# Classification of samples with order restricted discriminant rules

David Conde, Miguel A. Fernández
Bonifacio Salvador and Cristina Rueda

**Summary**

In recent years, mass spectrometry techniques have helped proteomics to become a powerful tool for the early diagnosis of cancer, as they help to discover protein profiles specific to each pathological state. One of the questions where proteomics is giving useful practical results is that of classifying patients into one of the possible severity levels of an illness, based on some features measured on the patient. This classification is usually made using one of the many discrimination procedures available in statistical literature. We present in this chapter recently developed restricted discriminant rules that use additional information in terms of orderings on the means, and we illustrate how to apply them to mass spectrometry data using `R` package `dawai`. Specifically, we use proteomic prostate cancer data, and we describe all steps needed, including data preprocessing and feature extraction, to build a discriminant rule that classifies samples in one of several disease stages, thus helping diagnosis. The restricted discriminant rules are compared with some standard classifiers that do not take into account the additional information, showing better performance in terms of error rates.

**Key words:** Mass Spectrometry, Preprocessing, Feature extraction, Mean spectrum, Supervised classification, Order restrictions, Restricted discriminant rules, `R` `dawai` package

## 1. Introduction

Proteomics has become a powerful tool for the early diagnosis of cancer, allowing to characterize proteins and therefore to identify diagnostic biomarkers from tissues and body fluids ***(1)***. Recent advances in mass spectrometry (MS) techniques have made it possible to discover protein profiles specific to each pathologic state from high-dimensional MS data ***(2)***. In association with other approaches, MS has

become the central technique used by most proteomic biomarker discovery platforms ***(3)***. In cancer, one important proteomic issue is the identification of a panel of biomarkers suitable for discriminating different pathological states, which helps to improve the early detection of cancer ***(4)***.

Once the biomarkers are identified, the classification of the patients is done using discrimination techniques. The general discrimination problem deals with the prediction of the group a patient belongs to, based on some features measured on the patient. In supervised classification, the discriminant rule is built using a training sample, that is, a set of patients for which both the features and the group membership are known. When the underlying distribution of the data is known, the optimal classification rule is the so called Bayes rule. When it is assumed that the measurements from each population are normally distributed, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are two of the most commonly used in applications discriminant analysis methods. LDA and QDA are obtained replacing in the Bayes rule the unknown parameters, that is, the means vectors and the covariance matrix, by their usual estimators, assuming for LDA that the covariance matrices of the populations are identical. The relaxation of this assumption leads to the QDA rule, where no equality of covariance matrices is assumed.

These two classical discriminant rules were followed by a great deal of classification algorithms such as nearest neighbors ***(5)***, classification trees ***(6)***, neural networks ***(7)***, support vector machines (SVM) ***(8)*** or random forests (RF) ***(9)***. All of them build the corresponding classification rules solely from the information in the training sample.

In this chapter we describe and show the usefulness of recently developed discriminant rules ***(10; 11; 12; 13)*** that use not only the training sample but also other information called additional information. For instance, in a certain application it can be known that certain features of the patients take, on average, higher values in some groups than in others. Let us suppose that we want to classify patients in one of the following groups: $G_1$ - healthy, $G_2$ - early-stage disease and $G_3$ - advanced-stage disease. We know from previous studies that the mean of variable $V_1$ increases with the severity of the disease and that the mean of variable $V_2$ decreases. This additional information can be expressed in terms of restrictions on the model parameters: if $\mu_{i,j}$ represents the mean of variable $V_i$ in group $G_j$, $i = 1,2, j = 1,2,3$, then the additional information can be expressed as $\mu_{1,1} \leq \mu_{2,1} \leq \mu_{3,1}$, $\mu_{1,2} \geq \mu_{2,2} \geq \mu_{3,2}$.

Restricted linear and quadratic discriminant rules ***(10; 11; 12; 13)*** are obtained plugging into the respective Bayes rules the estimators of the unknown parameters, defined from the training sample and the restrictions on the means, via an iterative procedure ***(10; 11; 13)*** to ensure that the estimators fulfill the restrictions.

In applications as cancer diagnosis, patients are intended to be classified into

one of various diagnosis groups based on gene expression or proteomic data. Often, some of the predictors are known to take higher values in some groups with respect to the others. Taking into account this underlying order could potentially result in significantly lower misclassification rates with respect to the standard classification methods that do not take it into account. In this chapter, we use proteomic prostate cancer data from patients with the following diagnosis groups: normal prostate, benign prostate conditions, prostate cancer and prostate-specific antigen (PSA) between 4 and 10, and prostate cancer and PSA levels above 10.

