Logistic Discrimination with Many Variables

T. Fearn*, P. J. Brown† and M. S. Haque‡

Abstract

Motivated by problems in near infrared spectroscopy, we study the discrimination problem with several groups and very many predictor variables. A Bayesian version of logistic regression with a ridge-type prior distribution on the coefficients is shown to give realistic group membership probabilities in a spectroscopic example. We compare two versions of these probabilities, one using plug-in estimates of regression parameters and the other a Laplace approximation to the true predictive probabilities.

KEYWORDS: Bayes; discriminant analysis; Laplace approximation; logistic regression; near infrared spectroscopy; ridge regression

1 Introduction

The problem studied here is discriminant analysis or supervized pattern recognition. Given a training set of n cases, each with a measured $q \times 1$ vector of variables Y and a known assignment to one of g groups, the aim is to derive a rule for assigning future cases to groups on the basis of their measured Y. McLachlan (1992) provides a comprehensive review of the statistical methodology. We are particularly interested in the case where q is not small compared to n, which causes problems for most approaches. The application

*University College London, UK

†University of Kent at Canterbury, UK

‡University of Kent at Canterbury, UK

that motivated this work, and that of some other researchers, was near infrared (NIR) spectroscopy, Osborne *et al.*(1993). NIR spectrophotometers measure the transmission or reflectance of radiation at multiple wavelengths in the NIR region of the spectrum, from around 800 to 2500nm. This can result in a $q = 100$ or even 1000 point spectrum per sample. Sample here is used in the analytical chemistry sense: one sample is one case. Training sets rarely exceed a few hundred samples for reasons of cost. NIR spectra are not simple to interpret, and it is not usually possible to select a small number of relevant wavelengths on *a priori* grounds. The calibration of such instruments to derive prediction equations for the sample chemistry from the highly multivariate measurement has been extensively studied in the chemometric literature, see Stone and Jonathan (1994) for a review. NIR spectra have been used in a variety of discrimination problems, for example to discriminate between orange juice from different sources, Evans *et al.*(1993), and to check the identity of raw materials or products in the pharmaceutical industry.

Section 2 describes a Bayesian treatment of logistic discrimination for this problem. Bayesian analyses of logistic regression have recently been presented by Malec *et al.*(1996) and Kahn and Raftery (1997), but with very different applications to that considered here. In both cases the emphasis was on using hierarchical prior structures with small numbers of predictor variables and a structured sampling scheme. In Section 3 we apply our methods to an example involving the identification of wheat varieties from NIR spectra, and Section 4 contains some brief discussion.

2 Logistic Discrimination

Many of the standard approaches to discriminant analysis are based, explicitly or implicitly, on a multivariate Gaussian model for the distribution of Y conditional on group membership, so that

$$Y \sim N(\mu_j, \Sigma_j) \tag{1}$$

for cases in group j. We have explored this formulation with large q in Brown *et al.*(1999). Here however we employ the diagnostic paradigm, Dawid (1976), which involves direct modelling of the probability of group membership given Y. There is an analogy here with the so-called 'classical' and 'inverse' approaches to calibration, Brown (1982), although there is considerably more divergence between the two approaches in the discrimination context than in regression. In choosing this route we trade efficiency for robustness: the multinomial likelihood we shall use is more robust since it is valid for a range of non-normal distributions for the q responses conditional on group. The drawback is that the multinomial likelihood is less efficient if normality really does hold, see Efron (1975).

2.1 Likelihood

If we define $\tau_j(Y)$, for $j = 1, \ldots, g$, to be the probability that a case with observed vector Y belongs to group j, then the standard approach to logistic discrimination, see Anderson (1982) or McLachlan (1992), chapter 8, models

$$\ln\left(\frac{\tau_j}{\tau_g}\right) = \beta_{0j} + \beta_j^T Y \quad j = 1, \ldots, g-1 \tag{2}$$

where the gth group has been arbitrarily chosen as the reference. One motivation for this form is that for the Gaussian model in (1) with equal covariance matrices $\Sigma_j = \Sigma$ Equation (2) holds with

$$\beta_j = \Sigma^{-1}(\mu_j - \mu_g)\,. \tag{3}$$

If Y_i is the observation vector for the ith case in the training set and we define

$$\eta_{ij} = \beta_{0j} + \beta_j^T Y_i \quad i = 1, \ldots, n;\ j = 1, \ldots, g-1$$

