A Bayesian Decision Theory Approach to Variable Selection for Discrimination

T. Fearn[*], P. J. Brown[†] and P. Besbeas[‡]

October 5, 2000

Abstract

Motivated by examples in spectroscopy, we study variable selection for discrimination in problems with very many predictor variables. Assuming multivariate normal distributions with common variance for the predictor variables within groups, we develop a Bayesian decision theory approach that balances costs for variables against a loss due to classification errors. The approach is computationally intensive, requiring a simulation to approximate the intractable expected loss and a search, using simulated annealing, over a large space of possible subsets of variables. It is illustrated by application to a spectroscopic example with 3 groups, 100 variables, and 71 training cases, where the approach finds subsets of between 5 and 14 variables whose discriminatory power is comparable with that of linear discriminant analysis using principal components derived from the full 100 variables. We study both the evaluation of expected loss and the tuning of the simulated annealing for the example, and conclude that computational effort should be concentrated on the search.

KEYWORDS: Bayes; decision theory; discriminant analysis; near infrared spectroscopy; simulated annealing; variable selection.

1 Introduction

The problem studied here is that of discriminant analysis, sometimes called supervised pattern recognition, with many variables. For each case in some training set we have a $q \times 1$ vector observation X, and a known assignment to one of g groups. On the basis of these data, we wish to derive a classification rule that will assign further cases to their

[*]University College London, UK
[†]University of Kent at Canterbury, UK
[‡]University of Kent at Canterbury, UK

correct groups using the information in their X vectors. McLachlan (1992) gives a good overview of statistical approaches to the discriminant problem.

If the distributions of X conditional on group membership are assumed to be multivariate normal, with common variance matrix but different means across groups, then the appropriate statistical methodology is that of linear discriminant analysis (LDA). This is simple to implement in either classical (McLachlan, 1992, Chapter 3) or Bayesian (Brown, 1993, Chapter 8) frameworks when the dimensionality q of X is small, and there is evidence that it often performs well even when the multivariate normality is a best an approximation (McLachlan, 1992, Chapter 5). There are many interesting cases though when q is not at all small. The example in Section 4 is typical of the spectroscopic problems that motivated this work. It has $q = 100$, and a training set of 71 cases split into three groups. In problems like this the pooled sample within-groups covariance matrix, of dimension 100×100 and based on 68 degrees of freedom, will be singular and LDA is no longer straightforward. In practice, as q increases, we run into problems of unstable variance estimates long before the matrix actually becomes singular.

In this situation there are a number of ways to proceed, see for example Stone & Jonathan (1994), Krzanowski *et al* (1995) and Hastie *et al* (1995). Brown *et al* (1999a) use proper and highly structured prior distributions in Bayesian analyses that are able to cope with high dimensions for X. An approach often taken for spectroscopic examples like the one in Section 4 is to reduce the dimensionality of X by a principal components analysis of the total covariance matrix before carrying out LDA. This method is described in more detail in Section 4.4, where it is applied to the example. See also Jolliffe (1986), Chapter 9, for a critique of this approach. Another way of reducing dimensionality prior to an LDA is to select a small subset of the q variables, and it is this method that we explore here.

Brown *et al* (1999b) study the variable selection problem for normal theory multivariate linear regression, using a Bayesian decision theory approach that trades off expected prediction losses against costs for variables. In Sections 2 and 3 a similar approach is developed for the discrimination problem, balancing expected classification loss against costs for variables. The discrimination problem is less tractable than regression, because the expected classification loss, unlike the prediction loss in the regression setup, cannot be found analytically and has to be approximated numerically.

The methods developed are applied to a spectroscopic example in Section 4, where various methods of estimating the expected classification loss are compared, and the setting up of a search over subsets by simulated annealing is explored, and there is some general discussion in Section 5.

2 Theory

First we will describe the multivariate normal model for the data, then we set up the variable selection as a decision theory problem and develop the Bayesian solution.

The training data for the jth of g groups is represented as the $n_j \times q$ matrix X_j for $j = 1, \ldots, g$. The ith row of this matrix is a random vector from a q-variate multivariate normal with mean $\mu_j(1 \times q)$ and covariance matrix Σ. We denote the total sample size by $n = \sum_1^g n_j$. Here for simplicity we have considered a common covariance matrix across groups, but having it depend on j is a straightforward generalisation (Brown, 1993, Chapter 8), and further elaboration may be desirable (Brown *et al.*, 1999a). Taking all the jth group observations together, using the matrix-variate notation of Dawid (1981) (see also Brown (1993), Appendix A), we have

$$X_j - 1_{n_j}\mu_j \sim \mathcal{N}(I, \Sigma) \tag{1}$$

independently for $j = 1, \ldots, g$, with I the $n_j \times n_j$ identity matrix, indicating independence across rows.