The layout of the chapter is as follows. In Section 2 we start describing briefly the data set and the software with which we show the use of the restricted discriminant rules. In Section 3 we detail the standard methods used to perform the preprocessing and feature extraction, in order to obtain the data matrix (patients in rows, features in columns) needed for classification. Also in this section, a brief summary of the restricted discriminant rules is presented. The `R` library `dawai` is used in Section 4 with the purpose of showing how the restricted discriminant rules can be used in practice when additional information is present. These rules, applied on the mentioned data set, show a significantly better performance than other usual discrimination rules not considering the additional information such as LDA, RF or SVM.

## 2. Materials

MS techniques allow the identification of the amount and type of proteins present in a sample by measuring the mass-to-charge ratio ($m/z$) and abundance of gas-phase ions. A mass spectrum is a plot of the ion signal (intensity) versus $m/z$. These $m/z$ ratios can be used to calculate the molecular weights of protein or peptide. Two broadly used MS techniques for proteome screening are Matrix-Assisted Laser Desorption and Ionisation (MALDI) and Surface-Enhanced Laser Desorption and Ionisation (SELDI) with Time-Of-Flight (TOF) tubes.

In this chapter we will apply our restricted discriminant rules ***(11; 12; 13)*** to the proteomic prostate cancer data of JNCI Data 7-30-02.zip ***(14)***, which are publicly available at `http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`. The data consist of 322 serum spectra measuring peak amplitudes at 15154 $m/z$ values in the range 0-20000 Da. Serum samples provided by patients have the following frequencies: normal prostate (63), benign prostate conditions (190), prostate cancer and PSA levels between 4 and 10 (26), and prostate cancer and PSA levels above 10 (43). Samples were applied to a C16 hydrophobic interaction protein chip (Ciphergen Biosystems, Freemont, CA) and analyzed as described in Petricoin et al. ***(14)***. Data were generated using the SELDI-TOF MS techniques and are provided with baseline subtracted. In Figure 1, the mean spectrum, com-

puted averaging over all raw spectra, is shown.

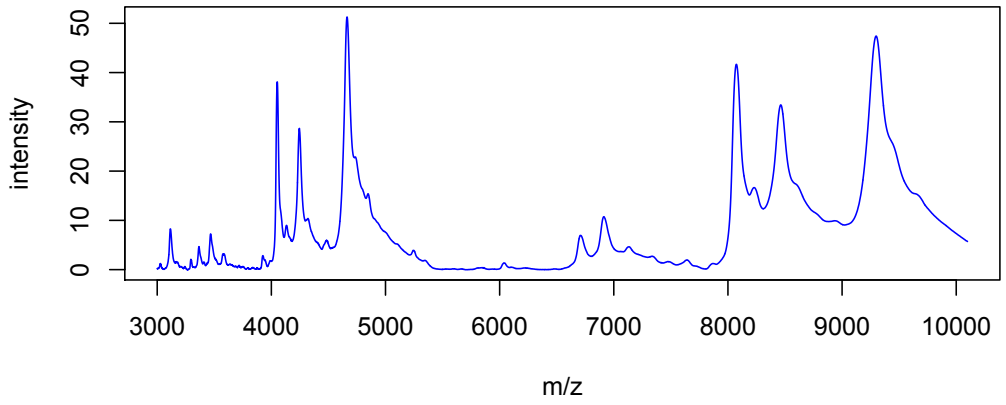


Figure 1: Mean spectrum.

We will illustrate how, when additional information on the means of the populations (normal prostate, benign prostate conditions and prostate cancer) is known, our restricted discrimination rules can be applied to MS data using the R package dawai ***(13)***, which can be downloaded from http://cran.r-project.org/web/packages/dawai/.

# 3. Methods

## 3.1. Preprocessing

Data from MS measurements usually contain a substantial amount of noise and show large inter-measurement variation ***(15)***. A preprocessing stage is needed, including, as a first task, noise reduction.

### 3.1.1. Baseline reduction

Commercial mass spectrometers implement basic noise reduction methods. Some studies explore methods for reducing noise, especially baseline noise ***(16; 17)***. Most of the studies employing a baseline reduction method estimate the baseline noise and substract the estimated baseline from the original mass spectrum.

The data we will use to illustrate our methods are already provided with baseline substracted, so we do not need to care about baseline reduction ***(14)***.

### 3.1.2. Normalization

A peak in mass spectra indicates the relative abundance of a protein, but mass spectra cannot be directly compared with each other as MS spectra of similar samples are not always quantified within the same amplitude range. A second task in the preprocessing stage is normalization, needed to compare the real intensities, converting all the spectra to the same intensity ranges.

