Statistical methods for mass spectrometry-based clinical proteomics

Kakourou, A.A.

2

Combination approaches improve predictive performance of diagnostic rules for mass-spectrometry proteomic data

Abstract

We consider a proteomic mass spectrometry case-control study for the construction of a diagnostic rule for patients disease status allocation. We propose an approach for combining a collection of classifiers for the construction of a "combined" classification rule in order to enhance calibration and prediction ability. In a first stage this is achieved by building individual classifiers separately, each one using the entire proteomic data set. A double leave-one-out cross-validatory approach is used to estimate the class-predicted probabilities on which the combination method will be calibrated. The performance of the combination approach is examined both through a breast cancer proteomic data set and through simulation studies. Our experimental results indicate that in many circumstances gains in classification performance and predictive accuracy can be achieved.

This chapter has been published as: Alexia Kakourou, Werner Vach and Bart Mertens (2014). Combination approaches improve predictive performance of diagnostic rules for mass-spectrometry proteomic data. *Journal of Computational Biology* 21(12), 898-914.

2.1 Introduction

Based on the most recent statistics, there is an estimate of 12.7 million cancer cases around the world which is expected to rise to 21 million by 2030. There is therefore an urgent need to develop effective and reliable diagnostic tools for early detection of this disease. Mass spectrometry based clinical proteomics has emerged as a powerful analytical technology towards this objective. Proteomic methods are used for protein profiling and identification of cancer-associated markers in biological fluids which offer the opportunity to understand better the specific disease and also to improve detection ability and diagnostic accuracy. The statistical analysis of protein profiles collected in mass-spectrometry based case-control study, compares the protein expression patterns between the two different groups. This process allows the construction of discriminating rules which will potentially facilitate early diagnosis and prognosis.

In this paper we consider a case-control study, the design of which is described in detail in van der Werff et. al (2008). The data set we analyze is the same data set which was used in the context of the International Competition on Proteomic Diagnosis, the structure of which is fully described in Mertens (2008). The experiment involves a total of 231 individuals, consisting of 116 anonymous healthy controls and 115 breast cancer patients from each of which a serum sample was obtained and stored according to a standardized protocol. The available samples from both groups were randomly distributed across 3 plates in roughly equal proportions and spotted in four replicates on the corresponding plate. A single mass spectrum was generated from each spot using MALDI-TOF spectrometry. The experiment was conducted in 3 consecutive days, by assigning and processing each single plate separately on a distinct day. The data corresponding to the first two plates was used as the calibration set and thus consisted of 153 samples of which 76 were cases and 77 were controls. The data from the last plate was used as the validation set and consisted of 78 samples, 39 of which were cases and 39 were controls. The four replicate spectra from each individual were processed and combined into a single output spectrum, prior to subsequent analysis, as described by Mertens (2008). The output spectra were stored at a fixed grid of 11205 mass/charge (m/z) values ranging from 960 to 11.168 Dalton. Figure 2.1 shows the mean mass-spectrum of the cases (top) and the controls (bottom) groups for the calibration set. In addition to the preprocessed calibration and validation sets, we have the corresponding case-control labels (1 for cases and 0 for controls) for each of the included individuals.

The objective of the competition was to provide a comparative evaluation on different technics of classifying mass spectrometry data. The competition participants, coming from different areas of expertise such as statistics, (bio)informatics, proteomics, chemometrics etc., were asked to construct a diagnostic classification rule, based on the proteomic data which would predict the disease status of future patients. In the first stage of the competition the participants were given only the calibration set together with the corresponding case labels as well as the protocol containing a summary description of the data, the study design and a description of the prepossessing steps applied to the spectral

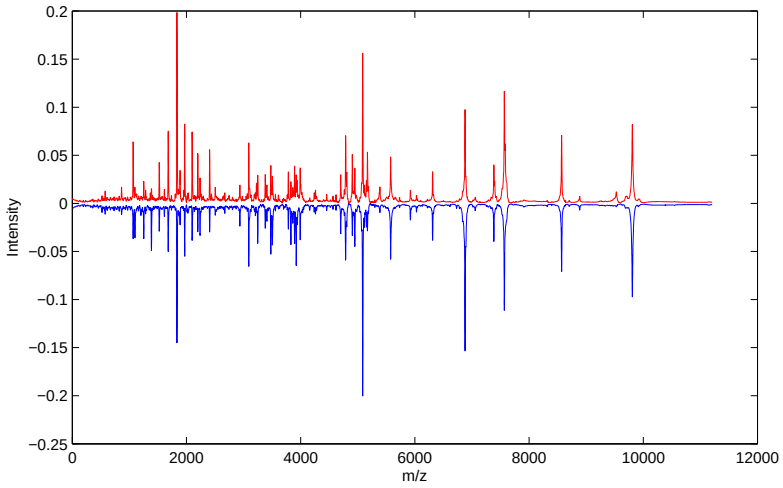

Figure 2.1: A graphical representation of the mean mass-spectrum for cases (top) and controls (bottom) separately (negative intensity values are plotted for the control group).

data. After sending in a first report containing the description of their chosen approach and the calibration results, the validation set was forwarded to the participants, without the corresponding case labels so that the diagnostic assignments for this data set would be based solely on the previously calibrated discriminating rule. Evaluation of the predictive performance of the calibrated allocation rules was based on the true class information on the validation set while the comparative discussion and evaluation of those rules is presented in Hand (2008).

A key result of the competition was that the majority of the participants chose linear methods to analyze the proteomic data, i.e. methods of which the classification decision is based on the calibration of a prognostic score using a linear combination of the input variables. Their preference for linear classifiers seemed reasonable since such methods are relatively simple and provide more easily interpretable classifiers, which is a key advantage in applied biomedical research. Furthermore, several authors have argued in favor of linear methods because of their ability to calibrate more stable and accurate estimates compared to more complex and sophisticated methods, particularly in high dimensions (Rendell and Seshu, 1990; Shavlik et al., 1991; Hand, 2006). Indeed, the linear approaches used in the context of the competition produced similar classification results while they were all close to the “optimum" (using error-rate as classification criterion). This “optimum" solution was achieved with a random-forest classification method applied to a set of detected peaks. In this paper we investigate an alternative approach to the calibration of a discriminating rule. That is to use a collection of distinct classification procedures to construct a combination-based classifier. This combination approach

was initially recommended in Hand (2008) as “an effective way to improve classification predictions" while the key idea of combining several predictors instead of selecting one, goes back to Stone's (1992) “model-mix" proposal, which was later adapted and applied by Wolpert (1992) to the neural network context and by Breiman (2001) to the regression context. The aim of the paper is to explore the potential of improving linear predictors, both in terms of prediction performance and predictive accuracy, by combining the class estimates obtained using different linear classification methods. Moreover, we investigate the extend of improvement over using only an individual classifier.

