OXFORD

Gene expression

Differential protein expression and peak selection in mass spectrometry data by binary discriminant analysis

Sebastian Gibb[1] and Korbinian Strimmer[2,*]

[1]Anesthesiology and Intensive Care Medicine, University Hospital Greifswald, Ferdinand-Sauerbruch-Straße, D-17475 Greifswald, Germany and [2]Epidemiology and Biostatistics, School of Public Health, Imperial College London, Norfolk Place, London, W2 1PG, UK

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Abstract

Motivation: Proteomic mass spectrometry analysis is becoming routine in clinical diagnostics, for example to monitor cancer biomarkers using blood samples. However, differential proteomics and identification of peaks relevant for class separation remains challenging.

Results: Here, we introduce a simple yet effective approach for identifying differentially expressed proteins using binary discriminant analysis. This approach works by data-adaptive thresholding of protein expression values and subsequent ranking of the dichotomized features using a relative entropy measure. Our framework may be viewed as a generalization of the 'peak probability contrast' approach of Tibshirani *et al.* (2004) and can be applied both in the two-group and the multi-group setting. Our approach is computationally inexpensive and shows in the analysis of a large-scale drug discovery test dataset equivalent prediction accuracy as a random forest. Furthermore, we were able to identify in the analysis of mass spectrometry data from a pancreas cancer study biological relevant and statistically predictive marker peaks unrecognized in the original study.

Availability and implementation: The methodology for binary discriminant analysis is implemented in the R package binda, which is freely available under the GNU General Public License (version 3 or later) from CRAN at URL http://cran.r-project.org/web/packages/binda/. R scripts reproducing all described analyzes are available from the web page http://strimmerlab.org/software/binda/.

Contact: k.strimmer@imperial.ac.uk

1 Introduction

Mass spectrometry, a high-throughput technology commonly used in proteomics, enables the measurement of the abundance of proteins, metabolites, peptides and amino acids in biological samples. The study of changes in protein expression across subgroups of samples and through time provides valuable insights into cellular mechanisms and offers a means to identify relevant biomarkers, e.g. to distinguish among tissue types, or for predicting health status. In practice, however, there still remain many analytic and computational challenges to be addressed, especially in clinical diagnostics (Leichtle *et al.*, 2013).

A recent overview of statistical issues in the analysis of proteomics mass spectrometry data is Morris (2012) who discusses a wide range of methods ranging from data preprocessing, i.e. removal of systematic bias, peak identification, peak alignment and quantification and calibration of relative peak intensities, to methods for high-level statistical analysis, such as peak ranking and classification. Of particular importance is the problem of differential protein expression and the identification of peaks informative for group separation and class prediction.

A special characteristic of mass spectrometry data is their dual-valued nature, i.e. they contain both continuous as well as discrete

information. Specifically, a protein may be differentially expressed if its intensity of expression varies among groups and is relatively up- or down-regulated, or if a corresponding peak is either absent or present in a specific group. Consequently, mass spectrometry intensity matrices typically contain very large amounts of missing values, which renders application of standard statistical methodology from other omics platforms, such as regularized *t*-scores, difficult and potentially suboptimal. Accordingly, this has initiated the development of new statistical methodology (Tibshirani *et al.*, 2004; Wang *et al.*, 2012).

Two main strategies to address this issue in the high-level analysis of mass spectrometry data have emerged:

1. All data are treated as continuous, with missing intensity values set to zero or imputed. Subsequently, standard omics methods are employed, such as *t*-scores for feature selection (e.g. Datta and DePadilla, 2006).
2. The absence–presence data are used for data analysis in conjunction with the intensity values. Tibshirani *et al.* (2004) propose peak probability contrasts (PPCs), the absolute difference in frequency of occurrence of a peak, for ranking and feature selection, and also use PPC to improve absence–presence data by dichotomization of intensity values. Wang *et al.* (2012) propose a test based on the PPC statistic and propose to apply joint false discovery rate control of the union of intensity-based and PPC-based rankings.

Here, we follow the second route and propose a novel coherent model for differential protein expression and prediction based on binary discriminant analysis (BinDA). Our approach may be viewed as a generalization of Tibshirani *et al.* (2004) and comprises the following:

- The binary absence–presence data are explicitly modeled by a multivariate Bernoulli (MVB) distribution.
- Multi-group binary discriminant analysis (BinDA) is employed for feature ranking, variable selection and prediction.
- For ranking of peaks the natural relative entropy variable importance measure coherent with BinDA is used, rather than PPC.
- Likewise, for dichotomization of the intensity data matrix containing missing values we employ the same entropy-based criterion.