Many different approaches have been proposed and used to handle this issue. Normalization of mass spectra typically involves subtracting an offset and dividing by a scaling factor. Such offset and scaling parameters can be defined and applied globally or locally using a window. One of the most frequently used approaches is normalization with respect to the total ion current (TIC), i.e., dividing each intensity by the sum of all the intensities in a mass spectrum ***(18; 19; 20)***. This is equivalent to normalization with respect to the mean of the intensities in the spectrum ***(17)***. One alternative to TIC normalization is scaling by the sum of the squares of the intensities so the spectrum forms a unit vector ***(21)***. Other studies perform normalization with respect to the largest peak ***(16)*** or linear scaling with the largest and the smallest peak intensities ***(22; 14)***, known as min range.

Meuleman et al. ***(23)*** compare 8 such normalization procedures for both global or local normalization, according to two objectives: inter-spectra variance minimization and classification performance maximization. They state that in general it is better to use a local than a global normalization method. As for the distinct methods, they show that global mean SD (substracting the global mean and dividing by the standard deviation) is the best method attending the classification maximization objective and one of the best regarding the variance minimization objective. Here, we follow their advice and normalize our data with this method.

### 3.1.3. Smoothing

An ion peak may be spread across many data points, so each $m/z$ data point should not be regarded as the record of a distinct peptide. To reduce this noise, we smooth the normalized spectra using a Gaussian kernel with a full width at half maximum (FWHM) of 11 $m/z$ values ***(17; 24; 25)***, with the convention FWHM $\approx 2.355\sigma$. In this way, for each $m/z$ value, the normalized intensity $x$ is replaced by a weighted average of the form $\Sigma_t\ t\ N(t;x,\sigma)$ where the summation is over all the 11 $m/z$ values around $x$, 5 each side, and $N(t;x,\sigma)$ is a Gaussian kernel with mean $x$ and variance $\sigma^2$.

Other common approaches are smoothing filters ***(26)***, wavelet transform (WT) ***(27)***, deconvolution filter ***(28)*** and moving average filter ***(29)***.

## 3.2. Feature extraction

Feature extraction is the process of selecting small sets of relevant features. Because of the high measurement variation in MS data, peaks are the most suitable biomarkers ***(24)***. A peak is defined as an $m/z$ value with higher intensity than the nearby values around it and than the average intensity at those nearby values.

Peak detection deals with identifying peaks in a mass spectrum, which is not simple due to variability between samples in intensity and location. Peak alignment is the process of matching peaks that represent the same protein species in distinct spectra.

Morris et al. ***(30)*** propose a method for performing feature extraction that uses the average spectrum for peak detection. The mean spectrum is computed averaging over all raw spectra. After the mean spectrum is normalized and smoothed, peaks are detected. Then the peaks are quantified in the individual spectra. We perform peak detection in this way, using the mean spectrum, and obtain the following 14 $m/z$ peaks: 3116, 3468, 4052, 4133, 4245, 4483, 4662, 4847, 6709, 6912, 8073, 8228, 8462 and 9297.

Once the peaks are identified, they usually do not correspond to local maxima in the individual processed spectra. Wagner et al. ***(16)*** consider the local maximum within 30 measurements points of the peak mass from the processed spectra as the features. Petricoin et al. ***(22)*** also consider the maximum peak height, but other metrics as average or median peak height ***(19)*** can be considered too. We proceed as in ***(16)***, looking for the maximum heights within 30 measurements points of the mentioned 14 identified peaks for each of the processed spectra. These 14 values and the disease stage for each patient serum sample are rendered to a text file.

## 3.3. Restricted classification rules

Let us consider $k$ disease stages $G_1,\ldots,G_k$, also called groups or populations throughout this chapter, so that each patient belongs to one and only one of them. Let $\pi_1,\ldots,\pi_k$ be the a priori probabilities of the groups, with $\sum_{j=1}^{k}\pi_j = 1$. Let us suppose that for each patient a $p$-dimensional vector $\boldsymbol{X}$ of features is measured, and that $\boldsymbol{X}$ is normally distributed with mean $\boldsymbol{\mu}_j$ for group $G_j, j = 1,\ldots,k$, and common covariance matrix $\boldsymbol{\Sigma}$. Then, if $\boldsymbol{U}$ is the vector of features measured for a patient whose disease stage is unknown, the optimal classification rule is the Bayes rule:

$$\begin{aligned}\text{Classify } \boldsymbol{U} \text{ in } G_j \text{ iff } &\log\pi_j + (\boldsymbol{U}-\boldsymbol{\mu}_j)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{U}-\boldsymbol{\mu}_j) \leq \\ &\log\pi_l + (\boldsymbol{U}-\boldsymbol{\mu}_l)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{U}-\boldsymbol{\mu}_l), l = 1,\ldots,k.\end{aligned}$$

