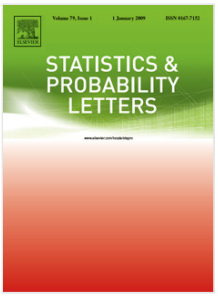
On the use of double cross-validation for the combination of proteomic mass spectral data for enhanced diagnosis and prediction

B.J.A. Mertens, Y.E.M. van Der Burgt, B. Velstra, W.E. Mesker, R.A.E.M. Tollenaar, A.M. Deelder

Accepted Manuscript



On the use of double cross-validation for the combination of proteomic mass spectral data for enhanced diagnosis and prediction

B.J.A. Mertens, Y.E.M. van der Burgt, B. Velstra, W.E. Mesker,
R.A.E.M. Tollenaar,
A.M. Deelder

PII: S0167-7152(11)00083-6
DOI: 10.1016/j.spl.2011.02.037
Reference: STAPRO 5935

To appear in: *Statistics and Probability Letters*

Please cite this article as: Mertens, B.J.A., van der Burgt, Y.E.M., Velstra, B., Mesker, W.E., Tollenaar, R.A.E.M., Deelder, A.M., On the use of double cross-validation for the combination of proteomic mass spectral data for enhanced diagnosis and prediction. *Statistics and Probability Letters* (2011), doi:10.1016/j.spl.2011.02.037

On the use of double cross-validation for the combination of proteomic mass spectral data for enhanced diagnosis and prediction.

Mertens, B. J. A.[1,*] Burgt, Y.E.M. van der[3], Velstra, B.[2], Mesker, W.E.[2], Tollenaar, R.A.E.M.[2] and Deelder, A.M.[3].

February 24, 2011

Departments of Medical Statistics[1], Surgery[2] and Biomolecular Mass Spectrometry[3], Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, The Netherlands

Keywords: Clinical proteomics, mass spectrometry, predictive data fusion, diagnosis, double cross-validation, classification, model combination.

Abstract

We consider a proteomic mass spectrometry case-control study for the calibration of a diagnostic rule for the detection of early-stage breast cancer. For each patient, a pair of two distinct mass spectra is recorded, each of which derived from a different prior fractionation procedure on the available patient serum. We propose a procedure to combine the distinct spectral expressions from patients for the calibration of a diagnostic discriminant rule. This is achieved by first calibrating two distinct prediction rules separately, each of which on only one of the two available spectral data sources. A double cross-validatory approach is used to summarize the available spectral data using the two classifiers to posterior class probabilities, on which a combined predictor can be calibrated.

1 Introduction

The need for novel tools for early diagnosis in cancer is widely acknowledged. Clinical proteomics has emerged as powerful strategy to develop such tools. This applies particularly for body fluid protein profiling based on mass spectrometry (MS), which holds great promise for personalized medicine (Aebersold *et al.* 2003). The problem of calibrating such an early diagnosis tool based on MS-prowling data obtained from patient serum samples has been considered by de Noo *et al.* (2006) and Mertens *et al.* (2006, 2008), in an application to breast cancer. These studies describe what is now a standard MS-based proteomics case-control protocol which collects both patient and healthy control samples that are pre-processed to obtain a sub-class of proteins prior to the spectral measurement. This strategy of sub-sampling is often referred to as fractionation or sample clean-up. Such a prior reduction to a protein sub-class is necessary

*Correspondence to: Bart J. A. Mertens, b.mertens@lumc.nl, http://www.lumc.nl

to obtain good quality spectrometry data. The exact approach for fractionation depends on the type of mass analysis further used, in this case matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) MS (Bruker Daltonics).

For the data presented in this paper, fractionation is performed by mixing prepared serum samples with magnetic beads which have specific physicochemical binding properties. First the proteins attach to the beads, then the mixture is washed and finally the proteins are eluted to yield a subset of proteins suitable for MALDI-TOF MS. For this purpose, the eluate is spotted onto plates after mixing with matrix solution. The original above mentioned papers discussed MS-based protein profiling using hydrophobic WCX (Weak Cation Exchange) beads for pre-processing and provide more detail on the procedures and protocols (de Noo 2006 and Mertens *et al.* (2006, 2008).

One way to address the problem of sub-sampling is to obtain repeated (spectral) measurements for each patient, each time changing the fractionation step prior to measurement so that a larger fraction of the proteome is covered by the joint set of measurements. The objective of the analysis is then to combine the distinct spectra from individuals in such a manner that allows for improved predictions or diagnostic classifications - should such be achievable - and subsequently verify the extent of improvement over using individual single-fractionation spectra only.