As a result, we obtain simple principled framework for analyzing dual-valued mass spectrometry data without the need for imputation, with a natural measure for variable ranking and for differential protein expression, and with coherent prediction rules. In contrast to many other methods this approach also allows multiple groups as response variable, and thus extends beyond simple pairwise comparisons.

The remainder of the paper is structured as follows. Next, we describe in detail the statistical methodology underlying BinDA. Then, for validation we investigate the performance of the proposed approach in comparison with a random forest on a large-scale chemometric dataset. Subsequently, we present a detailed case study analyzing mass spectrometry data from a pancreas cancer study. For reproducibility, we provide the R package binda implementing our approach and R scripts for all analyzes described. Finally, we discuss applicability of the BinDA approach to other molecular data as well as further extensions.

2 Methods

2.1 Setup and notation
Our analysis starts after the raw mass spectrometry data have been adequately preprocessed, i.e. transformed, smoothed, background-removed, calibrated, aligned and peak-extracted (e.g. Morris, 2012; Gibb and Strimmer, 2012).

We denote the resulting peak intensities by $z_{ij} \geq 0$, with spectrum index $i \in \{1, \ldots, n\}$ and peak index $j \in \{1, \ldots, d\}$. The data matrix z_{ij} typically contains missing values as not all of the registered d peaks will be present in all of the n spectra. In a classification setting each spectrum i also carries a class label $y_i \in \{1, \ldots, K\}$ that assigns it to one of K different groups, for instance health status, tissue type or treatment outcome. The label is known for training data and unknown for test data. The sample size in group with label y is n_y with $n = \sum_{y=1}^{K} n_y$.

From the continuous data z_{ij} we obtain binary peak intensities x_{ij} by thresholding at peak-specific levels $\boldsymbol{w} = (w_1, \ldots, w_d)$. Specifically, we set $x_{ij} = 1$ if the peak is present in sample i and $z_{ij} \geq w_j$. Conversely, if the peak is absent or $z_{ij} < w_j$ then $x_{ij} = 0$. The methodology we present here uses the binary matrix x_{ij} for prediction and variable ranking, rather than the original data z_{ij}, and it also estimates the thresholds $\boldsymbol{w}$.

2.2 Modeling binary data
Stochastic models for multivariate binary data are well established (e.g. Cox, 1972; Dai *et al.*, 2013). A univariate binary random variable $X \sim Be(\mu)$ with two states $x = 0$ and $x = 1$ is completely described by a Bernoulli distribution $Be(\mu)$ with expectation $\mathrm{E}(X) = \mu$ and variance $\mathrm{Var}(X) = \mu(1 - \mu)$.

In the multivariate case this generalizes to $\boldsymbol{X} = (X_1, \ldots, X_d) \sim Be_d(\boldsymbol{\mu}, \boldsymbol{\zeta})$ where d denotes the dimension of the MVB distribution, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ is the vector of expectations $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$, and $\boldsymbol{\zeta}$ contains the $2^d - d - 1$ interaction parameters. As in the univariate case the variances $\mathrm{Var}(X_j) = \mu_j(1 - \mu_j)$ are fully determined by the means.

In many cases it is useful to ignore the dependencies among the individual variables X_j in order to reduce the number of parameters in the model. Despite, or perhaps because, of its simplicity the independence 'naive Bayes' assumption can be very effective, especially for prediction in high dimensions and small sample size, see Hand and Yu (2001) and Park (2009). For MVB with independent predictor variables the joint probability mass function is given by

$$\Pr(\boldsymbol{x}) = \prod_{j=1}^{d} \begin{cases} 1 - \mu_j & \text{if } x_j = 0, \\ \mu_j & \text{if } x_j = 1. \end{cases}$$

with diagonal covariance matrix $\mathrm{Var}(\boldsymbol{X}) = \mathrm{diag}\{\mu_j(1 - \mu_j)\}$.

2.3 Discriminant analysis with binary predictors
For prediction of the class associated with an unlabeled spectrum we need to construct a prediction rule. Here, we employ a Bayesian prediction rule similar as in diagonal discriminant analysis (DDA) that is routinely and successfully used, e.g. in transcriptomics (Tibshirani *et al.*, 2003).

We first define group-specific models for each group with label y,

$$\Pr(\boldsymbol{x}|y) = \prod_{j=1}^{d} \begin{cases} 1 - \mu_{yj} & \text{if } x_j = 0, \\ \mu_{yj} & \text{if } x_j = 1. \end{cases} \tag{1}$$

For each group y we also specify a prior probability $\Pr(Y = y) = \pi_y$ with $\sum_{y=1}^{K} \pi_y = 1$. By $\mu_{0j} = \sum_{y=1}^{K} \pi_y \mu_{yj}$ we denote the pooled mean for each variable j, i.e. the mean we would assign if there was only a single category.

