Molecular BioSystems

Cite this: *Mol. BioSyst.*, 2012, **8**, 2339–2346

www.rsc.org/molecularbiosystems

PAPER

Published on 13 June 2012. Downloaded by Pennsylvania State University on 18/05/2016 05:44:05.

Thresholding for biomarker selection in multivariate data using Higher Criticism†

Ron Wehrens* and Pietro Franceschi

Received 29th March 2012, Accepted 7th June 2012
DOI: 10.1039/c2mb25121c

Biomarker selection is an important topic in the omics sciences, where holistic measurement methods routinely generate results for many variables simultaneously. Very often, only a small fraction of these variables are really associated with the phenomena of interest. Selection and identification of these biomarkers is essential for obtaining an understanding of the complex biological processes under study. Finding biomarkers, however, is a difficult task. Even if a relative order can be established, *e.g.*, on the basis of *p* values, it is usually hard to determine where to stop including candidates in the final set. Higher Criticism is an approach for finding data-dependent cutoff values when comparing two distinct groups of samples. Here, we extend its use to multivariate data, providing a principled approach to compromise between not selecting too many variables and catching as many true positives as possible. The results show a marked improvement in biomarker selection, compared to the standard settings available for some methods. Interestingly, HC thresholds can differ considerably from what has been suggested in literature before, again showing that it is not possible to use the same cutoff value for all data sets. The data-specific cutoff values provided by HC also open the way to more fair comparisons between biomarker selection methods, not biased by unlucky or suboptimal threshold choices.

1 Introduction

In many areas in the life sciences one is comparing measurements on two populations, with the specific goal to identify those variables that allow to distinguish the two – these are commonly indicated by the term "biomarkers". Examples can be found in transcriptomics,[1,2] proteomics[3,4] and metabolomics,[5] to name but a few. The question is what difference is significant. In early days of microarrays, the typical approach was to look for at least a two-fold difference, but this kind of approach ignores differences in variability. Currently, several different forms of *t* tests (see, *e.g.*, Zuber and Strimmer[6]), and approaches specifically developed for particular data types are often used. An example in the area of expression data is the Significance Analysis of Microarrays, SAM.[7] They have the advantage that they are very simple to interpret and require little or no tuning. Just like for any other form of statistical testing, however, applying these tests many times will, by definition, lead to cases where low *p* values are obtained purely by chance. Often, multiple-testing corrections are employed to decrease the number of False Positive (FP) decisions – a popular example is the False Discovery Rate (FDR) correction.[8,9]

Whether or not a multiple-testing correction is employed, a cutoff for the *p* values needs to be chosen above which the null hypothesis is rejected. Historically, $\alpha = 0.05$ has become synonymous with a significant difference, but this is a rather arbitrary convention, and especially in fields where the number of variables is large, such as in microarray analysis, other *p* values are used regularly. In practice, multiple-testing corrections are often too strict[10] and, although they successfully decrease the number of false positives (FPs), the number of true positives (TPs) usually suffers. Adaptive forms of multiple-testing correction (see, *e.g.*, Reiner *et al.*,[9] Storey and Tibshirani[10] and references therein) tackle this by at least roughly estimating the number of non-null hypotheses. However, this estimate is not the main goal of these techniques and serves primarily to calculate more accurate corrected statistics; the final selection of putative biomarkers will still be made using a standard cutoff value like $\alpha = 0.05$ or $\alpha = 0.01$.

Recently, a more principled approach, called Higher Criticism,[11,12] has been suggested, choosing the cutoff on the basis of the data at hand. The method has shown to work well in a number of applications and has some attractive properties, amongst which the ease of use stands out: there is only one tuning parameter which affects the results very little. It is based on the assumption that the real differences between the classes are rare – the number of biomarkers is relatively small – and

Centre for Research and Innovation, Fondazione Edmund Mach., Via E. Mach 1, San Michele all'Adige (TN), Italy. E-mail: ron.wehrens@fmach.it; Fax: +39 0461 615200; Tel: +39 0461 615563

† Electronic supplementary information (ESI) available: Complete results from simulation experiments. See DOI: 10.1039/c2mb25121c

Published on 13 June 2012. Downloaded by Pennsylvania State University on 18/05/2016 05:44:05.

weak, indicating that class differences are small and hard to detect individually.

The problem of selecting a suitable threshold also arises in multivariate approaches, such as regularized discriminant analysis,[13] Principal Component Linear Discriminant Analysis (PCLDA),[14] Partial Least Squares Discriminant Analysis (PLSDA)[15] and the related Variable Importance of Projection values (VIP).[16] These multivariate methods, popular in many omics sciences, in principle may lead to better biomarker selections since they benefit from explicitly incorporating correlation structure. Putative biomarkers then are indicated by those variables that have a large influence on the model, *i.e.* show large absolute values of regression coefficients (PCLDA and PLSDA) or large VIP values. Selecting a threshold in these cases can be even more difficult than in the univariate case: whereas p values at least have one common range, from zero to one, and a more or less clear interpretation, regression coefficients depend on the units and scaling of the data, as well as on the number of variables and their correlations, so providing one universal cutoff value is impossible. For the VIP, an empirical rule of thumb has been to include variables with a VIP value larger than one,[16] but there is no theoretical justification for this. Even though one study indeed has found optimal values close to one for VIP cutoffs,[17] these values depend strongly on the data and the data pretreatment, and different data sets may require very different thresholds.