Whereas we assume vague priors for $\mu_j, j = 1, \ldots, g$, the prior distribution of Σ is given as

$$\Sigma \sim \mathcal{IW}(\delta; Q) \tag{2}$$

where the $q \times q$ scale matrix Q is such that $E(\Sigma) = Q/(\delta - 2)$ for $\delta > 2$ and the shape parameter $\delta > 0$ and Q need to be specified, with larger values of δ being more informative. For the example in Section 4 we will take $\delta = 3$, thus tempering the effect of the strong choice $Q = kI$.

The joint posterior distribution of μ_j, Σ is by conjugacy also multivariate normal-inverse Wishart and depends on sufficient statistics $\hat{\mu}_j, S$, where

$$\begin{aligned}
\hat{\mu}_j &= 1'_{n_j} X_j / n_j \\
S_j &= (X_j - 1_{n_j}\hat{\mu}_j)'(X_j - 1_{n_j}\hat{\mu}_j) \\
S &= \sum_{j=1}^{g} S_j.
\end{aligned}$$

The updated hyperparameters are $Q^* = S + Q$ and $\delta^* = \delta + n$.

The predictive distribution of a future observation X^f, $(1 \times q)$ from group j given all the training data is a key ingredient in Bayesian discrimination. This is multivariate Student T

$$X^f - \hat{\mu}_j \sim \mathcal{T}(\delta^*; a_j, Q^*), \tag{3}$$

with $a_j = 1 + 1/n_j$, and has probability density function proportional to

$$p_j = a_j^{\delta^*/2}\{a_j + (X^f - \hat{\mu}_j)(Q^*)^{-1}(X^f - \hat{\mu}_j)'\}^{-(\delta^*+q)/2},$$

(Brown, 1993).

We now consider variable selection. Let γ $(q \times 1)$ be a binary vector indicating selected variables, where the ith variable is selected if the ith entry of γ is a 1 and 0 if not. A typical selected model will have $p = |\gamma| < q$ variables. The indicator can serve to pick out appropriate parts of our data. Let X^f_γ be the $1 \times p$ sub-vector of X^f, and $\hat{\mu}_{j\gamma}, \mu_{j\gamma}, Q^*_{\gamma\gamma}$

the corresponding parts of $\hat{\mu}_j, \mu_j, Q^*$. Then the appropriate predictive distribution derived from (3) is

$$X^f_\gamma - \hat{\mu}_{j\gamma} \sim \mathcal{T}(\delta^*; a_j, Q^*_{\gamma\gamma}), \tag{4}$$

and the corresponding probability density function is proportional to

$$p_j(\gamma) = a_j^{\delta^*/2}\{a_j + (X^f_\gamma - \hat{\mu}_{j\gamma})(Q^*_{\gamma\gamma})^{-1}(X^f - \hat{\mu}_{j\gamma})'\}^{-(\delta^*+p)/2}.$$

The problem at hand is that of discrimination. Let $y^f \in \{1, \dots, g\}$ be the true group for the item with observation X^f. Let π_j be the prior probability that a future observation comes from group j. The plug-in estimate $\hat{\pi}_j = n_j/n$ is appropriate under an uninformative Dirichlet prior distribution for π and training data exchangeable with future data, in particular arising in the same proportions from the groups. Then the estimated conditional probability that the future observation comes from group j is

$$\pi_j(\gamma) = p(y^f = j \mid X^f_\gamma, \mathcal{X}) \propto p(X^f_\gamma \mid j, \mathcal{X})\hat{\pi}_j \tag{5}$$

where $\mathcal{X} = \{X_1, \dots, X_g\}$. By the special structure of the $T-$distribution the conditioning on $\mathcal{X}$ really only conditions on $\mathcal{X}_\gamma$ (Brown, 1993, section 6.4) and

$$\pi_j(\gamma) = \frac{p_j(\gamma)\hat{\pi}_j}{\sum_{i=1}^g p_i(\gamma)\hat{\pi}_i}. \tag{6}$$