The structure of the paper is as follows. In a first stage we present two different frameworks for combining the estimates obtained using different linear classification methods which allows for combined prediction and performance evaluation. The first approach is based on the convex combination of the posterior class probabilities from each separately calibrated classifier. The second method is based on fitting a model using cross-validated predictions of the distinct classifiers as predictor variables. Subsequently, we present a comparative analysis between the single-classifier analyses and the combination approaches and we demonstrate how classification performance and predictive accuracy improves for the latter. We then present a post-hoc analysis in which on one hand we explore how the two combination methods work and allow for improved predictions and on the other hand we explore different ways to fit the model-based combination approach. We next present a simulation study based on reusing the proteomic mass spectra data analyzed before. In order to simulate the class difference, we impute differentially expressed signal between cases and controls, generated under some specific conditions and with a known structure. We finish with a discussion.

2.2 Combination Method

Suppose there are K different classification procedures available, giving rise to distinct classifiers $p^1(x), ..., p^K(x)$ to predict the true class y of an observation in terms of an input vector x. Assume furthermore that these classifiers were constructed using the same learning set $\mathcal{L} = \{(y_i, x_i), i = 1, ..., n\}$ where $x_i = (x_{i1}, ..., x_{ip})$. We restrict our discussion to the case where each fit $p^k(x)$ yields estimates of the posterior class probabilities $\hat{p}^k(x)$. Our objective is to combine $p^1(x), ..., p^K(x)$ in a way which allows for joint calibration and assessment of the discriminating rule. We present a combination approach based on replacing the original set of predictors $\{x_i, i = 1, ..., n\}$ with the set of estimated class probabilities $\{\hat{p_i}^k(x_i), k = 1, ..., K\}$. We consider two different combination approaches. The first approach is based on convex combinations of the estimated class probabilities while the second approach is based on fitting a model to the set of the estimated class probabilities. Both approaches have been considered in the literature (Wolpert, 1992; Breiman, 2001; Leblanc and Tibshirani, 1996).

2.2.1 Convex Combination via Linear Mixtures

One of the simplest and most straightforward methods to combine the class predicted probabilities functions $p^k(x)$ is to consider linear mixtures (Leblanc and Tibshirani, 1996) of the posterior class probabilities functions

$$p_C(x) = \sum_{k=1}^{K} w_k p^k(x)$$

where $p_C(x)$ are the new combined class probabilities and w is a vector of weights which take values in the interval $[0, 1]$ such that $\sum_{k=1}^{K} w_k = 1$. Each choice of $w_k > 0$ gives rise to a different classifier. We may seek to optimize the parameter vector w in order to construct the final prediction rule. In that way, w represents the relative contribution of each classifier, in determining the overall value of $p_C(x)$. However, it has been observed in applications that the predictive power of linear rules is often insensitive to the precise values of their coefficients due the flat maximum effect (Hand 1997, Hand 2006). This may be particularly true when all classifiers in the sum are of linear form and highly correlated. In this case we expect *a priori* that classification performance of linear classifiers is likely to be similar (Hand, 2006) while their estimated class probabilities may be variable. This expectation was confirmed by the proteomics competition outcome, since all linear classifiers provided similar classification results, all getting close to the optimal.

Taking the above into consideration, for the remaining of this paper we will restrict to the choice $w_k = 1/K$ for $k = 1, ..., K$ and thus enforcing equal weights on all classifiers in the sum. In a Baysian sense, the choice of taking equal weights reflects our prior expectation that all classifiers would have similar contribution to the derivation of the linear combination rule. Averaging across their class predicted probabilities accounts for any variability between the calibrated class probabilities and derives improved estimates.

2.2.2 Model-based Combination

An alternative approach to convex combination is based on fitting a (semi)parametric model such as logistic regression to the set of posterior class probabilities functions. Replacing the original set of predictors with the set of class probabilities may be viewed as a dimension reduction approach which reduces the predictor data to a low-dimensional space. In this low dimensional space the newly calibrated class probabilities $p_C(x)$ can be combined by fitting the logistic model

$$\log(\frac{p_C(x)}{1 - p_C(x)}) = \alpha + \sum_{k=1}^{K} \beta_k \operatorname{logit}(p^k(x))$$

However, fitting this model based on the re-substitution estimates $\hat{p}_i^k(x_i)$ can lead to

serious bias and overfitting (Leblanc and Tibshirani, 1996). For this reason it is crucial that each distinct classifier is "pre-validated" (Tibshirani and Efron, 2002) so that biased estimates will not occur. Pre-validation implies the use of cross-validation in order to avoid overfitting resulting from estimation using the same training set for the construction of the individual predictors. Therefore, we choose to use leave-one-out cross-validation to calibrate the predicted class probabilities of the classifiers to be combined. In this way we let $p^k_{-i}(x_i)$ denote the leave-one-out cross-validated fit for $p^k(x_i)$ without the i^{th} observation. The estimates from each such fit are cross-validated class probabilities, evaluated at $x = x_i$ and denoted by $\hat{p}^k_{-i}(x_i)$. We can then use the set of cross-validated class probabilities as our new input variables for the construction of the combined classifier, as suggested by Wolpert (1992) and Breiman (2001). The estimates $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ can be obtained by maximizing the cross-validated log-likelihood function of the logistic model

$$l(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \left(\alpha + \sum_{k=1}^{K} \beta_k \operatorname{logit}(p^k_{-i}(x_i)) \right) - \log \left(1 + \exp(\alpha + \sum_{k=1}^{K} \beta_k \operatorname{logit}(p^k_{-i}(x_i)) \right)$$

The above likelihood may still be ill-conditioned as a consequence of the high correlation between the new covariates $\hat{p}^k_{-i}(x_i)$ introduced by the fact that these were calibrated using the same training set. This may lead to excessively variable model parameter estimates and poor predictions. One approach to address the problem is to augment the above likelihood with a penalty term or some other form of regularization. For our first implementation we chose the quadratic penalty, which leads to logistic ridge regression, which is known to perform well in the calibration of prediction and classification rules. This method shrinks the regression coefficients towards zero and provides parameter estimates with reduced variance. We will return to the issue of model choice for the combination of classifiers in the post-hoc analysis.