In this paper, we will again consider spectra generated from the WCX beads and augment the data with one additional spectrum for each individual which is generated from a distinct fractionation step using C18 beads. Details on the C18 bead fractionation have been reported previously (Villanueva *et al.* 2004)(Nicolardi *et al.* 2010). A case-control sample of 307 individuals was obtained consisting of 105 breast cancer cases and 202 healthy controls from each of which a serum sample was obtained. Fractionation was achieved in the above described manner using the two distinct beads, each time processing part of the serum with each bead type separately, resulting in two processed serum samples from each of which a MALDI-TOF spectrum is obtained. For the purpose of this paper, we restrict attention to spectra reduced to a list of identified peaks, each of which is summarized to its integrated value under the peak curve. This reduction is however not fundamental to the methodology described further on. This gives for each i^{th} patient, $i = 1, ..., n$ a paired sets of spectral measurements $\boldsymbol{X}^1$ and $\boldsymbol{X}^2$ consisting of the spectral data generated from the first (WCX) and second (C18) bead processing respectively such that

$$\boldsymbol{X}^1 = \begin{pmatrix} \boldsymbol{x}_1^1 \\ . \\ . \\ . \\ \boldsymbol{x}_n^1 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{X}^2 = \begin{pmatrix} \boldsymbol{x}_1^2 \\ . \\ . \\ . \\ \boldsymbol{x}_n^2 \end{pmatrix}$$

where $\boldsymbol{x}_i^1 = (x_{i1}^1, ..., x_{il}^1)$ and $\boldsymbol{x}_i^2 = (x_{i1}^2, ..., x_{im}^2)$ with $n = 307$ and $l = 48$ and $m = 42$ represent the dimensionality of the peak list for the WCX and C18 based spectra respectively. To complete the observed information on individuals, we have the binary case-control outcome Y which equals 1 for cases or 0 for controls.

The structure of the paper is as follows. In the first instance we repeat the double cross-validatory analyses proposed by Mertens *et al.* (2006) for both the WCX and C18 spectra separately. We then explain how these 'within-bead'

double cross-validatory results may also be used to allow predictive combination and evaluation of the distinct spectra. Next we present a comparative analysis between the single-bead analyses and the double cross-validation combination and demonstrate how predictive performance improves for the combination. For consistency with existing results and research, we focus on analysis and combination of linear discrimination results as presented in the original papers of Mertens *et al.* (2006) in the first instance and then also introduce an alternative analysis via Random Forests for comparison. Finally, we show that naïve combinations of the two spectral data sets do not improve on single-spectrum based approaches for these data.

2 Single-spectrum based analysis

Mertens *et al.* (2008) and de Noo *et al.* (2006) describe a double cross-validatory implementation of linear discriminant analysis for the calibration of a diagnostic rule based on a single spectrum per patient (and for a single fractionation). The key idea in double cross-validation is to embed a (leave-one-out) 'inner' cross-validatory loop within a secondary 'outer' cross-validatory loop, where the inner loop is used to identify an optimal tuning parameter (in some sense) for calibration of diagnostic performance of some chosen classifier rule and the outer loop is used to obtain an unbiased estimate of the predictive performance of the approach across all observations by applying the chosen optimized rules from the inner layer to the left-out datum. The final calibration of the prediction rule may then be obtained from one last application of a single leave-one-out cross-validation to the chosen classifier scheme - or alternatively - investigating the chosen tuning parameters for each individual left-out datum within the outer loop, as chosen based on cross-validated assessment within the inner loop. Yet another alternative for the prediction of new observations is to use the full set of n predictors calibrated in the inner loop for new observations. Mertens *et al.* (2006) describe the procedures and application for spectrometry in detail, which goes back to suggestions in Stone's seminal paper on cross-validation (1974, pages 126-7) whose ideas were further developed by Wolpert (1992) and subsequently applied by Breiman to the regression context (Breiman 1996).

For consistency with the original papers, we will focus on simple linear discrimination based on prior principal component dimension reduction, such that the only tuning parameter is the number of components to keep from the first component onwards. We note however that this choice is not crucial and any classifier could be used, such as a ridge shrinkage based calibration, a Random Forest or any other method. The presented approach has been thoroughly validated on past spectral data and experiments (see above and related papers such as Alagaratnam *et al.* 2008) and proved effective in the International Competition on Proteomic Diagnosis (Mertens 2008). We also refer the interested reader to Hand (2006), for discussion on the relative merit of simple linear classifiers, which arguments apply generally and certainly for relatively small sample sizes as in this experiment.