Parameters $\boldsymbol{\mu}_j, j = 1,\ldots,k$, and $\boldsymbol{\Sigma}$ are usually unknown, but they can be estimated from a training sample (a set of patients for which the features values and

the group they belong to are known) by the sample vector means $\overline{\boldsymbol{X}}_j, j = 1,\ldots,k$, and the pooled sample covariance matrix $\boldsymbol{S}$:

$$\overline{\boldsymbol{X}}_j = \frac{1}{n_j}\sum_{l=1}^{n_j} \boldsymbol{X}_{j_l}, j = 1,\ldots,k,$$

$$\boldsymbol{S} = \frac{1}{n-k}\sum_{j=1}^{k}\sum_{l=1}^{n_j} \left(\boldsymbol{X}_{j_l} - \overline{\boldsymbol{X}}_j\right)\left(\boldsymbol{X}_{j_l} - \overline{\boldsymbol{X}}_j\right)^\top,$$

where $n_j$ is the sample size of group $G_j$, $j = 1,\ldots,k$, and $n = \sum_{j=1}^{k} n_j$.

The linear discriminant rule (LDA) or Fisher's rule is obtained plugging estimators $\overline{\boldsymbol{X}}_j, j = 1,\ldots,k$, and $\boldsymbol{S}$ into the Bayes rule.

In applications, it is usual that some additional information is available, often through order restrictions on the populations means. Let us suppose that we want to classify patients in one of the following groups: $G_1$ healthy, $G_2$ early-stage disease and $G_3$ advanced-stage disease ($k = 3$), and that two variables $V_1$ and $V_2$ are measured for each patient ($p = 2$).

If we know that the patients from $G_1$ (the control group) take, in mean, lower values than those coming from any of the other groups for all variables, in the usual statistical terminology we can say that there is a "tree order" among the means of the variables: $\mu_{1,1} \leq \mu_{i,1}, i = 2,3$, $\mu_{1,2} \leq \mu_{i,2}, i = 2,3$.

Another common situation appears when it is known that there is an increase in the means of the variables. We can say now that there is a "simple order" among the groups means: $\mu_{1,1} \leq \mu_{2,1} \leq \mu_{3,1}$, $\mu_{1,2} \leq \mu_{2,2} \leq \mu_{3,2}$.

Let us denote as $C$ the subset of the parameter space where the restrictions on the means are fulfilled.

The family of restricted linear classification rules ***(11)*** that we apply here to the proteomics data considers estimators for $\boldsymbol{\mu}_j, j = 1,\ldots,k$, that take into account the additional information known about the parameters. When the sample means do not verify the restrictions, an iterative procedure starting from vector $\overline{\boldsymbol{X}} = (\overline{\boldsymbol{X}}_1^\top,\ldots,\overline{\boldsymbol{X}}_k^\top)^\top$ is used to obtain an estimator of the means that verifies the additional information contained in set $C$.

A first approach would be to consider the value in $C$ closest to $\overline{\boldsymbol{X}}$, that is, the projection of $\overline{\boldsymbol{X}}$ onto $C$. We call this value $\widehat{\boldsymbol{\mu}}^0$ (see Figure 2). However, it is known that that estimator lacks good statistical properties as it is not admissible ***(31)***. For this reason, we consider an estimator that is inside the set $C$ (parameter $\gamma \in [0,1]$ will control how much inside $C$ is the estimator considered). However, as it can be seen in Figure 2, we need an iterative procedure to ensure that the final estimator is inside the set $C$, as it might happen that trying to put the estimator inside $C$ takes it to the other side of $C$. For each iteration, if $v$ is the vector obtained in the previous iteration and $p$ is the projection of $v$ over $C$, the new vector is defined

as $p-\gamma(v-p)$. The procedure ends when the new vector verifies the restrictions, i.e., when $p-\gamma(v-p)$ belongs to $C$. We denote as $\widehat{\boldsymbol{\mu}}^{\gamma}$ the limit of the procedure. Figure 2 illustrates this process for a set $C$ and an initial estimator $\overline{\boldsymbol{X}}$ not belonging to $C$ for three different values of $\gamma$ $(0, 0.5, 1)$. The results in ***(10; 11)*** ensure the convergence of this scheme and the good properties of these estimators.

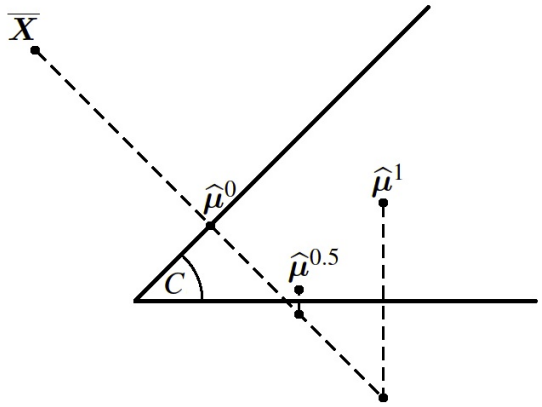


Figure 2: Set $C$ and estimators $\widehat{\boldsymbol{\mu}}^0$, $\widehat{\boldsymbol{\mu}}^{0.5}$ and $\widehat{\boldsymbol{\mu}}^1$.

