paper.

Conflict of Interests Statement
The authors have declared no conflict of interest.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X. *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511.

Anderson, J. A. and Blair, V. (1982). Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika* **69**, 123–136.

Bild, A. H., Yao, G., Chang, J. T., Wang, Q. Potti, A., Chasse, D., *et al.* (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353–357.

Boxer, L. M. and Dang, C. V. (2001). Translocations involving c-myc and c-myc function. *Oncogene* **20**, 5595–5610.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2002). Multiple hypothesis testing in microarray experiments. *The Berkeley Electronic Press*, 2002, paper 110.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32**, 407–499.

Furnival, G. M. and Wilson, R. W. (1974). Regressions by leaps and bounds. *Technometrics* **16**, 499–511.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58** 453–467.

Goeman, J. J. and Mansmann, U. (2008). Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* **24** 537–544.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning; data mining, inference, and prediction*. Springer, New York.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

Hommel, G. and Kropf, S. (2005). Tests for differentiation in gene expression using a data-driven order of weights for hypotheses. *Biometrical Journal* **47**, 554–562.

Hummel, M., Bentink, S., Berger, H., Klapper, W. Wessendorf, S., Barth, T.F.E.,*et al.* (2006). A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. The NEW ENGLAND JOURNAL of MEDICINE **354**, 2419–2430.

Kropf, S. (2000). Hochdimensionale multivariate Verfahren in der medizinischen Statistik. Shaker, Aachen.

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. (2007). Hochdimensionale Statistik, Anwendung in der Genexpressionsanalyse. Leipzig Bioinformatics **15**, ISSN 1860–2746.

Läuter, J., Glimm, E. and Eszlinger, M. (2005). Search for relevant sets of variables in a high-dimensional setup keeping the familywise error rate. *Statistica Neerlandica* **59**, 298–312.

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. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**, 1436–1462.

Rosenwald, A., Wright, G., Chan, W. C., Connors J. M. Campo, E., Fisher, R.I., Gascoyne, R.D., ler-Hermelink, H.K., Smeland, E.B., Giltnane, J.M., *et al.* (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. The NEW ENGLAND JOURNAL of MEDICINE **346**, 1937–1947.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*, **58**, 267–288.

Tibshirani, R. and Wasserman, L. (2006). Correlation-sharing for detection of differential gene expression. arXiv:math/0608061v1 [math.ST] 2 Aug 2006. http://dx.doi.org/10.1002/bimj.200800207.

Wasserman, L. and Roeder, K. (2007). Multi-stage variable selection: Screen and clean. arXiv:0704.1139v1 [math.ST]. http://dx.doi.org/10.1002/bimj.200800207.

Westfall, P. H., Kropf, S. and Finos, L. (2004). Weighted FWE-controlling methods in high-dimensional situations, in: Benjamini, Y., Bretz, F. and Sarkar, S. K. (eds.), *Recent Developments in Multiple Comparison Procedures, IMS Lecture Notes– Monograph Series* **47**, 143–154.

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