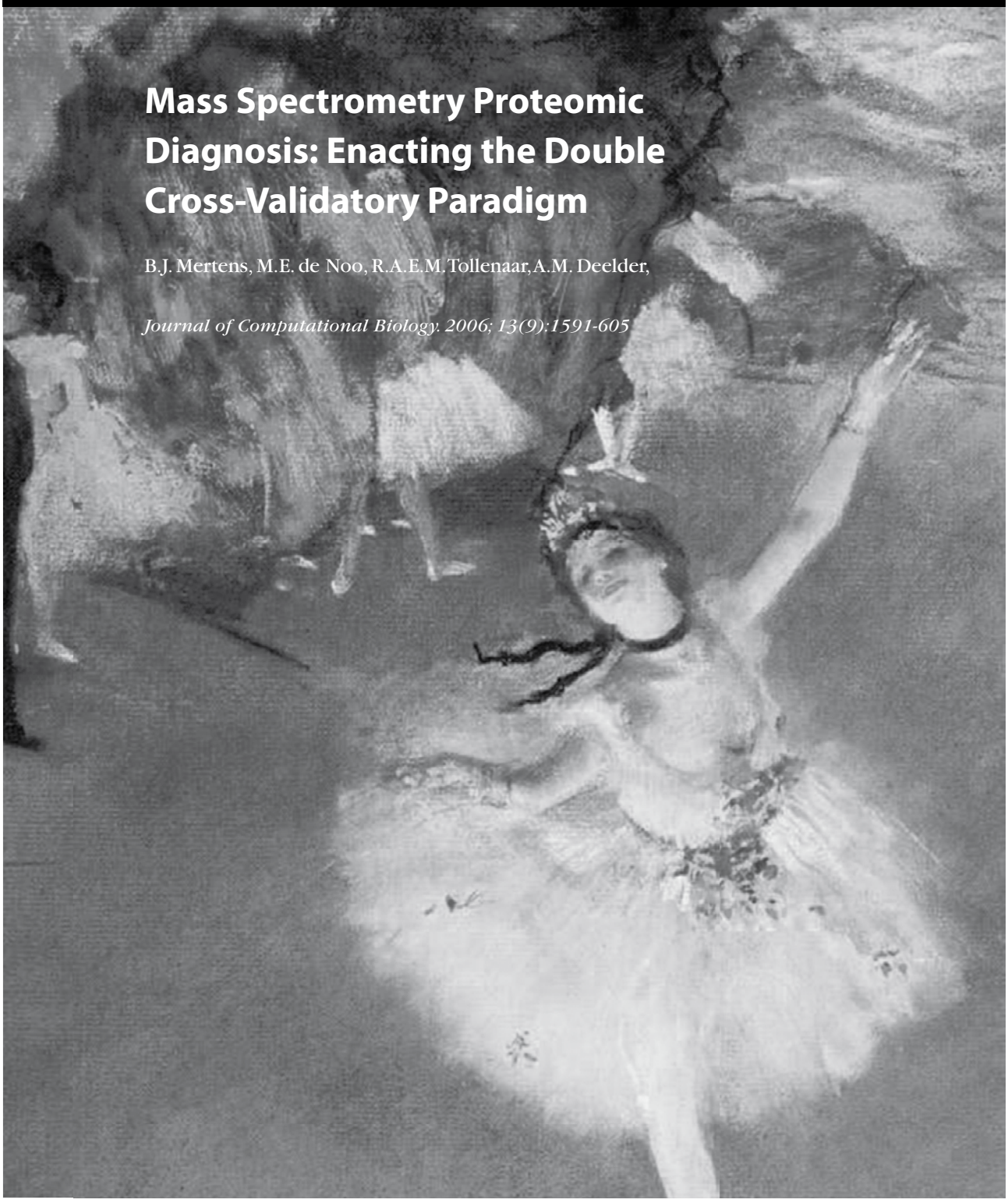
Clinical proteomics in oncology : a passionate dance between science and clinic

Noo, M.E. de

Chapter 5



Mass Spectrometry Proteomic Diagnosis: Enacting the Double Cross-Validatory Paradigm

B.J. Mertens, M.E. de Noo, R.A.E.M. Tollenaar, A.M. Deelder,

Journal of Computational Biology. 2006; 13(9):1591-605

ABSTRACT

This paper presents an approach to the evaluation and validation of the diagnostic potential of mass spectrometry data in an application on the construction of an 'early warning' diagnostic procedure. Our approach is based on a full implementation and application of double cross-validatory calibration and evaluation. It is a key feature of this methodology that we can jointly optimize the classifiers for prediction while simultaneously calculating validated error rates. The methodology leaves the size of the training data nearly intact. We present application to data from a designed experiment in a colon-cancer study. Subsequent to presentation of results from the double cross-validatory analysis, we explore a post-hoc analysis of the calibrated classifiers to identify the markers that drive the classification.

INTRODUCTION

There is currently much interest in application of mass spectrometry for the construction of new diagnostic proteomic approaches for the early detection of disease. This is particularly the case in oncology, where there is need for new and reliable diagnostic tests. In this paper, we discuss the problem of ascertaining the presence of discriminatory information in mass spectra of serum samples in a case-control study for the detection of colorectal cancer. In other words, we describe -in essence -an early-stage feasibility study for subsequent construction of a diagnostic test based on proteomic mass spectra. A crucial objective of such research is to provide information which allows researchers to make informed decisions as to the continuation of the research effort (which may involve experiments of much greater cost and complexity in comparison to the first-stage evaluation). Hence, it is essential to get a fully validated and unbiased assessment of predictive error rates that may be achieved, based on the proteomic data. At the same time, in a high-dimensional setting such as mass spectrometry, it is desirable that construction of the diagnostic classifier would involve calibration of the predictive potential of the allocation rule itself.

Mass spectrometry proteomics, sample size and clinical science

In problems such as these and related settings (*e.g.*: microarray diagnostics, chemometric discriminant studies), a key difficulty is often the collection of a sufficient number of samples. In oncology applications this may tend to happen, due to logistical and ethical reasons. Our example is a typical one, as our study is a first-stage evaluation within the context of an academic center, which has a typical patient population with more advanced disease. This limits the number of patients available for research. On the other hand, clinicians and biomedical researchers who wish to explore application of proteomic mass spectrometry for the construction of new diagnostic procedures, will be interested first to get an indication of whether there is information in the spectra to allow groups to be separated and what the likely error rates of misclassification will be. This is particularly the case since ethical review boards (or funding authorities) are not inclined to give permission for large-scale collaborative trials between hospitals, which would ease the patient recruitment problem, without preliminary evidence from smaller within-center trials. Both these reasons may conspire to cause proteomic studies to be of small sample size initially.