If $\widehat{\boldsymbol{\mu}}^{\gamma} = (\widehat{\boldsymbol{\mu}}_1^{\gamma\top}, \ldots, \widehat{\boldsymbol{\mu}}_k^{\gamma\top})^{\top}$, the restricted linear classification rules are:

$$\text{Classify } \boldsymbol{U} \text{ in } \Pi_j \text{ iff } \log \pi_j + (\boldsymbol{U} - \widehat{\boldsymbol{\mu}}_j^{\gamma})^{\top} \boldsymbol{S}^{-1} (\boldsymbol{U} - \widehat{\boldsymbol{\mu}}_j^{\gamma}) \leq \log \pi_l + (\boldsymbol{U} - \widehat{\boldsymbol{\mu}}_l^{\gamma})^{\top} \boldsymbol{S}^{-1} (\boldsymbol{U} - \widehat{\boldsymbol{\mu}}_l^{\gamma}), l = 1, \ldots, k.$$

When the covariance matrices are not assumed to be equal, an analogous family of restricted quadratic classification rules is proposed in ***(13)***.

# 4. Notes

Now we go back to our data set and detail how to use the restricted discriminant rules in a case like this. As it was seen in previous section, after normalization and feature extraction, the maximum heights within 30 measurements points of 14 identified peaks for each patient serum sample were rendered to a text file. It consists of a matrix with 322 rows and 15 columns, the 14 features and the group label, i.e., the group each patient belongs to: 1 - normal prostate, 2 - benign prostate conditions, 3 - prostate cancer and PSA between 4 and 10, and 4 - prostate cancer and PSA levels above 10.

Atlhough the restricted discriminant rules have shown good performance under normality assumptions ***(10; 11; 12; 13)***, these rules have also good robustness properties against different types of contamination ***(32)***.

## 4.1. Results

In this section we illustrate the restricted linear discriminant rules on the mentioned data set using the R package dawai that we have developed. R is a free software environment for statistical computing that runs on a broad variety of platforms including UNIX, Windows and MacOS. It is widely used for developing and sharing statistical software, which makes it highly extensible through user-created packages, and can be easily installed without any cost. R package dawai depends on boot *(33)*, ibdreg *(34)* and mvtnorm *(35)* R packages, that must be installed before loading dawai.

As we lack the experts knowledge regarding additional information for our data set, in this example we will consider as restrictions the order restrictions verified by the total sample. Then we split the sample into training and test sample, and use the training sample to build the restricted linear discriminant rules and the test sample to evaluate the accuracy of the rules.

We first load package dawai and read the data file, called data.txt.

```
R> library(dawai)
R> data <- read.table("data.txt", header=TRUE)
```

We separate the first variable (data$Class), with the group label for each patient, from the other ones with the 14 peaks heights.

```
R> class <- as.factor(data$Class)
R> dataset <- data[, 2:15]
```

Variable class contains the group label for each patient.

```
R> table(class)

  1   2   3   4
 63 190  26  43
```

These are the number of patients in each of the 4 groups. The groups correspond to increasingly advanced levels of the disease. We join the original groups 3 and 4 into one (the ones with prostate cancer), relabelling them, so that the groups have enough elements to split the sample into training and test data sets of reasonable sizes.

```
R> levels(class) <- c(1, 2, 3, 3)
R> table(class)

  1   2   3
 63 190  69
```

Let us have a look to the order restrictions verified by the total sample.

```
R> means <- colMeans(dataset[class == 1, ])
R> means <- rbind(means, colMeans(dataset[class == 2, ]))
R> means <- rbind(means, colMeans(dataset[class == 3, ]))
R> rownames(means) <- 1:3
R> t(means)

         1    2    3
P3116 0.06 0.60 0.26
P3468 1.64 0.22 0.03
P4052 3.33 3.35 3.65
P4133 0.40 0.47 0.65
P4245 0.88 3.03 2.18
P4483 0.02 0.28 0.09
P4662 3.14 5.76 4.38
P4847 0.54 1.50 0.93
P6709 1.09 0.30 0.11
P6912 2.31 0.59 0.35
P8073 3.35 4.02 3.82
P8228 0.93 1.32 1.44
P8462 1.19 3.58 2.74
P9297 2.50 5.30 3.98
```

Except for variables 2, 9 and 10, i.e., P3468, P6709 and P6912, all means are lower for group 1 than for groups 2 and 3. As for variables 2, 9 and 10, means are higher for group 1 than for groups 2 and 3. Group 1 (normal prostate) can be regarded as the control group in a decreasing tree order among the mean values of variables 2, 9 and 10, and an increasing tree order among the rest of the variables. We change the sign of these three variables so that there is the same tree order on all predictors.

```
R> dataset[, c(2, 9, 10)] <- -dataset[, c(2, 9, 10)]
```

In this way, restrictions in the training sample can now be easily specified by just `restext = "t<1,2,3,4,5,6,7,8,9,10,11,12,13,14"`. See `dawai` package ***(13)*** help files (`help(rlda)`) for advise.

