JOURNAL OF CHEMOMETRICS, VOL. 8, 1-20 (1994)

REVIEW PAPER

STATISTICAL THINKING AND TECHNIQUE FOR QSAR AND RELATED STUDIES. PART II: SPECIFIC METHODS

MERVYN STONE
Department of Statistical Science, University College London, Gower St., London WC1 E6BT, U.K.

AND

PHILIP JONATHAN
Shell Research Ltd., Sittingbourne, Kent ME9 8AG, U.K.

SUMMARY

Twenty-two contrasting statistical methods are reviewed for their applicability to QSAR studies and similar prediction-oriented fields. Each method is concisely specified prior to explanatory or critical comment.

KEY WORDS Discriminant analysis Least squares Prediction Regression Relationship Structure

1. INTRODUCTION

The two parts of this paper form a critique of a variety of statistical techniques of actual or potential use in quantitative structure–activity relationship (QSAR) studies as well as in related fields. Part I (henceforth referred to as S&J[1]) explored the statistical thinking underpinning those techniques. Here in Part II we offer a concise account of the most widely used methods – and some that are less well-known or practised. In the interests of interdisciplinary communication we have tried to keep technical mathematics and chemistry to a minimum consistent with the objectives of the paper. Neither this restraint nor the necessarily small volume of our unified treatment compared with that of a recently published comprehensive volume[2] should be inimical to exposition of the current status of statistical methods in QSARs. Conciseness has, however, dictated the exclusion of important but rather diffuse topics such as treatment of outliers or the use of robust alternatives to least squares.[3] Conciseness also increases the risk of not giving due weight to work whose value we have not yet recognized – we apologise in advance to any authors thus slighted or misrepresented.

The reader should refer to S&J[1] for fuller details of the terminology and notation used.

0886-9383/94/010001-20$15.00

Received June 1992
Accepted May 1993

2. MAKING A BED FOR LEAST SQUARES

Statistical aids for QSARs have to recognize openly that

(i) most QSAR studies are based on a fixed number of compounds, n, which is rarely in excess of 100 and usually less than 50
(ii) there is often effectively no limit to the value of the number of variables, p.

Hence the prerequisite for ordinary least squares, i.e. $n \geqslant p+1$, is often unsatisfied. For the case $n < p+1$ several techniques may be viewed as strenuous efforts to 'bring the situation back to one where MR (multiple regression) is appropriate' though potentially 'self-deceptive'.[4] Either by the selection of a small number of descriptors or by the construction of a small number of new variables that are usually linear combinations of $x(1), \ldots, x(p)$, a basis is prepared for least-squares prediction of y with a reduced complement of variables, $t(1), \ldots, t(A)$ say, where $A \leqslant n-1$ (or $n-2$ for cross-validation to be applicable). The least-squares predictor would then have the form $\tilde{a} + \tilde{b}_1 t(1) + \cdots + \tilde{b}_A t(A)$, say. When $t(1), \ldots, t(A)$ are linear combinations of $x(1), \ldots, x(p)$, this predictor expands to the familiar $a + b_1 x(1) + \cdots + b_p x(p)$ when the particular forms of $t(1), \ldots, t(A)$ have been inserted. (The variables $t(1), \ldots, t(A)$ may be called *component variables* or *regressors.*)

The potential self-deception of this approach may be greatly reduced by cross-validation of the whole procedure, for which a heavy computational cost may be unavoidable.

2.1. Stepwise regression[5]

At any stage of this particular procedure the next descriptor introduced into the current selection of components is the one with the highest F-value provided that it exceeds a cut-in value F_{in}. Any descriptor is eliminated from the current selection if its F-value then falls below a cut-out value F_{out}.

Comments

(i) To reduce self-deception, the control parameters F_{in} and F_{out} could be chosen by cross-validation by the method indicated in Section 6.6 of S&J[1] – but this would involve heavy computation. Usually F_{in} and F_{out} are taken to be equal, in which case the cross-validatory choice becomes a feasible one-dimensional optimization.

(ii) The use of the F-statistic for cross-validatory control is fairly arbitrary. The fact that F is justifiably used for hypothesis testing in a fixed normal model has no proven relevance to the question of sequential control. The introduction and elimination of descriptors could (with equal lack of justification) be based on fixed percentage changes in s or in the statistics s^{+}, s^{++} and RMSPE of Section 6.4 of S&J.[1]

2.2. Principal components[6]

The *score* of a compound for a variable that is a linear combination $t = c_1 x(1) + \cdots + c_p x(p)$ is simply the value of the linear combination for that compound.

The *variance* of a linear combination is the sample variance of the corresponding n scores of the compounds in $\mathscr{C}$.

Two linear combinations are *uncorrelated* if the Pearson correlation of the corresponding scores is zero, calculated for the compounds in $\mathscr{C}$.

The *first principal component* is the linear combination of maximum variance when the condition

$$c_1^2 + \cdots + c_p^2 = 1 \tag{1}$$

is imposed on the coefficients of the combination.

The *second principal component* is defined in the same way, with the extra condition that it be uncorrelated with the first.

The *third principal component* is required to be uncorrelated with the first two — and so on until there are no more linear combinations with non-zero variance.

With all the principal components in hand, there are two procedures for fixing the set of regressors for least-squares prediction.

(a) In *(standard) principal component regression* (PCR) the first A principal components are the regressors to be used.

(b) For what might be called *reordered principal component regression* the components are placed in decreasing order of their individual Pearson correlations with y before taking the first A.

For both procedures the single control parameter A may be chosen by cross-validation.

Comments

(i) The requirement that the components be mutually uncorrelated is not fundamental but does ensure that they express different aspects of the variation between the descriptors for the n compounds.

(ii) The condition (1) is also not fundamental, although it is made to appear so by the routine presentation of principal components as directions in p-dimensional vector space. The condition is more mathematically and computationally convenient than alternatives such as $|c_1| + \cdots + |c_p| = 1$ or $\max\{|c_j|\} = 1$.

(iii) The maximum value of A is the number of linearly independent vectors among $\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}$. For $n \leqslant p + 1$ this number is usually $n - 1$.

(iv) In choosing between (a) and (b), the standard ordering might be preferred if there is some confidence that information about y lies in the large-variance components.

2.3. Partial least squares (PLS)[7,8]

The *covariance* between the activity y and a linear combination t of the descriptors $x(1), \ldots, x(p)$ is the sample covariance between the scores $t_1, \ldots, t_n$ (say) and $y_1, \ldots, y_n$ respectively:

$$\sum_i (t_i - \bar{t})(y_i - \bar{y})/(n - 1)$$

The description of PLS closely parallels that just given for PCR.

The *first PLS component* is the linear combination of maximum covariance when the condition $c_1^2 + \cdots + c_p^2 = 1$ is imposed on the coefficients of the combination.

The *second PLS component* is definable in the same way, with the extra condition that it be uncorrelated with the first.

The *third PLS component* is required to be uncorrelated with the first two – and so on until there are no more linear combinations with non-zero variance (and hence, it may be supposed, non-zero covariance.)

The cross-validatory PLS predictor is then the least-squares predictor based on the first A PLS components, with A chosen by cross-validation.