Classical statistics will often use a separate validation set to optimize a chosen diagnostic classifier for prediction first. Assessment of the rule is then carried out on yet another test set, which must often be set-aside from the available data [1]. Unfortunately, when the amount of experimental data is small to begin with, the

training set left over may be too small to allow researchers to apply this paradigm fully. In this paper, we present a double cross-validatory approach which allows for simultaneous predictive calibration and assessment of the allocation rule, without (substantial) reduction of the size of the calibration data.

Mass spectrometry data

The experiment and data discussed and analysed in this paper are derived from a MALDI-TOF (Matrix Assisted Laser Desorption Ionisation Time-Of-Flight) mass spectrometer (Ultraflex TOF/TOF, Bruker Daltonics, equipped with a SCOUT ion source which was operated in linear mode). The spectrometer produces a sequence of intensity readings for each sample on an ordered set of contiguous bins in the m/z range from 960 to 11,160 Dalton. Bin sizes (length) of the unprocessed spectra gradually increase with increasing m/z values, ranging from 0.07 Dalton at the lower end of the mass/charge scale up to 0.24 Dalton at the upper end of the scale. This gives intensity readings on a fixed grid of 4483 bins within the mass-charge range across all samples. We refer to an earlier paper by our group for detailed information on experimental setup and measurement protocols.[2]

We will discuss the essential aspects of the study design first, followed by a description of the discriminant method and the double cross-validatory approach to joint predictive estimation (calibration) and validation of the allocation rule, which allows for validated error rate evaluation. Subsequent to description of the methodological approach, we consider application to the colon cancer data and present a *post hoc* exploratory data analysis to interpretation of the results. While we will focus on our example to structure the discussion, the issues apply quite generally to similar problems in proteomics and many other related problems in bioinformatics, chemometrics, statistical prediction and beyond. We will assume that the reader has some knowledge of standard leave-one-out cross-validation.

DESIGN AND SAMPLE REPLICATION

Design

A characteristic problem of proteomic mass spectrometry design is the need to cope with the presence of what we may loosely refer to as so-called 'batch effects'. Examples are plate-to-plate variability, day-to-day variation and so on, whose presence is in reality unavoidable. To accommodate these effects, we identify each plate by day combination as a block and employ standard *randomised block design* by randomly distributing the available samples from each group (colon cancer and controls) across the blocks such that proportions are (as near as) equal within and across

blocks for each group. For colon cancer, we randomised samples to plates in such a manner that the distribution of disease stages is in approximately equal proportions within and across plates. The position on the plates of samples allocated to each plate was also randomised. Each plate was then assigned to a distinct day, which completes the design. Table 1 summarizes the design as executed on the first week, which provides mass spectra on 63 colon cancer patients and 50 healthy controls.

Table 1. Design as executed on the first week. A replicate of the entire experiment was run on the subsequent week using plate duplicates. 'Stage' refers to the distribution of cases across the four respective disease stages.

	TNM stage	Plate 1	Plate 2	Plate 3	Total
Cases	I	4	4	3	
	II	10	10	8	63
	III	4	4	4	
	IV	4	4	4	
		22	22	19	
Controls		17	17	16	50
Total		39	39	35	113

In our case, it was decided to carry out the experiment in a single week using three plates only, each of which was assigned to a consecutive day in the middle of the week - Tuesday to Thursday. We refer the reader to the statistical literature on design of experiments for further discussion and details of the issues involved, as well as many other examples of these basic design principles.[3-6]

Sample replication

We can exploit design to augment cross-validatory analysis. This is because while sample sizes may be small (*i.e.* it is difficult to get new independent samples), the amount of sample material available for each sample may be more abundant. This allows the introduction of so-called replicate samples into the design. Since the samples are pre-arranged on rectangular plates, a second 'copy' of any plate can be made provided sufficient sample material is available from each sample. (In our case, sufficient sample material was available for a second copy only). Thus, we can duplicate the entire design from the first week and remeasure the replicate plates through the same design on the second subsequent week, using new sample material from each sample (but of course not new samples themselves). With this

approach, we thus generally have available from each i^{th} sample an observation $x^1_i = (x^1_{i1}, \ldots, x^1_{ip})$ of the associated recorded mass spectrum in the first week, where the vector elements refer to the measured mass/charge intensities on a predefined and ordered grid of mass/charges of dimensionality p. In addition, we have for each sample a duplicate measurement $x^2_i = (x^2_{i1}, \ldots, x^2_{ip})$ obtained from the corresponding replicate on the corresponding plate measured on the same day one week later. We may denote the associated class label from each i^{th} observation as $c(i)$ which takes value in the set of group indicators $\{1,\ldots,G\}$, where G is the number of groups. [Note we will drop use of the suffixes 1,2 when the context makes clear to which week the data relates.] Unfortunately, due to a technical malfunction which occurred on the last day of the second week the replicate measurements from the third plate are unavailable. As a consequence we only have available the 78 replicates from the first 2 plates in week 2 for further analysis.

INTEGRATED CALIBRATION AND VALIDATION FOR CLASSIFICATION BY DOUBLE CROSS-VALIDATION

We restrict attention to double cross-validated linear discrimination for joint calibration and validation.[7] First we discuss shrinkage-based estimation and the need for it in linear discrimination. Then we explain the double cross-validatory implementation.

Linear classification and shrinkage estimation methodology

We base classification on Fisher linear discrimination. There is voluminous literature on the method, which is well established in the applied sciences, such as biology and medicine.[1;8-10] An article by Hastie et al. contains an up-to-date account of many new applications which demonstrate the continuing success of the approach.[1]

Fisher linear discriminant allocation may be defined as assigning a new observation with feature vector $\mathbf{x}$ to the group for which the distance measure

$$Dg(\mathbf{x})=(\mathbf{x}-\boldsymbol{\mu}_g)\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)^T$$

is minimal, where g denotes the group indicator with $g \in \{1,\ldots,G\}$, $\boldsymbol{\mu}_g$ the population means and Σ the population within-group dispersion matrix which is assumed equal across groups. In practice, the population means and dispersion matrix will be unknown and hence must be estimated from the data. In a high-dimensional problem such as in mass spectrometry proteomics, this leaves us with a difficulty in estimating the dispersion matrix as we will typically not be able to achieve a full rank estimate.

At the risk of some oversimplification of the discussion, there are basically two ways in which we may remedy the problem so that the above methodology may again be applied. The first is through either selection or construction (or a combination of both) of a set of features which is reduced in dimensionality, while capturing most of the variability in the data. In essence, this is the approach which is currently applied in most of the mass spectrometry proteomics literature. Typical examples are found in papers by Baggerly, Yasui, Sauve and Morris, among others.[11-14] We do not consider this approach to be fundamentally flawed for mass spectrometry proteomic data. On the contrary, it is self evident that mass spectra consist of mixtures of possibly overlaid intensity peaks corresponding to substances present in the analyte. Thus, to elucidate this structure (first) is in principle of interest.