Table 1 shows classification results from separate double cross-validatory analyses for both the WCX and C18 based data using the linear discriminant approach (second and fourth column LIN). Shown are the error rate, Brier score (B), deviance and area under the ROC curve (AUC), all of which are based on

the double cross-validated predicted class probabilities. Our definitions for Brier score, deviance and AUC are as below.

$$\mathrm{B} = \frac{1}{n}\sum_i [1 - p(c(i) \mid \mathbf{x}_i)]^2,$$

$$\mathrm{deviance} = -2\sum_i \log(p(c(i) \mid \mathbf{x}_i))$$

where $p(c(i) \mid \mathbf{x}_i)$ is the double cross-validated predicted a-posteriori class probability for the correct class $c(i)$ for each i^{th} sample and n is the total sample size.

$$\mathrm{AUC} = \frac{1}{n_1 n_2}\sum_{i\in G_1}\sum_{j\in G_2}[I(p(1 \mid \mathbf{x}_i) > p(1 \mid \mathbf{x}_j)) + 0.5 * I(p(1 \mid \mathbf{x}_i) = p(1 \mid \mathbf{x}_j))],$$

where G_1 and G_2 refer to the sample index labels for the cases and controls group respectively. For class assignments, we use a threshold of 0.5 here and throughout the remainder of the paper, which is arbitrary but sufficient for evaluation and comparison of diagnostic potential. It can be seen that both WCX and C18 have comparable classification performance, though the deviance for the C18 data is slightly larger.

Table 1: Double cross-validatory classification performance measures.

	WCX		C18		WCX and C18		
	LIN	RF	LIN	RF	MIX	RF	LG
error	0.15	0.20 (0.20)	0.14	0.26 (0.33)	0.091	0.091 (0.091)	0.098
Brier	0.11	0.15 (0.15)	0.11	0.18 (0.18)	0.084	0.088 (0.088)	0.080
deviance	242.4	288.5 (287.7)	266.8	333.9 (335.7)	186.9	221.1 (220.9)	175.9
AUC	0.91	0.85 (0.85)	0.89	0.78 (0.77)	0.94	0.92 (0.92)	0.93

For comparison, we also calculated Random Forest classifiers (Breiman 2001) for both the WCX and C18 data, using standard calibration settings, sampling 5000 trees per Forest, each calibrated from bootstrap samples of size 307 drawn with replacement and using a random input selection of 6 measurements for splitting at each node (the default value - floor of $\sqrt{42}$ and $\sqrt{48}$). Optimal splits were selected based on the Gini index (Hand 1997). Within each Forest, trees were grown till node purity was obtained for all final nodes and without further pruning. Calculations were carried out in Matlab (Matlab 2010) using a MEX wrapper file which interfaces to Andy Liaw's C code which is used in the R package randomForest (Cran 2010).

Rather than reporting the out-of-bag classification results in the first instance, we embedded the Random Forest calibrations in a single leave-one-out outer loop, such that for each left-out datum the entire Random Forest calibration was repeated and then applied to the left out datum, for consistency and in analogy with the doubly cross-validated linear discriminant approach. Results are presented in columns three and five of table 1 (RF) and show the performance to be consistently worse than achieved with the linear approach. The results from a single Random Forest calibration using the out-of-bag predictions are shown in brackets for comparison. The median out-of-bag error

rates for this single classifier were 0.20 and 0.27 for WCX and C18 respectively. Clearly, the RF-based calibrations are not competitive with those from the linear approach and hence the remainder of the paper will be based on the double cross-validation linear discriminant class probabilities.

3 Double cross-validatory combining of spectra for prediction and classification

Double cross-validation is not restricted to the joint calibration and assessment of predictive rules. An alternative view of double cross-validation relevant to predictive data combination or fusion is that it replaces the original sets of predictors $\boldsymbol{X}^1$ and $\boldsymbol{X}^2$ with the sets of double cross-validated predicted probabilities $\boldsymbol{p}^1 = (p_1^1, ..., p_n^1)^T$ and $\boldsymbol{p}^2 = (p_1^2, ..., p_n^2)^T$. These may be used as joint new input variables for the construction of a combined classifier, as has already been recognized by Wolpert (1992) and Breiman (1996) for example. Nevertheless, the focus of these authors was primarily on the combination of *distinct* models (or different re-calibrations of the same basic model or classification method) on the *same* data. In principle however, there is no restriction within the basic idea to prevent application to the combination of the same basic model calibrated from distinct datasets, which will be the focus and approach for the remainder of this paper. For this reason, we will in the following on occasion loosely refer to 'model' and 'dataset' interchangeably, as we are effectively combining the data through the combination of the distinct models calibrated on them. Model combination methods have seen many applications and publications since these seminal papers, the paper by Datta and Datta (2010) being a recent addition for example.