Estimates for α and $\boldsymbol{\beta}$ can thus be obtained by maximizing the penalized cross-validated log-likelihood function

$$l_\lambda(\alpha, \boldsymbol{\beta}) = l(\alpha, \boldsymbol{\beta}) - \lambda \Omega(\boldsymbol{\beta})$$

where λ is the ridge penalty which controls the amount of shrinkage on the parameter vector $\boldsymbol{\beta}$ and $\Omega(\boldsymbol{\beta}) = \sum_{k=1}^{K} \beta_k^2$ is the quadratic penalty. To select the regularized parameter λ we choose to use a (leave-one-out) cross-validatory approach. The optimal value of λ is then defined as the one minimizing the cross-validated deviance

$$CVD(\lambda) = -2 \sum_{i=1}^{n} y_i \log \hat{p}^C_{-i}(x_i) + (1 - y_i)(\log (1 - \hat{p}^C_{-i}(x_i))$$

where $\hat{p}^C_{-i}$ is the estimated combination class probability of the i^{th} sample with regression parameters estimated by maximizing $l_\lambda(\alpha, \boldsymbol{\beta})$ without the i^{th} observation.

An essential difference between the convex combination and the model-based combination approach is that there is an implicit "re-calibration" phenomenon involved in the

latter, since the class probabilities $p^k(x)$ are not only combined, as in the case of convex combination, but also re-calibrated as suggested by Cox (1958). Fitting a model based on the set of the calibrated estimates from a single classification rule, instead of the set of predictors on which this classification rule was based, could alter the final predictions. Thus, one must be careful when interpreting the results from this combination approach and when comparing them directly to the performance measures of the separately calibrated − yet not re-calibrated − models themselves. We can assess this problem by re-calibrating the individual classifiers i.e. by fitting a logistic model using each time only the cross-validated predictions of a single classifier as input variables such that

$$\log(\frac{p_R^k(x)}{1-p_R^k(x)}) = \alpha + \beta_k \operatorname{logit}(p_{-i}^k(x))$$

for $k = 1, ..., K$, with $p_R^k(x)$ the re-calibrated probabilities for the k^{th} classifier. A comparison between the cross-validated predictions of the ridge logistic model combination and the cross-validated predictions of the re-calibrated logistic models would give us insight in whether the improvement in classification performance is due to combining the cross-validated estimates of the individual classifiers or due to a “re-calibration" effect one must account for.

2.3 Application and Analysis

2.3.1 Model Choice

Because of both the proteomics competition method submissions and results (Fearn, 2008; Strimenopoulou and Brown, 2008; Hoefsloot et al., 2008; Hand, 2008), we decided to consider classifier combinations for linear base classifiers ensembles. We therefore use five related linear classifiers, well established in applied sciences and classification literature and effective in high dimensions, which were also proven effective in the proteomics competition. The first three methods correspond to three different forms of regularized logistic regression. This method is commonly used in situations where the number of covariates exceeds the number of observations and/or when there are high correlations between them. Regularized logistic regression shrinks the coefficients towards zero by imposing a penalty on their size. We consider three different penalty functions, giving rise to three distinct regularization approaches. The first penalty function we use is the quadratic penalty leading to Ridge logistic regression (RLR) and shrinkage of the parameter vector (Le Cessie and van Houwelingen, 1992). The second penalty function is the absolute penalty leading to Lasso logistic regression (LLR) which allows for both shrinkage and variable selection (Tibshirani, 1996). We finally use the convex combination of the quadratic and absolute penalties leading to the Elastic Net logistic regression (ENLR) (Zou and Hastie, 2005). The next two methods we use are based on Linear discriminant analysis. This method finds a direction in space which maximizes the between-group variations while minimizing the within-group variations. Since we work in the

high-dimensional setting, Linear discriminant analysis can not be applied directly, as the pooled within-group sample dispersion matrix is singular and therefore can not be inverted to give the required estimates. There are various publications in the literature which propose ways to deal with the problem of singularity. One of the most common approaches to tackle this problem is to use a shrinkage-based estimation for the pooled within-group covariance matrix. Here we use two different forms of regularization. The first one is based on principal component decomposition (LDAP) (Krzanowski et al., 1995; Mertens, 2003) while the second one is based on ridge regularization (LDAR). In the first case, the tuning parameter which controls the amount of shrinkage of the covariance matrix is the number of principal components to keep from the first component onwards where in the second case the tuning parameter is a ridge penalty.

The above methods to be used in the combination approach, require the tuning of a regularized parameter while at the same time their predicted class probabilities need to be "pre-validated" as previously explained. This implies the use of a double cross-validatory strategy for the calibration of each individual classification rule. The outcome of this cross-validatory procedure are double-cross-validated class probabilities $\hat{p}^k_{-i}(x_i)$ which can then be used as new input variables for the construction of the combined classification rule. Details of the above double-cross-validatory procedure and its specific application in the mass-spectrometry context is presented in Mertens et.al (2006), on the basis of Stone's original paper on cross-validation (Stone, 1974). Ideas about the use of cross-validation for the combination of predictions can be found in Breiman (1996) for the regression context, and also in Wolpert (1992). We should note that the above described implementation is essentially a *sequential* cross-validatory approach, in which first the double-cross-validation predictions are generated, followed by a repeat single cross-validatory step for the calibration of the combination tuning parameter λ on the individual double-cross validation predictions, "as if given".

To optimize the tuning parameter of each individual classification method we use a grid search with a nested leave-one-out cross-validation. For RLR we use a grid of 100 λ-values, with equal space on the log scale, varying from 0.0005 to 1000. To speed up computations we use an approximated version of leave-one-out cross-validation in which the real leave-one-out regression coefficients are approximated rather than fully calculated (Meijer and Goeman, 2013). For LLR and ENLR we choose the optimal λ-value among 100 equidistant points, varying from 0.0001 to 0.1. We should mention here that in the case of ENLR, parameter estimation involves combined optimization of λ and a parameters, where a is the elasticnet mixing parameter with $0 < a < 1$. To optimize a we use a linear grid of 11 equidistant points between 0 and 1. In the case of LDAP, the maximum model size is restricted to the first few 35 principal components. Finally, for LDAR the optimal penalty is chosen among 1000 equidistant points on the log scale between 0 and 1. All the computations for RLR, LDAP and LDAR were carried out in Matlab, using programs written by the authors. In the case of LLR and ENLR, the GLM net algorithm implemented in package "glmnet" in R was used (Friedman et al., 2013).

	Classification Methods					Combination Methods		
	RLR	LLR	ENLR	LDAP	LDAR	LINC	RLGC	LGC$_{eq_restr}$
Error-rate	0.231	0.359	0.359	0.244	0.269	0.166	0.205	0.154
Brier score	0.143	0.212	0.208	0.141	0.206	0.133	0.120	0.108
Sensitivity	0.744	0.462	0.487	0.769	0.872	0.821	0.795	0.846
Specificity	0.795	0.821	0.795	0.744	0.589	0.846	0.795	0.846
AUC	0.894	0.769	0.765	0.878	0.871	0.905	0.915	0.933
Deviance	67.40	94.50	92.97	67.25	143.24	65.73	57.78	53.12

Table 2.1: Predictive performance measures of single classifiers, convex combination (LINC), ridge logistic combination (RLGC) and equality-restricted logistic combination (LGC$_{eq_restr}$) for the validation set.