The alternative is not to select in the first instance, but instead explicitly utilize the correlations which are induced between intensities on the mass-charge bins through the associated discretisation of the continuous signal (peaks). The simplest approach is through principal components decomposition [15], which has a long history of successful application in classical spectroscopy such as in near infrared spectroscopy for example Krzanowski et al. [16]

Within this approach, we leave the dimensionality of the data intact and instead introduce a regularised estimation of the dispersion matrix to cope with the singularity of the sample dispersion matrix, based on the component decomposition. We explore two distinct forms of regularization, both of which may be expressed in terms of the spectral decomposition of the 'observed' (or sample) pooled dispersion matrix $\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathrm{T}}$ where $\mathbf{Q}$ and $\mathbf{\Lambda} = \mathrm{diag}\,(\lambda 1, \ldots, \lambda r)$ are the matrices of principal component weights (or loadings) and variances respectively, with $\lambda 1 > \ldots > \lambda r > 0$ respectively (r is the rank of the pooled covariance matrix). The within-group covariance matrix is re-estimated by only retaining the first $1 \leq k \leq r$ components only, which gives an estimate

$$\mathbf{S}(k) = \mathbf{Q}_{(k)} \Lambda_{(k)} \mathbf{Q}^{\mathrm{T}}{}_{(k)},$$

where $\Lambda_{(k)} = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_k)$ and $\mathbf{Q}_{(k)}$ denotes the corresponding reduced matrix of component loadings. The associated linear discriminant allocation rule hence assigns observations to the group for which the smallest sample-based distance estimates

$$\hat{D}_g(\mathrm{x}) = (\mathrm{x} - \bar{\mathrm{x}}_g)\, \mathbf{S}^{-1}{}_{(k)}\, (\mathrm{x} - \bar{\mathrm{x}}_g)^{\mathrm{T}}$$

are observed, with $\bar{\mathrm{x}}_g$ the sample group means for $g \in \{1, \ldots G\}$. In the two-group case, this is also equivalent to least-squares regression analysis using the Moore-Penrose inverse of the pooled covariance matrix when $k=r$ (all components kept, also known

as shortest least squares regression), or else ($k<r$) is equivalent to so-called shrunken least-squares regression.[8;17] Alternatively, we may employ ridge regularization

$$\mathbf{S}(\gamma) = \mathbf{Q}[(1 - \gamma)\Lambda + \gamma\mathbf{I}]\mathbf{Q}^{T},$$

where $0<\gamma\leq 1$ is the ridge regularization or 'tuning' parameter, in which case the sample distance measures are $(x - \bar{x}_g)\ \mathbf{S}^{-1}_{(\lambda)}(x - \bar{x}_g)^T$.

Double cross-validatory estimation and validation

Application of the above described classification approaches still require choice of the tuning parameters k or γ involved. As we are specifically interested in an evaluation of predictive performance of any diagnostic allocation rule, it becomes crucial that any optimization -such as the choice of the tuning parameters - does not take place on the same data used for validation. On the other hand, predictive tuning is clearly highly desirable if diagnosis is of interest, so we would not wish to base the choice of tuning parameters on the full calibration data itself (and thus effectively drop predictive tuning from the analysis), but use a truly validatory choice instead. This implies we either set aside a so-called separate 'tuning set' from the available calibration data prior to validation of predictive performance itself or appeal to some form of cross-validation. Good predictive optimization or tuning becomes particularly important in a high-dimensional setting, such as proteomics, as it provides an opportunity to safeguard model choice against over-fitting (in other words: over-interpreting the data). Meanwhile, even if we were able to effectively choose good tuning parameters, the predictive performance (in our case essentially the error rates) of any implied allocation rule should again be validated, which again introduces a need for yet another set-aside validation set or cross-validation.

We may solve both problems by carrying out a so-called double cross-validatory approach, which avoids the need to introduce separate test (tuning) and validation sets. The method has been first proposed and investigated by Stone[7] and integrates predictive optimization and unbiased validated error rate estimation in a single validatory procedure. While the principle of the methodology is sound and well described, this procedure has until recently not been applied in practice due to the considerable computational cost and (algebraic) complexity of the method.[18] This paper describes a first full implementation in the related setting of discriminant allocation on microarray data. Other papers by our group give further details on computational background, application for leave-one-out in spectroscopy and further references.[19;20]

Similar to with ordinary leave-one-out cross-validation, double cross-validation removes each individual (sample) in turn from the data, after which the discriminant

rule is fully recalibrated (and optimised for prediction) on the leftover data and using the same procedure in each case. The resulting classification rule is then applied to the left-out datum to obtain an unbiased allocation for this sample. This procedure is then repeated across all individuals and for each person separately, after which misclassification rates are calculated on the basis of the thus validated classifications. The double-validatory aspect results from the fact that the discriminant rule constructed to classify each left-out datum is optimised through a secondary cross-validatory evaluation within the first cross-validatory layer (i.e. full cross-validation again on each 'leftover' set after removal of an observation). In this manner, we are able to combine predictive optimization and predictive unbiased validation in the same procedure, without loss of data -which is an important requirement to get realistic estimates of error rate with high-dimensional data.

APPLICATION AND EVALUATION

Preprocessing of mass spectra

Some pre-processing can be beneficial when it removes variation from the data which does not relate to the group separation and might obscure an existing group separation. We describe the pre-processing steps carried out prior to the double cross-validatory classification analysis.

First, we calculated for each sample the average intensity within each bin across the four mass spectra from the associated spots on the target plate. Then, we aggregated contiguous bins on the m/z scale, such that the new aggregated bin size spans approximately one Dalton at the left side of the spectrum and gradually increases to a width of approximately 3 Dalton at the right hand side. For each of these new aggregated bins, we calculated for each spectrum the associated aggregate intensity by summing the intensities across the bins being aggregated. Subsequently, spectral baseline was removed from each of the thus aggregated spectra separately using an asymmetric least squares algorithm.[21]

Suppose $\mathrm{x}_{bi} = (x_{bi1}, \ldots, x_{bip})$ denotes the ordered sequence of baseline corrected m/z intensity values for the i^{th} sample at this stage of preprocessing. We then correct the spectrum for the typical intensity and variability across the spectrum by calculating the standardised values

$$x_{sbij} = \frac{x_{bij} - \text{medain}(x_{bi})}{(\mathrm{q}_{0.75}(\mathrm{x}_{bi}) - \mathrm{q}_{0.25}(\mathrm{x}_{bi}))} \ ,$$

where $\mathrm{q}_{0.25}(\mathrm{x}_{bi})$ and $\mathrm{q}_{0.75}(\mathrm{x}_{bi}))$ denote the 25$^{\text{th}}$ and 75$^{\text{th}}$ percentiles of the baseline corrected intensity values for the i^{th} sample. These steps bear close resemblance to the

preprocessing procedure proposed by Satten et al , although our cruder version does not employ local estimates.[22] The final preprocessing step is a log-transformation

$$x_{ij} = \log(x_{sbij} + \alpha)$$

of each spectrum, where α is a real constant. We chose $\alpha = 100$. The main purpose of the log-transform is to ensure numerical stability of calculations. The above preprocessing steps were applied for each sample and within each week separately, which thus gives us the observations x^1_i and x^2_i from the first and second weeks. It is important to stress that the preprocessing of the data of any i^{th} sample does not involve use of any information based on the remaining samples $\{k \mid k \neq i\}$, nor of the duplicate replicate measured spectrum of the same sample on another week. This is an important requirement to ensure the validity of the cross-validatory evaluation described subsequently.