3.1 Linear mixture combination

One of the simplest ways to combine the prediction $\boldsymbol{p}^1$ and $\boldsymbol{p}^2$ is to consider mixed versions of the separate prediction probabilities

$$\boldsymbol{p} = w\boldsymbol{p}^1 + (1-w)\boldsymbol{p}^2$$

with $\boldsymbol{p} = (p_1, ..., p_n)^T$ the newly calibrated class probabilities vector and w some number in the interval $[0, 1]$. Each choice of w provides a combined classifier and thus we may seek to optimize the final prediction rule by estimating the parameter w in some sense. Viewed this way, the parameter w expresses the optimal balance between the WCX and C18-based predictions, with the two extreme choices with w either 0 or 1 excluding the WCX or C18 set completely. The latter could also give rise to a testing problem on the relative merit of the WCX or C18 measurements (see further).

To start the analysis, we will take an even simpler view of the problem by focusing solely on the predictive aspect, by taking $w = 0.5$ and thus enforcing equal weight on both the WCX and C18 sets. This latter choice may be defended by the *a priori* expectation that both the WCX and C18 sets must likely have some complementary predictive power - or at least predictive ability in their own right - and hence combination should allow for the derivation of improved classifiers, as well as the flat maximum effect (Hand 1997) which says that the

predictive ability of classifiers tends to be relatively insensitive to the precise choice of relative weights about some optimal value, hence the choice of equal weights here. The first assumption could be supported by the simple fact that the two sets were measured at all, implying the expectation.

The advantage of this approach is that no further optimization of any kind is required and - most importantly - that the double cross-validatory nature of the predicted probabilities is preserved entirely, which allows us to directly obtain the unbiased evaluation of predictive performance. These are presented in the sixth column of table 1 (MIX). It is clear that the combination improves upon both the WCX and C18-only based results, and this both in terms of error rate and accuracy, as evident from both the Brier and deviance scores.

Figure 1 shows a scatterplot representation of both the WCX and C18 data for cases and controls separately to give an insight in how the combination allows for improved predictions. Vertical axes are case probabilities based on

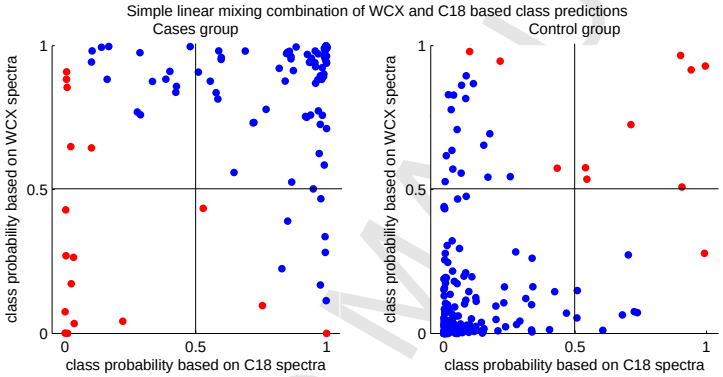


Figure 1: Separate scatter plots for cases and controls versus the double cross-validatory posterior class probabilities calculated from the WCX spectral data only (on the y-axis, $\boldsymbol{p}^1$) and from the C18 spectral data only (x-axis, $\boldsymbol{p}^2$). Symbols are plotted blue when correctly classified by the linear mixture combination and red otherwise.

the WCX data ($\boldsymbol{p}^1$) and the corresponding case probabilities based on the C18 data are on the horizontal axis ($\boldsymbol{p}^2$). Plotting symbols are shown as blue if correctly classified based on the mixture combination and red otherwise. The figure shows that large discrepancies can occur between the WCX and C18 based assignments, which are found in the first and fourth quadrant. The picture shows how the combination works, as data above the first diagonal within these quadrants is assigned to the correct class for the cases data and likewise for observations below the first diagonal for control data. There are 57 observations in the first and fourth quadrants in total, 29 of which are cases of which 21 are classified correct as yet by the combination and 28 are controls, of which 24 are recovered by the combination. Data within second and third quadrant will be

classified correctly always by the combination for the cases and control groups respectively, as both methods agree in these quadrants. There are 234 such observations, of which 67 in the cases group and 167 in the control group. All observations must be misclassified by the combination within the third and second quadrant for cases and controls respectively, as both methods jointly calibrate incorrect assignments here, which can not be recovered by the linear mixture combination. There are 16 such observations of which 9 cases and 7 controls. The classification improvement is thus mainly due to 45 (21+24) out of 57 observations in the first and fourth quadrant shifting to the correct assignment (here implicitly counting any incorrect assignment by either - but not necessarily both - the WCX or C18 single-spectrum based methods as a 'false' try) as well as improvement in precision of the calibrated posterior class probabilities.