Comments

(i) Our description of PLS (which is here specialized for prediction of a single y-variable) is non-standard but fully equivalent to the standard one. For each component the standard algorithm constructs instead linear combinations of *residuals* of descriptors (from their least-squares regression on the combinations already constructed). This avoids the *requirement* for the conditions of uncorrelation, because they are then automatically satisfied.

(ii) Conveniently for the PLS acronym, partial least squares is increasingly referred to as 'projection to latent structure', which is more informative.

(iii) The power of PLS can be no more and no less than what is suggested by our description: it does not have any other magical properties (see the caution in Section 2.4, comment (iv)).

(iv) A numerical–analytical account of the standard PLS algorithm for the single-y case ('PLS1', to distinguish it from the multiple-y case, 'PLS2', of Section (6) is given by Manne.[9] Another algorithm for PLS1 is provided by de Jong:[10] it parallels our description of PLS.

2.4. Continuum regression (CR)[11]

Specification of CR requires an additional control parameter, but otherwise follows that of PCR or PLS precisely. The only change needed is that the criterion to be maximized at each stage, which was 'variance' for PCR and 'covariance' for PLS, is now generalized to

$$(\text{covariance})^2\ (\text{variance})^{\alpha/(1-\alpha)-1} \qquad (2)$$

where α is restricted to lie between zero and unity.

Comments

(i) For $\alpha \approx 1$ the criterion (2) is effectively equivalent to 'variance' and at $\alpha = 0{\cdot}5$ it is equivalent to 'covariance'. Thus CR generalizes PCR and PLS.

(ii) For $\alpha = 0$ the criterion (2) is equivalent to 'Pearson correlation', so that CR delivers ordinary least squares (OLS) in the non-singular case. For $n < p + 1$ CR is definable at $\alpha = 0$ as a limit – which happens to deliver the (unique) *shortest least-squares predictor* (SLS) (see Section 3.1).

(iii) Details of the cross-validated implementation of CR with applications are given by Stone & Brooks,[11] from whom a computer programme is obtainable.

(iv) It has yet to be established that a mixture, such as CR, of two extreme methods is predictively superior *in practice* to a simple cross-validatory choice between them.

2.5. Intermediate least squares (ILS)[12]

This generalization of PLS constructs components at each stage that are linear combinations of maximal covariance, with the additional constraint that they be combinations of no more than α of $x(1), \ldots, x(p)$. As for CR, the control parameters α and A are chosen by cross-validation.

Comments

(i) PLS is given by $\alpha = p$.

(ii) If $x(1), \ldots, x(p)$ are scaled to have unit standard deviations over $\mathscr{C}$, then ILS with $\alpha = 1$ is stepwise regression without backward elimination of descriptors.

(iii) The restriction to just α variables is potentially beneficial in cutting out the 'noise' from uninformative variables.

(iv) 'ILS' has also been used[13] to acronymize 'inverse least squares'.

2.6. Continuum powering[14]

We do not have an account of this method of constructing components which is both concise and definitive. In broad terms the method is a generalization of its authors' own PLS algorithm, involving

(a) singular value decomposition of a key $n \times p$ matrix at each stage
(b) replacement of the singular values by their Nth power, where N is a control parameter in the *continuum* $(0, \infty)$.

Comments

(i) $N = 1$ is PLS and N very large gives PCR.

(ii) It is claimed[13] that for $N = 0$ in cases with $n \geqslant p + 1$ the method gives the least-squares (multiple-regression) predictor. We would also like to know the limiting behaviour as N tends to zero in the singular case.

(iii) The technique was described as 'unnamed' in the discussion of Stone and Brooks[11] but has since[13] been christened 'continuum regression'. We have here suggested 'continuum powering' to avoid confusion with Section 2.4.

2.7. Principal covariate regression (PCovR):[15] univariate y

Like CR, this is a continuum method mixing least squares and PCR. As formulated by de Jong and Kiers,[15] the method is equivalent in output to replacement of the construction stage criterion (2) by

$$(1-\alpha)\left(\frac{\Delta\text{FSS for } y}{\text{TSS for } y}\right) + \alpha\left(\frac{\sum_1^p \Delta\text{FSS for } x(j)}{\sum_1^p \text{TSS for } x(j)}\right)$$

where ΔFSS is the increment in the fitted sum of squares for the specified variable in its least-squares regression on the so far constructed regressors and TSS is the ordinary total sum of squares of the analysis of variance.

Comments

(i) For $\alpha = 1$ PCovR delivers PCR.

(ii) For $\alpha = 0$ it gives OLS in the non-singular case with $n \geqslant p + 1$. For $n < p + 1$ PCovR (as formulated here) is definable as a limit which, as for CR, is the SLS predictor of Section 3.1.

(iii) The continuum does not include PLS.

(iv) The method handles the case of several y-variables $\{y(j)\}$ by extending the first term of the criterion to match the second term in $\{x(j)\}$ (see also Section 6).

3. AFFILIATED TECHNIQUES

The following techniques do not depend on an initial selection of descriptors or reduction of descriptor dimensionality prior to least-squares application. Instead they are free to involve all the available descriptors, perhaps recklessly (as cross-validation might or might not reveal).

3.1. Shortest least squares (SLS)[16]

With $n < p + 1$, i.e. with fewer compounds than parameters to be fitted, the least-squares values of $a, b_1, \ldots, b_p$ for fitting $\hat{y} = a + b_1 x(1) + \cdots + b_p x(p)$ are not unique. There is an infinity of solutions and therefore an infinity of contenders for the prediction formula.

For some new compounds all these predictors may yet agree about the prediction to be made: these are the compounds whose vector of descriptors fortuitously lies in the space spanned by the vectors of the n measured compounds. There would be uniqueness of prediction for *any* new compound provided that a unique choice could be made from the above infinity of solutions for $a, b_1, \ldots, b_p$. The SLS choice is the one that makes the vector $\mathbf{b} = (b_1, \ldots, b_p)$ the shortest, i.e. $b_1^2 + \cdots + b_p^2$ a minimum.

Comments

(i) The SLS solution is otherwise known as the *minimum norm* solution. Its algebraic expression is

$$\mathbf{b} = y_1 \mathbf{S}^+ (\mathbf{x}_1 - \bar{\mathbf{x}}) + \cdots + y_n \mathbf{S}^+ (\mathbf{x}_n - \bar{\mathbf{x}}) \tag{3}$$

where $\mathbf{S}^+$ is the Moore–Penrose generalized inverse of the sum-of-squares-and-products matrix, which is necessarily singular for $n < p + 1$. (Formula (3) would give the usual OLS coefficients if $\mathbf{S}$ *were* non-singular.)

(ii) As noted in Sections 2.4 and 2.7, the SLS predictor is delivered as the special case of CR and PCovR at $\alpha = 0$.

(iii) The SLS predictor does not have any control parameters to be determined by cross-validation. Provided that the descriptors have not been autoscaled (Section 6.9 of S&J[1]), the cross-validatory RMSPE criterion (Section 6.4(c) of S&J[1]) may be speedily calculated with the aid of the following leave-one-out result of Dunne and Stone:[16]

> when the ith compound is omitted, the $\mathbf{b}$ of (3) is reduced to its component orthogonal to $\mathbf{S}^+(\mathbf{x}_i - \bar{\mathbf{x}})$. (Marbach and Heise[17] offer an incorrect version of this result: in their equation (44) the denominator $1 - h_{mm,+}$ has to be zero. However, they make no essential use of this version in an otherwise interesting paper.)