Double cross-validatory error rates

First, we restrict ourselves to the data from the first week. Table 2 displays the estimated recognition rates and performance measures from an analysis of the first week data (leftmost 3 columns). All of the estimates are based on double cross-validation. We used the average of sensitivity (Se) and specificity (Sp) as our estimate of the total recognition rate (T), which implies we assume prior class probabilities to equal 0.5. A threshold of 0.5 was also used to assign observations on the basis of the a-posteriori class probabilities within the cross-validatory calculations. B denotes the Brier distance defined

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

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. Likewise, AUC is a double cross-validation estimate of the area under the empirical ROC curve defined as

$$AUC = \frac{1}{n_1 n_2}\sum_{i \in G1}\sum_{i \in G2} [I(p(1 \mid \mathrm{x}_i) > p(1 \mid \mathrm{x}_i)) + 0.5 * I(p(1 \mid \mathrm{x}_i) = p(1 \mid \mathrm{x}_j))],$$

where $G1$ and $G2$ refer to the sample index labels for samples from the first and second group respectively. Use of the threshold at 0.5 is appropriate and sufficient for an evaluation of diagnostic potential only. Application in *e.g.* a screening type application would require a more careful choice of prior probability, which is how-

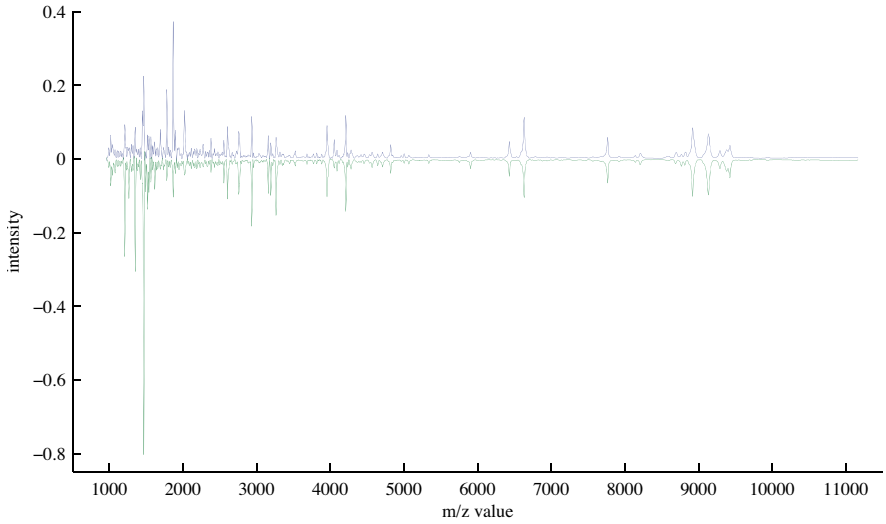


Figure 1. Mean spectra for each group separately, after preprocessing. We plot negative intensity value for the control group (bottom mean spectrum).

ever a subtly different and also subsequent research question and not the focus of this paper.

The rightmost three columns of the table refers to a repetition of this entire double cross-validatory exercise, which replaces each sample feature vector x^1_i with the corresponding replicate measurement x^2_i immediately prior to classification of that i^{th} sample (*i.e.* replacing the feature vectors with the data from week 2 in the outermost layer (only!) of the double cross-validatory calculation). Crucially and importantly, construction of the corresponding discriminant rule for the classification of each such i^{th} sample in the internal 'calibration' layer of the double cross-validatory procedure does of course remain based on the data from week 1. Note that as the replicate data from the third plate are not available, these results are based on the double cross-validated predictions for the remaining 78 replicate samples from week 2 only.

Table 2. Double cross-validated classification results for the colon cancer data. T is the total recognition rate. Se and Sp are sensitivity and specificity, respectively. B is the Brier distance and AUC is the estimated area under the ROC curve.

Method	First week			Second week		
	T (Se, Sp)	B	AUC	T (Se, Sp)	B	AUC
Moore-Penrose $S_{(r)}$	92.6 (95.2, 90.0)	0.0618	97.6	94.4 (91.7,97.1)	0.0600	97.4
PCA Selection $S_{(k)}$	92.6 (95.2, 90.0)	0.0606	97.3	88.8 (80.6,97.1)	0.0914	96.8
Moore-Penrose Euclidian $S_{(r)}$ $\lambda_{(r)} = I_{(r)}$	89.4 (88.9,90.0)	0.0829	96.0	87.2 (86.1,88.2)	0.0770	97.0
PCA Selection Euclidian $S_{(k)}$ $\lambda_{(k)} = I_{(k)}$	88.7 (87.3, 90.0)	0.0865	96.0	90.0 (88.9,91.2)	0.0795	97.0
Ridge $S_{(r)}$	92.0 (95.2,88.0)	0.0602	98.4	95.8 (91.7,100.0)	0.0469	97.9

At first sight, the Moore Penrose implementation (top line of the table, both weeks one and two) would seem to be the best performing and most consistent method. In week 1, Moore-Penrose, PCA-selection (both using the Mahalanobis distance) and ridge estimation perform equally well, but there seems to be an increase in error rate for week 2 for both the PCA-selection and ridge implementation. The Euclidean distance based implementations are worse in the evaluation on the first week, but recognition rates are consistent across both weeks when compared to the other methods. These results should be interpreted with some caution and require some explanation. First of all, the 'plain' Moore-Penrose is leave-one-out only as it does not involve choice of shrinkage or data reduction parameter (k or λ). The deterioration of the PCA-selection implementations is partly due to the uncertainty in estimating the shrinkage terms or choice which is introduced by the double-cross-validatory estimation. For the ridge implementation, performance is comparable to that from Moore-Penrose in week 1, which is not surprising since the chosen ridge shrinkage parameter $\lambda < 0.0001$ for most observations. The effects of uncertainty in the determination of the shrinkage term become particularly apparent for PCA-selection using Mahalanobis distance (second line in the table) in week 2. The two Euclidean distance based implementations on the other hand seem more consistent across both weeks. The reason is that component selection is much more stringent for these two implementations, which selects only the first 2 components for nearly all observations (with exception of two observations out of 113 for which only the first principal component is retained). This explains the reduced performance but also the greater consistency of the classification results. It is precisely because of this reason that these results (from the Euclidean based implementations) are more credible and may well turn out to be more repeatable if the classifier were applied in the future to data from a new repeat experiment. For comparison, component selection in the Mahalanobis distance based PCA implementation is much less stringent and selects (k = 23 for 53 observations, k = 28 for 28 observations and the remainder of

the samples uses even more components). There is thus some evidence of insufficient shrinkage for this method, and similarly for the ridge implementation.