Here, we extend the Higher Criticism (HC) approach for application with multivariate methods like PCLDA, PLSDA and the VIP. We propose a permutation approach to obtain p values for model coefficients (PCLDA, PLSDA) and VIP statistics, after which the usual HC thresholding can be applied. We demonstrate the potential of the HC thresholding approach in these cases using experimental metabolomics data of spiked apple extracts,[18] where the differences between the groups are known beforehand. The use of this kind of spike-in data is essential: especially in complex biological data, there always a risk of overinterpreting differences that are found by a selection method, and therefore one must rely on real experimental data where the true differences are unambiguous. For the metabolomics data described in this paper, and simulations that are based on them, the results are very convincing – the HC threshold is shown to be able to identify cutoff points that are very close to the optimal ones. For data with more unfavourable sample-to-variable ratios, such as microarrays or next-generation sequencing data, the technique should still be evaluated, preferably using real-life spike-in data with several replicates per class.

Although we are not comparing different biomarker selections methods in the current paper, it is important to mention that the identification of data-dependent cutoff levels is essential in making such comparisons. The HC cutoffs provide a consistent way of fine-tuning different biomarker selection methods so that differences between the methods then are due to method characteristics and not to an unlucky choice of a threshold for one particular method. The final conclusion of the paper is that HC thresholds can be easily combined with other methods to identify biomarkers: HC threshold selection is basically independent of the primary biomarker selection method.

2 Theory

We focus on the situation that relevant differences need to be found between two classes, where the number of features is much larger than the class sizes. This situation is typical for many of the omics sciences, perhaps most obviously so in transcriptomics where the number of samples typically is two to three orders of magnitude smaller than the number of genes. In fields like metabolomics and proteomics the sample-to-variable ratio is less unhealthy but typically less than 0.1.

In the next paragraphs we present background on multivariate biomarker selection approaches and HC thresholding. The novel contribution of this paper is described in paragraph 2.3, explaining how to apply the HC approach in cases where no explicit p values are obtained from the primary selection method.

2.1 Multivariate biomarker selection

Since the covariance matrices of most omics data sets are singular (the number of variables usually exceeds the number of samples by a fair amount), Linear Discriminant Analysis, the simplest possible classification method, cannot be directly applied. One approach, taken in both PLSDA and PCLDA, is to compress the information in the data into a small number of latent variables (LVs), and perform the discriminant analysis on these. Usually, the problem is tackled in one step by formulating it in a regression context, with the dependent variable Y taking on values of either 0 or 1 (in a two-class problem):

$$Y = XB + \mathscr{E} \approx TP^{\mathrm{T}}B + \mathscr{E} \tag{1}$$

where $\mathscr{E}$ is the matrix of residuals. Matrix X is decomposed into a score matrix T and a loading matrix P, both consisting of a very low number of latent variables, typically less than ten or twenty. The coefficients for the scores, $A = P^{\mathrm{T}}B$, can easily be calculated, since the cross product matrix of the scores is not singular, by least mean squares regression:

$$A = (T^{\mathrm{T}}T)^{-1}\, T^{\mathrm{T}}Y \tag{2}$$

which by premultiplication with P lead to estimates for the overall regression coefficients B:

$$B = PA \tag{3}$$

Large values in the regression vector B indicate those variables that are important for the discrimination between the two classes. The question is where to put the cutoff to find which of the coefficients should be classified as putative biomarkers, and which ones should not.

The difference between PCLDA and PLSDA lies in the decomposition of X. In PCLDA, T and P correspond to the scores and loadings, respectively, from Principal Component Analysis (PCA). That is, the class of the samples is completely ignored, and the only criterion is to capture as much variance as possible from X. In PLSDA, on the other hand, the scores and loadings are taken from a PLS model and the decomposition of X *does* take into account class information: the first PLS components by definition explain more, often much more, variance of Y than the first PCA components. Tuning the PLSDA and PCLDA models, *i.e.* choosing the optimal number of LVs, is usually done by methods like crossvalidation.

This journal is © The Royal Society of Chemistry 2012

Published on 13 June 2012. Downloaded by Pennsylvania State University on 18/05/2016 05:44:05.

The VIP statistic is derived from the PLS components and basically summarizes the (squared) loadings for the p variables, weighted with the amount of variance explained by each LV, v_a:

$$\mathrm{VIP}_A = \sqrt{p\left(\sum_a^A P_a^2 v_a\right) \Big/ \sum_a^A v_a} \qquad (4)$$

where VIP_A is the vector of VIP values for all variables, obtained with A LVs, and P_a is the ath loading vector, normalized to a length of one.[17] Note that because of the weighting with v_a, VIP values tend not to change too much after the first few LVs. As a result, VIP-based biomarker selections can be expected to be more robust than selections relying on the size of regression coefficients.