(iv) An extremely artificial example reveals why SLS may be rewarding when the descriptors are not autoscaled and the predictively useful descriptors are those of high variation in the training set. Suppose $n = 2$, $p = 2$ and that compound C_1 has $x(1) = x(2) = 0$ while C_2 has $x(1) = 1$ and $x(2) = 10$. It may be verified that the SLS predictor is

$$\hat{y} = y_1 + [(y_2 - y_1)/101]\,[x(1) + 10x(2)]$$

which is dominated by the high-variation descriptor $x(2)$.

(v) SLS is given by *any* least-squares prediction from $n-1$ regressors whose combination vectors – the $(c_1, \ldots, c_p)^{\mathrm{T}}$ of Section 2 – span the same space as $\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}$. CR does this for any α and hence so do PLS and PCR.

3.2. Ridge regression[18]

This method produces a unique predictor by moving outside the class of least-squares solutions for $\mathbf{b}$ (Section 3.1)1 at the cost of introducing one control parameter $k > 0$. The value of $\mathbf{b}$ for the choice k is $\mathbf{b}(k) = (\mathbf{S} + k\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

Comments

(i) $b_1(k)^2 + \cdots + b_p(k)^2$ decreases as k increases, *shrinking* the predictor coefficients towards zero, perhaps beneficially.

(ii) As k goes to zero, $\mathbf{b}(k)$ goes to the SLS value.

(iii) Sundberg[19] has shown that when S is non-singular, $\mathbf{b}(k)$ is proportional (with a multiplier less than unity) to the value of $\mathbf{b}$ given by least-squares regression on the first CR component for a value of the CR control parameter α lying between zero and $\frac{1}{2}$ and increasing with k. Sundberg's proof may be seen to carry over to the present case in which S is singular.

(iv) $\mathbf{b}(k)$ is derivable as the minimizer of $\| \mathbf{y}^{\mathrm{c}} - \mathbf{X}^{\mathrm{c}}\beta \|^2 + \mathbf{k} \| \beta \|^2$, where $\mathbf{y}^{\mathrm{c}}$ is the centred vector of y-values and $\mathbf{X}^{\mathrm{c}}$ is the centred $n \times p$ matrix of x-values.

(v) Frank and Friedman[20] highlight ridge regression in a wide-ranging analysis, including simulations.

3.3. Regression trees (RT) in CART[21]

The methods of Sections 2.1–2.6 all have as output a linear predictor

$$\hat{y} = a + b_1 x(1) + \cdots + b_p x(p) \tag{4}$$

whose form is the same at all points in the space of descriptors. This global uniformity means that either great care has to be taken in the definition of $x(1), \ldots, x(p)$ (e.g. by the inclusion of descriptors for plausible non-linearities or interactions) or else the class $\mathscr{C}$ of congeneric compounds to which (4) is to be applied must be suitably restricted. Wold[22] has emphasized that justification of the use of (4) as a Taylor expansion linearization of a complex function of $x(1), \ldots, x(p)$ requires that the variables be 'measured on processes with a limited variation'. All of this is an impediment to bold and potentially informative choice of the compounds in $\mathscr{C}$. Admittedly there is the Free–Wilson additivity justification for the use of indicator variables ($x(j) = 0$ or 1), but the use of such regressors as a formal device for putting several regressions into the same formula requires the questionable supposition that the individual regressions are parallel. The flexibility in (4), given by quadratic and interaction terms and the like, has to be built in prior to analysis, when knowledge of how to do it may not be available. What is needed is a method that builds in flexibility as required in the analysis of the actual data and that maintains control over the potential excesses of adaptivity.

'Regression trees' (RT), as refined and controlled in CART, provide such flexibility. The output of an RT analysis is simply a *partitioning* of the space of $\mathbf{x} = (x(1), \ldots, x(p))^{\mathrm{T}}$ into generalized rectangles of the type

$$\mathscr{R} = \{\mathbf{x} \colon a_j < x(j) < b_j,\ j = 1, \ldots, p\}$$

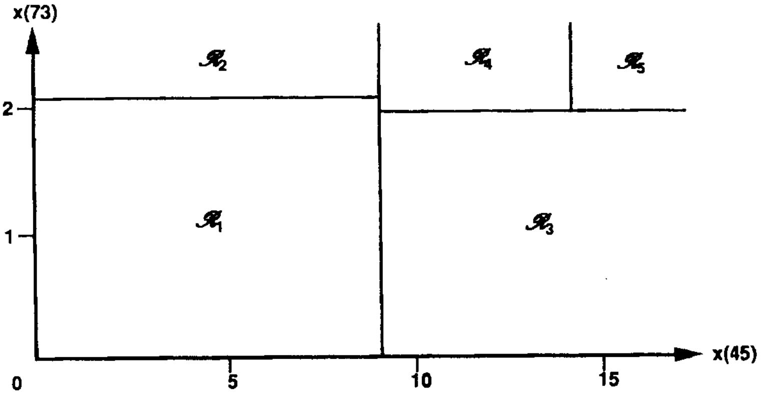

Figure 1. Illustration of regression tree rectangles

and an *associated predictor* $\hat{y}$. The value of $\hat{y}$ for a new compound at $\mathbf{x}_C$ is taken to be the average of the values of y for compounds in $\mathscr{C}$ whose $\mathbf{x}$-values are in the same rectangle as $\mathbf{x}_C$. Typically, when p is large compared with n, the RT analysis will give $a_j = -\infty$ and $b_j = \infty$ for all but a few values of j, while for those few, usually $a_j = -\infty$ or $b_j = \infty$. Thus, for example, we may find that the basis of the output predictor is defined by $x(45)$ and $x(73)$ with five rectangles as in Figure 1. (It is the calculation of the *averages* of the y_i-values in each of the rectangles to define $\hat{y}$ that puts the term 'regression' into RT.)

The RT procedure is a combination of *sequential splitting on individual descriptors until further splitting is impossible* (producing a 'regression tree' with many fine terminal branches beyond scientific justification) and *pruning* of the tree back to a point usually chosen by cross-validation. At each stage of the tree-growing procedure the split in CART is made on the variable and associated (a_j, b_j) that give the maximal reduction in the splitting criterion, taken to be the residual sum of squares of the predictor at that stage.

Comments

(i) The 'AID' method of Sonquist *et al.*[23] was defective as a precursor of CART because it lacked cross-validatory control of its tree growing.

(ii) CART permits splits to be made using linear combinations of descriptors.

(iii) The partitioning shown in our five-rectangle illustration must have been produced as the pruned tree in Figure 2.

(iv) The method is flexible enough to produce modal partitions of the sort illustrated in Figure 3. These would be useful in applications calling for Wold *et al.*'s 'asymmetry'.[4]

3.4. Nearest neighbours[2]

There is a plausible simplicity in the following approach to the problem of finding a predictor of the form $\hat{y} = f(\mathscr{S})$, where $\mathscr{S}$ denotes chemical structure, on the basis of the information $(y_1, \mathscr{S}_1), \ldots, (y_n, \mathscr{S}_n)$ in the construction set.