Investigating bias: a permutation exercise

We have proposed double cross-validatory integrated estimation and assessment of statistical diagnostic rules on the basis of the argument that it should protect against optimistically biased evaluations. We may check this property by 'removing' the class labels $c(i)$ from the samples $i \in \{1, \ldots, n\}$, randomly permute and then reassign them to the samples. We then carry out the double cross-validatory procedure again for any of our classification methods. Repeating this procedure several times will give an indication of the biases involved, as the typical recognition rate -for example -should equal 50% across a large number of permutations for an unbiased method.

Table 3. Permutation-based evaluation of double cross-validatory calculations for linear discrimination using principal component selection. DBCV refers to the actual double cross-validatory results (see table 2). $q2{:}5$ and $q97{:}5$ are the 2.5 and 97.5 percentiles. B is the Brier distance and AUC is the estimated area under the ROC curve.

		Permutation results		
Measure	DBCV	median	$q_{2.5}$	$q_{72.5}$
Misclassification rate	7.4	50.0	36.3	72.7
AUC	97.3	49.4	24.8	64.2
B	0.0606	0.324	0.200	0.446

Table 3 shows results from such an exercise for the pca-selection based algorithm across more than 600 such permutations. The results, both for misclassification rate as we find median rates and areas of 50% exactly. Table 3 also includes 95% confidence intervals for the permutation-based performance measures. These give an indication of the variability which can be expected with purely random data and can be compared with the actually observed double-cross-validation results in our study (second column of the table). Clearly, the distance between the validated measures actually observed and even the extreme bounds of the random permutation confidence intervals is considerable, demonstrating the presence of discriminating information in the mass spectra.

Data reduction and post-hoc exploratory analysis

We wish to get an indication of which markers drive the classification. To explore these aspects, we can complement the double cross-validatory analysis with post-hoc exploratory analyses. We consider two analyses, the first of which is based on a very ad hoc algorithmic approach through pre-selection of a small set of adjacent

bins which together account for most of the variation in the spectra. The second explores the linear discriminant weights from a post-hoc fit on the full data.

Data reduction

Initialize $I = \{1,\ldots,p\}$ as the ordered set of bin indices and $V = \{v_1,\ldots,v_p\}$ the associated set of variances for all p bins in the preprocessed spectra and across all n samples, such that $v_j = \sum_i [(x_{ij} - \bar{x}_j)^2]/(n - 1)$, where $\bar{x}_j = \sum x_{ij}/n$ is the sample mean and j is the bin index number. Calculate the constant $vref = q_{0.95}(V)$ as the 95% percentile of all p bin variances. Now initialize the bin selection set B as the set containing the bin indicator j for which the maximum variance vj is observed in the set V. Initialize the set of intensity readings $X_s = \{x_{|j|} \mid j \in B\}$ corresponding to the set B, where $x_{|j|} = (x_{1i}, \ldots, x_{nj})^T$. We write $\mathbf{m} = (m_1,\ldots,m_n)^T$ as the set of means $m_1 = \text{mean}(\{x_{ij} \mid j \in B\})$, i :1, ... ,n. Define cor(**a**, **b**) to be the coefficient of correlation between two vectors **a** and **b**.

Now run the following algorithm.

{*Start of outer loop*}

{Start of inner loop}

Set k=1, $I = I - \{j\}$ and $V = V - \{v_j\}$

Now iterate the following procedure until termination.

Calculate $\rho_{lower} = \text{cor}(\mathbf{m}_i, \mathbf{x}_{[j-k]})$ and $\rho_{upper} = \text{cor}(\mathbf{m}_i, \mathbf{x}_{[j+k]})$

If $\rho_{lower} > 0.9$ and $\rho_{upper} > 0.9$ then

1. Add $j - k$ and $j + k$ to the bin selection set: $B = \{j - k\} \cup B \cup \{j + k\}$.
2. Update the means $m_i, i : 1, \ldots, n$.
3. Remove indices $j - k$ and $j + k$ from the index set I, such that $I = I - \{j - k, j + k\}$. Similarly update $V = V - \{v_{j-k}, v_{j+k}\}$
4. set k=k+1

Else

k=k-1

End iteration.

Now select the bin index j for which $v_j = \max(V)$.

If $v_j > v_{ref}$ then

Update the index set $B = B + \{j\}$ and likewise X_s and m.

Go to {Start of inner loop}

Else End algorithm.

The algorithm identifies a set of 'clusters' of bins. There is no assumption on either shape of the signal or of monotonicity involved (a single cluster may span mixture of underlying peaks). Running this algorithm on the data from the first week finds the set of indices B that corresponds to the bins which account for most of the variation

in the data. Applying this to our data results in a subset of 330 bins (in 32 bin clusters -but note it is possible that we visit the same contiguous region of bins several times). Repeating the entire double cross-validatory procedure using the principal component selection shrinkage procedure on this reduced set yields recognition rates as described in table 4, which are not inconsistent with those from the full double cross-validatory evaluation shown in table 2. (Note however, "double-cross" error rates from this algorithmic approach will be biased as they are based on feature selection from the full data.)

Table 4. Results from re-running double cross-validatory calculations after bin-selection for the colon cancer data (week 1 data only). T is the total recognition rate. Se and Sp are sensitivity and speci¯city respectively. B is the Brier distance and AUC is the estimated area under the ROC curve.

Method	T (Se,Sp)	B	AUC
PCA-selection $S_{(k)}$	90.0 (92.1, 88.0)	0.0807	96.5
PCA-selection Euclidisch $S_{(k)}$ $\lambda_{(k)} = I_{(k)}$	89.0 (92.1, 86.0)	0.0824	95.4

Post-hoc data exploration

The second aspect which is of interest is a post-hoc exploration of the (linear) discriminant coefficients β $(\beta_1, \beta_2, \ldots, \beta_p)^T = \mathbf{S}^{-1}_{(k)}(\bar{x}_1 - \bar{x}_2)^T$ [see (Seber 1984) or (Hand 1997)], where $\bar{x}_1$ and $\bar{x}_2$ are the two sample group means (for cases and controls). [9;17] An appropriate and convenient way to summarize the information contained in these coefficients is via the associated correlations of the measured intensities for each j^{th} bin with the class indicator, which are easily calculated as $\rho_j = s_{xj}\beta_j / s_g$, for $j = 1, \ldots, p$ where $s_{xj} = \sqrt{v_j}$ is the standard deviation at the j^{th} bin and s_g the standard deviation of class indicators. We will base this investigation on the linear discriminant fit using the Euclidean distance on the first two principal components (use $\mathbf{S}_{(k)}$, with $k = 2$ and $\Lambda_{(k)} = \mathbf{I}_{(k)}$), as the double validatory assessment of this classifier clearly identifies the first 2 components as containing the discriminatory information.