We split the data set into a randomly selected training set and a test set, fixing a seed in order to get the same results as the reader. We are doing it here in order to have a test set and to show in an easy way that our rules outperform the usual ones.

```
R> set.seed(4100)
R> values <- runif(dim(dataset)[1])
R> trainsubset <- (values < 0.5)
R> testsubset <- (values >= 0.5)
```

Now we can build the restricted linear discriminant rules on the training sample, using $\gamma = 0, 0.5, 1$.

```
R> obj <- rlda(dataset, class, subset = trainsubset,
               restext = "t<1,2,3,4,5,6,7,8,9,10,11,12,13,14",
               gamma = c(0, 0.5, 1))
```

We have not specified `prior` parameter, so the group proportions of the training set have been used as the prior probabilities of group membership.

```
R> obj$prior
   class1    class2    class3
0.1863354 0.5838509 0.2298137
```

Now, let us consider the test set and classify its observations. We know the groups that the observations in the test set belong to, so we can estimate the true error rates as the proportion of observations in the test set wrongly classified by the restricted discrimination rules. The first command below classifies the observations in the test set. The second command yields the percentages of wrong classification of these observations.

```
R> pred <- predict(obj, newdata = dataset[testsubset, ],
                   grouping = class[testsubset])
R> pred$error
                      gamma=0 gamma=0.5  gamma=1
True error rate (%): 11.80124  11.18012 10.55901
```

These results can also be compared with the error rates for some standard classiffiers that do not take into account the additional information considered such as LDA in R `MASS` package ***(36)***, RF in R `randomForest` package ***(37)*** and SVM in `e1071` package ***(38)***, packages that have to be loaded first. The following commands load these 3 packages:

```
R> library(MASS)
R> library(randomForest)
R> library(e1071)
```

Now we build the rule and compute the LDA error as we did with the restricted rules:

```
R> lda_out <- lda(dataset, class, subset = trainsubset)
R> lda_error <- mean(predict(lda_out, newdata = dataset[testsubset, ])
                     $class != class[testsubset])*100
R> lda_error
[1] 13.04348
```

Also for RF:

```
R> rf_out <- randomForest(class ~ ., data = dataset,
                          subset = trainsubset)
R> rf_error <- mean(predict(rf_out, newdata = dataset[testsubset, ])
                    != class[testsubset])*100
R> rf_error
[1] 14.28571
```

And also for SVM, using always the default parameters:

```
R> svm_out <- svm(class ~ ., data = dataset,
                  subset = trainsubset)
R> svm_error <- mean(predict(svm_out, newdata = dataset[testsubset, ])
                     != class[testsubset])*100
R> svm_error
[1] 13.04348
```

We can see that, for $\gamma = 1$, the test error rates for the restricted linear rules are 19.04% lower than for LDA and SVM, and 29.09% lower than for RF.

## 4.2. Variable selection

We finish showing the behavior of the rules when variable selection is performed. We ask ourselves if we can dispense with redundant or irrelevant variables and if we can reduce possible overfitting by performing variable selection. We have searched for the variables maximizing the Mahalanobis ***(39)*** distances among $G_1$, $G_2$ and $G_3$ means. For $k = 6$, variables selected are P4052, P4133, P4847, P6709, P6912 and P8462. For $k = 10$, variables selected are P3468, P4052, P4245, P4483, P4847, P6709, P6912, P8228 and P8462. For these two reduced sets of variables we perform the same analysis we have just detailed for the full 14 variables set. The corresponding error rates for $k = 6, 10, 14$ are represented in Figure 3.

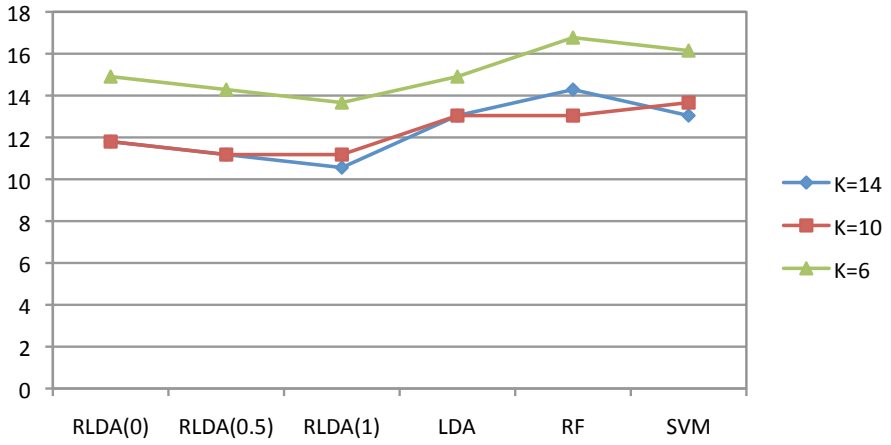


Figure 3: Error rates for $k = 6, 10, 14$.