We assume that the misclassification loss is that of a logarithmic scoring rule (Savage, 1971; Bernardo and Smith, 1994), with

$$L(\gamma, X^f_\gamma, y^f) = -\ln\{\pi_{y^f}(\gamma)\}.$$

The overall loss is then this averaged over the predictive distribution of (X^f_γ, y^f) conditional on the training data. As above it suffices to condition on $\mathcal{X}_\gamma$ because of the properties of the multivariate T. As in Brown *et al* (1999b) we combine this inferential loss with a cost function for variables selected. Thus we have

$$\begin{aligned} C(\gamma) &= E\{L(\gamma, X^f_\gamma, y^f) \mid \mathcal{X}_\gamma\} + g(\gamma) && (7) \\ &= \mathcal{L}(\gamma) + g(\gamma). && (8) \end{aligned}$$

In our example later we will adopt the simplest form of $g(\gamma) = cp$ with $p = |\gamma|$ and c a constant cost per variable. In Brown *et al* (1999b) c was interpreted, both in the development and in an example involving the specification of an instrument, as a real cost. To understand the exchange rate between variable costs and misclassification losses, suppose that the addition of an extra variable was able to increase the probabilities given for true assignments by a factor of $(1 + \epsilon)$, i.e. $\pi_{y^f}(\gamma') = \pi_{y^f}(\gamma) \times (1 + \epsilon)$ for all future y^f if we add the variable to γ to give γ'. Then $C(\gamma') - C(\gamma) = -\ln(1 + \epsilon) + c$ which is approximately $c - \epsilon$ for small ϵ, so the added variable reduces total cost if $\epsilon > c$. Thus a variable costing c is tradable for a relative improvement of $100c\%$ in the classification probabilities. An alternative way to think of c, which may be more appropriate to some problems, is as a (somewhat arbitrary) penalty function on the number of variables. Either way, the use of more general forms for $g(\gamma)$ would add very little to the computational burden.

3 Implementation

In summary we require to minimise over γ the cost $C(\gamma)$ given in (8). The first difficulty is the computation of the expectation implicit in the first term of (8). There is no analytical expression for this and the distribution of X^f_γ has dimension p which is as high as 100 in the example of Section 4. We use simulation to evaluate the integral. Our approach involves creating a once-and-for-all master matrix of evaluation data, comprising N vectors $X^f(1 \times q)$ and associated y^f, and approximating the first term of (8) by

$$\mathcal{L}(\gamma) \approx \frac{1}{N} \sum_{j=1}^{g} \sum_{i=1}^{N_j} L(\gamma, X^f_{j\gamma}(i), j). \tag{9}$$

Here $X^f_{j\gamma}(i)$ represents the γ variable selection for the ith row of the master matrix from the jth group, which has N_j observations. The evaluation data are created either (i) by using a subset of cases taken from the training data and set aside for this purpose or (ii) by simulating from (3), with $N_j = \hat{\pi}_j N$. In the example of Section 4 we use both $N = 200$ and $N = 1000$ so as to investigate sensitivity to the choice of N.

The second difficulty concerns the search over the large number of possible γ. The obvious forward and backward stepwise algorithms do not work for the example described below (see Section 4.3), each getting trapped in a local minimum. To overcome this we use simulated annealing as in Brown *et al* (1999b) where it seemed to work well, one motivation for this approach being that it is easy to implement.

The search starts at a randomly chosen γ. At each step the algorithm constructs γ^{new} from γ^{old} by choosing at random between 3 types of move, as follows.

A : Add a variable by choosing at random a 0 in γ^{old} and changing it to a 1.

D : Delete a variable by choosing at random a 1 in γ^{old} and changing it to a 0.

S : Swap two variables by choosing independently at random a 0 and a 1 in γ^{old} and changing both of them.

The moves are chosen randomly with probabilities P_A, P_D and $P_S = 1 - P_A - P_D$. At the boundaries, with all variables included or no variable present, only deletion or addition is possible respectively, and we choose this move with probability 1.

At each step $d = C(\gamma^{new}) - C(\gamma^{old})$ is calculated. If $d < 0$, γ^{new} is always accepted. Otherwise it is accepted with probability $\exp(-d/T)$ where T is a control parameter, the temperature. Allowing moves to 'worse' subsets may help to avoid being trapped in local minima. We begin with a high temperature, allowing almost unrestricted movement, and reduce the temperature sequentially, using for most of our computations a cooling schedule of the form $T_i = \rho T_{i-1},\ 0 < \rho < 1$. We stop when the temperature becomes so low that the system essentially stops moving. In practice this means stopping the search after M steps with no accepted move, for some large M.

The aim in setting all this up is to strike the right balance between the computational effort devoted to evaluating (9) and that devoted to the search. Both the size of the set of observations used to evaluate (9) and the choice of cooling schedule and stopping criterion for the annealing have a major impact on the time taken for the search. How to get the balance right in general is not clear, but we study this for a particular example below, and try to make some general comments in Section 5 in the light of our experience with this example.