At this point, we can carry out the analysis starting from a linear discriminant fit based on the full data. Alternatively, we may equally well base the evaluation on a recomputation of the linear discriminant fit on the reduced data described in previous subsection (in both cases we use the data from the first week). Figure 2 (middle section) shows a plot of the correlation coefficients, subsequent to data reduction (previously described selection of 330 bins, but of course now using all 113 samples from the first week). We only show results within the m/z region between 1200 and 2200 Dalton, as the correlations are effectively zero in the remainder of the m/z range. Evidently, this immediately implies that the separating information is to be found within the 1200 to 2200 m/z range.

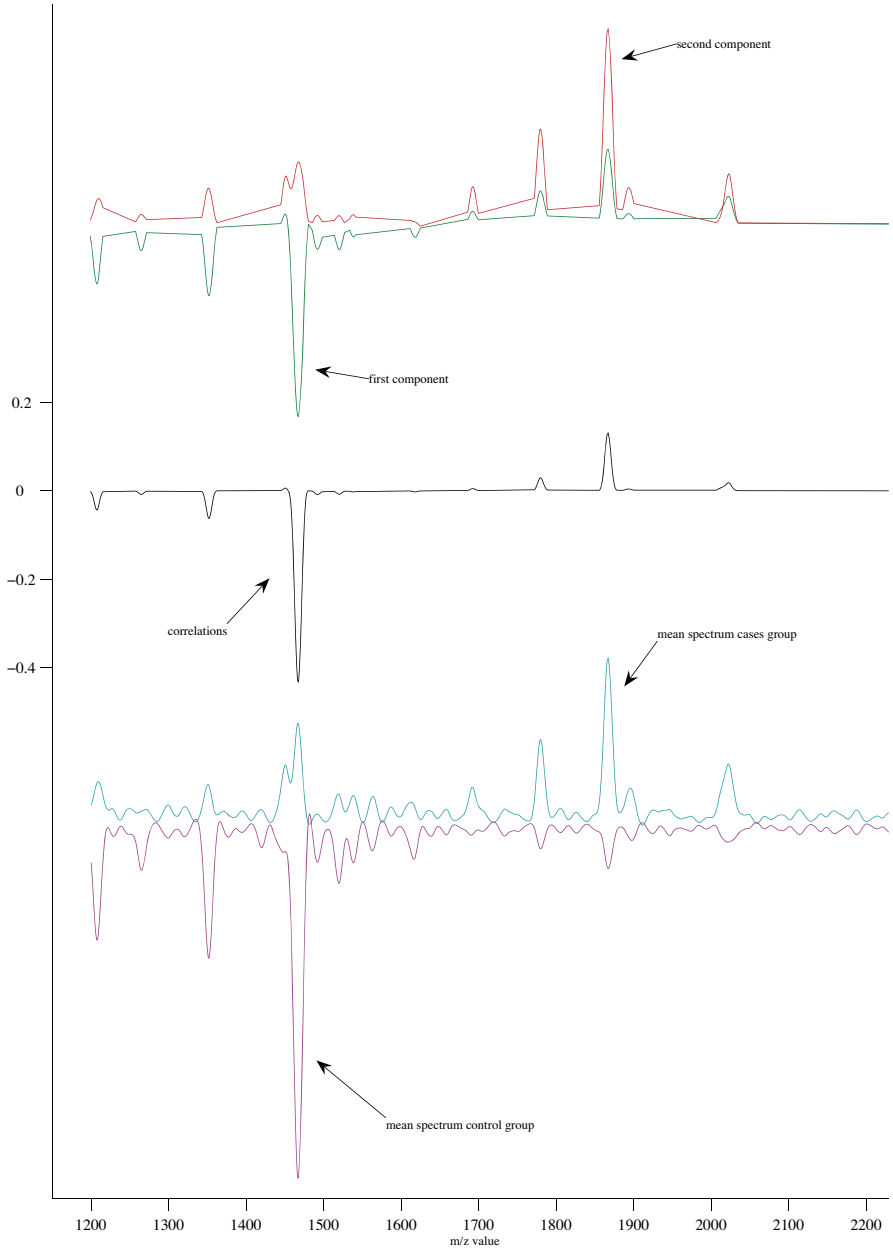


Figure 2. Discriminant correlation coefficients $\rho_j = s_{xj}\beta_j / s_g$ of observed intensity values with the class indicators in the m/z range from 1200 up to 2200 Dalton. We have plotted the first two principal components above these correlations for visual comparison and interpretation. Below the correlations, we plot mean spectra per group (*i.e.*, the vectors x1 and x2, as in figure 1). The y-axis is only relevant to the correlation coefficient, while we have vertically offset and rescaled both components and mean spectra to aid visual comparison across the m/z range.

We note that the picture shown is virtually indistinguishable by eye from that which results from an analysis of the full data (not shown to save space). The reason for this is that the data reduction restricts attention to the dominant sources of variation, which is not very different from what is achieved through principal component reduction. Immediately above the correlation coefficients graph, figure 2 displays the first two principal components (vertically offset and rescaled to aid visual interpretation) and again based on the reduced data. In this case, the distinct bin subsets selected by the previous data reduction step are clearly visible in the two components, and display the characteristic 'peaks' we would expect to identify. Disjoint neighbouring bin sets are connected with straight lines. The thus calculated components are a close approximation to those which would result from an analysis of the full data, as we should expect (results not shown). As for the correlation coefficients, any conclusions are therefore identical whether we use the reduced data or not, although the data reduction step perhaps makes the component plot easier to 'read'. At the bottom of the graph we give the mean spectrum again for each group separately and from the original data within the m/z range of interest, as shown in figure 1 also, along the complete m/z range.

From this graphical analysis, it is clear how the linear discriminant correlation coefficients identify two major discriminating contributions, the first of which is centered at 1467.7 Dalton and the second at 1867.7 Dalton. Furthermore, the correlations have opposite signs at these locations, which would indicate that the discriminating information can be summarised through a contrast effect between corresponding measured intensities in the spectra. An investigation of the principal components plots above learns that the contribution at 1467.7 Dalton is primarily accounted for by the first component, which also already contains the contrast with intensities recorded at 1867.7 Dalton. This contrast is then further amplified by the second component which identifies a second orthogonal source of variation relative to the first component, centered predominately at the already identified peak at 1867.7 Dalton. Note how each component identifies several other smaller contributions, which could also be of interest for further investigation. Comparing these graphs with the within-group mean spectra, the resemblance with the principal components plots at the top of the figure are striking and would suggest that the first component may be primarily explained through variation within the control group at 1467.7 Dalton. Likewise, the second component accounts for a substantial intensity peak at 1867.7 Dalton within the colon cancer group.