Fixing $\mathscr{S}$, i.e. thinking of a totally specified new compound C, select those compounds in $\mathscr{C}$ whose structures are *close* in some sense to $\mathscr{S}$. Then take $\hat{y}$ to be a reasonable function of the y-values of the selected compounds, for instance their average (as in RT). The selection

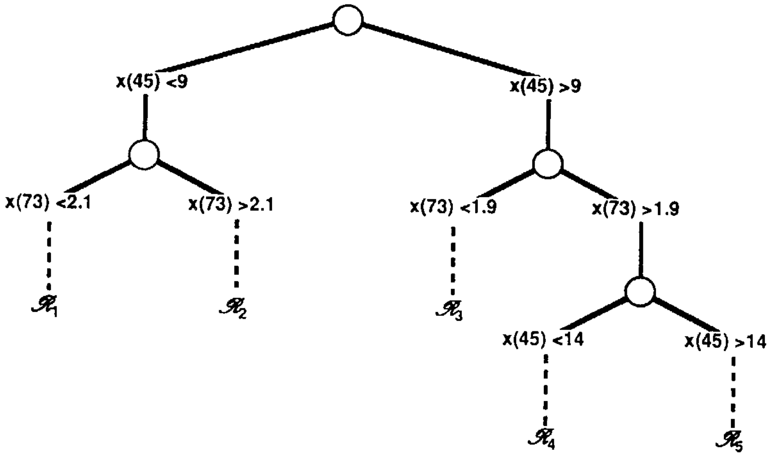

Figure 2. The regression tree for Figure 1

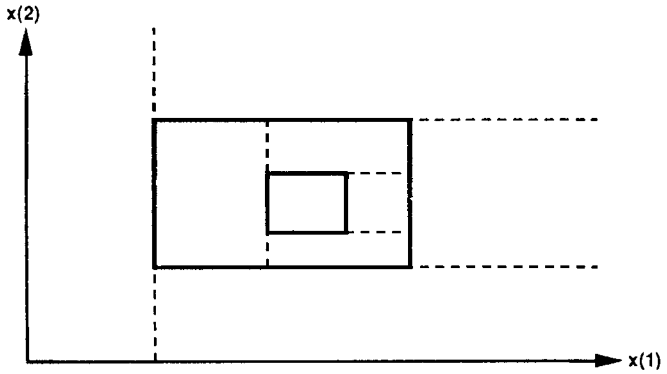

Figure 3. Modal regression tree output

could be done in two ways, given that we can acquire by prior insight a function $d(\mathcal{S}, \mathcal{S}_i)$ whose smallness is thought to encapsulate the 'closeness' of structures $\mathcal{S}$ and $\mathcal{S}_i$ for $i = 1, \ldots, n$.

(a) For fixed k select the k *nearest neighbours* of C, i.e. the k compounds with the smallest values of d.

(b) Fix $\Delta > 0$ so that for all structures $\mathcal{S}$ for which $\hat{y}$ is required there are to-be-selected compounds in $\mathcal{C}$ with $d(\mathcal{S}, \mathcal{S}_i) < \Delta$.

Comments

(i) The high dimensionality of descriptors of $\mathcal{S}$ reduces the feasibility of arriving at a prior choice of d that will be predictively informative. The automatic adoption of Euclidean distance defined for all the available (autoscaled?) descriptors cuts the Gordian knot, of course.

(ii) In comparison with the choice of d, the choice of the control parameter, either k or Δ, may be a relatively straightforward but computationally heavy exercise in cross-validation.

(iii) The usual applications of nearest neighbours have been for $n \gg p$.

4. CLASSIFICATION: TWO CLASSES

So far in this account of different techniques, no conditions have been put on the character of y. It can be either a smoothly variable measure of activity or, as we now suppose, a measure of maximal discreteness taking only two values corresponding to 'active' and 'inactive'.

4.1. Linear discriminant analysis (LDA)

It may be shown that the square of the correlation coefficient, calculated over $\mathscr{C}$ between y and a potential predictor $\hat{y}$, is an increasing function of Student's t-statistic comparing the two sets of values of $\hat{y}$ for active and inactive compounds. At the same time the objective of least-squares regression may be stated as the maximization of the correlation coefficient, while the square of Student's t is the criterion that is maximized in Fisher's LDA.[24] Thus we see that when the predictor $\hat{y}$ is constructed by linear least-squares regression on component variables $t(1), \ldots, t(A)$, there is equivalence between the regression approach and the LDA approach using those variables. The use of $\hat{y}$ to classify a new compound according to whether or not $\hat{y}$ exceeds some chosen value is the same as using the linear discriminant in LDA.

Comments

(i) The variables $t(1), \ldots, t(A)$ may, for example, be principal components constructed as in Section 2.2 – but, whatever it is, their provenance must be taken to be part of the whole procedure when it comes to assessment.

(ii) Wold *et al.*[4] curiously describe the regression formulation as an 'inefficient' variant of LDA.

(iii) LDA is often presented as though it were peculiarly dependent on a very specific probability model, namely that

(a) for each set of compounds (active or inactive) the A-component variables have a multivariate normal distribution
(b) the two covariance matrices are equal.

For QSARs these are off-putting requirements unlikely to be satisfied and therefore LDA is dismissed as a 'parametric' method in contrast to the more flexible, less assumption-ridden, 'non-parametric' methods. (The term 'parametric' refers to all those parameters in the multivariate normal distributions.) The distinction is far-fetched, since, as the regression connection shows, LDA is no less 'non-parametric' than many other methods where linearity is accepted without question. LDA does, however, inherit one specific feature from part (b) of its multivariate normal pedigree: the equality in (b) expresses itself as the *pooling* of the two within-class sums of squares in the denominator of Student's t. (The corresponding feature in the regression formulation is the use of the *total*-sum-of-squares and-products matrix in the 'normal' equations.) Unless the numbers of active and inactive are appreciably unequal, this pooling does not strongly implicate any assumption of equality of variance, since the denominator of t still robustly serves to standardize the unequal variance case.

(iv) 'Quadratic discriminant analysis' (QDA), although derivable as a 'parametric' method,[25] may be regarded as a restriction of the special case of LDA in which the regressors $t(1), \ldots, t(A)$ are specified as a set of s selected variables together with their s squares and $s(s-1)/2$ cross-products, whence $A = s(s+3)/2$. The quadratic variables allow classification boundaries that are curved in the space of the selected variables and may therefore be particularly applicable in the ‘asymmetric case’[4] where the active compounds occupy a roughly ellipsoidal subregion.

(v) Procedures intermediate between LDA and QDA are

(a) ‘proportional covariances’ in which, with multivariate normal modelling, the covariance matrices are supposed proportional rather than equal

(b) ‘common principal components’ in which the two classes share predictively important principal components.

The book by Flury[26] contains a full account of the necessary technicalities.

4.2. Logistic regression (LR)[27]

As a predictor, the LDA discriminant is of unbounded magnitude, whereas the quantity it is predicting takes only two finite values, which, without loss of generality, may be taken to be unity for active compounds and zero for inactives. There is then something to be said for logistic regression, which uses a predictor of the form

$$\hat{y}=\frac{\exp[\tilde{a}+\tilde{b}_1t(1)+\cdots+\tilde{b}_At(A)]}{1+\exp[\tilde{a}+\tilde{b}_1t(1)+\cdots+\tilde{b}_At(A)]} \tag{5}$$

since for all values of the constants $\tilde{a},\tilde{b}_1,\ldots,\tilde{b}_A$ the value of $\hat{y}$ lies between zero and unity.