The posterior probability of each group is then given by Bayes' theorem $\Pr(y|\boldsymbol{x}) = \Pr(\boldsymbol{x}|y)\pi_y/\Pr(\boldsymbol{x})$ which after taking the logarithm yields the discriminant function

$$d_y(\boldsymbol{x}) = \log \Pr(y|\boldsymbol{x}) = \log \pi_y + \log \Pr(\boldsymbol{x}|y) + C. \tag{2}$$

As the purpose of $d_y(x)$ is only to compare among different groups we can drop all terms that do not depend on y, such as $\Pr(x)$, represented above by the constant C. Prediction of a label for test data x is carried out by choosing the group y that maximizes the discriminant function,

$$\hat{y} = \arg\max_{y} \ d_y(x)$$

This MVB independence prediction rule has shown to be highly effective (e.g. Park, 2009), even if there is correlation among predictors.

Typically, the parameters of discriminant function are unknown themselves and have to be learned themselves from training data, i.e. from spectra with known group labels. The training is done by estimating the means μ_{yj} and the group probabilities π_y in Equations (1) and (2). We suggest a flexible estimation strategy by employing maximum-likelihood estimation for large sample size, and otherwise using regularized estimation. For instance, to estimate the group probabilities we use observed frequencies $\hat{\pi}_y = n_y/n$ if n is large, and for small n the Stein-type shrinkage estimator of proportions described in Hausser and Strimmer (2009).

2.4 Variable ranking and selection

Closely tied in with prediction is the question which variables are most important for successful assignment of a class label, and conversely, which variables are irrelevant. Especially in large-dimensional problems it is very important to remove the null features as the build-up of random noise from these variables can substantially degrade the overall prediction accuracy (cf. Ahdesmäki and Strimmer, 2010).

For ranking features in discriminant analysis with binary variables there have been many, in part contradictory, propositions. For the case of $K = 2$ groups the following criteria, among others, have been used:

- The χ^2 statistic of independence between response and predictors (An *et al.*, 2013),
- PPCs $|\mu_{y1} - \mu_{y2}|$ (Tibshirani *et al.*, 2004),
- Quinlan's information gain measure (Bender *et al.*, 2004), and
- ratio of between-group and within-group covariance (Wilbur *et al.*, 2002).

See Tan *et al.* (2004) for many other proposals for measuring associations between categorical outcomes and binary variables. Only some of the criteria above can also be applied to the multiple group case ($K > 2$).

We use a principled approach to variable ranking relying on predictive information, see Gelman *et al.* (2014) for an overview. Conceptually, we use the expected log-predictive density as measure of model fit, and compare the fully specified joint model containing all predictors and the response with a 'no-effects' model where the response is independent of the predictors. The difference of expected log-likelihood between full and 'no-effects' model is given by the relative entropy or Kullback–Leibler divergence $D = KL(F_{\text{full}} || F_{\text{no-eff}})$. The relative contributions of each individual predictor to D then provides a measure of variable importance. This procedure applied to linear regression with independent predictors results in squared marginal correlations, and applied to DDA it yields squared t-scores, both of which are optimal measures for variable ranking in their respective settings (Zuber and Strimmer, 2011).

For independent binary predictors the joint full model is

$$\Pr(x, y)_{\text{full}} = \prod_{j=1}^{d} \begin{cases} (1 - \mu_{yj})\pi_y & \text{if } x_j = 0, \\ \mu_{yj}\pi_y & \text{if } x_j = 1 \end{cases}$$

whereas the no-effects model is

$$\Pr(x, y)_{\text{no-eff}} = \prod_{j=1}^{d} \begin{cases} (1 - \mu_{0j})\pi_y & \text{if } x_j = 0, \\ \mu_{0j}\pi_y & \text{if } x_j = 1. \end{cases}$$