4 Example

We have applied the methodology described in Sections 2 and 3 to a spectroscopic example which is typical of the sort of application that motivated this work. Classification using spectra is becoming increasingly important in areas such as remote sensing, as well as in authenticity applications like this one.

4.1 Wheat variety data

The data consist of NIR transmission spectra on 117 samples of wheat. The spectra were measured on samples of ungrounded wheat using a Tecator Infratec Grain Analyzer which measures transmission through the wheat sample of radiation at $q = 100$ wavelengths from 850 to 1048nm in steps of 2nm. Each wheat sample was classified into one of $g = 3$ named varieties, on the basis of known provenance. The investigation from which these data are taken had a number of aims, one of which was to explore the possibility of using NIR spectra to assign unknown samples to the correct variety.

The 117 samples were split, randomly within groups, into training, tuning and validation sets containing as close to 60%, 20% and 20% as possible. The tuning set was used as evaluation set in one approach to the computation of (9), and as the basis for simulation in another. The resulting numbers of samples are shown in Table 1 and graphs of the training spectra for all three varieties are given in Figure 1. The plotted spectra have been centered by subtraction of the mean spectrum for the 71 samples in the training set.

Variety	1	2	3	Total
Training	32	22	17	71
Tuning	10	7	6	23
Validation	10	7	6	23

Table 1: Wheat data: numbers of samples in the training, tuning and validation sets by variety.

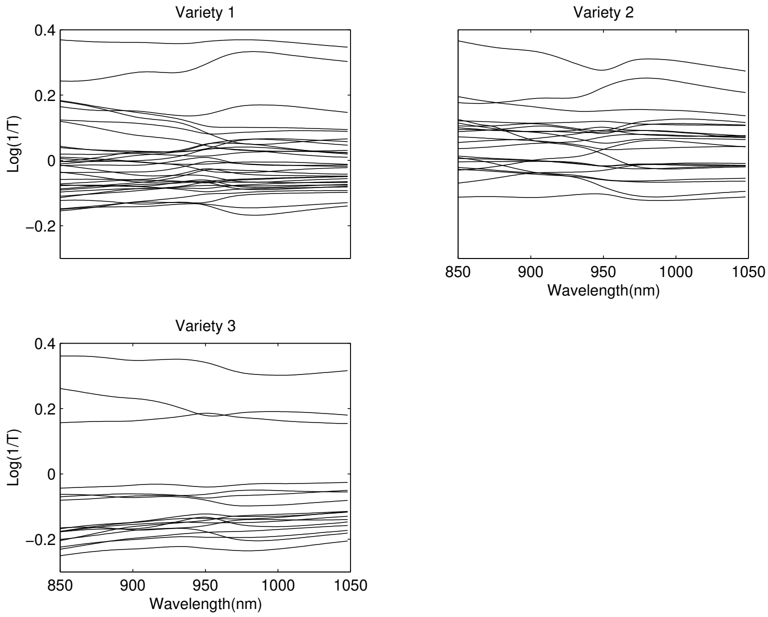

Figure 1: Transmission spectra of three wheat varieties in the training data, centered by subtraction of mean spectrum

4.2 Application of variable selection methodology

To apply the methodology we need to specify the inverse Wishart prior distribution (2) for Σ and the cost function $g(\gamma)$ in (8). We use the simple choices $\delta = 3$, $Q = kI$ and $g(\gamma) = cp$ suggested in Section 2. This leaves us with two scalar parameters, k, which controls the shrinkage in the estimation of the common covariance matrix, and c, which is the cost per variable used. We have taken an empirical approach to the choice of both of these constants.

The shrinkage parameter, $k = 10^{-8}$, was chosen to have the same magnitude as the lower quartile of the nonzero eigenvalues of the pooled sums of squares and products matrix S based on all 100 variables and calculated using the training data. The reasoning behind this was that such a choice should provide enough shrinkage to prevent numerical problems occurring when the search investigates subsets with many variables, but would be likely to have little effect on the relatively small subsets of wavelengths we were interested in finding. To some extent, shrinkage and variable selection are alternative strategies for coping with the problem of large numbers of variables. In investigating variable selection, we have tried to use minimal shrinkage.