2.2 Higher Criticism

Higher Criticism, a term first used by John Tukey and later formalized and generalized by Jin and Donoho,[11,12] is a second-level form of significance testing, basically comparing the number of significant differences found with the number expected under the null hypothesis. Under this joint null hypothesis of not a single difference, p values are uniformly distributed. For a given number of p values, the distributions of the smallest, the one-but-smallest, *etc.*, are given by beta distributions which for a large number of samples can be conveniently described by normal distributions: the order statistic p_i (the ith value out of N sorted p values) is approximately normally distributed with mean i/N and variance $i/N(1 - i/N)$. From this, the HC statistic can be defined:[12]

$$\mathrm{HC} = \max \frac{\sqrt{N}(i/N - p_i)}{\sqrt{i/N(1 - i/N)}}$$

where the maximum is taken over i values between 0 and a fraction α_0 of N. Typically, $\alpha_0 = 0.1$. The form of the statistic is familiar: the difference between the actual and expected value, divided by the square root of the variance. In other words, the HC objective is the "z score of the p value", indicating the largest deviation of the expected behaviour of the p values. The test will select the first k variables, where k corresponds to the location of the maximum of the HC statistic in the chosen range of $[0, \alpha_0]$, and is specifically devised to detect a small fraction of true differences in the presence of a large number of no-difference situations. Note that HC threshold selection should not be performed on p values corrected for multiple testing: their distribution under the null hypothesis is different.

2.3 HC cutoffs for model coefficients and VIP values

It has been shown that HC thresholding works very well in combination with t testing on a number of "standard" data sets.[12] However, the p values on which the HC statistic is based need not necessarily come from t tests,[11] but can be applied to all cases where p values are obtained which under the null hypothesis of no difference show a uniform distribution. In this paper, we further extend the applicability of the approach for those cases where no p values are obtained by the primary variable selection method. This is achieved by generating approximate p values using permutation testing, so that HC thresholding can also be applied to statistics like regression coefficients. In particular, we show that cutoff values can be obtained for PCLDA and PLSDA model coefficients as well as for the VIP measure, all of which are popular biomarker selection methods in the omics sciences.

Basically, there is no distribution theory on model coefficients from PLSDA or PCLDA, nor for VIP values, which could be used to derive p values – the distribution of such coefficients depends strongly on the distribution of the original data. Therefore we resort to simulation. We use permutations of the class response variable to set up null distributions for the regression coefficients of PLSDA and PCLDA on one hand, and VIP coefficients on the other hand. Then, p values are determined by comparing the experimentally found coefficients with the null distributions, for each variable separately. These are simply the fraction of values of the null distribution that is larger in size (in absolute sense). As an example: if 210 out of 100 000 permutations lead to a regression coefficient whose absolute value is larger than the one found for the real, experimental data, then the p value for this variable is $210/100\,000 = 0.0021$. It should be noted that for a meaningful application of the HC criterion the p values should have sufficient granularity. If the number of permutations would be too small, too many variables would have p values that are exactly the same. In this paper, a number of 100 000 permutations is used for the analysis of the real data, and 10 000 permutations are used for the simulated data sets. An advantage of this simulation approach is that it is independent of the scaling of the data, and can also be applied to data that have not been standardized to zero mean and unit variance. For example, in metabolomics, one popular form of scaling is Pareto scaling,[19] where every column is mean-centered and then divided by the square root of the standard deviation, rather than by the standard deviation itself.

Note that the permutation strategy is not applicable for the case of PCLDA with only one latent variable: since the compression step only takes into account X, the LDA step will always use the same latent variable in which the original variables have exactly the same ratios. In practice, PCLDA with one latent variable will not work well, and we will not consider it further here. For PLSDA and the VIP using one latent variable, the simulation of the null distribution *will* work since the compression takes into account class information as well, and scrambling the class information will lead to other loadings. The p values found in this way can be handled by Higher Criticism just like the p values from, *e.g.*, t tests. This enables us to find which regression coefficients or VIP values correspond to variables that are associated with real differences between groups.

3 Data

3.1 Spike-in apple data

In order to test biomarker selection algorithms, it is crucial that the data fulfill two conditions: they should be realistic, *i.e.*, they should be completely similar to true, experimental data, and it should be known exactly and in advance which differences are "true" differences. The first condition should

Published on 13 June 2012. Downloaded by Pennsylvania State University on 18/05/2016 05:44:05.

provide a reality check: even if a particular method performs well for data from a certain theoretical distribution, it does not mean that it will yield good results in practice. The second condition should guard against overinterpretation: it is quite easy to see meaning in differences that are present only by chance. This is especially true when the number of variables is high, as currently is the case in most omics data sets. In practice, the only way to obtain such data is to create spike-in data sets, in which specific differences have been introduced *prior to the physical measurement*. Several such data sets are available in the microarray field.[20,21] For other omics data, far fewer data sets have been published.

Here, we use a metabolomics data set from spiked apples.[18] In total, nine chemical compounds have been spiked in extracts from ten Golden Delicious apples – three different sets of concentration were used. Extracts from ten other Golden Delicious apples were used as controls. Raw MS data, measured in positive ionization mode (measuring the positively charged ions produced in the electrospray ionization of the sample) as well as in negative ionization mode (for the negative ions), have been exported to CDF format, and analyzed with XCMS[22] and CAMERA[23] according to the settings given elsewhere.[18] In total, 1632 features have been identified in the positive-ionization mode data, and 995 in the negative ionization mode. The spike-in compounds have been measured separately as well, and these data have been treated in exactly the same way as the data from the apple extracts. In total, 22 features in both positive and negative ionization modes could be related to features from the spiked-in compounds. One spike-in compound, quercetin, could not be identified due to both its low concentration and unfavorable measurement conditions.[18] Note that even when not all features of the spike-in compounds are selected as biomarkers, it may very well be possible that all compounds themselves are represented in the list of significant differences between the two classes.