This results in

$$\begin{aligned} D &= \sum_{j=1}^{d}\sum_{y=1}^{K} \left(\mu_{yj}\pi_y \log\left(\frac{\mu_{yj}}{\mu_{0j}}\right) + (1 - \mu_{yj})\pi_y \log\left(\frac{1 - \mu_{yj}}{1 - \mu_{0j}}\right) \right) \\ &= \sum_{j=1}^{d}\sum_{y=1}^{K} \pi_y KL\Big(Be(\mu_{yj}) || Be(\mu_{0j}) \Big) \qquad (3) \\ &\approx \sum_{j=1}^{d}\frac{1}{2}\sum_{y=1}^{K}\pi_y\left(\frac{\mu_{yj} - \mu_{0j}}{\sigma_j}\right)^2 = \sum_{j=1}^{d} S_j \end{aligned}$$

where $\sigma_j^2 = \mu_{0j}(1 - \mu_{0j})$ is the variance of $Be(\mu_{0j})$. For the special case of $K = 2$ groups S_j simplifies to

$$S_{j(K=2)} = \frac{\pi_1\pi_2}{2}\left(\frac{\mu_{1j} - \mu_{2j}}{\sigma_j}\right)^2.$$

By construction, the score S_j is a measure of variable importance of feature j where S_j is a weighted sum of the squared z-scores that compare each group mean with the overall pooled mean. This is precisely analogous to the pooled-mean formulation of discriminant analysis described in Ahdesmäki and Strimmer (2010). If the variances σ_j^2 are similar across features, then $S_{j(K=2)}$ is apart from a scale factor the squared PPC.

As above for learning the discriminant function, we use for estimation of the entropic ranking scores S_j either maximum-likelihood or shrinkage estimates of proportions, depending on sample size.

Noting that $n\hat{\pi}_y/(1 - \hat{\pi}_y) = (1/n_y - 1/n)^{-1}$ we may also introduce squared t-scores

$$t_{y0}^2 = n\frac{\hat{\pi}_y}{1 - \hat{\pi}_y}\left(\frac{\hat{\mu}_{yj} - \hat{\mu}_{0j}}{\hat{\sigma}_j}\right)^2$$

that are properly scaled to allow to contrast the individual contributions of each class relative to each other, similar as in the methods described in Ahdesmäki and Strimmer (2010) and Tibshirani *et al.* (2003). The estimated ranking score may also be expressed in terms of a weighted sum of the t-scores via

$$\hat{S}_j = \sum_{y=1}^{K}(1 - \hat{\pi}_y)t_{y0}^2/(2n).$$

After ranking variables according to the estimated scores $\hat{S}_j$, with highest scores indicating the most relevant predictors, we use cross-validation to evaluate prediction accuracy for different numbers of included predictors to determine a suitable cutoff.

2.5 Dichotomization

With the above setup for BinDA it is straightforward to perform dichotomization. Specifically, we choose thresholds $w = (w_1, \ldots, w_d)$ to discretize the continuous data z_{ij} to maximize the entropy score D (Equation 3). As in our model assume the predictors are assumed to be independent we can optimize each threshold w_j independently by maximizing the individual S_j. Note that the same entropy measure is used both for determining the thresholds and for ranking the predictors, thus ranking and discretization is done in an integrative fashion.

3 Results

3.1 Implementation and reproducible research

We have implemented our approach for multi-class discriminant analysis using binary predictors including functions for variable ranking and dichotomization in the R package binda that is freely available under the GNU General Public License (version 3 or later) from URL http://cran.r-project.org/web/packages/binda/. For reproducibility of the analyzes presented in this paper we provide corresponding R scripts at http://strimmerlab.org/software/binda/.

3.2 Validation of binda

BinDA has been studied extensively and is well established in the literature (e.g. Cox, 1972). More recently, it was demonstrated that BinDA with naive Bayes assumption can yield high rates of predictive accuracy even if the underlying assumption of independence of predictors is not met (Wilbur *et al.*, 2002; Bender *et al.*, 2004; Park, 2009; An *et al.*, 2013).

For validation of our implementation of BinDA in the R package binda we analyzed a large-scale chemometric test dataset. Specifically, we investigated the 'Dorothea' drug discovery dataset from the NIPS 2003 feature selection challenge (Guyon *et al.*, 2005). The dataset contains $d = 100\,000$ binary features describing the 3-D properties of chemical compounds that either bind (response label $+1$) or not (label -1) to thrombin, an enzyme involved in blood clotting.

We used the 'Dorothea' training data with $n_{\text{train}} = 800$ samples and corresponding class labels to learn a classifier with binda. Subsequently, we applied the resulting classification rule to the validation dataset with $n_{\text{val}} = 350$ samples and predicted the sample labels of the validation data. As for the validation data the true labels are known we were then able to compute the actual prediction accuracy, i.e. the proportion of correctly identified labels.

Due to its algorithmic simplicity training the classifier and ranking variables with binda was computationally inexpensive. Applying class-balanced 5-fold cross-validation with 20 repetitions using the R package crossval on the training data alone we determined that the response can be predicted well with only very few top ranking predictors included in the classification rule. For instance, a binda classifier with three predictors yielded prediction accuracy on the validation set of 0.9371 and of 0.9429 if 10 predictors were used. The predictive accuracy without any variable selection including all 100 000 predictors was 0.9057.