The two possible interpretations of the cost parameter c have been discussed earlier, in Section 2. In this particular application the interpretation of c as a true cost would be difficult. The 100 points are delivered in one scan, for the same cost as a subset. One might conceive of building a cheaper instrument that measured at some wavelengths only, but we are not in a position to quantify the costs per wavelength for this. Thus we have treated cp as a penalty function, and chosen, by trial and error, some values of c (0.05, 0.025, 0.0125) that result in the selection of small to moderate numbers of variables for this problem.

We investigated five different ways of estimating the expected loss as described in Section 3. The sum in (9) was calculated using samples of size 200 and 1000 simulated from the predictive distribution (3) conditional on first the training and then the tuning data sets with group sizes directly proportional to those of the actual training/tuning data. The fifth variant was to use the 23 actual observations in the tuning data over which to calculate the sum.

4.3 Setting up the search

The other decisions to be made concern the search process. We investigated this by running a number of searches for one choice of the cost parameter, $c = 0.05$, and the cheapest option for estimating the loss, which was to use the 23 observations in the tuning set.

The lowest cost subset found for this case included variables (13, 19, 30, 35, 45, 47) with associated cost $C = 0.6107$. This cost is reported in the bottom row of Table 3. It is not possible to be certain that this is truly the best subset, but in over 400 stochastic searches it was regularly identified as the best visited and its cost was never beaten. There are a number of similar subsets with similar costs. In particular, (14, 17, 30, 35, 45, 47), the next best seen, has $C = 0.6170$ and (14, 17, 30, 36, 44, 47), the third best, has $C = 0.6181$.

Neither forward nor backward stepwise searches find any of these subsets. A step-up search identifies variable 51 as the single variable with the lowest cost, $C = 1.0557$, but cannot reduce this cost by adding any one of the other 99 variables to this one. A step-down search, starting with all 100 variables, stops at the subset (2, 13, 17, 28, 36, 44, 46, 67, 93, 100) with $C = 0.7674$. Deletion of any one of these 10 variables increases the cost. Thus some more sophisticated search procedure is needed. We chose to use simulated annealing as described in Section 3.

Simulated annealing is attractively easy to implement, but requires a number of choices, starting temperature, cooling schedule, and stopping rule being the key ones. Prescriptive rules for setting these are hard to find in the literature. One can however learn a lot by guessing some sensible settings and then watching the progress of a few searches, which is what we did for the case described above.

For all our searches the three types of move were chosen with equal probabilities of 1/3. It seems highly unlikely that it is worthwhile expending effort on tuning this choice. Similarly we always began with 50 randomly selected variables. The starting temperatures were always chosen high enough so that the early movement of the search was almost random, and the influence of the starting point rapidly lost.

To get some idea of an appropriate temperature scale and cooling rate, note that with a cost per variable of $c = 0.05$, $\exp(-c/T) = 0.6$ for $T = 0.1$ and 0.007 for $T = 0.01$. Thus the cost penalty on additional variables changes from mild to severe over this temperature range, and the cooling schedule needs to be chosen so that a reasonable amount of time is spent there.

From the many types of cooling schedule that have been proposed,(see for example Dowsland (1995)) we chose to try one of the simplest, a geometric schedule in which the temperature is reduced by a factor of $\rho = 0.999$ after each accepted move. This choice of ρ means that $-1/\log_{10}\rho = 2300$ steps are taken as the temperature reduces by a factor of 10. This represents a first guess at 'a reasonable amount of time'.

Running the search at fixed temperatures of $T = 0.1$, $T = 1$ and $T = 10$ showed that at $T = 1$ most, and at $T = 10$ essentially all proposed steps were accepted. Using $T = 10$ as a starting temperature we then ran the search with the geometric cooling, tracking the progression of numbers of variables and cost, C. A few such runs established a clear pattern of wild movement above $T = 1$ involving subsets with very many variables, a tendency for the number of variables to decrease quite steeply around $T = 0.1$, and much slower (as moves are rejected) movement around plausible looking subsets in the temperature range $T = 0.05$ to $T = 0.005$, with very slow movement indeed at temperatures below this. This visibly slow movement corresponded to sequences of rejected moves of order 1000, so we took $M = 1000$ tries with no accepted move as the stopping rule.