3.2 Random Forest combination

Calibration

The above linear mixture restriction may be overly restrictive as it enforces the first diagonal as the decision surface. To evaluate this, we ran a Random Forest classifier on the set of double cross-validation predictions $\{\boldsymbol{p}^1, \boldsymbol{p}^2\}$. Because the Random Forest classifier involves additional data-based estimation and to maintain the double cross-validatory estimation of classification performance, we must again embed the Random Forest calibrations within an additional single leave-one-out cross-validatory loop as explained before for the single-spectrum based RF analysis, this time leaving out each double cross-validated pair (p_i^1, p_i^2) in turn and calculate the Forest on the remainder. We used identical settings as given for the single-spectrum analysis, except that the number of randomly selected input measurements for splitting at each node was set to 1, such that either the WCX-based or C18-based probabilities were used for splitting any node. The seventh column of table 1 shows the results (RF), which indicate that although the RF result comes very close to the mixture-based results no further improvement is obtained in a predictive sense above those already given by the simple mixture-based analysis. The double cross-validated deviance grows larger indicating some loss of accuracy. The classification results from a single Random Forest calibration using the out-of-bag based predictions are shown in brackets for comparison. The median out-of-bag error rate for this single classifier was 0.091.

Figure 2 shows two graphical representations of the obtained data combination. Two graphs are shown with the axes defined as in figure 1. The left figure shows the fitted decision surface, plotting light blue for regions assigned to control and light red for the assigned cases region. The actual observations are superimposed with blue plotting symbol for controls and red for cases. The right plot shows the contours of the fitted class probability surface.

Model importance measures

Model importance measures may be derived based on a single Random Forest calibration by randomly permuting the (in this case two, one for the C18 data, and the other for the WCX data) input variables separately after the model fit and running the new permuted input matrices down the Forest for each such

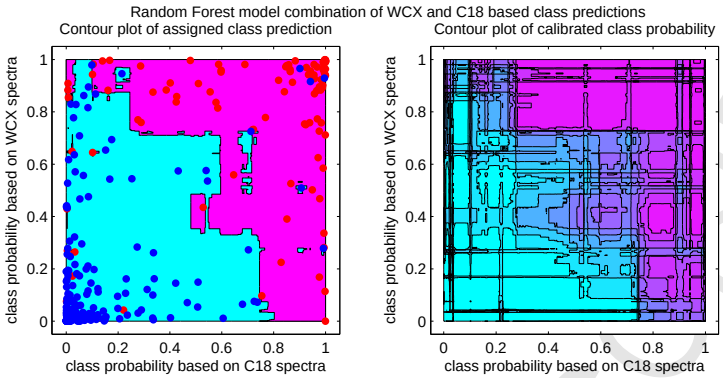


Figure 2: Random Forest decision surface (left plot) and posterior probability surface (right plot) versus single-spectrum double cross-validated posterior class probabilities as in figure 1. Light red indicates assignment to the cases class and blue for the control group.

permutation, each time monitoring changes in predictive performance. Since we are using the double cross-validatory calibrated probabilities $\{\boldsymbol{p}^1, \boldsymbol{p}^2\}$, these importance measures are not affected by correlations between peaks within the spectra (see Barrett *et al.* 2009 for discussion on Random Forest model importance in the presence of extreme correlation in proteomic mass spectrometry application). Table 2 shows importance measures (imp) defined as the mean percentage reduction in classification accuracy for cases and controls separately as well as for total accuracy. Associated standard deviations and t-values are also shown. Results suggest that mean classification performance reductions are about equal for both WCX and C18 using this measure and much larger than the associated standard errors. The total mean reduction in Gini index was 74.2 and 78.7 for the C18 and WCX data respectively. An alternative comparison is

Table 2: Importance measures based on a single Random Forest calibration.

	cases		controls		total	
bead	imp (s.e.)	t	imp (s.e.)	t	imp (s.e.)	t
C18	0.18 (0.0093)	19.7	0.14 (0.0093)	14.6	0.15 (0.012)	12.3
WCX	0.17 (0.0091)	18.8	0.13 (0.0088)	14.8	0.14 (0.012)	12.3

between the median out-of-bag error rates of a single Random Forest on the joint WCX and C18 data (0.091, Sens=0.84, Spec=0.95) and those from a RF calculated on the WCX-based probabilities $\boldsymbol{p}^1$ only (0.22, Sens=0.69, Spec=0.83) and likewise, when only using the C18-based probabilities $\boldsymbol{p}^2$ (0.19, Sens=0.73, Spec=0.84). This leads to somewhat smaller but still substantive differences as compared to the importance measures calculation.