For comparison we also trained a random forest (Breiman, 2001), a tree-based machine learning approach that emerged as one of the overall best performing methods for classification in a recent systematic study (Fernández-Delgado *et al.*, 2014). Due to the high-dimensionality the running time for learning the random forest from the training data was two magnitudes slower than binda, taking 652 seconds on our workstation in comparison to 5 seconds for binda. The random forest yielded an accuracy of 0.94 for prediction of the labels of the independent validation dataset. This analysis confirms that BinDA, though very simple, is able, at least for this data, to perform prediction as accurately as random forest. Correspondingly, if variables in the random forest were ranked according to the Gini variable importance measure the top-ranking features were mostly identical to those ranked best by BinDA, which indicates that BinDA is indeed able to select the most relevant variables.

3.3 Analysis of pancreas cancer proteomics data

3.3.1 Pancreas cancer study

For illustration of our proposed approach to classification and peak ranking in mass spectrometry data we also reanalyzed experimental proteomics data from a pancreas cancer study conducted in Leipzig and Heidelberg (Fiedler *et al.*, 2009). For the training dataset of this study 40 patients with diagnosed pancreas cancer as well as 40 healthy controls were recruited. Each participant of the study donated serum samples which provided the basis for MALDI/TOF measurements. For each sample four technical replicates were
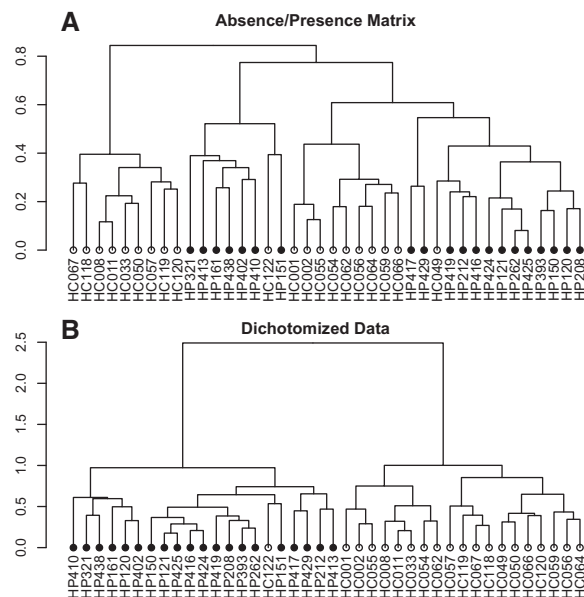

Fig. 1. Clustering of samples from the pancreas cancer study of Fiedler *et al.* (2009) using (**A**) the original absence–presence data and (**B**) the optimized binary matrix. Filled circles indicate pancreas cancer samples, empty circles healthy controls. For clustering, we employed Ward's agglomerative hierarchical clustering based on a Jaccard distance matrix computed using R standard functions hclust() with method='ward.D2' and dist() with method='binary'
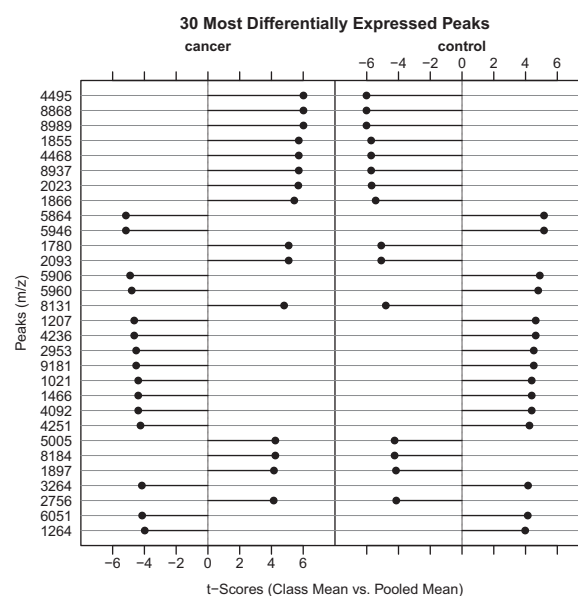

Fig. 2. Ranking of 166 peaks in the preprocessed spectra from the pancreas cancer study according to BinDA framework

obtained. Due to the presence of strong batch effects in our analysis below we restrict ourselves to patients and controls from Heidelberg, leading to a raw dataset containing 160 spectra for 40 probands,

The aim of the study was to determine biomarkers to discriminate patients with pancreas cancer from healthy persons. Fiedler *et al.* (2009) found marker peaks at m/z 3884 (double charged) and 7767 (single charged) and correspondingly supposed platelet factor 4 (PF4) as potential marker, arguing that PF4 is down-regulated in blood serum of patients with pancreatic cancer.