In fitting the linear LDA predictor by least squares, the use of squared error is a mathematical convenience. There is no analogous simplification in fitting the logistic regression. The standard choice (inspired by the statistical theory of maximum likelihood in a model in which $\hat{y}$ is interpreted as a probability) is the logarithmic error function: $\log(1/\hat{y})$ for an active compound and $\log[1/(1-\hat{y})]$ for an inactive. The criterion to be minimized to determine $\tilde{a},\tilde{b}_1,\ldots,\tilde{b}_A$ is then

$$\sum_{y_i=1}\log(1/\hat{y}_i)+\sum_{y_i=0}\log[1/(1-\hat{y}_i)] \tag{6}$$

Comments

(i) Performed iteratively in computer packages, minimization of (6) has non-trivial complications.

(ii) The criterion gives *great* weight to the avoidance of values of $\hat{y}$ near zero for actives or near unity for inactives. This feature reflects the maximum likelihood pedigree but is hardly justifiable for prediction.

(iii) In place of (6), the modulus error criterion

$$\sum_{y_i=1}(1-\hat{y}_i)+\sum_{y_i=0}\hat{y}_i$$

would avoid complications arising from the logarithms.

(iv) Putting $(\tilde{a},\tilde{b}_1,\ldots,\tilde{b}_A)=\lambda(\alpha,\beta_1,\ldots,\beta_A)$ in (5) and letting $\lambda\to\infty$ makes $\hat{y}$ go to unity or zero according to whether $\alpha+\beta_1t(1)+\cdots+\beta_At(A)$ is positive or negative. Thus LR will respond to the ‘linear separability’ of Section 4.3 and is therefore subject to the associated comments.

4.3. The linear learning machine (LLM)[28]

The active and inactive compounds in $\mathscr{C}$ are said to be *linearly separable* for the variables

$t(1), \ldots, t(A)$ if there are constants $\tilde{a}, \tilde{b}_1, \ldots, \tilde{b}_A$ such that the value of the *linear separator* $\tilde{a} + \tilde{b}_1 t(1) + \cdots + \tilde{b}_A t(A)$ is positive for all active compounds in $\mathscr{C}$ and negative for the inactives. If there is linear separability, LLM theory[28] provides an iterative algorithm that will find such constants. The idea is that $\hat{y} = \tilde{a} + \tilde{b}_1 t(1) + \cdots + \tilde{b}_A t(A)$ would then be useful to predict the activity class of a new compound.

Comments

(i) Linear separability may be a more useful concept for deterministic pattern recognition by AI than for QSARs. Consider the case of $n = 12$ and $A = 2$ in Figure 4, where '×' and '○' denote active and inactive compounds respectively.

(ii) Non-uniqueness of the LLM separator is typical, which may leave appreciable ambiguity in the prediction for a new compound.

4.4. The Ho–Kashyap algorithm[29]

A feature of the LLM algorithm almost fatal to its applicability in QSARs is that when the compounds in $\mathscr{C}$ are not separable, the algorithm does not terminate. The alternative algorithm of Ho and Kashyap[29] is therefore, on this count, preferable.

1. *Construct* (i) the $n \times (A + 1)$ matrix $\mathbf{A}$ whose ith row is $\pm(1, t_i(1), \ldots, t_i(A))$ with $\pm$ according to whether $y_i = 1$ or 0, (ii) the n-vector $\boldsymbol{\beta}(0) = (1, \ldots, 1)^{\mathrm{T}}$ and (iii) the $(A + 1)$-vector $\boldsymbol{\alpha}(0) = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\beta}(0) = \mathbf{A}^{+}\boldsymbol{\beta}(0)$.
2. *Iterate* on $k = 0, 1, 2, \ldots$: $\mathbf{e}(k) = \mathbf{A}\boldsymbol{\alpha}(k) - \boldsymbol{\beta}(k)$, $\boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) + \mathbf{A}^{+}[\mathbf{e}(k) + |\mathbf{e}(k)|]/2$, $\boldsymbol{\beta}(k+1) = \boldsymbol{\beta}(k) + [\mathbf{e}(k) + |\mathbf{e}(k)|]/2$, where $|\mathbf{e}(k)|$ is the vector whose components are the absolute values of those of $\mathbf{e}(k)$.
3. *Terminate* iteration when either $\mathbf{e}(k) \equiv \mathbf{0}$ or else all $e_j(k) \leqslant 0$ with some inequality.
4. *Interpret* $\boldsymbol{\alpha}(k)$ as $(a, b(1), \ldots, b(A))^{\mathrm{T}}$. In the case $\mathbf{e}(k) \equiv \mathbf{0}$, $\boldsymbol{\alpha}(k)$ defines a linear separator of actives and inactives; otherwise the compounds are not separable.

Comments

(i) $\boldsymbol{\alpha}(0)$ gives the coefficients of the LDA discriminant.

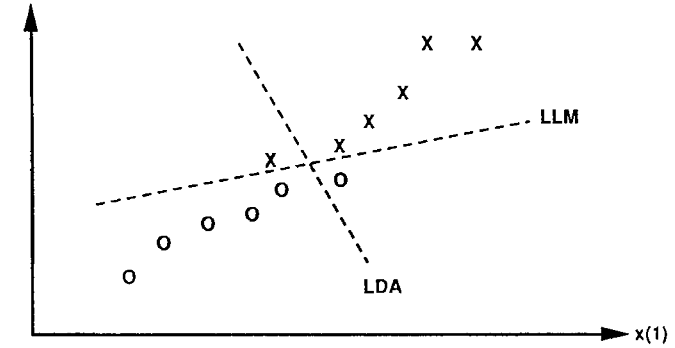

Figure 4. Comparison of LLM with LDA

(ii) The algorithm terminates in a finite number of iterations, but that number may be large and has no specifiable bound.

4.5. Neural networks[30]

Neural network predictors that take $x(1), \ldots, x(p)$ as their 'input' have 'outputs' $\hat{y}$ whose forms have been thought of as models of activity of networks of neurons. A general example[31] is

$$\hat{y} = \beta_1 \psi[\gamma_{11}x(1) + \cdots + \gamma_{1p}x(p)] + \cdots + \beta_q \psi[\gamma_{q1}x(1) + \cdots + \gamma_{qp}x(p)] \qquad (7)$$

where ψ is some sigmoidal function. The estimation of $\{\beta_i\}$ and $\{\gamma_{ij}\}$ ('connection strengths') from the training set has usually been designed with real-time applications in mind, with sequential updating ('back propagation') as new training items come on stream. Although the applications do admit large values of p, they do not have the limitation on n that is a feature of QSARs. Formula (7) may be seen as a generalization of logistic regression: indeed, by choice of ψ and q, it can approximate practically any function of $x(1), \ldots, x(p)$. In particular, for the case of $y = 0$ or 1 and a very large training set which is a random sample from a population of (y, x) values, (7) may approximate[31,32] the Bayes posterior probabilities that would be calculable if the underlying probability distribution of (y, x) were known.

Comment

For QSARs the hard statistical problems of exploiting the attractive generality of formulae such as (7) cannot be bypassed. The current high excitement of the neural network literature may not be particularly relevant to the problems of QSARs.