To investigate the effectiveness of these choices and compare them with some alternatives we ran a small factorial experiment, trying five cooling schedules, each with two starting temperatures $T = 10$ and $T = 1$ and two stopping rules $M = 1000$ and $M = 500$. Three of the cooling schedules were of the continuous geometric type, with a temperature reduction after each accepted step and $\rho = 0.995$, 0.999 and 0.9995, ie our proposed schedule is compared with faster and slower ones. The fourth, originally suggested by Lundy and Mees (1986), is also continuous but cools at a variable rate using

$$T_i = \frac{T_{i-1}}{1 + \beta T_{i-1}}$$

so that the cooling rate is slower at lower temperatures. The chosen rate parameter $\beta = 0.027$ cools from $T = 10$ to $T = 0.005$ in 7400 steps, comparable with the 7600 steps for $\rho = 0.999$. The final schedule is of the more traditional type, reducing the temperature in steps. We hold the temperature fixed until 500 moves have been accepted, then reduce it by a factor of $\rho = 0.6$. This schedule takes 7500 steps in cooling from $T = 10$ to $T = 0.005$.

At each of these 20 combinations of five cooling schedules with four starting and stopping conditions, we ran 10 searches. The results were disappointing. Although 'good' subsets were identified by most searches, the best one (with $p = 6$ and $C = 0.6107$) was identified in only 106 of the 200 attempts, and none of the combinations could be judged really adequate. We then reran all the combinations with the addition of a reheating stage. After the search had stopped, the temperature was raised to 1/10 of the original starting value, ie $T = 1$ for the runs that had started at $T = 10$ and $T = 0.1$ for the runs that had started at $T = 1$, and the search run again starting from the best subset found so far and using the same stopping rule. This gave much better results. Table 2 shows the number of times out of 10 that the best subset was found (i.e. visited, not necessarily stopped at) in each case. Many of the 'failures' identified one of the other two subsets listed above, which are very close both in cost and in the variables involved to the best one. Table 2 also shows the number of times out of 10 that one or more of these three best subsets was found.

Cooling schedule	No. times best subset found				No. times one of best 3 found			
	$M = 1000$		$M = 500$		$M = 1000$		$M = 500$	
	$T = 10$	$T = 1$	$T = 10$	$T = 1$	$T = 10$	$T = 1$	$T = 10$	$T = 1$
Geometric, $\rho = 0.995$	5	5	4	1	6	8	4	2
Geometric, $\rho = 0.999$	9	7	7	5	10	10	9	9
Geometric, $\rho = 0.9995$	9	10	7	8	10	10	10	10
Variable rate	10	10	9	8	10	10	10	10
Blocks of 500	9	7	6	6	10	10	7	8

Table 2: Numbers of times out of 10 searches that the best subset and that one of the best three subsets was found, for five cooling schedules each run with two starting temperatures and two stopping rules

Two fairly clear conclusions from these results are that the fast geometric cooling with $\rho = 0.995$ is indeed too fast, and the stopping rule with $M = 500$ tends to stop the searches too soon. The other four schedules all perform well, with the variable rate schedule and the slow geometric schedule appearing to perform best. Both of these are, however, considerably slower to run than the other two. In the latter case the reason is obvious: the cooling is slower. The variable rate schedule is slow because although it cools at the same average rate as $\rho = 0.999$ it takes many more of its steps at low temperatures where the rejection rate for proposed moves is high.

For the main searches reported in Section 4.4 we elected to stick with the geometric cooling using $\rho = 0.999$, $M = 1000$ and, perhaps overinterpreting the difference between the 9 and 7 at the start of row two in Table 2, a starting temperature of $T = 10$. We included a reheating stage, starting at $T = 1$. This gives acceptable results and is nearly twice as fast as the two schedules that are arguably a little more reliable. In the case studied, using a set of 23 observations to evaluate loss, this does not matter. For the larger evaluation sets it does. Of course the other searches are carried out on slightly different problems, but one can reasonably hope that what has been learned for this case will carry over. One attraction of the geometric cooling is that it takes about 2000 steps to reduce the temperature by one order of magnitude anywhere across the scale, so if the area of interest shifts a little (because c changes for example) this should not affect the performance of the search.

4.4 Results and discussion

Table 3 gives the values of cost, $C(\gamma^\star)$, and loss $L(\gamma^\star)$, and the numbers of variables $n(\gamma^\star)$ in the best subset visited by the search for each of the 15 combinations of method of expected loss calculation $\times$ choice of c that were investigated. These costs and losses (the only difference is the cp term) are calculated using (9) applied to the appropriate one of the five specified simulated or real sets described above. Since the cost is, in each case, the lowest one found by an extensive search of possible subsets, it might be expected to be somewhat optimistic.