3.3.2 Preprocessing and dichotomization

For preprocessing of the raw mass spectrometry data we employed the standard analysis pipeline implemented in the R package MALDIquant (Gibb and Strimmer, 2012). Specifically, the raw data were variance-stabilized, smoothed, baseline-corrected, TIC-standardized and aligned. Technical replicates were then averaged, peaks were identified and corresponding intensities extracted from each averaged spectrum. Precise details on the preprocessing can be found in the R script. As a result a protein expression matrix of size 40 patients times 166 peaks was obtained. In total 26% of intensities in the matrix were missing, corresponding to about 44 missing peaks per spectrum.

Subsequently, we performed dichotomization of the intensity matrix using the relative entropy criterion of Equation (3). To illustrate the improvement of the resulting binary data matrix over the original absence–presence matrix we conducted hierarchical clustering on the samples. As can be seen in Figure 1 the clustering based on the optimized binary matrix almost perfectly separates pancreas cancer samples from healthy samples, indicating that there is a strong signal in the data.

Fig. 3. Top ranking peak group containing five differentially expressed peaks: (**A**) double- and (**B**) single-charged peaks

3.3.3 Peak ranking and differential expression thresholds

In order to identify features responsible for the separation of cancer versus healthy samples in Figure 1B we applied peak ranking for binary data according to BinDA. The resulting ranking of the 30 best discriminating peaks is shown in Figure 2. As a consequence of the discrete data, the first three top-ranking peaks with m/z values 4495, 8868 and 8989) achieved the same maximum score, followed by the next three peaks 1855, 4468 and 8937, that also achieved an identical score.

We note that none of these peaks was identified in the original study. The two PF4 peaks with m/z 3884 and 7768 (the slight difference is due to the MALDIquant alignment procedure) rank on places 148–151 and 157–163, respectively.
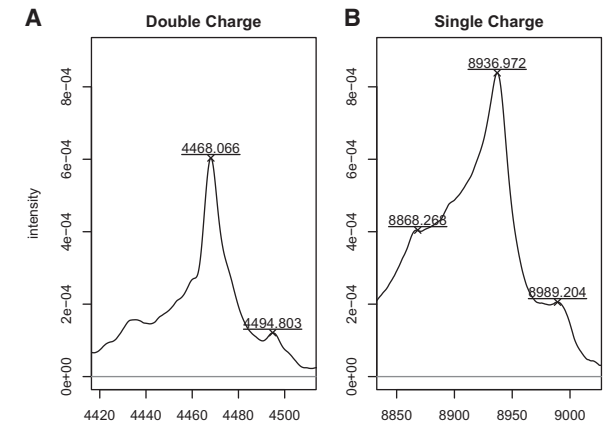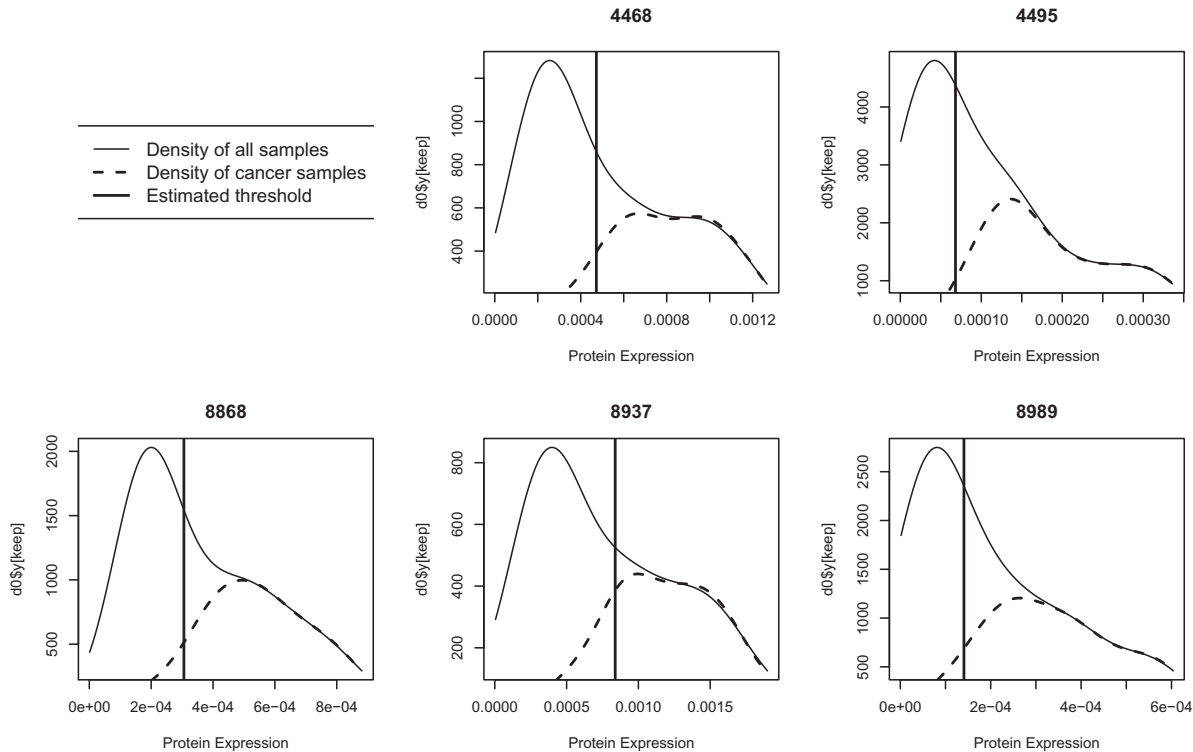