with $\eta_{ig} = 0$ for all i, then straightforward calculations give the multinomial log-likelihood for the training set as

$$\ell = \sum_{i=1}^{n} \eta_{iG(i)} - \sum_{i=1}^{n} \ln\{\sum_{j=1}^{g} \exp(\eta_{ij})\}\,, \tag{4}$$

where $G(i)$ is the index of the correct group for case i. The relevance of this likelihood is clear when the training set is obtained by sampling cases conditionally on Y, or by sampling randomly from the mixture of groups. When the sampling is conditional on group membership various arguments have been put forward to justify the use of the same likelihood, see Farewell (1979) and McLachlan (1992). The main practical effect of the way the training set is sampled concerns the intercepts β_0. The estimate of β_{0j} will reflect the distribution of group membership in the training set. Under mixture sampling this will be relevant for prediction. Under other types of sampling it may or may not be.

2.2 Prior and posterior distributions

With q large the estimation of $\beta = (\beta_1^T, \ldots, \beta_{g-1}^T)^T$ is problematic. We overcome this difficulty by putting a prior distribution on these parameters in the spirit of the Bayesian formulation of ridge regression, see Lindley and Smith (1972).

We take diffuse prior distributions for the intercepts $\beta_{0j}, j = 1, \ldots, g-1$. The form of the prior distribution for β is motivated by (3). If in the Gaussian model

$$\mu_j \sim N_q(\mu, \sigma_\mu^2 I)$$

independently for $j = 1, \ldots, g$, and Σ is proportional to an identity matrix then the implied prior distribution for β has the form

$$\beta \sim N_{q(g-1)}(0, k^{-1}(I_{g-1} + J_{g-1}) \otimes I_q)\,, \tag{5}$$

where k is a scalar and J_n is an $n \times n$ matrix of 1's.

The log posterior density of β_0 and β given the training data and the hyperparameter k is (up to an additive constant)

$$\ell - \frac{1}{2}k \left(\sum_{j=1}^{g-1} \beta_j^T \beta_j - \frac{1}{g} \sum_{j=1}^{g-1} \sum_{l=1}^{g-1} \beta_j^T \beta_l \right)\,. \tag{6}$$

Thus the prior distribution adds to the log likelihood a quadratic penalty on the coefficients. This penalty is invariant to the choice of reference group, something that is not true for the more obvious penalty function, used by Duffy and Santner (1989) and Le Cessie and Van Houwelingen (1992), that omits the cross product terms.

2.3 Classification probabilities

To classify a new case on the basis of its observed Y we can calculate, for given β_0 and β, a set of g membership probabilities $\tau_1, \ldots, \tau_g$ using (2). In a fully Bayesian analysis these probabilities need to be averaged over the posterior distribution of β_0 and β given the training data. The density in (6) is not analytically tractable, so we need either to adopt a sampling approach, or to approximate the expectation. We have looked at two such approximations: a plug-in version, and a Laplace approximation. Khan and Raftery (1996) argue for the Laplace approximation in preference to Gibbs sampling in another application of logistic regression.

When $g = 2$ there is another approximation available, see Aitken (1978) or McLachlan (1992), p266. If a probit approximation to the logistic form of τ is combined with a normal approximation to the posterior density in (6) the expectation of τ can be found analytically. We have not been able to generalise this to the case $g > 2$ and so have not studied it further.

2.4 Plug-in estimates

The simplest approach is to maximize (6) to obtain posterior modal estimates of β_0 and β, and plug these estimates into (2) to classify new cases. This corresponds closely to the standard non-Bayesian approach of plugging in maximum likelihood estimates. We used the MATLAB quasi-Newton routine fminu to carry out the maximization for the example reported below. The derivatives of (6) with respect to β_{0j} and β_j are easily found analytically and were used in the maximization. An alternative approach to the computations would be to use a modification of the iterative scheme for the corresponding generalised linear model. The algebra for this seems straightforward, but we have not pursued this option.