Expected loss		$C(\gamma^\star)$			$L(\gamma^\star)$			$n(\gamma^\star)$		
estimate based on	c	.05	.025	.0125	.05	.025	.0125	.05	.025	.0125
200, sim, training		0.58	0.38	0.25	0.23	0.16	0.07	7	9	14
1000, sim, training		0.66	0.44	0.30	0.41	0.21	0.14	5	9	13
200, sim, tuning		0.71	0.48	0.34	0.36	0.26	0.16	7	9	14
1000, sim, tuning		0.96	0.55	0.40	0.66	0.38	0.24	6	7	13
23, real, tuning		0.61	0.42	0.28	0.31	0.20	0.13	6	9	12

Table 3: Optimised cost function, $C(\gamma^\star)$, and corresponding loss, $L(\gamma^\star)$, and number of variables selected, $n(\gamma^\star)$, for five methods of estimating the expected loss and three choices of the cost per variable c. The first column gives the size of the evaluation set, whether it is simulated or real data, and which subset of the wheat data it is based on

The extent of this optimism is shown by the results in Table 4, where the costs and losses for each of the 15 chosen subsets are evaluated over the validation set, and are substantially larger than those in Table 3.

Expected loss		C_{val}			L_{val}			No. correct		
estimate based on	c	.05	.025	.0125	.05	.025	.0125	.05	.025	.0125
200, sim, training		1.53	1.23	1.21	1.18	1.01	1.03	18	19	21
1000, sim, training		1.18	1.16	1.17	0.93	0.93	1.01	19	19	20
200, sim, tuning		1.25	1.12	0.94	0.90	0.90	0.77	19	18	19
1000, sim, tuning		1.11	1.07	0.97	0.81	0.89	0.81	18	19	20
23, real, tuning		1.34	1.10	1.03	1.04	0.87	0.88	18	18	20

Table 4: Value of cost function, C_{val}, loss, L_{val}, and number of cases correctly classified out of 23 when the 15 optimised classification rules are applied to the validation set

Also shown in Table 4 are the numbers of correctly classified cases out of the 23 in the validation set. For comparison, we carried out the following 'standard' analysis. A principal components analysis was performed using the covariance matrix of the spectral data for the 71 cases in the training set, ignoring group membership. Then the first s scores were retained, for $s = 1, \ldots, 40$, and a linear discriminant analysis rule derived on the training set and applied to the tuning set. The best performance on the tuning set

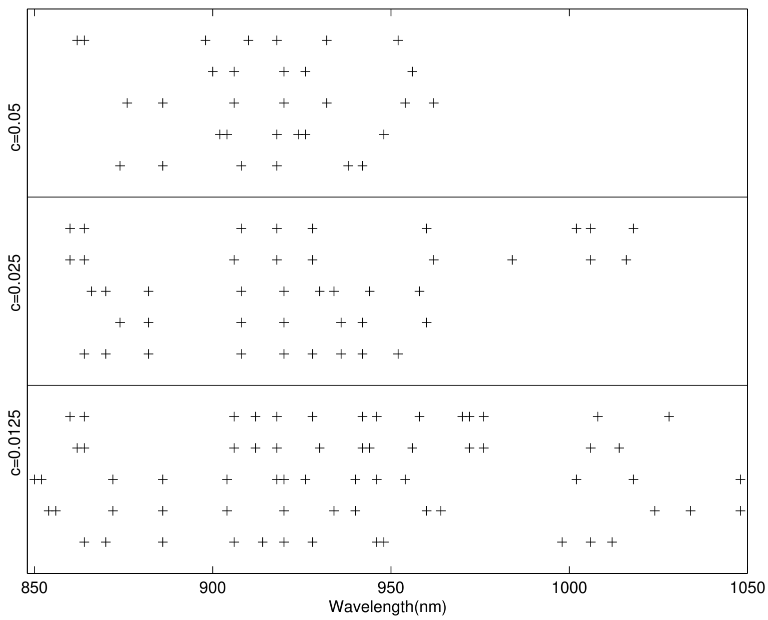

Figure 2: Wavelengths in the selected subsets. Within the blocks of five the ordering corresponds to that in Tables 2 and 3.

was 19/23 correct, achieved by any s in the range 14 to 18. All these choices of s also gave 19/23 correct on the validation set. Thus the performance of the selected subsets in the final block of Table 4 is very similar to that achieved when principal components analysis, rather than variable selection, is used for dimension reduction. The subsets that might reasonably be selected from Table 3 as 'best buys', the larger ones corresponding to $c = 0.0125$, do generally beat 19/23, although one should be cautious about over interpreting small differences.

Comparing the rows within Tables 3 and 4 it appears that the choice of method for estimating the expected loss is not critical. Specifically, using simulations of size 200 seems adequate and is a good deal cheaper than using 1000, while even the use of a small real tuning set gives comparable results.