To investigate this further, figure 3 provides scatter plots of cases and controls versus the first 2 components (left plot) and between intensities at 1467.7 and 1867.7 Dalton respectively (right plot). The resemblance between both graphs is striking as the right plot can be obtained (virtually) after clockwise rotation of the left plot. As

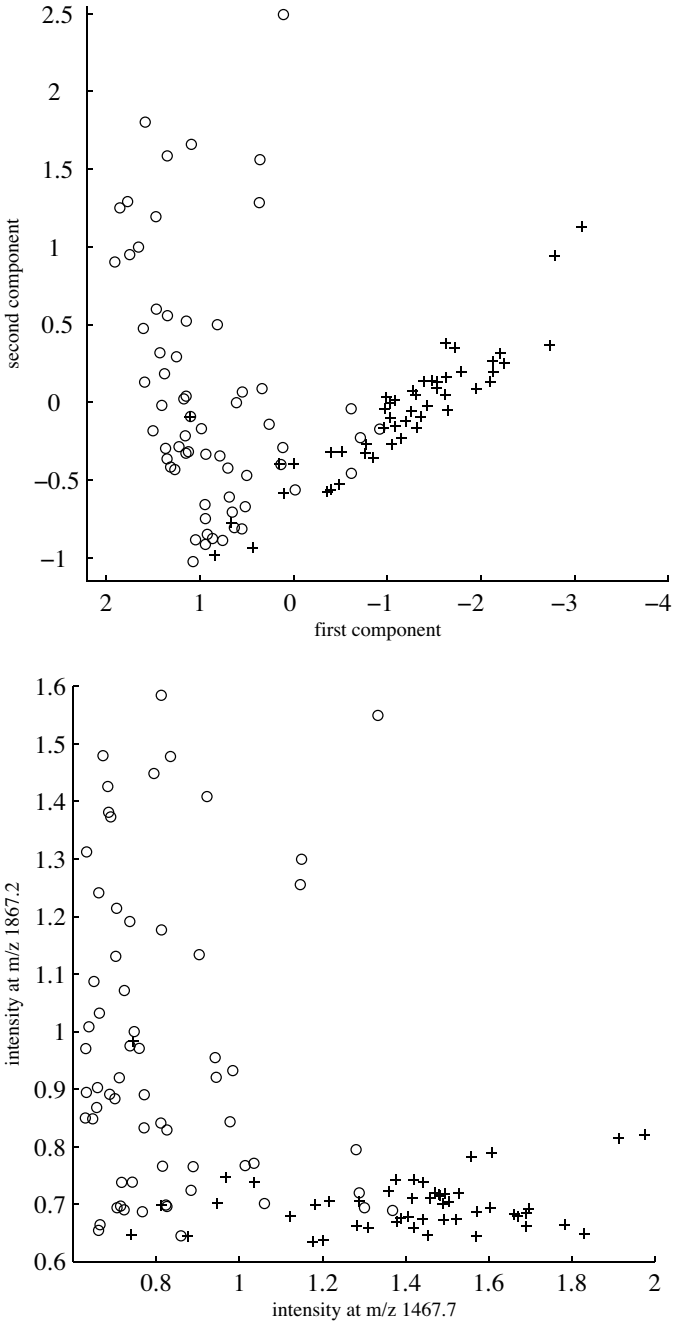


Figure 3. Scatter plots distinguishing cases (o) from controls (+). On the left we plot the second versus the first principal component. The right plot shows intensity values at 1867.2 m/z versus those at 1467.7 m/z.

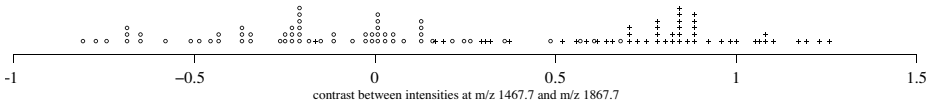


Figure 4. Plot of the diferences between intensities at 1467.7 m/z and 1867.7 m/z across all observations, using distinct plotting symbols for each group: cases (o) and controls (+).

we can see, an increase in intensity at 1467.7 Dalton separates controls from cases. Similarly, an increase in intensity at 1867.7 Dalton separates cases from controls. The same interpretation applies to the principal components scatter plot, which confirms our interpretation of the data in figure 2. Figure 4 provides a concise summary graphical illustration of the results. We calculate the contrast (difference) for all 113 individuals participating in the study between the measured intensities at 1467.7 and 1867.7 Dalton and display the differences in a dot plot using distinct plotting symbols for cases and controls respectively, which demonstrates the separation between both groups.

For further discussion of the clinical background, study rationale, setup, execution and interpretation of results from a substantive clinical perspective, we refer to (Noo 2006) and subsequent papers from these authors.

DISCUSSION

Double validatory analysis

Use of a separate validation or test set is often precluded in high dimensional problems, due to sample size restrictions. In our case, this arises because the experiment was carried out in an academic medical center, which implies (colon cancer) cases are restricted to a maximum of about 50 patients yearly and with more advanced disease. Selection of appropriate control samples may be more difficult still, even if we use surrogate serum samples -as in this experiment. Larger numbers of cases may be recruited by setting up multi-center trials and using longer recruitment periods. However, researchers may need some justification in the form of a small feasibility study before setting up such complex trials. It is in such situations that double cross-validatory analysis can be most useful to help researchers make the maximum use of the scarce data available. The other option of reducing the available calibration data prior to optimization of any discriminant rule by setting aside data (perhaps for both a 'predictive tuning' as well as 'validation' set) is not as innocent as appears at first sight. This is because it will often reduce the calibration set beyond what is

needed for reasonable calibration. Moreover, reducing the size of the calibration data changes the condition of the estimation itself. To put this simply: we are not only reducing the data by setting-aside data from the calibration set, but also changing the discriminant problem itself. This is again particularly the case in high-dimensional cases such as in proteomics where the problem will typically be ill-conditioned.

The approach we have described in this paper avoids these difficulties through application of double cross-validation to combine the two aspects of *predictive optimization* and *validation*. Subsequent to this basic evaluation of the discriminatory potential of the spectral data, a more exploratory analysis can be carried out, provided we are carefully to interpret results cautiously without contradicting the primary validated evaluation. We discuss a number of issues related to application of (double) cross-validation.