3.2 Simulated data

Further insight in the behaviour of the HC threshold selection in a multivariate context can be obtained by simulation. Not only does this give an idea of the robustness of the method and the spread that can be seen in the results, it also prevents one lucky combination of data and method to lead to overly optimistic assessments.

Two approaches can be found in the literature to generate such data sets. The first is to make assumptions on the (joint) distributions of the variables, such as multivariate normality. Different correlation scenarios can be investigated. Unfortunately, it is unclear how close such simulations are to real data: testing multivariate normality is not a realistic option given the low object-to-variable ratio usually encountered in the omics sciences.

Here, simulated spiked data sets are generated from the spike-in apple data by randomly selecting a set of variables that will act as biomarkers and multiplying the corresponding columns in the data matrices with predefined constants. Put differently, the matrix X of experimental data (without spiked variables) is postmultiplied by a diagonal matrix S, where certain elements of the diagonal of S have values larger than one:

$$X' = XS$$

These artificially spiked data are then compared with the control set in exactly the same way as is done with the real data. This strategy has been applied in literature before (see, *e.g.*, Meinshausen and Bühlmann,[24] where this kind of data is indicated with the term "real" data sets) and has the advantage that no assumptions of multivariate normality are made.

Simulated concentration levels are again at 120, 140, and 200% of the original concentrations, just like in the experimental data; before the simulations, the features associated with the real spiked-in compounds were removed from the data matrices. The number of artificially increased variables is, again, twenty-two. Again, the comparison in each case is between the control group (unchanged from the experimental set, save the removal of the real spike-in features) and one of the artificially spiked-in sets. Biomarkers were chosen randomly so that the differences between groups one, two and three only consist of the differences in the starting data. One hundred data sets were simulated for each situation.

4 Software

The techniques described in this paper as well as the spike-in apple data are part of the *BioMark* package[25] for R,[26] available from the CRAN repository (http://cran.r-project.org). The package implements the extraction of biomarkers on the basis of a variety of *t*-statistics (from the *st* package[6]), PLSDA regression coefficients and VIP values (using package *pls*[27]), as well as methods to find optimal cut-off values for them, both based on the Higher Criticism approach (this paper) and on an alternative approach, called stability selection, assessing the coefficient stability under perturbation of the data.[24,28]

5 Evaluation of results

For the real apple data as well as for every individual simulated set, six different comparisons are made: comparing groups one, two and three against the controls for both positive and negative ionization mode data. In total, twenty-eight models are considered for each combination of spike-in data and controls. As a univariate approach, the usual two-sample *t* test assuming equal variances is employed. In addition, PCLDA, PLSDA and VIP models, each taking into account two to ten latent variables are considered. For brevity, we will focus on the results of two, four and ten latent variables only – these will show the trends with sufficient clarity. In every case, putative biomarkers are identified by the HC-thresholding methodology. For the student-*t* and VIP statistics, we also calculate putative biomarkers given by the usual cutoffs of 0.05 and 1, respectively.

The results of all biomarker selections will be evaluated in the familiar terms of True and False Positives, and True and False Negatives (TNs and FNs). Graphically, this can be conveniently visualized in a Receiver-Operator Characteristic (ROC) curve. For one particular selection, one plots the fraction of false-positive features FP/(FP + TN) on the *x* axis, and the fraction of true biomarkers that is found by the selection method (TP/(TP + FN)) on the *y* axis. Another quality assessment is made by the "selection efficiency", also

View Article Online

Published on 13 June 2012. Downloaded by Pennsylvania State University on 18/05/2016 05:44:05.

known as the "positive predictive value" (PPV). This is defined as the fraction of selected features that corresponds to TPs: TP/(TP + FP). Especially if follow-up experiments are costly or time-consuming, and one cannot afford many false positive selections, this is an important criterion.

6 Results and discussion

6.1 Real apple data

6.1.1 Standard cutoff values. Fig. 1 shows the results of *t* testing and VIP analysis using the "standard" cutoff values of 0.05 and 1, respectively. In interpreting this type of figures, one should keep in mind that the *x*-axis corresponds to the fraction of the total number of variables: a value of 0.1 therefore corresponds to 100 variables for negative ionization data and 162 variables for positive ionization data. In contrast, the *y*-axis range from zero to one corresponds to the twenty-two true biomarkers. Several observations can be made immediately. The VIP consistently selects more features than the *t* tests, somewhere around 30% of the total number of features. In return, it usually also finds more true positives: for the negative mode around twenty (out of twenty-two, more than 80%), and for the positive mode between fifteen and twenty. The number of latent variables shows no consistent pattern and is of minor importance, which is in agreement with expectations – since the VIP is a weighted sum of loadings with the weights given by the explained variance, the addition of later latent variables will only cause minor changes in VIP. Clearly, the default cutoff value of one for the VIP in this case is not strict enough: selecting one third of all variables as biomarkers in many cases is not useful.

The regular *t* test selects approximately ten percent of all features, which is a lot better than the VIP – however, the number of TPs is also lower. The difference in the latter is not so big, however, and overall one probably could say that the *t* test is more efficient than the VIP using the standard cutoffs. Performing a multiple testing correction on *p* values (using the same $\alpha = 0.05$ cutoff) leads to a fraction of selected features less than five percent, but this time at the expense of a drastic decrease in the number of true positives, in agreement with observations from literature.[10] Again, one feels that it may be possible, with a more suitable selection of the cutoff level, to strike a better balance between FPs and TPs. From these results, it follows that comparing the *t* test and the VIP as biomarker selection methods is difficult since in both cases the threshold is chosen arbitrarily.