It is of interest to use the Random Forest to obtain a measure of proximity between observations, defined as the number of times any pair is classified into the same end-node, which provides a symmetric matrix of proximity measures (Breiman 2004). Unfortunately, multi-dimensional scaling representations of this matrix have been found to be of less use in practical data analyses as the graphs tend to have similar 'star-like' configurations across distinct datasets. We therefore derived an alternative summary by calculating for all case-to-case pairs the percentage of available pairs having a proximity value at least as great as $\lambda \in [0, 1]$ and monitoring the reduction in the fraction as λ increases. This calculation is then repeated for the set of all control-to-control pairs and likewise for the set of cases-to-control pairs. The resulting summary is automatically adjusted for the size differences in the available pairs sets. Figure 3 shows the graph which shows the decreasing percentages of pairs being similar when the desired proximity level is increased. Controls seem to have higher similarity levels as compared to the cases group, which may correspond to heterogeneity due to distinct cancer subtype classifications within this group. Cases-to-controls have greatest dissimilarity at any level of proximity, as would be expected in a classification problem where discriminant information is available to (partly) distinguish cases from controls.

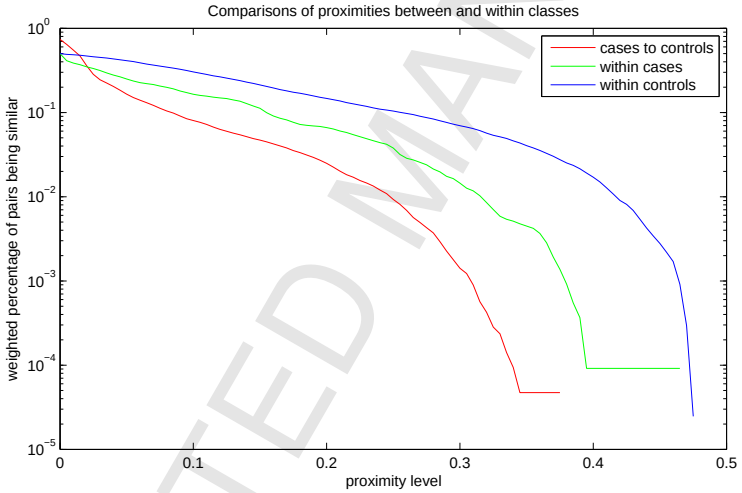


Figure 3: Fractions of pairs sharing a specified level of proximity (or higher) for the case-to-case, control-to-controls and cases-to-controls sets of pairs.

3.3 Logistic regression calibration combination

Calibration

An alternative to linear mixture combination or a non-parametric approach such as Random Forest is to calibrate the final class probabilities through a (semi) parametric model such as logistic regression, based on the set $\{\boldsymbol{p}^1, \boldsymbol{p}^2\}$. As double cross-validation reduces the predictor data to a low-dimensional space,

there is no need to use special optimization tools such as shrinkage regression and simple maximum likelihood fitting can be used - as long as we embed the calibration within a single leave-one-out cross-validatory layer as for the RF calibration. Strictly speaking, a more apt description of such an approach would be as logistic regression calibration, since the double cross-validatory summaries are not only combined (as in the linear mixture method) but also recalibrated, as suggested by Cox (1958). The last column of table 1 shows the cross-validation based performance results in the last column (LG), which are comparable with those from the linear mixture and Random Forest except for a further reduction in deviance.