We can see that error rates are quite similar for $k = 10$ and $k = 14$, being significantly lower for both sets of variables than for $k = 6$. This means that we can reduce the number of variables from $k = 14$ to $k = 10$ without a significant loss of prediction accuracy, but not from $k = 10$ to $k = 6$. In all cases, the restricted rules perform better than those procedures that do not consider additional information and the lowest test error rates correspond to the restricted linear rules for $\gamma = 1$ (RLDA(1)).

# References

[1] Toss, A., DeMatteis, E., Rossi, E., Casa, L.D., Iannone, A., Federico, M., and Cortesi, L. (2013) Ovarian cancer: can proteomics give new insights for therapy and diagnosis?. *International journal of molecular sciences* 14(4), 8271–8290.

[2] Yasui, Y., Pepe, M., Thompson, M.L., Adam, B.L., Wright, G.L., Qu, Y., Potter, J.D., Winget, M., Thornquist, M., and Feng, Z. (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* 4(3), 449–463.

[3] Paul, D., Kumar, A., Gajbhiye, A., Santra, M.K., and Srikanth, R. (2013) Mass spectrometry-based proteomics in molecular diagnostics: discovery of cancer biomarkers using tissue culture. *BioMed research international 2013*.

[4] Khadir, A. and Tiss, A. (2013) Proteomics Approaches towards Early Detection and Diagnosis of Cancer. *J Carcinogene Mutagene* 14.

[5] Cover, T. and Hart, P. (1967) Nearest Neighbor Pattern Classication. *IEEE Transactions on Information Theory* 13(1), 21–27.

[6] Buntime, W. (1992) Learning classification trees. *Statistics and Computing* 2(2), 63–72.

[7] Bishop, C.M. (1995) *Neural networks for pattern recognition*. Oxford University Press.

[8] Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Machine learning* 20(3), 273–297.

[9] Breiman, L. (2001) Random Forests. *Machine Learning* 45(1), 5–32.

[10] Fernandez, M., Rueda, C., and Salvador, B. (2006) Incorporating Additional Information to Normal Linear Discriminant Rules. *Journal of the American Statistical Association* 101(474), 569–577.

[11] Conde, D., Fernandez, M.A., Rueda, C., and Salvador, B. (2012) Classification of Samples into Two or More Ordered Populations with Application to a Cancer Trial. *Statistics in Medicine* 31(28), 3773–3786.

[12] Conde, D., Salvador, B., Rueda, C., and Fernandez, M.A. (2013) Performance and Estimation of the True Error Rate of Classification Rules built with Additional Information. An Application to a Cancer Trial. *Statistical Applications in Genetics and Molecular Biology* 12(5), 583–602.

[13] Conde, D., Fernandez, M.A., Salvador, B., and Rueda, C. (2014) dawai: An R Package for Discriminant Analysis With Additional Information. *Journal of Statistical Software. To appear*.

[14] Petricoin, E.F., Ornstein, D.K., Paweletz, C.P., Ardekani, A., Hackett, P.S., Hitt, B.A., Velassco, A., Trucco, C., Wiegand, L., Wood, K., Simone, C.B., Levine, P.J., Linehan, W.M., Emmert-Buck, M.R., Steinberg, S.M., Kohn, E., and Liotta, L.A. (2002) Serum Proteomic Patterns for Detection of Prostate Cancer. *Journal of the National Cancer Institute* 94(20), 1576–1578.

[15] Semmes, O.J., Feng, Z., Adam, B.-L., Banez, L.L., Bigbee, W.L., Campos, D., Cazares, L.H., Chan, D.W., Grizzle, W.E., Izbicka, E., Kagan, J., Malik, G., McLerran, D., Moul, J.W., Partin, A., Prasanna, P., Rosenzweig, J.,

Sokoll, L.J., Srivastava, S., Srivastava, S., Thompson, I., Welsh, M.J., White, N., Winget, M., Yasui, Y., Zhang, Z., and Zhu, L. (2005) Evaluation of Serum Protein Profiling by Surface-Enhanced Laser Desorption/Ionization Timeof-Flight Mass Spectrometry for the Detection of Prostate Cancer: I. Assessment of Platform Reproducibility. *Clinical Chemistry* 51(3), 102–112.