2.3.2 Results

The above linear methods as well as the combination methods were fitted to the calibration set and the resulting discriminating rules were evaluated on the validation set. Table 2.1 shows the predictive performance measures for Ridge logistic regression (RLR), Lasso logistic regression (LLR), Elastic Net logistic regression (ENLR), LDA with PCA (LDAP) and LDA with ridge penalty (LDAR). For each of these procedures we calculate the error-rate, the sensitivity and specificity and the area under the ROC curve (AUC). To evaluate the accuracy of each calibrated classifier we also calculate the Brier score and the deviance, the definitions of which are given below.

$$Brier\ score = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}(x_{i_{val}}) - y_{i_{val}})^2$$

$$\begin{aligned} Deviance &= -2\sum_{i=1}^{n} y_{i_{val}} \log \hat{p}(x_{i_{val}}) + (1 - y_{i_{val}})(\log(1 - \hat{p}(x_{i_{val}})) \\ &= -2\sum_{i=1}^{n} \log(1 - |\hat{p}(x_{i_{val}}) - y_{i_{val}}|) \end{aligned}$$

where $\hat{p}(x_{i_{val}})$ is the estimated posterior class probability of the i^{th} validated sample, $y_{i_{val}}$ is the true class of that sample and n is the total validation sample size. For class assignments, we use a threshold of 0.5 and thus we assign an observation as a disease case if the estimated class probability $\hat{p}(x_{i_{val}})$ is greater than 0.5, otherwise we assign it as control.

The performance measures of the linear mixture combination approach (LINC) and the ridge logistic combination approach (RLGC) are presented in the right part of Table

	Re-calibrated Models					Combination Methods		
	RLR	LLR	ENLR	LDAP	LDAR	LINC	RLGC	LGC_{eq_restr}
Error-rate	0.218	0.372	0.369	0.210	0.269	0.166	0.205	0.154
Brier score	0.136	0.220	0.219	0.140	0.173	0.133	0.120	0.108
Sensitivity	0.769	0.410	0.436	0.769	0.872	0.821	0.795	0.846
Specificity	0.795	0.846	0.821	0.821	0.589	0.846	0.795	0.846
AUC	0.894	0.769	0.765	0.878	0.871	0.905	0.915	0.933
Deviance	63.53	98.57	98.61	67.13	91.52	65.73	57.78	53.12

Table 2.2: Predictive performance measures of re-calibrated logistic models based on cross-validated class probabilities of single classifiers and original performance measures of convex combination (LINC), ridge logistic combination (RLGC) and equality-restricted logistic combination (LGC_{eq_restr}) for the validation set.

2.1. It can be seen that there is an improvement both in error-rate and predictive accuracy, as indicated by the Brier and the Deviance scores.

In Table 2.2 we report performance measures after re-calibrating each individual classifier. We observe that predictive performance measures improve for RLR, LDAP and LDAR after re-calibration while performance measures for LLR and ENLR remain the same. Such changes in predictive performance indicate the presence of a “re-calibration" effect which could have high impact on the comparison between the individual classification methods and the combination methods. Despite these changes in the performance measures of most of the individual classifiers after re-calibration, we can still see that both combination methods outperform any single re-calibrated model.

2.3.3 Post-hoc Analysis

Figure 2.2 shows a scatter-plot representation for cases and controls separately to give an insight into the way improved predictions may occur from the linear mixture combination. Horizontal axes represent the proportion of correct assignments of an observation to a class across all classifiers while vertical axes represent the variance among the predicted class probabilities across the classifiers. Plotting symbols are shown as dots if correctly classified by the linear combination and as crosses otherwise. All observations above the 0.5 proportion of correct classification are assigned to the correct class except for one observation in the cases group. The gain of using the linear combination approach becomes more clear by looking at the data points below the 0.5 proportion of correct classification, which is where the majority of classifiers calibrates incorrect assignments. There are 5 observations in the controls group and 2 observations in the cases group for which, although three out of five classifiers do not assign them to the correct class, linear combination is

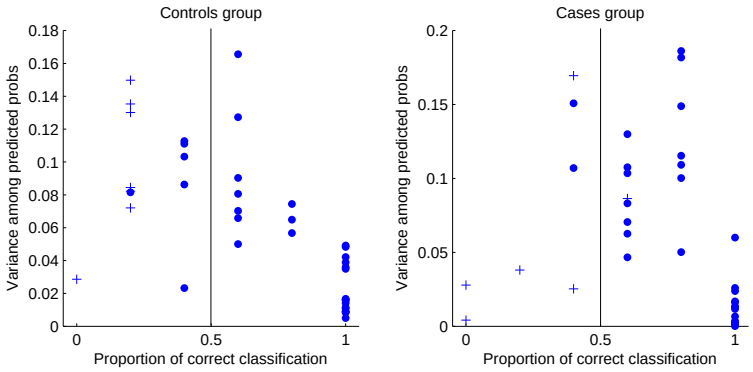

Figure 2.2: Variance between predicted probabilities versus proportion of correct classification for cases and controls separately (symbols are plotted as dots when correctly classified by the linear mixture combination and as crosses otherwise).

able to recover a correct assignment. Moreover, there is one observation in the controls group for which, while four out of five classifiers are not assigning it to the correct class, linear combination manages to recover. The estimated class probabilities for this particular observation are $0.55, 0.61, 0.59, 0.72$ and $1.5e-0.6$ for RLR, LLR, ENLR, LDAP and LDAR respectively. Thus, recovering of this observation by the linear combination is due to the class probability of LDAR which pulls the average of the class predicted probabilities towards zero. Finally, the most difficult scenario for the combination occurs when all classifiers calibrate incorrect assignments. In this case, observations will always be misclassified by the linear combination method.

In contrast to the convex combination approach, fitting the ridge logistic model combination implies that one allows the relative contribution of the individual classifiers to the combined classification rule to vary, as represented by their corresponding regression weights. The estimates for these weights can be found in the upper part of Table 2.3 in which we can see that the largest contribution is provided by RLR with estimated regression coefficient much larger in absolute terms, than any other's classifier. Moreover we can see that LDAP is associated with a negative regression coefficient which complicates the interpretation of the estimated weights. To facilitate interpretation, we refit the ridge logistic model, constraining the coefficients of all penalized covariates to be non-negative. The estimated coefficients for this fit are given in Table 2.3. Comparing the new estimates of the regression weights with the previous ones, we observe that LDAP attributes zero effect to the combined classification rule, when we make the non-negativity constraint.