Model importance measures

Table 3 shows the maximum likelihood estimates from a fit of the logistic model

$$\log \frac{p_i}{1-p_i} = \alpha + \beta_1 p_i^1 + \beta_2 p_i^2$$

on the full data. The cross-validated deviance of this full model equals 175.9

Table 3: Logistic regression maximum likelihood estimates.

	coef	s.e.	t
Intercept (alpha)	-3.62	0.38	-9.47
WCX (β_1)	3.85	0.57	6.76
C18 (β_2)	3.48	0.57	6.06

(error rate=0.098, B=0.080, AUC=0.93), which increases to 226.60 on removing the WCX data (difference of 50.7 and error rate=0.14, B=0.11, AUC=0.86) and to 216.2 on removing the C18 data (difference of 40.3 and error rate=0.14, B=0.11, AUC=0.90). It can be seen how both the C18 and WCX estimated regression effects are of comparable magnitude and highly significant, judging from both the deviance changes as well as from the sizes of estimated regression effects in comparison to standard errors. An alternative comparison is directly via the calibrated class assignments of the models themselves by comparing the error rates between the cross-validated predictions from a recalibration of the logistic model using the WCX data only ($\boldsymbol{p}^1$) and the cross-validation predictions based on the combination logistic model (using a threshold of 0.5). We score the cross-validation classifications as either correctly or incorrectly identified (0 or 1) and then compare these scores with a McNemar test, which gives highly significant test outcome (P=0.024), indicating difference between the error rates when leaving out the WCX data. Repeating this procedure for a comparison using classifications based on C18 data only ($\boldsymbol{p}^2$) and the logistic model-based combination approach gave a similarly significant result (P=0.024).

The observed difference between estimated regression coefficients is 0.37 with a 95% confidence interval estimate of (-1.31, 2.04). We next refit the logistic model with the additional restriction that both regression coefficients are equal $\beta_1 = \beta_2 = \beta$ such that

$$\log \frac{p_i}{1-p_i} = \alpha + \beta p_i^1 + \beta p_i^2 = \alpha + \beta(p_i^1 + p_i^2)$$

such that we restrict the decision surface to be parallel to the first diagonal (see figure 1). The fitted regression coefficients are -3.59 and 3.66 for intercept and regression term respectively with a cross-validated model deviance of 174.2 (difference of 1.6 relative to the full model and error rate=0.097, B=0.078, AUC=0.93). Class assignments are identical between both models for all observations except two (using the McNemar test procedure on cross-validated scores as explained above gives P$\approx$ 1).

These results provide some confirmation on the validity of the linear mixture assumption using equal weights between both WCX and C18 spectral data. They are however also in line with the Random Forest 'model importance' measures which suggest similar strength predictive contributions from both the WCX and C18 data and that the combination of data sources improves on using a single-source spectrum only.

4 Discussion

While this paper is concerned with predictive combination of multiple proteomic spectra, it could form a general template for the problem of predictive calibration when distinct 'omics' data sources must be combined. Such combinations are difficult because they are easily affected by systematic differences in either scaling or batch effects from set to set, which causes problems for most standard shrinkage methods, such as ridge, lasso or dimension reduction based approaches. To illustrate this problem, we carried out a simple double cross-validatory linear discriminant analysis on the combined original WCX and C18 data matrix $\boldsymbol{X} = [\boldsymbol{X}^1 \boldsymbol{X}^2]$. This gave a total error rate of 0.16 (Sens=0.76, Spec=0.88), B=0.12, deviance=250.8 and AUC=0.90 which is close to but nevertheless slightly worse than the WCX or C18-only based results reported in table 1. Since Random Forests are based on variable selection, the calibration should be insensitive to scaling and hence we estimated leave-one-out posterior class probabilities using RF on the joined dataset. The cross-validated performance measures were 0.18 for the total error rate (Sens=0.63, Spec=0.92), B=0.15, deviance=291.2 and Auc=0.84. Hence, a naïve data combination approach based on simply combining the measures data does not seem to work for this data, at least for classical linear discriminant methods or an off-the-shelf classifier such as RF.

One of the advantages of the double cross-validatory combination approach is that it automatically adjusts for such effects as differences in scaling between sets, simply because it replaces the original data with the calibrated posterior class probabilities within each set separately prior to the combination. The prior calculation of the double cross-validatory summaries is computationally highly intensive. However, significant gains in computing can be achieved through application of some rank downdating matrix algebra as explained by Mertens (1998, 2001). Once these cross-validatory summaries have been obtained, the predictive combination is almost for free (mixture with fixed weights) or requires an additional single leave-one-out loop at most as explained for the RF or logistic regression model combinations.

A second advantage is that standard variable importance measures or model comparison techniques can be used to judge the relative strength of the contributions to the combined predictor, as we have shown for both the RF and logistic

regression combination. Furthermore, these combinations can be executed with any standard statistical software package (such as SPSS *e.g.*) - once the double cross-validatory summaries have been obtained or provided - which in turn implies that any standard statistical classification approach could in principle be used to combine the summary predictors $\{\boldsymbol{p}^1, \boldsymbol{p}^2\}$, besides the methods used in this paper. This can be important in practical medical sciences consultation settings where such data must be routinely combined, as such work could in principle be done by clinicians or biomedical researchers after some basic training and with supervision in statistical analysis, after the double cross-validatory summaries have been provided.