6.1.2 HC-thresholding. The efficiency of the HC thresholds in picking up the true biomarker features is summarized in Fig. 2 for a number of models. Especially for the VIP models, but also for the *t* tests, the number of selected variables has decreased dramatically upon the application of the HC threshold. The VIP selections have decreased from a massive 30% – corresponding to a selection containing hundreds of variables – to less than 5%, *i.e.*, several tens of variables. The size of the selection based on *t* statistics decreases from approximately 10% to 5%.

Using the HC cutoffs, PCLDA is the method with the most generous selection. Indeed, in most cases the maximal number of variables is selected, corresponding to 10% of the variables (determined by α_0, the one parameter of the HC method). Nevertheless, its number of TPs is not higher than that of the other methods. Overall, PCLDA seems to be the worst of the methods considered. It also shows the largest influence of the number of latent variables, not surprising given that the latent variables are constructed without taking class information into account. For the subsequent LDA step to be able to pick up relevant information, one would expect more components to lead to more meaningful models, and this is clearly the case here: for PCLDA, more components means a better performance. Selection on the basis of PLSDA coefficients leads to very low numbers of selected variables, especially in the negative ionization data. Both the VIP and PLSDA show a

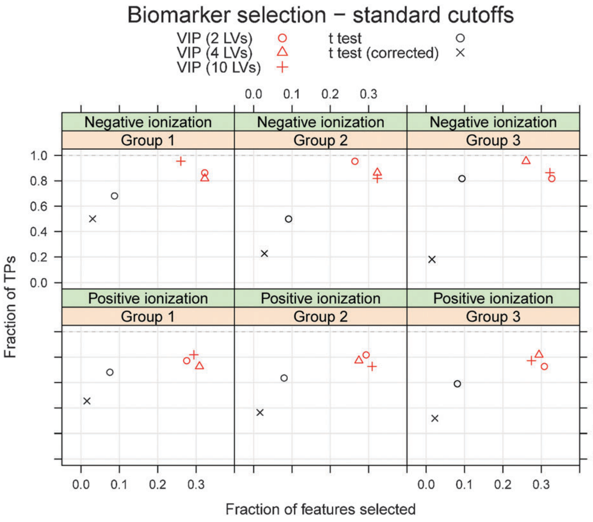

Fig. 1 Biomarker selection results for VIP (red symbols) and *t* tests (black symbols), using standard cutoff values of 1 and 0.05, respectively. The *x* axis shows the fraction of features selected; the *y* axis the fraction of true biomarkers that is found by the selection method.

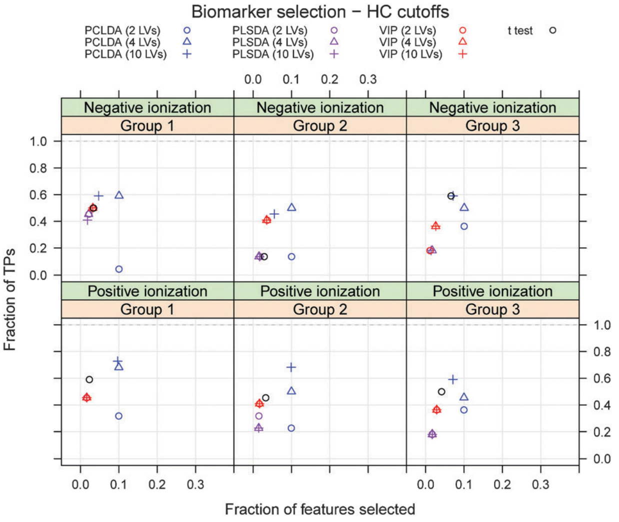

Fig. 2 Biomarker selection results using HC cutoffs. Four methods are compared: the *t* test (in black), the VIP statistic (in red), and the PLSDA (purple) and PCLDA (blue) regression coefficients. Scales are equal to those of Fig. 1.

Published on 13 June 2012. Downloaded by Pennsylvania State University on 18/05/2016 05:44:05.