4.6. Classification trees (CT) in CART[21]

The method of Section 3.3 is applicable without modification to the case of two-valued y.

Comments

(i) Take $y = 1$ and 0 for active and inactive compounds respectively and suppose a split is made at some stage of the tree-growing procedure that divides a node with r active and s inactive compounds $(r + s = t)$ into two nodes with the frequencies r_1, s_1 and r_2, s_2. The reduction in the residual sum-of squares splitting criterion is rs/t^2 multiplied by the chi-squared statistic for the 2×2 table

$$\begin{array}{cc} r_1 & r_2 \\ s_1 & s_2 \end{array}$$

(ii) The simple case where all descriptors are also two-valued (corresponding to the presence or absence of medical symptoms) was independently developed with full cross-validation by Mabbett *et al.*[33] using the chi-squared statistic itself for choice of split in an example with $n = 237$ and $p = 28$. For QSARs we would be more interested in the possibility $n = 28$ and $p = 237$!

(iii) The book by Breiman *et al.*[29] gives an excellent account of this promising flexible technique.

4.7. Nearest neighbours (NN) again

The application of NN to classification for two classes follows Section 3.4, classifying by 'majority vote' among the k nearest neighbours.

Comments (additional to those of Section 3.4)

(i) A pioneering application of NN to QSARs was that of Kowalski and Bender,[34] in which n was 200 and p was 20 (after some selection) and in which the choice of the function d was allowed to depend on the between-class differences of the individual descriptors. The application was critically reanalysed by Mathews[35] using a precursor of RT (Section 3.3).

(ii) The statistical theory,[36] which shows the predictive performance of NN to be agreeably comparable with the (unrealizable) optimal Bayes performance, is asymptotic for large n and fixed p (which is not the QSAR case at all).

4.8. SIMCA (soft independent modelling of class analogy)[22,37]

In contrast with LDA (which used a *single* set of regressors $\{t(1), \ldots, t(A)\}$ to discriminate between actives and inactives), this pioneering technique of Wold uses two sets, $\{t^1(1), \ldots, t^1(A_1)\}$ and $\{t^0(1), \ldots, t^0(A_0)\}$, independently fitted to the active and inactive compounds respectively. These regressors are the first A_1 and first A_0 principal components in the p-dimensional space of predictor variables, defining hyperplanes of dimensions A_1 and A_0 respectively. The predictive classifiability of a new compound as, say, active is then assessed by the ratio of its distance from the hyperplanes for active to the root-mean-square distance of the actives from the hyperplane (with degrees of freedom for the fitting of the principal components); likewise for the classifiability of the new compound as inactive.

Comments

(i) The SIMCA approach is flexible, allowing the new compound to be declared an outlier to both classes or even classifiable in either.

(ii) The values of the control parameters A_1 and A_0 are chosen by a sort of cross-validation which, as Frank and Friedman[25] note, is not directly related to predictive performance.

(iii) Frank and Friedman[25] consider SIMCA in relation to multivariate normal modelling and are critical of the then non-Bayesian character of the ratio classification criterion. The case for *one* modification may be apparent in Figure 5 for $p = 2$ and $A_1 = A_0 = 1$: the new compound C would be classified as active by SIMCA.

4.9. DASCO (discriminant analysis with shrunken covariances)[38]

This development of SIMCA by Frank[38] uses multivariate normal modelling to modify SIMCA in two ways.

(a) The ratio criterion of Section 4.8 is supplemented to take account of the distances of the new compound from the means of the classes in addition to its distance from the class hyperplanes.

(b) The criterion is further adjusted to take account of the effect on classification inference of any difference in the dispersion about their mean for the two classes of compounds. This is illustrated for $p = 2$ and $A_1 = A_0 = 0$ in Figure 6, in which compound C has the

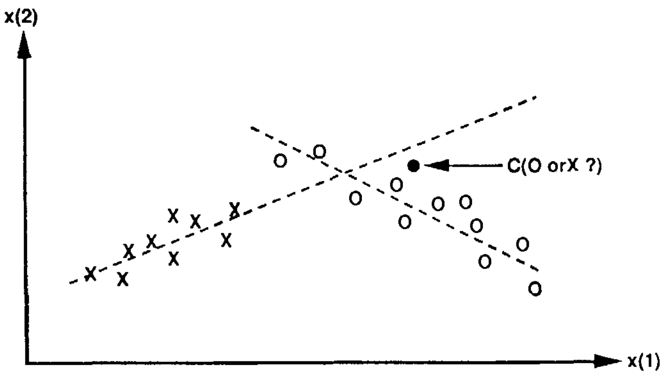

Figure 5. A query for SIMCA

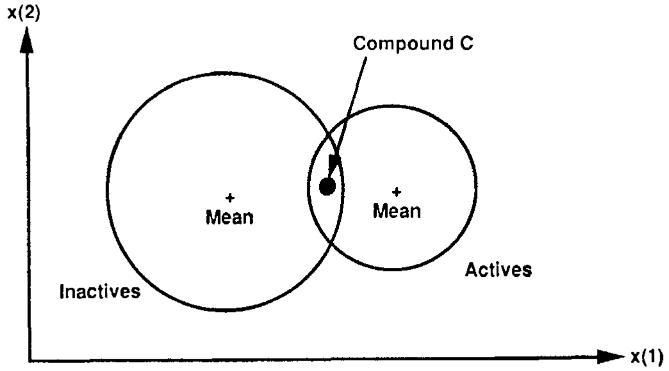

Figure 6. Unequal covariances

same scaled relationship to actives and inactives but the smaller dispersion of the actives enhances the probability that C is an active.

In a further refinement of SIMCA, DASCO reduces the two control parameters A_1 and A_0 to one by choosing A_1 (or A_0) to be the minimum number of principal components that gives a fraction F of the total 'variance' of the actives (or inactives). The value of the single control parameter F is then chosen by standard leave-one-out cross-validation.

Comments

(i) Acronyms are rarely reinterpretable – PLS is a notable exception as observed in Section 2.3, comment (ii) – but we do suggest that the S in DASCO might better expanded as '(partly) sphericized'.

(ii) Frank[38] shows that DASCO outperforms SIMCA on a number of simulated and real data examples. However, these do not put the techniques to the test for the singular case with $n \ll p$.

4.10. Regularized discriminant analysis (RDA)[39]

Like DASCO, RDA is based on approximate multivariate normal modelling: the classification of a new compound is the class with the higher likelihood approximation, in which the two covariance matrices are 'estimated' as

$$(1-\gamma)\mathbf{C}_k(\lambda)+\gamma[\text{trace }\mathbf{C}_k(\lambda)/p]\mathbf{I}$$

where $\mathbf{C}_k(\lambda)$ is $(1-\lambda)\mathbf{C}_k+\lambda\mathbf{C}$, the $\mathbf{C}_k$ ($k=1,2$) are the standard covariance estimates, $\mathbf{C}$ is the standard pooled estimate and $\mathbf{I}$ is the $p\times p$ identity matrix. The control parameters λ and γ are chosen by cross-validation.

Comments

(i) For $n>p$ the calculation of λ and γ is facilitated by leave-one-out matrix algebra.

(ii) The 'regularized' in RDA refers to the regularization of the individual covariance matrices $\mathbf{C}_1$ and $\mathbf{C}_2$ towards $\mathbf{C}$ as λ approaches unity and further possible regularization towards multiples of $\mathbf{I}$ as γ approaches unity.