Fig. 4. Densities of expression values and estimated thresholds for the top ranking peaks. Each column shows the single charged (bottom row) and the corresponding double charged variant (top row)

Using cross-validation, we estimated prediction errors for group separation from the binary data matrix. As for the 'Dorothea' dataset we employed class-balanced 5-fold cross-validation with 20 repetitions. Interestingly, using only five predictors was sufficient to achieve an accuracy of 0.96, sensitivity of 0.96, specificity of 0.97, positive predictive value of 0.97 and negative predictive value of 0.95. This indicates that the observed clear separation between cancer and control samples in Figure 1B is attributable to only very few features of the data.

Visual inspection of the group of top-ranking differentially expressed peaks revealed a further pattern (Fig. 3). First, five of the peaks are all part of the same peak group. Second, the peak group appears both in a single charged (m/z 8868, 8937, 8989) version as well as in a mirrored double charge version (m/z 4468 and 4495). This affirms that there must an underlying biological marker driving the observed changes between cancer and control samples.

To study this further, we inspected the intensities for the peaks belonging to the differentially expressed peak group. Figure 4 shows the overall density, as well as the sub-density for the cancer samples, along with the dichotomization threshold estimated by binda. For all five peaks the expression respectively the underlying protein abundance is up-regulated in the cancer samples compared with the controls. In addition, the estimated thresholds provide an effective means to separate the two groups.

3.3.4 Biological relevance

Finally, we also tried to identify the biological molecules behind the differentially expressed peak group shown in Figure 3. Specifically, we used the TagIdent tool (Gasteiger *et al.*, 2005) with settings Mw 8936.97, Mw range 0.05% and organism *homo sapiens* to query the UniProtKB/Swiss-Prot database (The UniProt Consortium, 2015). This indicated a potential link of the central peak m/z 8937 to PDPFL_HUMAN, the pancreatic progenitor cell differentiation and proliferation factor-like protein, as well as to a fragment of C3adesArg, an acylation stimulating protein. The increased abundance of PDPFL_HUMAN in pancreas cancer tissue appears highly plausible, and the increased concentration of C3adesArg in serum of cancer patients has also been reported previously (e.g. Opstal-van Winden *et al.*, 2011).

Another biologically relevant result of our analysis based on BinDA is that the originally proposed PF4 marker is not differentially expressed and hence cannot be used to distinguish between cancer and healthy samples.

4 Discussion

We have presented a simple yet effective approach to differential expression and classification for mass spectrometry data using BinDA. Our approach may be viewed as generalization of Tibshirani *et al.* (2004) and can be applied also for multi-group discriminant analysis. A particular feature is the use of the same relative entropy criterion for peak ranking and selection and for dichotomization of the continuous protein intensity data. In addition, we obtain decision thresholds from the protein intensities that are biologically and diagnostically easy to interpret.

In illustrative analysis of high-dimensional drug discovery data we showed that our approach implemented in the R package binda is computationally effective and yet competitive with a random forest. Furthermore, in reanalysis of proteomics data from a pancreas cancer study we found statistically predictive marker peaks to tumor cell growth unrecognized in the original analysis. This confirms the importance of reproducible research in proteomics, where it is unfortunately still not common to provide analysis scripts and software openly.