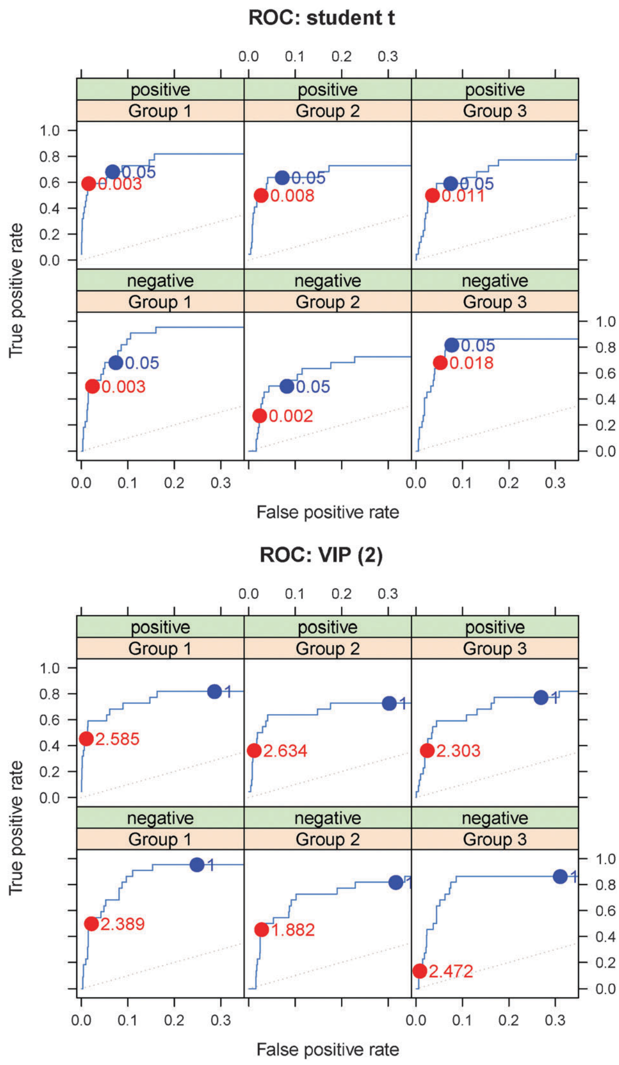

Fig. 3 ROC curves for the *t* and VIP statistics. The position of the standard cutoff values (0.05, 1) are indicated in blue; HC-thresholds in red.

much lower dependence on the number of latent variables than PCLDA, which is a good thing: it means that a misspecification of the optimal number of latent variables will not influence the selection to a great extent.

In Fig. 3, ROC curves are shown for the *t* test and the VIP using two latent variables. The corresponding plots for VIPs with four and ten latent variables are very similar. Standard cutoff values, for the *t* test corresponding to the 0.975 quantile of the *t* distribution with 18 degrees of freedom, and in case of the VIP the value of one, are indicated in blue, and HC-based thresholds in red. Note that the ROC curves for the VIP and the *t* test are very similar: both in this case select variables in almost exactly the same order, even though the numerical values of both statistics show differences. The optimal point in these plots is the top left corner. What is the best point on the curve, *i.e.*, the best threshold, depends on the application. In cases where all true differences need to be found, even if this means a large number of false positives, the optimal value will be less conservative than in cases where false positives need to be kept at a minimum. In all cases, the HC-selection is more conservative than the standard selection – the difference is particularly big for the VIP, where the standard selections are clearly far too generous.

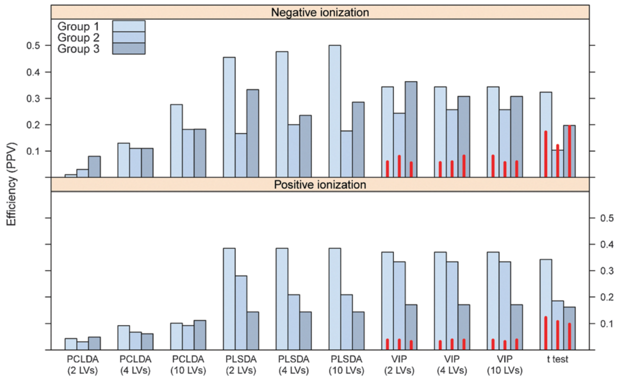

Fig. 4 Bar plots showing the selection efficiencies for all models considered using HC-based thresholds. Red bars indicate standard thresholds of 0.05 and 1 for the *t* test and the VIP measure, respectively.

Fig. 4 presents one way of evaluating the optimality of a solution, focussing on the efficiency, *i.e.* the ratio of the number of true positives and the number of selected variables. Gray bars indicate HC-based selection for the three comparisons (groups one to three *versus* controls); thick red line segments for the *t* test and VIP selections indicate selections with the default thresholds of 0.05 and VIP = 1, respectively. Again, we can compare the performance of the HC-based selections with the standard thresholds. For the VIP statistic, significant gains in efficiency can be achieved in all cases by choosing cutoffs with HC, caused, as we have seen, by the much lower number of selected variables. For the *t* test, HC threshold selection gives similar or better results than the standard threshold of 0.05, which in some cases, of course, can be very close to the optimal threshold identified by HC.

6.2 Simulated data

The simulated data confirm the conclusions from the experimental data. Fig. 5 shows the results of *t* test and VIP-based selections, both with their standard thresholds of 0.05 and 1, respectively, and the selections based on HC-suggested thresholds – the comparison is between group 1 and the controls for the positive ionization mode. Results for the other comparisons can be found in the ESI.† In each case, it is clear that the efficiency of the HC-based selections is much better than that of the standard thresholds, mostly caused by a drastic reduction in the number of selected variables: the standard thresholds are too generous. The price that is paid is that also the number of true positives suffers somewhat, but for the VIP in particular the gain in efficiency is quite extreme, indicating again that the standard threshold of one for these data is totally inappropriate.

Although the point of the current paper is not to compare different primary biomarker selection methods, the application of the HC statistic in this sense does provide an advantage: an unbiased and objective way to select optimal data- and method-dependent thresholds. Clearly, if "standard" thresholds are chosen, some methods, like the VIP in our apple data

Published on 13 June 2012. Downloaded by Pennsylvania State University on 18/05/2016 05:44:05.