4.11. Miscellaneous techniques

When y takes the value y_1 for the n_1 actives and y_2 for the n_2 inactives, with $y_1>y_2$, the SLS predictor of Section 3.1 turns out to be an increasing function of the linear discriminant

$$\mathbf{d}^{\mathrm{T}}\mathbf{S}^{+}\mathbf{x} \tag{8}$$

where $\mathbf{d}=\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}$ is the difference between the average $\bar{\mathbf{x}}^{(1)}$ of the descriptor vectors for actives and their average $\bar{\mathbf{x}}^{(2)}$ for inactives. In the terminology of the analysis of variance and covariance, $\mathbf{S}$ is the total-sum-of-squares-and-products matrix of the descriptors for all n compounds. For the non-singular case with $n\geqslant p+1$ the discriminant becomes $\mathbf{d}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{x}$, which for $n\geqslant p+2$ is an increasing function of the Fisher discriminant $\mathbf{d}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}^{-1}\mathbf{x}$, where $\mathbf{S}_{\mathrm{W}}$ is the within-sum-of-squares-and-products matrix. (This mathematical equivalence is required by the equivalence of LDA and regression, also used in Section 4.1.) For the singular case with $n<p+1$ there is no such relationship between (8) and the Fisher discriminant analogue

$$\mathbf{d}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}^{+}\mathbf{x} \tag{9}$$

Discriminant (8) is partially justified by the argument of Section 3.1, comment (iv), but use of (9) would appear to rest on the hope that predictive information can be extracted from the space spanned by the within-group descriptor vector deviations from $\bar{\mathbf{x}}^{(1)}$ and $\bar{\mathbf{x}}^{(2)}$, together with the orthogonal projections of $\bar{\mathbf{x}}^{(1)}$ and $\bar{\mathbf{x}}^{(2)}$ on to that space.

An almost 'orthogonal' hope is expressed in the 'zero-variance discriminant' $\mathbf{a}^{\mathrm{T}}\mathbf{x}$, which is definable algebraically by the $\mathbf{a}$ maximizing $[\mathbf{a}^{\mathrm{T}}(\bar{\mathbf{x}}^{(1)}-\mathbf{x}^{(2)})]^2$ subject to $a_1^2+\cdots+a_p^2=1$ and $\mathbf{a}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{a}=0$. (This definition is motivated as an attempt to preserve, in the singular case, the sense of Fisher's derivation of the LDA discriminant.)

These methods based on either $\mathbf{S}^{+}$ or $\mathbf{S}_{\mathrm{W}}^{+}$ may be regarded as making use (not necessarily with advantage) of all the eigenvectors of $\mathbf{S}$ or $\mathbf{S}_{\mathrm{W}}$ for non-zero eigenvalues (in number, $n-1$ for $\mathbf{S}$ and $n-2$ for $\mathbf{S}_{\mathrm{W}}$). Specifically, consider the eigendecomposition $\mathbf{S}=\lambda_1\mathbf{v}_1\mathbf{v}_1^{\mathrm{T}}+\cdots+\lambda_{n-1}\mathbf{v}_{n-1}\mathbf{v}_{n-1}^{\mathrm{T}}$. The just non-singular least-squares regression of y on the reduced set of variables $t(1)=\mathbf{v}_1^{\mathrm{T}}\mathbf{x},\ldots,t(n-1)=\mathbf{v}_{n-1}^{\mathrm{T}}\mathbf{x}$ delivers the predictor $\hat{y}$ corresponding to the discriminant $\mathbf{d}^{\mathrm{T}}\mathbf{S}^{+}\mathbf{x}$ (similarly for $\mathbf{S}_{\mathrm{W}}$).

The final technical twist in this miscellaneous bag is to use only a subset of $\{t(j)\}$. For S this brings us back to (standard) principal components or reordered principal components (Section 2.2). The reordering of the latter is here equivalent to selecting the $t(j)$ with the larger values of

$$[\mathbf{v}(j)^{\mathrm{T}}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})]^2/\lambda_j$$

5. CLASSIFICATION: MORE THAN TWO CLASSES

We now briefly indicate how some of the techniques of Section 4 extend to the case where 'activity' is categorized into $k > 2$ unordered classes. For example, for $k = 3$, compounds may be classed as 'inactive', 'active through biochemical mechanism 1' or 'active through biochemical mechanism 2'.

Following prior reduction to $t(1), \ldots, t(A)$ (via principal components maybe), LDA will now typically generate $k - 1$ linear discriminants for the user. As for $k = 2$, there is equivalence (now of the space spanned by the discriminants) between the multiple-regression (canonical variate) approach using dummy y-variates and that of analysis of variance.

For the canonical variate approach let $(y(1), \ldots, y(k))$ be the *multivariate indicator vector* of class membership, i.e. $y(j) = 1$ if the compound is in class j and $y(j) = 0$ if not. The first linear discriminant is the linear combination of $t(1), \ldots, t(A)$ of maximal correlation with some (unrestricted) linear combination of $y(1), \ldots, y(k)$. The second discriminant is likewise defined, subject to the side condition that it be uncorrelated (over $\mathscr{C}$) with the first – and so on.

In analysis of variance the discriminants are likewise sequentially constructed, but the criterion to be maximized is the F-statistic for one-way analysis of variance of the discriminant for the k classes, while the side condition requires that the discriminants be uncorrelated with respect to *within*-class variations.

For $k = 2$ the user of the single discriminant has to choose a critical value separating the two classes, usually corresponding either to the half-way point between the class means or to the value that maximizes the number of compounds in $\mathscr{C}$ that are (self-)classified correctly. For $k > 2$, however, there are no obvious analogous choices. This ambiguity is resolved (at perhaps some cost in feasibility and realism) by the Bayesian approach, which generates posterior probabilities of class membership.

The extension of neural network techniques to $k > 2$ uses a combination of multivariate indicator and an overall Euclidean squared error criterion. Subject to the severe qualifications mentioned in Section 4.5, this maintains the Bayesian connection.

The techniques SIMCA, DASCO and RDA cover the case $k > 2$ without difficulty, as they were designed to do. However, those techniques of Section 4.11 that involve $\mathbf{S}^+$ or $\mathbf{S}_{\mathbf{W}}^+$ do raise the same ambiguity of application to prediction that was noted for LDA.

6. JOINT PREDICTION OF ACTIVITIES

The question of the possible value of joint prediction was optimistically mooted in Section 4 of S&J[1] in rather general terms.

For the special case where two different activity measures $y(1)$ and $y(2)$ have been recorded for each of the compounds $C_1, \ldots, C_n$ of the congeneric series, the general question takes the form:

> For an unsynthesized or newly considered compound C for which only the structure $\mathscr{S}$ is known, is there any value in using a *joint technique* for the prediction of $y(1)$ and $y(2)$ apart from any gain in computational efficiency?

(By a 'joint technique' we mean one in which the prediction of $y(1)$, say, involves, and hopefully benefits from, the use of the values of $y(2)$ for $C_1, \ldots, C_n$, even though the value of $y(2)$ is not available for C.)

The intuitive argument of Section 4 of S&J[1] led us to think that there might be such value. However, we also believe that its realization may call for more scientific insight than is required if we merely adopt some automatic technique that happens to have the jointness property.