	Ridge logistic model combination							
	Intercept(α)	β_{RLR}	β_{LLR}	β_{ENLR}	β_{LDAP}	β_{LDAR}	Error-rate	Deviance
Coef.	-0.0230	0.5200	0.0902	0.1730	-0.1504	0.1215	0.205	57.78
	Positivity-restricted ridge logistic model combination							
	Intercept(α)	β_{RLR}	β_{LLR}	β_{ENLR}	β_{LDAP}	β_{LDAR}	Error-rate	Deviance
Coef.	0.0016	0.3800	0.0866	0.1161	0	0.1543	0.166	54.73
	Equality-restricted logistic model combination							
	Intercept(α)	β					Error-rate	Deviance
Coef.	0.0593	0.1911					0.154	53.12
St. Err.	0.2656	0.0290						
t	0.2232	6.5962						
	Equality-restricted logistic model combination (p^{RLR} and p^{LLR} as only inputs)							
	Intercept(α)	β					Error-rate	Deviance
Coef.	0.0119	0.5024					0.205	64.03
St. Err.	0.2709	0.0789						
t	0.0437	6.3707						

Table 2.3: Maximum likelihood estimates for ridge logistic, positivity-restricted ridge logistic, equality-restricted logistic and sub-equality-restricted logistic model combination.

The largest contribution to this positivity-constrained model, is provided by RLG as before, with estimated regression coefficient $\beta_{RLR} = 0.38$ while the smallest is provided by LLR with estimated regression coefficient $\beta_{LLR} = 0.086$. The cross-validated error-rate and cross-validated deviance of the positivity-restricted ridge logistic model combination is 0.166 and 54.73 respectively.

Recalling that all classifiers in our collection are linear and thus likely to have equal predictive contribution power in determining the combined predictions $p_C(x)$, we refit the logistic model with the additional restriction that the regression coefficients are equal

$\beta_1 = ... = \beta_K = \beta$ such that

$$\log(\frac{p_C(x)}{1-p_C(x)}) = \alpha + \beta \sum_{k=1}^{K} \text{logit}(p^k_{-i}(x))$$

Although the above model combination might look somewhat similar to the convex combination approach, an essential difference between them lies in the fact that in the first case, the final class probabilities $p_C(x)$ are being calibrated using model-based estimation while in the second case, $p_C(x)$ are simply the expectation of the individual calibrated class probabilities $p^k(x)$. The maximum likelihood estimates of this model for the intercept and the regression term are shown in Table 2.3. The cross-validated error-rate of the model-based combination drops from 0.205 to 0.1538 when we constrain the regression coefficients to be equal, getting close to that of the convex combination (0.166). Similarly, the cross-validated deviance of the model decreases from 57.78 to 53.12.

We calculate Kendall's τ on the pairs of cross-validated probabilities of the ridge logistic model combination and the convex combination as well as on the pairs of cross-validated probabilities of the equality-restricted logistic model combination and the convex combination. In this way we can explore how association between the estimated probabilities of the convex combination and the model-based combination is affected when we make the restriction of equal regression coefficients in the model combination approach.Kendall's rank correlation increases from 0.7403 to 0.8342 indicating a change

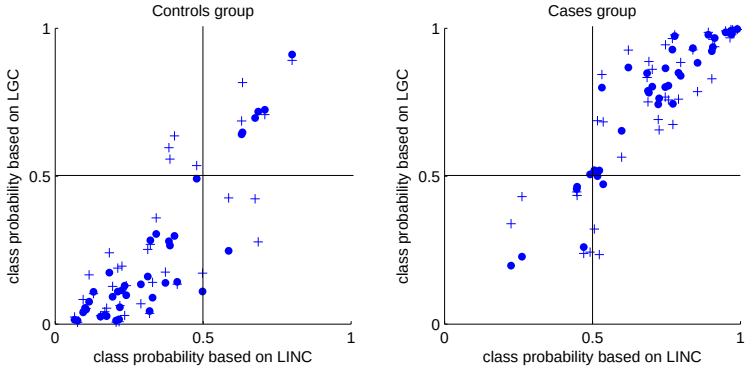

Figure 2.3: Separate scatter-plots for cases and controls versus the class probabilities of convex combination and ridge logistic model combination (plotted as crosses) and the class probabilities of convex combination and equality-restricted logistic model combination (plotted as dots).

	Individual classifiers being removed				
	RLR	LLR	ENLR	LDAP	LDAR
Error-rate	0.166	0.205	0.192	0.205	0.179
Deviance	57.30	52.38	52.10	56.43	63.75
p-value	1	0.134	0.248	0.134	0.724

Table 2.4: Validated classification results of equality-restricted logistic model combination removing a single classifier at a time and McNemar's test outcome for comparison between error-rates of full equality-restricted model and deletion models.

in the association between the class estimates of the two different combination methods. This change is represented in Figure 2.3 where we plot the class probabilities of the convex combination versus the probabilities of the ridge logistic model combination and the equality-restricted logistic model combination. The above results suggest that the assumption of similar predictive contribution across classifiers was valid, providing justification for enforcing equal weights on all classifiers in the convex combination approach. Estimates for the intercept and the regression term of the equality-restricted logistic model (LGC_{eq_restr}) are shown in Table 2.3. Predictive performance measures for this model can be found in the last column of Table 2.1.

We can further investigate the individual predictive contribution of each classifier to the equality-restricted logistic model combination by removing each one on it's own and refitting the model. This would provide some insight of the actual importance of each of those classifiers in deriving improved estimates when combined with the rest classifiers in the model-based combination. In addition, we gain insight into the robustness of the final combined classifications when deleting component methods from the ensemble of classifiers which is used. The cross-validated error-rate and the cross-validated deviance of this model is 0.166 and 57.30 respectively. Hence, removing the strongest classifier yields class assignments identical to the full equality-restricted logistic model, for all observations except one, while we observe a small deterioration in terms of predictive accuracy. We repeat the "deletion" procedure, this time removing the weakest classifier (LLR). The error-rate and deviance for this model is 0.20 and 52.38 respectively. In this case, the occurring discrepancies in classification between the full equality-restricted logistic model and the one removing the weakest classifier, are due to 4 observations in total, while predictive accuracy measures slightly improve for the latter model. We continue by removing ENLR, LDAP and LDAR at a time and we report the cross-validated error-rates and deviances of the fitted models in Table 2.4.