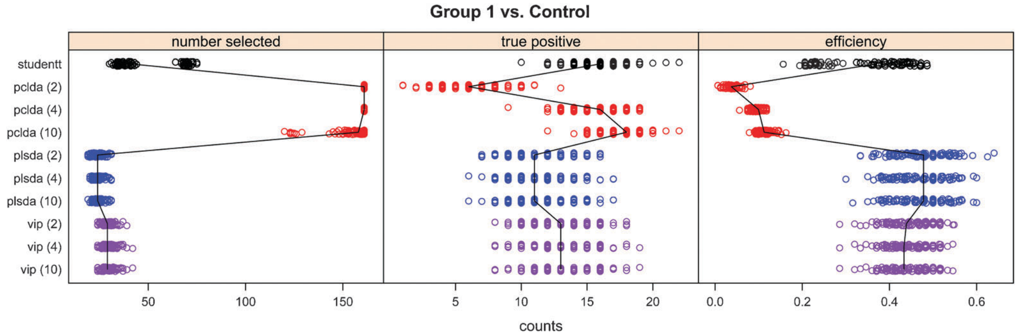

Fig. 5 Plots comparing the results for "standard" thresholds with HC-generated thresholds for 100 simulated data sets (group one *versus* control), positive ionization mode. The first column of plots presents the number of selected features; the dashed line indicates the upper selection limit of 10%, set by the one parameter for HC. The second column presents the number of true positives in the selections, and the third shows the efficiencies (PPVs).

examples, may be judged too harshly. In such a case it is not the primary selection method that is at fault, but rather the default selection threshold that is inappropriate. Fig. 6 shows an example of HC thresholding applied to the positive ionization data, group one *versus* controls. Similar plots for the other comparisons are available in the ESI.† The lowest number of variables is selected by PLSDA, with the VIP very close. These two types of models also lead to the highest selection efficiencies. Interestingly, and again in agreement with the experimental data, the choice of the number of latent variables does not seem to have any influence on the selection. This is a reassuring property, since it means that even a suboptimal choice does not lead to inappropriate selections. PCLDA shows a completely different picture: there, the only useful results are obtained with ten latent variables, the maximum number that is considered in this paper. In the other cases, the HC selection is just selecting as many points as possible, and the efficiency of these models is very bad indeed. The *t* test selects slightly more variables than PLSDA and the VIP. In general, it also retrieves more true positives, but is slightly worse in terms of efficiency. Should this behaviour be consistent over other data sets, one can use this information to determine which method is optimal in a particular case, depending on which kind of error, a false positive or a false negative, is more expensive.

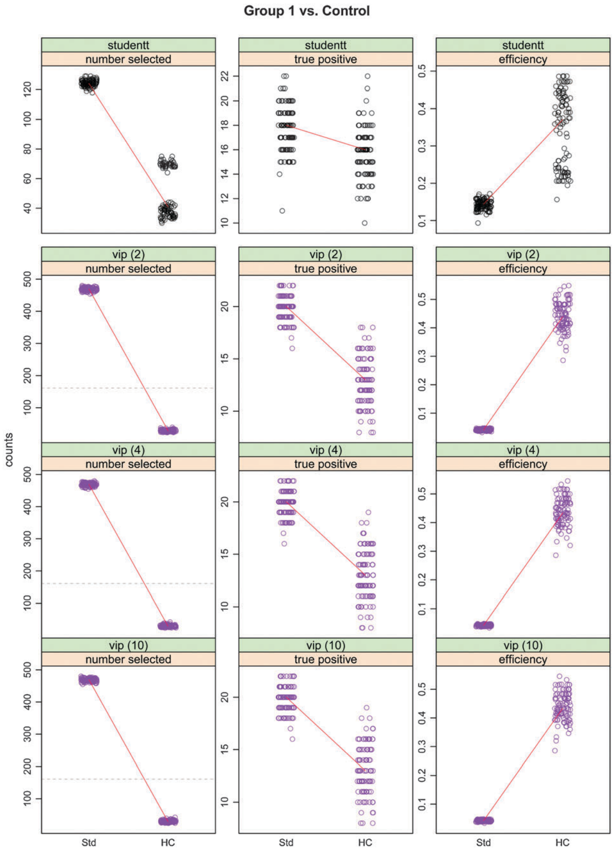

Fig. 6 Number of selected variables, number of true positives, and efficiencies (PPVs), respectively, of all ten methods considered, after selection based on HC thresholding. Comparison of simulated group 1 spiked data and controls, positive ionization. Solid lines, added for easier interpretation, connect the medians of the groups.

7 Conclusions

The question what difference is significant when comparing two data sets has become extremely important in this era of high-throughput measurements. Any method to answer this question will have to strike a balance between the costs of including too many unimportant variables and the cost of missing real differences. Even when having *p* values of statistical testing available, the cutoff point determining how many variables to accept as potential biomarkers must be chosen with care – the usual level of 0.05 is chosen rather arbitrary and can lead to sub-optimal selections. This is true regardless of whether the *p* values under consideration have been corrected for multiple testing or not. Because the widely different characteristics of data sets, leading to very different correlation structures, optimal cutoff points need to be determined on a case-by-case basis.

Higher Criticism tackles this task in a systematic fashion, and this paper shows the viability of such an approach using a spike-in data set from metabolomics for which the true differences are known. Obviously, a full evaluation of the merits of the approach would require application to many more of such data sets. Application to data sets of higher dimensionality, for instance from gene expression data or SNPs, needs additional research. Nevertheless, the benefits of using an adaptive choice of threshold, as compared to the

Published on 13 June 2012. Downloaded by Pennsylvania State University on 18/05/2016 05:44:05.

rather naive – but widely used – choices of 0.05 for *p* values and 1 for VIP coefficients should be clear. As an additional novelty, this paper extends the scope of HC thresholding to coefficients from PCLDA and PLSDA discriminant models, and to the VIP statistic, by using simulated null distributions from data permutation. In the case of the regression coefficients, no cutoffs have been suggested up to now, which is obvious because of the dependence on the scaling, size and structure of the data; the threshold of one, proposed earlier for the VIP, is shown to be also dependent on the data and in our case gives much inferior results than the thresholds suggested by the HC approach. The big advantage of HC threshold selection is that it does not rely on error estimates, which in small-sample situations can be highly variable.