The LDA method of Section 5 with $k = 3$ is an example of a joint prediction technique. (Since $y(1) + y(2) + y(3) = 1$, we may treat $y(3)$ as redundant and concentrate on $y(1)$ and $y(2)$.) Prediction of $y(1)$, say, using the LDA discriminants may be influenced by the values of $y(2)$ for $C_1, \ldots, C_n$. For the example mentioned in Section 5, this means that the prediction of the dichotomy 'inactive' versus 'active by either mechanism' may be influenced by knowledge of what the mechanisms of activity were for the active compounds among $C_1, \ldots, C_n$.

A more complex example of a joint technique is PLS2, referred to in Section 2.3,comment (iv), and widely recommended and adopted for QSAR studies. To appreciate the nature of PLS2 more easily, its standard algorithmic specification may be replaced by the following simple modification of Section 2.3, stated for the case of q activities:

> The activity y is generalized to a linear combination $l_1 y(1) + \cdots + l_q y(q)$ with $l_1^2 + \cdots l_q^2 = 1$, while the maximizations that define the sequentially constructed components are taken over $l_1, \ldots, l_q$ as well as over $c_1, \ldots c_p$.

Using this approach, Stone and Brooks[40] have devised an analogous generalization of CR in Section 2.4 – 'joint continuum regression' (JCR). There are two versions of JCR: JCR_1 is a continuum between PCR and PLS2, extrapolating to a type of least squares; JCR_2 is a continuum between PCR and canonical (shortest) least squares. Comparisons of JCR and CR on real data have so far uncovered no examples in which there is any significant gain in predictive performance from the introduction of 'jointness', although such gain may, with difficulty, be artificially simulated. Since this finding embraces the comparison of PLS2 and PLS1, we are led to question the adoption of PLS2 instead of the simpler, non-iterative PLS1 for the prediction of any particular activity. However, we would be delighted to receive evidence that our scepticism about jointness for this and related techniques is unjustified. Frank and Friedman[20] touch on the same doubts when they compare ridge regression with their multiple-y generalization of the technique.

7. DISCUSSION

We would have liked to end this review, if we could have managed it, with a finely discriminatory assessment of the pros and cons of the many techniques we have looked at. Instead we have to opine that any such assessment could be both premature and presumptuous. There is in QSARs such a plethora of unexplored possibilities that the only unifying idea seems to be that of realistic assessment, which we hope we have stressed sufficiently. Even in that matter, the question of selection bias in a fitted relationship (end of Section 6.5 of S&J[1]) requires further vigorous study.

What the area may need, above all else, is tolerance and open-mindedness for the new – tempered by informed criticism.

ACKNOWLEDGEMENTS

We thank Simon Belsten of Shell Research for the artwork and Mac McCarthy of Shell Research for his generous support.

REFERENCES

1. M. Stone and P. Jonathan, *J. Chemometrics*, **7**, 455, (1993).
2. C. Hansch, P. G. Sammes and J. B. Taylor (eds), *Comprehensive Medicinal Chemistry*, Vol. 4, Pergamon, Oxford (1990).
3. P. J. Rousseeuw, *J. Chemometrics*, **5**, 1 (1991).
4. S. Wold, C. Albano, W. J. Dunn, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg and M. Sjostrom, in *Chemometrics: Mathematics and Statistics in Chemistry*, ed. by B. R. Kowalski, Reidel, Dordrecht, p. 17 (1984).
5. N. R. Draper and H. Smith, *Applied Regression Analysis*, Wiley, New York (1981).
6. I. T. Jolliffe, *Principal Components Analysis*, Springer, New York (1986).
7. I. S. Helland, *Commun. Stat.* **17**, 581 (1988).
8. H. Wold, in *Encyclopaedia of Statistical Sciences*, Vol. 6, ed. by N. L. Johnson S. Kotz, p. 581, Wiley, New York (1984).
9. R. Manne, *Chemometrics Intell. Lab. Syst.*, **2**, 187 (1987).
10. S. de Jong, *Chemometrics Intell. Lab. Syst.* **18**, 251 (1993).
11. M. Stone and R. J. Brooks, *J. R. Stat. Soc. B*, **52**, 237 (1990).
12. I. E. Frank, *Chemometrics Intell. Lab. Syst.*, **1**, 233 (1987).
13. B. R. Kowalski and M. B. Seasholtz, *J. Chemometrics*, **5**, 129 (1991).
14. A. Lorber, L. E. Wangen and B. R. Kowalski, *J. Chemometrics*, **1**, 19 (1987).
15. S. de Jong and H. A. L. Kiers, *Chemometrics Intell. Lab. Syst.* **14**, 155 (1992).
16. T. T. Dunne and M. Stone, *J. R. Stat. Soc. B*, **55**, 369 (1993).
17. R. Marbach and H. M. Heise, *Chemometrics Intell. Lab. Syst.* **9**, 45 (1990).
18. A. E. Hoerl and R. W. Kennard, *Technometrics*, **12**, 55 (1970).
19. R. Sundberg, *J. R. Stat. Soc. B*, **55**, 653 (1993).
20. I. E. Frank and J. H. Friedman, *Technometrics*, **35**, 109 (1993).
21. L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Tress*, Wadsworth, Belmont, Calif. (1984).
22. S. Wold, in *Drug Design: Fact or Fantasy?* ed. by G. Jolles and K. R. H. Wooldridge, Academic, New York (1984).
23. J. A. Sonquist, E. L. Baker and J. N. Morgan, *Searching for Structure*, Institute of Social Research, Ann Arbor, MI (1973).
24. R. A. Fisher, *Ann. Eugen.* **7**, 179 (1936).
25. I. E. Frank and J. H. Friedman, *J. Chemometrics*, **3**, 463 (1989).
26. B. Flury, *Common Principal Components and Related Multivariate Models*, Wiley, New York (1988).
27. D. Collett, *Modelling Binary Data*, Chapman and Hall, London (1991).
28. N. J. Nilsson, *Learning Machines: Foundations of Trainable Pattern Classifying Systems*, McGraw-Hill, New York (1965).
29. Y. C. Ho and R. L. Kashyap, *J. SIAM: Control*, **4**, 112 (1966).
30. B. D. Ripley, *Statistical Aspects of Neural Networks*, Proc. SemStat, to be published by Chapman and Hall, London.
31. D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley and B. W. Suter, *IEEE Trans. Neural Netw.* **NN-1**, 296 (1990).
32. E. A. Wan, *IEEE Trans. Neural Netw.* **NN-1**, 303 (1990).
33. A. Mabbett, M. Stone and J. Washbrook, *Appl. Stat.* **29**, 198 (1980).
34. B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.* **96**, 916 (1974).

35. R. J. Mathews, *J. Am. Chem. Soc.* **97**, 935 (1975).
36. T. Cover and P. E. Hart, *IEEE Trans. Inf. Theory*, **IT-13**, 21 (1967).
37. S. Wold, *J. Pattern Recogn.* **8**, 127 (1976).
38. I. E. Frank, *Chemometrics Intell. Lab. Syst.* **4**, 215 (1988).
39. J. H. Friedman, *J. Am. Stat. Assoc.* **84**, 165 (1989).
40. M. Stone and R. J. Brooks, *Research Report 124*, Department of Statistical Science, University College London (1993).