[16] Wagner, M., Naik, D., and Pothen, A. (2003) Protocols for disease classification from mass spectrometry data. *Proteomics* 3, 1692–1698.

[17] Zhu, W., Wang, X., Ma, Y., Rao, M., Glimm, J., and Kovach, J.S. (2003) Detection of cancer-specific markers amid massive mass spectral data. *Proceedings of the National Academy of Sciences of the United States of America* 100, 14666–14671.

[18] Baggerly, K.A., Morris, J.S., Wang, J., Gold, D., Xiao, L.C., and Coombes, K.R. (2003) A comprehensive approach to the analysis of matrix assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics* 3(9), 1667–1672.

[19] Bhattacharyya, S., Siegel, E.R., Petersen, G.M., Chari, S.T., Suva, L.J., and Haun, R.S. (2004) Diagnosis of pancreatic cancer using serum proteomic profiling. *Neoplasia* 6, 674–686.

[20] Li, J., Zhang, Z., Rosenzweig, J., Wang, Y.Y., and Chan, D.W. (2002) Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clinical Chemistry* 48, 1296–1304.

[21] Alfassi, Z.B. (2004) On the normalization of a mass spectrum for comparison of two spectra. *Journal of The American Society for Mass Spectrometry* 15(3), 385–387.

[22] Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C.S., Fishman, D.A., Kohn, E.C., and Liotta, L.A. (2002) Use of proteomic patters in serum to identify ovarian cancer. *The Lancet* 359(9306), 572–577.

[23] Meuleman, W., Engwegen, J.Y.M.N., Gast, M.-C.W., Beijnen, J.H., Reinders, M.J.T., and Wessels, L.F.A. (2008) Comparison of normalisation methods for surface-enhanced laser desorption and ionisation (SELDI) time-of-flight (TOF) mass spectrometry data. *BMC Bioinformatics* 9(88).

[24] Bhanot, G., Alexe, G., Venkataraghavan, B., and Levine, A.J. (2006) A robust meta-classification strategy for cancer detection from MS data. *Proteomics* 6, 592–604.

[25] Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., and Le, Q.-T. (2004) Sample classification from protein mass spectrometry. *Bioinformatics* 20(17), 3034–3044.

[26] Wang, M.Z., Howard, B., Campa, M.J., Patz, E.F., and Fitzgerald, M.C. (2003) Analysis of human serum proteins by liquid phase isoelectric focusing and matrix-assisted laser desorption/ionization-mass spectrometry. *Proteomics* 3(9), 1661–1666.

[27] Taskin, V., Dogan, B., and Olmez, T. (2013) Prostate Cancer Classification from Mass Spectrometry Data by Using Wavelet Analysis and Kernel Partial Least Squares Algorithm. *International Journal of Bioscience, Biochemistry and Bioinformatics* 3(2), 98–102.

[28] Malyarenko, D.I., Cooke, W.E., Adam, B.-L., Malik, G., Chen, H., Tracy, E.R., Trosset, M.W., Sasinowski, M., Semmes, O.J., and Manos, M. (2005) Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques. *Clinical Chemistry* 51(1), 65–74.

[29] Liu, Q., Krishnapuram, B., Pratapa, P., Liao, X., A, A. Hartemink, and Carin, L. (2004) Identification of differentially expressed proteins using MALDI-TOF mass spectra. *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers* 2, 1323–1327.

[30] Morris, J.S., Coombes, K.R., Koomen, J., Baggerly, K.A., and Kobayashi, R. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics* 21(9), 1764–1775.

[31] vanEeden, C. (2006) *Restricted parameter space estimation problems: admissibility and minimaxity properties*. Springer.

[32] Salvador, B., Fernandez, M.A., Martin, I., and Rueda, C. (2008) Robustness of classification rules that incorporate additional information. *Computational Statistics & Data Analysis* 52(5), 2489–2495.

[33] Canty, A. and Ripley, B. (2014) boot: Bootstrap Functions (originally by Angelo Canty for S). R package version 1.3-13.

[34] Sinnwell, J.P. and Schaid, D.J. (2013) ibdreg: Regression Methods for IBD Linkage With Covariates. R package version 0.2.5.

[35] Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Bornkamp, B., Maechler, M., and Hothorn, T. (2014) mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-1.

[36] Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., and Firth, D. (2011) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-29.

[37] Breiman, L., Cutler, A., Liaw, A., and Wiener, M. (2014) randomForest: Breiman and Cutler's random forests for classification and regression. R package version 4.6-10.

[38] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., and Lin, C.-C. (2014) e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-4.

[39] Mahalanobis, P.C. (1936) On the generalised distance in statistics. *Proceedings of the National Institute of Science of India* 12, 49–55.