As such, HC aims at the same goals as adaptive forms of multiple testing correction[9,10] and as stability selection.[24,28] The latter approach uses sub-sampling of the data to assess the consistency of biomarker selection. The advantage of using permutations, as is done by HC, is that no sub-sampling of the data is needed, which in the small data sets typical for omics studies, where high replication is the exception rather than the rule, can be difficult. The disadvantage of the HC approach is the large number of permutations, needed to get *p* values with sufficient resolution (not just 0.01, 0.02, *etc.*). However, the calculations can easily be parallelized. Moreover, for PCLDA, PLSDA and the VIP, efficient dedicated functions can be utilized, completing the necessary calculations within an hour on an ordinary desktop computer. The HC approach requires no fine-tuning: the one parameter that needs to be set is related to the assumption that real biomarkers are rare, and its default value of 10% is shown to work well in practice.

A final important advantage of having one consistent threshold selection is that it allows for a fair comparison of different biomarker selection algorithms. For the spike-in data analyzed here, it could be shown that PCLDA is performing far worse than the other methods considered in almost all cases. Indeed, an evaluation of several biomarker selection methods, using spike-in data sets from different fields, is currently underway in our group.

References

1 H.-C. Huang, D. Jupiter and V. VanBuren, *PLoS One*, 2010, **5**, e9056.
2 M. Yousef, M. Ketany, L. Manevitz, L. Showe and M. Showe, *BMC Bioinf.*, 2009, **10**, 337.
3 Y. Araki, K. Yoshikawa, S. Okamoto, M. Sumitomo, M. Maruwaka and T. Wakabayashi, *BMC Neurol.*, 2010, **10**, 112.
4 J. Oh, J. Craft, R. Townsend, J. Deasy, J. Bradley and I. E. Naqa, *J. Proteome Res.*, 2011, **10**, 1406–1415.
5 M. Chadeau-Hyam, T. Ebbels, I. Brown, Q. Chan, J. Stamler, C. Huang, M. Daviglus, H. Ueshima, L. Zhao, E. Holmes, J. Nicholson, P. Elliott and M. D. Iorio, *J. Proteome Res.*, 2010, **9**, 4620–4627.
6 V. Zuber and K. Strimmer, *Bioinformatics*, 2009, **25**, 2700–2707.
7 V. Tusher, R. Tibshirani and G. Chu, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 5116–5121.
8 Y. Benjamini and Y. Hochberg, *J. R. Stat. Soc. Ser. B (Methodological)*, 1995, **57**, 289–300.
9 A. Reiner, D. Yekutieli and Y. Benjamini, *Bioinformatics*, 2003, **19**, 368–375.
10 J. Storey and R. Tibshirani, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 9440–9445.
11 D. Donoho and J. Jin, *Ann. Stat.*, 2004, **32**, 962–994.
12 D. Donoho and J. Jin, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 14790–14795.
13 Y. Guo, T. Hastie and R. Tibshirani, *Biostatistics*, 2007, **8**, 86–100.
14 S. Smit, M. van Breemen, H. Hoefsloot, A. Smilde, J. Aerts and C. de Koster, *Anal. Chim. Acta*, 2007, **592**, 210–217.
15 M. Barker and W. Rayens, *J. Chemom.*, 2003, **17**, 166–173.
16 S. Wold, M. Sjöström and L. Eriksson, *Chemom. Intell. Lab. Syst.*, 2001, **58**, 109–130.
17 I.-G. Chong and C.-H. Jun, *Chemom. Intell. Lab. Syst.*, 2005, **78**, 103–112.
18 P. Franceschi, D. Masuero, U. Vrhovsek, F. Mattivi and R. Wehrens, *J. Chemom.*, 2012, **26**, 16–24.
19 R. van den Berg, H. Hoefsloot, J. Westerhuis, A. Smilde and M. van der Werf, *BMC Genomics*, 2006, **7**, 142.
20 S. Choe, M. Boutros, A. Michelson, G. Church and M. Halfon, *Genome Biol.*, 2005, **6**, R16.
21 M. McCall and R. Irizarry, *Nucleic Acids Res.*, 2008, **36**, e108.
22 C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan and G. Siuzdak, *Anal. Chem.*, 2006, **78**, 779–787.
23 C. Kuhl, R. Tautenhahn and S. Neumann, *CAMERA: Collection of annotation related methods for mass spectrometry data*, 2010.
24 N. Meinshausen and P. Bühlmann, *J. R. Stat. Soc. Ser. B (Methodological)*, 2010, **72**, 417–473.
25 R. Wehrens and P. Franceschi, 2012, submitted.
26 R Development Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, 2012.
27 B.-H. Mevik and R. Wehrens, *J. Stat. Software*, 2007, **18**.
28 R. Wehrens, P. Franceschi, U. Vrhovsek and F. Mattivi, *Anal. Chim. Acta*, 2011, **705**, 15–23.