Full validation

One potential cause for concern is whether double cross-validation precludes the need for a completely separate validation set entirely. Is 'double-cross' also 'full' validation? The simple answer to this question is that it can not be, as any form of cross-validation must typically always remain 'within-study' validation and there can be factors beyond our knowledge which have influenced the study results. Good scientific practice requires that we replicate results in a separate repeat study. This caution applies particularly to the definition of the case and control group, as the impact of systematic effects due to measurement can be minimised through use of randomised block design. Repeat studies may help to detect such problems. Note however, that these criticisms would also have applied to the standard practice of using within-study set-aside test and validation sets. Meanwhile, double cross-validation should give reasonable protection against overfitting and unbiased estimates of error rate *at the time of study*. Double-cross represents the maximum usage we can make of the data for joint predictive optimization and validation *within a single experiment*. Even when separate test and validation sets are available however, researchers may still be interested to compare the thus validated re-search findings with those from a fully double cross-validated analysis on the combined data in order to evaluate whether the greater sample size would have allowed for better calibrations -possibly because of improved detection of the smaller signal sources in the spectra.[23] More generally, we could speculate where the validation process should stop. Typically, the performance of any decision rule or classifier has a tendency to 'decay' over time. To assess this, subsequent experiments are needed to verify the estimated error rates.

What classifier are we evaluating?

Two related questions to the previous discussion are 'What classifier does double cross-validation evaluate?' and 'How to assign a new observation?'. Indeed, each observation has its own classifier in the double cross-validatory evaluation. This seems to run counter to the intuition that we calibrate a discriminant rule first and only then evaluate. In that case, the estimated error rate is taken as a reflection of the diagnostic abilities of that particular classifier and the allocation of a new sample is immediate. There is however no logical inconsistency here. Double cross-validation estimates the error rate we would get 'if we were to apply leave-one-out' on the whole data. Once we know what the error rate is, we may choose the specific classifier (choice of k or λ in our case) for allocation of future samples (if required) through application of ordinary leave-one-out on the whole data (this is in line with the discussion presented by Mervin Stone.[7] With double cross-validation, there are however other options to allow allocation of new samples which have not yet been discussed in the literature. In our case for example, we may use the mode of the number of components selected (k) across all samples and then re-estimate the discriminant model with this choice from the full data. More adventurous still, we could retain each of the n classification rules which are calibrated within the double-cross procedure and use this ensemble (of classifiers) for allocation of any future new observation x. This could be done by calculating the associated a-posteriori class probabilities $p_i(g \mid \mathrm{x})$, for each $i \in \{1, \ldots, n\}$ and $g \in \{1, \ldots, G\}$, where p_i is obtained from the discriminant model calibrated in the double-cross procedure when the i^{th} datum has been removed from the data (in the outer shell of the double-cross procedure). Classification may then be based on the mean across these n a-posteriori class probabilities for any g^{th} class. We will not pursue these options further in this paper.

Validation and the future of (statistical) proteomics

Rigorous emphasis on validation and proper design can help to establish long-term credibility for proteomic research and more general bioinformatics applications. The double-cross approach with randomised block design described in this paper represents one contribution towards this goal. Many other steps may however be taken to enhance the quality of such research studies. One example is to promote use of 'truly' separate validation sets, as obtained from subsequent separate and additional sampling from the population of interest and measurement through identical protocols as applied in the first study. In practice, this will be particularly relevant for those studies which indicate potential from the first within-study verification of diagnostic ability. Editors of scientific journals can also contribute much to inspire a conservative attitude by careful scrutiny of the papers presented for publication. Perhaps simple check lists could be developed to help reviewers establish the degree

to which validatory evaluation did (or should) contribute to the research findings presented. This may also prevent mistakes from slipping through the net. Although this may cause considerable annoyance in some cases when we face the difficulties of establishing results in the short term, but may enhance scientific credibility of (proteomic) research as a whole in the long run. Results from the present study show that, with good designed experimentation, these precautions need not form insurmountable obstacles.

REFERENCES

1. Hastie,T., Tibshirani,R., and Friedman,J. (2001) *The elements of statistical learning*. Springer-verlag.
2. de Noo,M.E., Mertens,B.J., Ozalp,A., Bladergroen,M.R., van der Werff,M.P., van de Velde,C.J., Deelder,A.M., and Tollenaar,R.A. (2006) Detection of colorectal cancer using MALDI-TOF serum protein profiling. *Eur.J Cancer*, 42, 1068-1076.
3. Cox D.R. and Reid N. (2000) *The theory of the design of experiments*. Chapmann/Hall CRC.
4. Box,G.E.P., Hunter W.G., and Hunter J.S. (1978) *Statistics for experimenters*. John Wiley & Sons, Inc..
5. Neter,J.et.al. (1996) *Applied Linear Statistical Models*. McGraw-Hill/Irwin.
6. Fisher,R.A. (1935) *The Design of Experiments*. Oliver and Boyd: Edinburgh..
7. Stone,M. (1974) Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36, 111-147.
8. Ripley,B.D. (1996) *Pattern recognition and neural networks*. Cambridge University Press.
9. Seber,G.A.F. (1984) *Multivariate Observations*. Whiley Chichester.
10. McLachlan,G.J. (1992) *Discriminant analysis and statistical pattern recognition*.
11. Baggerly,K.A., Morris,J.S., Wang,J., Gold,D., Xiao,L.C., and Coombes,K.R. (2003) A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics.*, 3, 1667-1672.
12. Morris,J.S., Coombes,K.R., Koomen,J., Baggerly,K.A., and Kobayashi,R. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics.*, 21, 1764-1775.
13. Yasui,Y., Pepe,M., Thompson,M.L., Adam,B.L., Wright,G.L., Jr., Qu,Y., Potter,J.D., Winget,M., Thornquist,M., and Feng,Z. (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics.*, 4, 449-463.
14. Sauve, A. C. and Speed T. P. Normalization, baseline correction and alignment of high-throughput mass spectrometry data. *Proceedings Gensips*. 2004.
15. Jolliffe,I.T. (2002) *Principal Component Analysis*. Springer-Verlag, New York.
16. Krzanowski,W.J et al. (1995) Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Applied Statistics*, 44, 101-115.
17. Hand,D.J. (1997) *Construction and assessment of classification rules*. John Wiley and Sons; Inc.
18. Mertens,B.J.A. (2003) Microarrays, pattern recognition and exploratory data analysis. *Statistics in Medicine*, 22, 1879-1899.
19. Mertens,B.J.A. (1998) Exact principal component in?uence measures applied to the analysis of spectroscopic data on rice. *Applied Statistics*, 47, 527-542.
20. Mertens,B.J.A. (2001) Downdating: interdisciplinary research between statistics and computing. *Statistica Neerlandica*, 55, 358-366.
21. Eilers,P.H. (2004) Parametric time warping. *Anal.Chem.*, 76, 404-411.
22. Satten,G.A.et al. (2004) Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens. *Bioinformatics*, 20, 3136.
23. Ransohoff,D.F. (2004) Evaluating discovery-based research: when biologic reasoning cannot work. *Gastroenterology*, 127, 1028.