The reported results suggest that the equality-restricted logistic model combination is relatively robust to the deletion of an individual classifier − even if we remove the best (RLR) from the combination − while best performance for this model is achieved

when all classifiers are included in the set of predictors. We test the occurring differences between the class assignments of the full equality-restricted logistic model and the deletion-models with a McNemar's test which gives non-significant outcome for all comparisons (Table 2.4).

It is of interest to fit a "sub-model" of the equality-restricted logistic model combination which only uses RLR and LLR as predictors. This sub-model can be viewed as an alternative to the Elastic Net regularization, as instead of combining the quadratic and absolute penalties, we combine the estimated class probabilities derived from separately calibrating Ridge and Lasso. The regression estimates and the performance measures of this model are shown in the lower part of Table 2.3. Comparing the performance measures of this model to the ones obtained from fitting the Elastic Net, we see that fitting this alternative model yields a reduced error-rate of 0.20 and a reduced deviance of 64.03, as compared to the error-rate and deviance obtained with Elastic Net (0.35 and 92.97). We compare the measures between the two different approaches of combining the two regularization methods with a McNemar's test which gives highly significant outcome $(P = 0.00596)$, despite the insufficient power due to the relative small sample size. This outcome indicates a significant difference between the Elastic Net and the model combination which only uses RLR and LLR as predictors, in favor of the latter combination approach.

2.4 Simulation Study

In this section we perform a simulation study to assess performance of our proposed combination methods. Our aim is to obtain a simulation as realistic as the complex structure of this particular type of data may allow. We generate the data set based on reusing the breast cancer data set analyzed in the previous sections. In a first stage, this is achieved by keeping the class labels fixed and permuting the individuals' mass spectrums. We then add differentially expressed signals between the two different classes. The signals consist of gaussian shaped peaks, «added to the individual spectra. A peak at location a is defined as a function of the mass to charge (m/z) values x on the intensity scale»

$$Peak(x, c) = \begin{cases} r\Phi(\frac{x-a}{b}) & \text{if } c = 0 \\ (r + \kappa)\Phi(\frac{x-a}{b}) & \text{if } c = 1 \end{cases}$$

where $r \sim N(\mu, \sigma^2)$ with additional left-truncation such that $r \in (0, \infty)$, κ is a constant which controls the difference between the generated peaks for cases and controls, x is the mass/charge (m/z) value and b determines the width of the peak. In this way, we simulate exactly the same experimental setup as in our original experiment, except we add a known difference between the existing mass spectrums of the two different groups. This leaves us again with 153 samples for calibration and 78 samples for validation purposes.

We use the simulated data sets to evaluate the combination methods and compare them to each distinct classification method. For this purpose, we consider three different

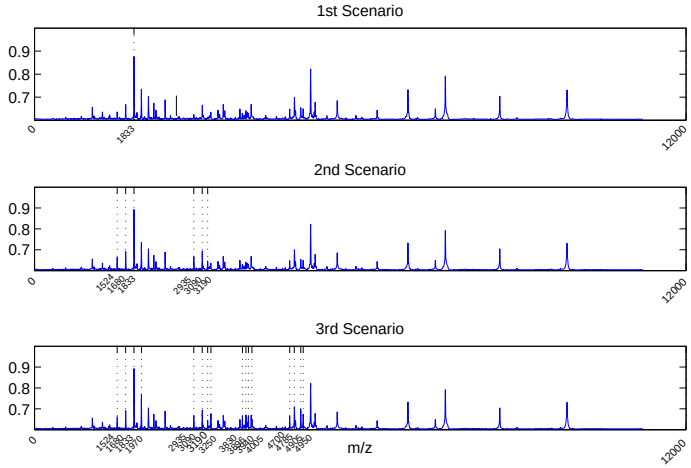

Figure 2.4: A simulated mass spectrum of a case for the three simulation scenarios (dashed lines indicate the location of the added peaks).

scenarios which correspond to different choices of the parameters μ, σ, κ and a while for all three scenarios b is fixed. We choose μ and σ so that the intensities of generated peaks match those of neighboring peaks. The choice for b is based on the typical peak-width in the breast cancer data. In the first scenario we base the discrimination between cases and controls upon a single peak differentially expressed between the two different groups. This is a situation which we expect *a priori* to favor Lasso logistic regression as this method tends to select low-dimensional models to explain the data. In the second scenario we impute 6 peaks differentially expressed between cases and controls in the same signal. Finally, in the third scenario we increase the number of differentially expressed peaks to 16 peaks in total, distributed along the individuals' mass spectrums. The simulated mass spectrums for a case for the three different scenarios are plotted in Figure 2.4. The chosen values for a, μ, κ and σ for each scenario are shown in Table 2.5.

	1st Scenario															
	p_1															
α	1833															
μ	0.03															
κ	0.15															
σ	0.2															
	2nd Scenario															
	p_1	p_2	p_3	p_4	p_5	p_6										
α	1524	1680	1833	2935	3090	3190										
μ	0.064	0.072	0.08	0.048	0.054	0.06										
κ	0.04	0.045	0.05	0.048	0.054	0.06										
σ	0.16	0.18	0.2	0.16	0.18	0.2										
	3rd Scenario															
	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	p_{16}
α	1524	1680	1833	1970	2935	3090	3190	3250	3830	3896	3940	4005	4700	4785	4905	4950
μ	0.064	0.072	0.08	0.072	0.048	0.054	0.06	0.054	0.056	0.063	0.07	0.063	0.056	0.063	0.07	0.063
κ	0.04	0.045	0.05	0.045	0.048	0.054	0.06	0.054	0.024	0.027	0.03	0.027	0.024	0.027	0.03	0.027
σ	0.16	0.18	0.2	0.18	0.16	0.18	0.2	0.18	0.16	0.18	0.2	0.18	0.16	0.18	0.2	0.18

Table 2.5: The chosen values for the parameters of the imputed peaks for the three different scenarios.

The value for b is set to 5.3 for all three scenarios. For each of the two first scenarios we repeat the simulation procedure 20 times while for the third scenario we repeat the simulation procedure 25 times. This gives rise to 20 different simulated calibration and validation data sets for the first and second scenarios and 25 simulated calibration and validation data sets for the third scenario.

In Table 2.6 we report the average error-rate, Brier score and deviance of each individual classification method and each combination method for the fist, second and third scenarios. In the first scenario we can see that the error-rate for both combination methods is lower than every individual method except LLR. The Brier score and the deviance for both combination methods are lower than those of RLR, ENLR, LDAP and LDAR while they are very close to the Brier score and the deviance of LLR. In the second and third scenarios all performance measures for both combination methods are lower than any individual classification method's.