This paper has focused on the methodological problem of predictive combination solely. Application and publication of the results in a clinical setting would however require further work in identifying the underlying peaks which drive the classification (see also comments by Breiman 2001 on this aspect). In principle however, no novel methodology is required for this as this could be easily achieved through repetition of the analyses described in Mertens *et al.* (2008) and then simply repeating some of the analyses presented in this paper, which is a common approach to the problem in many clinical applications papers. As this is beyond the purpose of this publication, we leave this work for presentation in a follow-up paper which will go into more depth on these aspects.

References

Aebersold, R., Mann, M. (2003) Mass Spectrometry-based Proteomics. Nature, **422**, 198-207.

Alagaratnam, S., Mertens B. J., Dalebout, J. C., Deelder, A. M., van Ommen G. J., den Dunnen J. T. and 't Hoen P. A. (2008) Serum protein profiling in mice: Identification of Factor XIIIa as a potential biomarker for muscular dystrophy Proteomics, **8**, 8, 1552-63.

Barrett, J. H. and Cairns, D. (2009) Random Forest classification and variable importance measures based on proteomic profiles from mass spectrometry. ISI World Statistics Congress Proceedings, 57^{th} Session.

Breiman, L. (1996) Stacked regressions. Machine Learning, **24**, 49-64.

Breiman, L. (2001) Random Forests. Machine Learning, **45**, 5-32.

Website describing the Random Forest package at http://oz.berkeley.edu/users/breiman/RandomForests/

Cox, D. R. (1958) Two further applications of a model for binary regression. Biometrika, **45**, 562-565.

The R Project on Statistical Computing. http://www.r-project.org/

Datta, S, Pihur, V. and Datta, S. (2010) An adaptive optimal ensemble classifier via bagging and rank aggregation with applications to high dimensional data, BMC Bioinformatics, **11**, 427 .

Hand, D. J. (1997) *Construction and Assessment of Classification Rules.* Chichester: Wiley.

Hand D.J. (2006) Classifier technology and the illusion of progress (with discussion). *Statistical Science*, 21, 1-34.

Matlab, version 7.10. The MathWorks, Inc.

Mertens, B. J. A. (1998) Exact principal component influence measures applied to the analysis of spectroscopic data on rice. *Applied Statistics*, **47**, 4, 527-542.

Mertens, B. J. A. (2001) Downdating: interdisciplinary research between statistics and computing. *Statistica Neerlandica. Statistica Neerlandica*, **55**(3), 358-366.

Mertens, B. J. A. (*ed.*) (2008) Competition On Clinical Mass Spectrometry Based Proteomic Diagnosis. *Statistical Applications in Genetics and Molecular Biology*, **7**, 2.

Mertens, B. J.A., De Noo, M.E., Tollenaar, R.A.E.M. and Deelder, A.M. (2006) Mass Spectrometry Proteomic Diagnosis: Enacting the Double Cross-Validatory Paradigm. *Journal of Computational Biology*, **13**, 9, 1591-1605.

Nicolardi S., Palmblad M., Dalebout H., Bladergroen M., Tollenaar R.A.E.M., Deelder A.M., van der Burgt Y.E.M. Quality Control Based on Isotopic Distributions for High-Throughput MALDI-TOF and MALDI-FTICR Serum Peptide Profiling. J Am Soc Mass Spectrom, **21**,1515-25.

de Noo, M.E., Deelder,A.M., Mertens, B.J. A., Ozalp, A., Bladergroen, M.R., van der Werff, M.P.J., Tollenaar, R.A.E.M.(2005) Detection of colorectal cancer using MALDI-TOF serum protein profiling. *Eur. J. Cancer*, **42**, 1068-1076.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). J. Roy. Statist. Soc. **36**, 111-147.

Villanueva, J., Philip, J., Entenberg, D., Chaparro, C. A., Tanwar, M. K., Holland, E. C., Tempst, P. (2004) Serum Peptide Profiling by Magnetic Particle-Assisted, Automated Sample Processing and MALDI-TOF Mass Spectrometry. Anal. Chem., **76**, 1560-1570.

Wolpert, D. H. (1992) Stacked generalization. Neural networks, **5**, 241-259.