In addition to mass spectrometry analysis, there are many bioinformatics applications in which binary data are collected, and hence, in which the present methodology and software will potentially be useful. Examples include meta-genomics, where the absence and presence of proteins and genes is compared with a pan-genome (Medini *et al.*, 2008), community analysis by DNA fingerprinting (Wilbur *et al.*, 2002), and chemometrics (Bender *et al.*, 2004).

Exploring additional applications may also lead to further methodological extensions of the procedures currently implemented in binda, such as modeling overdispersion, e.g. by employing the Beta-Bernoulli rather than Bernoulli distribution, and to take account of interactions among predictors, e.g. by modeling pair-wise correlation.

References

Ahdesmäki,M. and Strimmer,K. (2010) Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Ann. Appl. Stat.*, **4**, 503–519.

An,B. *et al.* (2013) Testing the statistical significance of an ultra-high-dimensional naïve bayes classifier. *Stat. Interface*, **6**, 223–229.

Bender,A. *et al.* (2004) Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *J. Chem. Inf. Comput. Sci.*, **44**, 170–178.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Cox,D.R. (1972) The analysis of multivariate binary data. *J. R. Stat. Soc. C*, **21**, 113–120.

Dai,B. *et al.* (2013) Multivariate Bernoulli distribution. *Bernoulli*, **19**, 1464–1483.

Datta,S. and DePadilla,L.M. (2006) Feature selection and machine learning with mass spectrometry data for distinguishing cancer and non-cancer samples. *Stat. Method*, **3**, 79–92.

Fernández-Delgado,M. *et al.* (2014) Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.*, **15**, 3133–3181.

Fiedler,G.M. *et al.* (2009) Serum peptidome profiling revealed platelet factor 4 as a potential discriminating peptide associated with pancreatic cancer. *Clin. Cancer Res.*, **15**, 3812–3819.

Gasteiger,E. *et al.* (2005) Protein identification and analysis tools on the ExPASy server. In: Walker, J.M. (ed.) *The Proteomics Protocols Handbook*. Humana Press Totowa, New Jersey, pp. 571–607.

Gelman,A. *et al.* (2014) Understanding predictive information criteria for Bayesian models. *Stat. Comput.*, **24**, 997–1016.

Gibb,S. and Strimmer,K. (2012) MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*, **28**, 2270–2271.

Guyon,I. *et al.* (2005) Result analysis of the NIPS 2003 feature selection challenge. In: Saul, L.K. *et al.* (eds.) *Advances in Neural Information Processing Systems*, **vol. 17**. MIT Press, Cambridge, MA, pp. 545–552.

Hand,D.J. and Yu,K. (2001) Idiot's Bayes—not so stupid after all? *Int. Stat. Rev.*, **69**, 385–398.

Hausser,J. and Strimmer,K. (2009) Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.*, **10**, 1469–1484.

Leichtle,A.B. *et al*. (2013) Potentials and pitfalls of clinical peptidomics and metabolomics. *Swiss Med. Wkly.*, **143**, w13801.

Medini,D. *et al*. (2008) Microbiology in the post-genomic era. *Nat. Rev. Microbiol.*, **6**, 419–430.

Morris,J.S. (2012) Statistical methods for proteomic biomarker discovery based on feature extraction or functional modeling approaches. *Stat. Interface*, **5**, 117–135.

Opstal-van Winden,A.W.J. *et al*. (2011) Searching for early breast cancer biomarkers by serum protein profiling of pre-diagnostic serum; a nested case-control study. *BMC Cancer*, **11**, 381.

Park,J. (2009) Independent rule in classification of multivariate binary data. *J. Multi. Anal.*, **100**, 2270–2286.

Tan,P.-N. *et al*. (2004) Selecting the right objective measure for association analysis. *Inf. Syst.*, **29**, 293–313.

The UniProt Consortium. (2015) UniProt: a hub of protein information. *Nucleic Acids Res.*, **43**, D204–D212.

Tibshirani,R. *et al*. (2003) Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Sci.*, **18**, 104–117.

Tibshirani,R. *et al*. (2004) Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics*, **17**, 3034–3044.

Wang,X. *et al*. (2012) A hybrid approach to protein differential expression in mass spectrometry-based proteomics. *Bioinformatics*, **28**, 1586–1591.

Wilbur,J.D. *et al*. (2002) Variable selection in high-dimensional multivariate binary data with application to the analysis of microbial community DNA fingerprints. *Biometrics*, **58**, 378–386.

Zuber,V. and Strimmer,K. (2011) High-dimensional regression and variable selection using CAR scores. *Stat. Appl. Genet. Mol. Biol.*, **10**, 34.