Figure 2.5 shows separate boxplots of the distribution of error-rates, Brier scores and deviances of the five individual classifiers and the two combination methods, for the three scenarios. The top three boxplots correspond to the first scenario. As expected, LLR outperforms every single classifier both in terms of classification performance and predictive accuracy. Despite these differences between LLR and the other classifiers, we can see that error-rates for both combination methods tend to be lower than every individual method except LLR, while their distributions are similar to the distribution of LLR. This is also true for the Brier score and deviance distributions. In the second scenario all five inde-

		Classification Methods					Combination Methods	
		RLR	LLR	ENLR	LDAP	LDAR	LINC	LGC
Error-rate	1st Scenario	0.27	0.19	0.27	0.29	0.27	0.21	0.20
	2nd Scenario	0.26	0.27	0.27	0.28	0.26	0.23	0.23
	3rd Scenario	0.25	0.24	0.24	0.24	0.23	0.19	0.18
Brier score	1st Scenario	0.18	0.14	0.20	0.21	0.23	0.15	0.14
	2nd Scenario	0.20	0.18	0.18	0.21	0.22	0.16	0.16
	3rd Scenario	0.18	0.17	0.16	0.18	0.18	0.13	0.12
Deviance	1st Scenario	88.92	72.72	93.82	119.17	195.47	72.70	72.32
	2nd Scenario	109.70	86.47	84.97	112.29	103.57	76.82	80.74
	3rd Scenario	122.87	80.93	77.29	104.72	171.26	67.25	71.78

Table 2.6: Average error-rates and deviances of the individual classification methods and the combination methods for the three different scenarios.

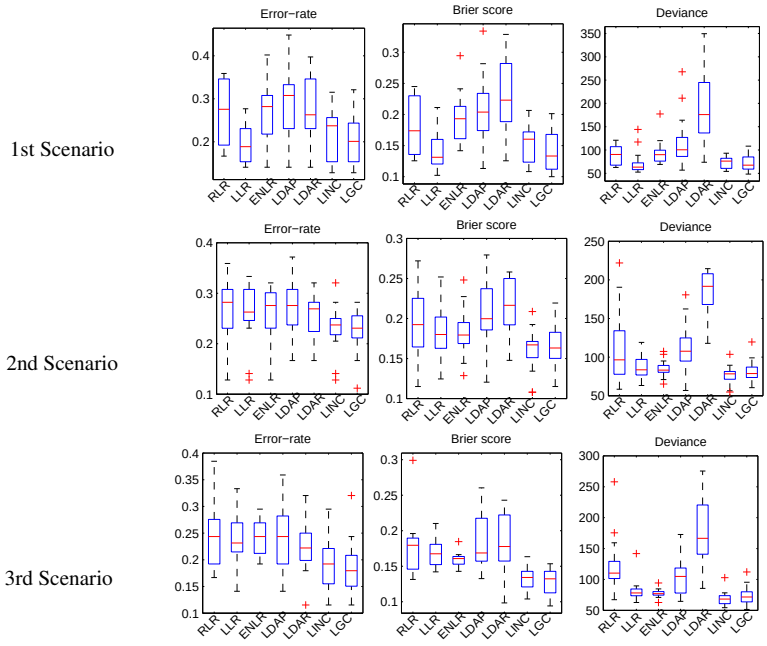

Figure 2.5: Boxplots of validated misclassification rates, Brier scores and deviances for the 1st, 2nd and 3rd simulation scenarios.

pendent classification methods perform almost equally well, as we can see from the three middle plots of Figure 2.5. We observe, on average, a slight improvement in classification performance and prediction accuracy when using LIN or LGC. The lower plots of Figure 2.5 report the results from the third scenario. Combining the different classification methods which yield similar predictive performance, accounts for the variability between the calibrated class probabilities, allowing for the derivation of improved predictions. This is particularly true for the Brier score and the deviance, as their entire distributions for both combination approaches is lower than every other classification method distribution.

Table 2.7 reports relative frequencies of improvement of each combination method on each individual classifier, as well as frequencies of each combination method having the first rank (being best classifier) as compared to all individual classifiers, according to the misclassification rate and deviance criteria. These frequencies are defined as $P(E_{p_k} > E_{LIN})$ and $P(E_{p_k} > E_{LGC})$ for the error-rate and $P(D_{p_k} > D_{LIN})$ and $P(D_{p_k} > D_{LGC})$ for the deviance, for convex and model-based combination respectively, with p_k denoting the individual classifier. The relative frequency of improvement

	$P(E_{p_k} > E_{LINC})$						$P(E_{p_k} > E_{LGC})$					
Individual p_k	RLR	LLR	ENLR	LDAP	LDAR	ALL	RLR	LLR	ENLR	LDAP	LDAR	ALL
1st Scenario	0.93	0.29	0.93	0.86	0.93	0.29	0.93	0.50	1	0.93	0.93	0.50
2nd Scenario	0.88	0.88	0.75	0.81	0.88	0.63	0.81	0.81	0.81	0.81	0.88	0.69
3rd Scenario	0.87	0.83	0.91	0.96	0.78	0.57	0.87	0.87	0.87	0.83	0.91	0.61
	$P(D_{p_k} > D_{LINC})$						$P(D_{p_k} > D_{LGC})$					
Individual p_k	RLR	LLR	ENLR	LDAP	LDAR	ALL	RLR	LLR	ENLR	LDAP	LDAR	ALL
1st Scenario	0.83	0.36	0.86	0.93	1	0.29	0.86	0.50	0.79	0.93	1	0.43
2nd Scenario	0.81	0.75	0.94	1	1	0.56	0.63	0.81	0.69	0.94	1	0.44
3rd Scenario	0.96	0.91	0.86	1	1	0.82	0.95	0.73	0.73	0.95	1	0.65

Table 7: Relative frequency of improvement of the convex combination (left part) and model-based combination (right part) compare to each individual classifier and compare to all individual classifiers in terms of error-rate and deviance.

in performance measures of the two combination methods is always greater than 0.5 for the second and third scenarios. In the first scenario in which differential expression occurs in low dimension, the relative frequency of improving on LLR is lower than 0.5 for the convex combination method and exactly 0.5 for the model-based combination method. As the estimated frequencies are based on 20 repetitions for the first and second scenarios and on 25 repetition for the third scenario, a relative frequency of more or equal to 0.75 and 0.70 respectively is significantly different from 0.5 at the 5% level. Improvements in classification performance and predictive accuracy, increase as we increase the number of added peaks differentially expressed between the cases and controls. Such improvements are more clear in the third scenario which is the one we expect to be closest to what should be expected to happen in real-life proteomics pattern recognition, since differential signal is likely to be scattered across many peaks of different magnitudes, signal intensities and variabilities, corresponding to different break-down fragments in the mass-spectrometric procedure of proteins which are present in different amounts between the cases and controls. In contrast, the first scenario is highly unlikely to occur, as it would imply that only a single differentially expressed protein or peptide was present in differing concentrations between the cases and controls, and in addition, this molecule should not fractionate either during the mass spectrometry procedure (and the initial sample-preparation).