It is interesting to examine the wavelengths included in the selected subsets. These are shown in Figure 2. The ordering of the subsets, reading down the plot, corresponds to that in Tables 3 and 4 reading down the columns. It is clear that many quite different subsets of wavelengths can be successfully used to discriminate. Two general impressions gained from the plot are that the region from 900 to 970nm is well represented in all the chosen subsets, and that the wavelengths selected are often evenly spaced, as though they are trying to summarise information over a spectral region. This suggests trying an evenly spaced selection over this key range. The subset consisting of the eight variables numbered 25(5)60, which spans the wavelength range 898 to 968nm in steps of 10nm, correctly assigns 20/23 cases in the validation set, and thus is comparable with the best of the subsets found in our search.

5 General discussion

In the example studied it seems that the discriminatory information in the spectra can be captured equally well by principal component scores derived from all 100 variables or by selection of around 10 from the 100. Our proposed approach was able to find suitable subsets. It is a possible criticism of the application that, having derived our procedure in a rather formal decision theoretic framework, we then interpret c as an arbitrary penalty and essentially choose c to get the answers (in terms of subset size) we want. However the procedure does have some attractions over and above its theoretical basis. In particular, embedding the normal model in a Bayes framework with proper prior distributions means that we can evaluate any subset, up to and including the use of all 100 variables, with no computational difficulties. It is this fact, combined with the use of a variable cost to penalise large subsets rather than explicitly specifying, say, 10 variables that enables us to let the search run 'free' in a very simple way.

The exploration of simulated annealing in Section 4.3 suggests perhaps why there are so many alternative cooling schedules and conflicting recommendations in the literature: many alternatives will work perfectly well. What probably matters most is that the search spends an adequate amount of time in the key temperature range where there is just the right balance between pressure to optimise and freedom to move away from local optima. Exactly how it cools over this range appears to be unimportant. One can get a good feel for the key range by tracking the progress of a few searches, and then tuning parameters can be set appropriately. In our example the reheating stage helped considerably. Of course it may be that other ways of allocating the same total effort would have been equally successful.

On the question of how to allocate computational effort between estimating the misclassification loss and carrying out the search, the evidence from our example is that quite modest sized evaluation sets are perfectly adequate but that economising too much on the search can result in missing good subsets. We believe it is more generally true that effort in this type of problem should be concentrated on the search process.

Many of the choices we have made in developing and applying the method could be varied quite easily. In particular the logarithmic misclassification loss function and the linear cost per variable could be replaced by alternative choices with no difficulty, and the extension to separate within-group covariance matrices would be straightforward.

Yet another option for the approximation of the expected loss via (9) would be to use cross-validation, ie to base the sum on the training set but use leave-one-out predictive distributions in calculating the classification losses. This idea is discussed in the general context of model choice by Bernardo and Smith (1994), Chapter 6. We have not investigated the approach here as it would be computationally rather expensive. Both this suggestion, and the use of the actual tuning set, highlight the parallel between the empirical methods widely used for model selection and the Bayesian approach of minimising expected predictive loss.

Acknowledgement

This work was supported by the UK Engineering and Physical Sciences Research Council. We are grateful to the Flour Milling and Baking Research Association for providing the wheat variety example.

References

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* John Wiley, Chichester.

Brown, P. J. (1993). *Measurement, Regression,and Calibration.* Oxford: Clarendon Press.

Brown, P. J., Fearn, T. and Haque, M. S. (1999a). Discrimination with many variables. *Journal of the American Statistical Association*, **94,** 1320–1329.

Brown, P. J., Fearn, T. and Vannucci, M. (1999b). The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach. *Biometrika*, **86,** 635–648.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, **68,** 265–274.

Dowsland, K. A. (1995). Simulated annealing. In *Modern Heuristic Techniques for Combinatorial Problems* (ed. C. R. Reeves), pp. 377–419. McGraw-Hill, London.

Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis. *Annals of Statistics*, **23,** 73–102.

Jolliffe, I. T. (1986). *Principal component analysis.* Springer-Verlag, New York.

Krzanowski, W. J., Jonathan, P., McCarthy, W.V. and Thomas, M. R. (1995). Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Applied Statistics*, **44,** 105–115.

Lundy, M. and Mees, A. (1986). Convergence of an annealing algorithm. *Mathematical Programming*, **34,** 111–124.

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition.* Wiley, New York.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, **66,** 783–801.

Stone, M. and Jonathan, P. (1994). Statistical thinking and technique for QSAR and related studies. Part II. Specific methods. *Journal of Chemometrics*, **8,** 1–20.