2.5 Discussion

In this paper we considered the problem of construction of classification rules for mass spectrometry proteomic data, by using combinations of individual classifiers as opposed to calibrating a single predictor only. This work was motivated by the results, suggestions and discussion generated in the mass spectrometry proteomic competition (Mertens, 2008; Hand, 2008). We re-analyzed the competition proteomic breast cancer case-control data to evaluate the combination approach. In addition, we simulated proteomic mass spectrometry data - based on re-using existing spectrometry data - and used this to implement an extensive evaluation of the proposed combination methods. Results from both the breast cancer data analysis and from the simulation study show that gains in classification performance and predictive accuracy can be achieved with a combination approach.

We restricted the constituent classifiers used in the construction of the combined classifier to linear classification methods, in the first instance because these were found to work well in the proteomic competition and were the methods of choice for the majority of participants. Furthermore, these methods are regarded as reliable and stable generally for high-dimensional data problems with relatively small sample sizes in Omics research. Our results show that when combining linear base classifiers, as in this paper, consistent classification gains are achieved. This seems to be particularly relevant for mass-spectrometry proteomic data. To evaluate the simulations we applied a two-way ANOVA approach on the calculated deviance residuals which adjusts for simulation-to-simulation systematic differences in classification potential, while evaluating the method effect which accounts for possible systematic differences in classification accuracy between the classifiers considered. These calculations/analyses were carried out (not shown) but can be accessed in the supplementary files available with this paper. Results from both the data example analyses and the simulations show that the combination methods are clearly separated from and improving on the constituent individual classifier methods in terms of classification potential.

Since the constituent classifiers are all relatively simple discriminant methods found in most packages, one could imagine combination rules as discussed in this paper could relatively easily be calibrated in routine applications. However, this is not quite as straightforward for the model-based approach, since it requires calibration of the combination of classifiers through some form of optimization - as through maximum (penalized or otherwise constrained) likelihood estimation as in this paper - which tends to destroy the calibration of the individual constituents' class probabilities. We may solve this problem by calibrating cross-validated class probabilities for each base classifier and then replace the original predictor data with these cross-validated probabilities prior to subsequent combination. Since the base classifiers require calibration themselves, a double cross-validatory routine must then be employed, which allows for both choice of the penalty terms required for optimization of the constituent classifiers and cross-validated prediction of each individual datum in the calibration data.

Combinations through weighted sums may be more straightforward in application,

since they may not require optimization of the weights in the procedure. Whenever the weights are known, an unbiased evaluation of the predictive performance of such weighted sum discriminant combination rule is then relatively easy to derive, as the cross-validatory nature of the predicted probabilities of each separately calibrated classifier is preserved in the weighted combination. We would speculate such weighted combinations to be particularly competitive in many applications, as also found in our application, particulary with a carefully chosen ensemble of base classifiers as in this paper. Indeed the subsequent post hoc analysis carried out on the model-based approach shows that the results come very close to those obtained from the convex mixture using equal weights. If we are prepared to make such restrictive assumption *a priori*, then such mixture combinations could be calibrated with relative ease in practical application. Recalling that the estimates of each constituent classifier can be regarded as a feature, an alternative to the linear combination is the model-based combination which can be more flexible and can lead to a more elaborate classification rule.

Another aspect related to the above comparative discussion between the model-base combination approach versus the convex linear weighting method, is that the first involves an implicit re-calibration of the class probabilities. This has been much overlooked and ignored in practical applications and discussion of combination approaches, including the statistical literature. Comparisons of model-based combinations with single-classifier approaches must therefore address this effect. We have shown that even allowing for the re-calibration effect, the model-based method is still competitive above using a single constituent classifier only. Model-based approaches are clearly also attractive if one wishes to combine more general collections of base classifiers. For example, it could be of interest to explore the possibility of including non-linear classifiers (such as Random Forest) in our portfolio as well. This would be attractive when we would expect non-linearity of the decision surface. An argument against use of non-linear methods would be the typically small sample size - relative to the dimensionality of the problem - and the greater variability of the resulting calibrated class probabilities when using such more complex prediction methods. We have not explored the latter suggestion because we do not believe non-linearity to be an issue with the breast cancer data.

Model combination have seen many applications and publications within related research fields, such as in Neural networks, Machine learning and Pattern recognition. Classifier fusion in these fields tends to be based on majority voting or bagging technics which do not directly yield estimates of class probabilities. The idea of estimating a combination of predictors instead of optimizing a single predictor was first introduced in the neural network context by Wolpert (1992), referring to this idea as “stacked generalization for combining estimators”. His proposal was later applied and studied in the regression setting by Breiman (1995) on what he called “stacked regression". Wolpert’s and Breiman’s ideas about classifier ensembles are closely related to the “model-mix” proposal by Stone (1974). The problem of combing estimates in regression and classification to obtain improved predictive models is also considered in LeBlanc and Tibshirani (1996). Their paper also examines the relationship between the “stacking" and “model-mix" algorithms.

A recent paper by van der Laan et al. (2007) proposed a new estimator, introduced as the “super learner”, which is connected to the stacking idea and classifier fusion. This “super learner" is a prediction algorithm which applies a collection of candidate learners to the training data, in order to achieve better performance than any of the given candidate learners. An essential difference between super learner and our combination approach lies in the fact that given a collection of learning processes, the super learner selects the optimal learner based on cross-validated risk, discarding all others. Our combination approach retains use of all learning processes, whether through an *a priori* assumption of equal weighting or through a model-based combination on double cross-validated predictions.

Combination methods as discussed in this paper have many advantages, besides those of improved predictive ability and their relative ease of implementation, discussed above. Another property demonstrated in the post hoc analysis in this research is their relative robustness to deletion of any individual constituent classifier from which it is built. The predictive improvements themselves, though consistent, are relatively modest. This is however no criticism which should be taken against combination methods (or any other method to improve on a simple linear base classifier), as it is nothing else but confirmation of the wisdom formulated by Hand (2006) that most of classification potential can be calibrated using a simple linear classifier. Improvements beyond this point are much harder to obtain, and require either more data (larger sample sizes) or better (expert) knowledge. The latter suggestion is very interesting for spectrometry data generally, as it would seem there is much expert knowledge to be exploited in spectrometry applications. However, since the variation in spectrometry equipment and the data types they produce is huge, the usefulness of such specializations may be restricted to specific applications, or require careful and time-consuming re-tuning to each new situation, while combination methods may in principle be quickly applied across many applications in (spectrometry-based) Omics data.