2.5 Predictive probabilities

Let $\alpha_j^T = (\beta_{0j}, \beta_j^T)$, $\theta^T = (\alpha_1^T, \ldots, \alpha_{g-1}^T)$, and let $h(\theta)$ be the log posterior density given in 6. Then (Bernardo and Smith (1994), p340) the Laplace approximation to the posterior expectation of a scalar function $g(\theta)$ is

$$\hat{E}(g(\theta)) = \exp\{\frac{1}{2}\ln\det H(\hat{\theta}) - \frac{1}{2}\ln\det H^*(\theta^*) - h(\hat{\theta}) + h^*(\theta^*)\} \tag{7}$$

where $\hat{\theta}$ is the value of θ that maximises $h(\theta)$, θ^* is the value of θ that maximises $h^*(\theta) = h(\theta) + \ln(g(\theta))$, $H(\theta)$ is the matrix with ijth element

$$[H(\theta)]_{ij} = -\frac{\partial^2 h(\theta)}{\partial\theta_i \partial\theta_j},$$

and $H^*(\theta)$ is the corresponding matrix for h^*. The formula for H is

$$H(\theta) = \sum_{i=1}^{n} T_i \otimes (\tilde{Y}_i \tilde{Y}_i^T) + k(I_{g-1} - \frac{1}{g} J_{g-1}) \otimes \tilde{I}_{q+1} \tag{8}$$

where $\tilde{Y}_i^T = (1, Y_i^T)$, $\tilde{I}_{q+1}$ is the $(q+1) \times (q+1)$ matrix obtained by adding an initial row and column of zeros to I_q, and T_i is the $(g-1) \times (g-1)$ matrix with rsth element

$$[T_i]_{rs} = \delta_{rs}\tau_r(Y_i) - \tau_r(Y_i)\tau_s(Y_i).$$

When $g(\theta)$ is actually $\tau_j(Y)$, the probability that a new case with observation vector Y belongs to group j, the change from h to h^* is equivalent to adding the new case to the training data and assigning it to group j. Using this fact the computations are straightforward if rather time consuming. After maximizing h and calculating $h(\hat{\theta})$ and $H(\hat{\theta})$ we add each new case in turn to the training data, assign it to each group in turn, and repeat the maximization and the calculations of h and H. Since the τ_j must add to one over groups we could omit one of these calculations, although we have deliberately calculated all g in the example below as a check on the accuracy of the approximations and computations.

3 Example

We have applied the methodology described in Section 2 to a spectroscopic example where discrimination is quite difficult.

3.1 Wheat variety data

The data consist of NIR transmission spectra on 292 samples of wheat. The spectra were measured on a samples of unground 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 $g = 9$ named varieties, on the basis of known provenance. One of the aims of the experiment was to investigate whether the NIR spectra could be used to assign unknown samples from one of these varieties to the correct variety. Of course if such a system were to be used in practice it would need to cope with the possibility that the new sample belonged to none of the 9 groups, but we have avoided such complications. The 292 samples were split, randomly within groups, into training, tuning and validation sets containing as close to 60%, 20% and 20% as possible. The resulting numbers of samples are shown in Table

1 and graphs of the training spectra for all nine varieties are given in Figure 1. Given the overlaps, it is perhaps remarkable that one can discriminate at all. In fact there is relevant information in the shapes of the curves as well as in their overall levels so the picture is not as bad as it looks.

Variety	1	2	3	4	5	6	7	8	9	Total
Training	32	8	22	17	40	7	10	23	17	176
Tuning	10	3	7	6	14	3	3	7	5	58
Validation	10	3	7	6	14	3	3	7	5	58

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

3.2 Application of logistic methodology

With 9 groups and 100 variables the dimension of θ would be $8 \times 101 = 808$. With a proper prior for β there is no reason in principle why we should not proceed even though there are only 176 training samples. However a singular value decomposition of the 176×100 data matrix shows that 20 principal components account for all but $2 \times 10^{-6}\%$ of the variability. Thus it is possible to save a considerable amount of computation, essentially without losing any information, by reducing the data to scores on the first 20 principal components, and we have done this. Because the transformation to principal component scores using orthonormal eigenvectors is a rotation, it is appropriate to keep the same form (5) for the prior distribution for β. Since this argument ignores the fact that the transformation is data dependent it is less than wholly rigorous, but it seems good enough to justify using the same form of shrinkage in the transformed space. Even with only 20 variables, we still have $8 \times 21 = 168$ parameters for 176 cases, and several of the groups have fewer than 20 cases in the training set.

The parameter k in the prior distribution for β in (5) needs either to be specified or given a prior distribution. We have adopted the pragmatic approach of selecting a value for k that gives good performance on the tuning set of 58 samples, using a geometrically spaced series of values for k and the number of correct assignments out of 58 (using the

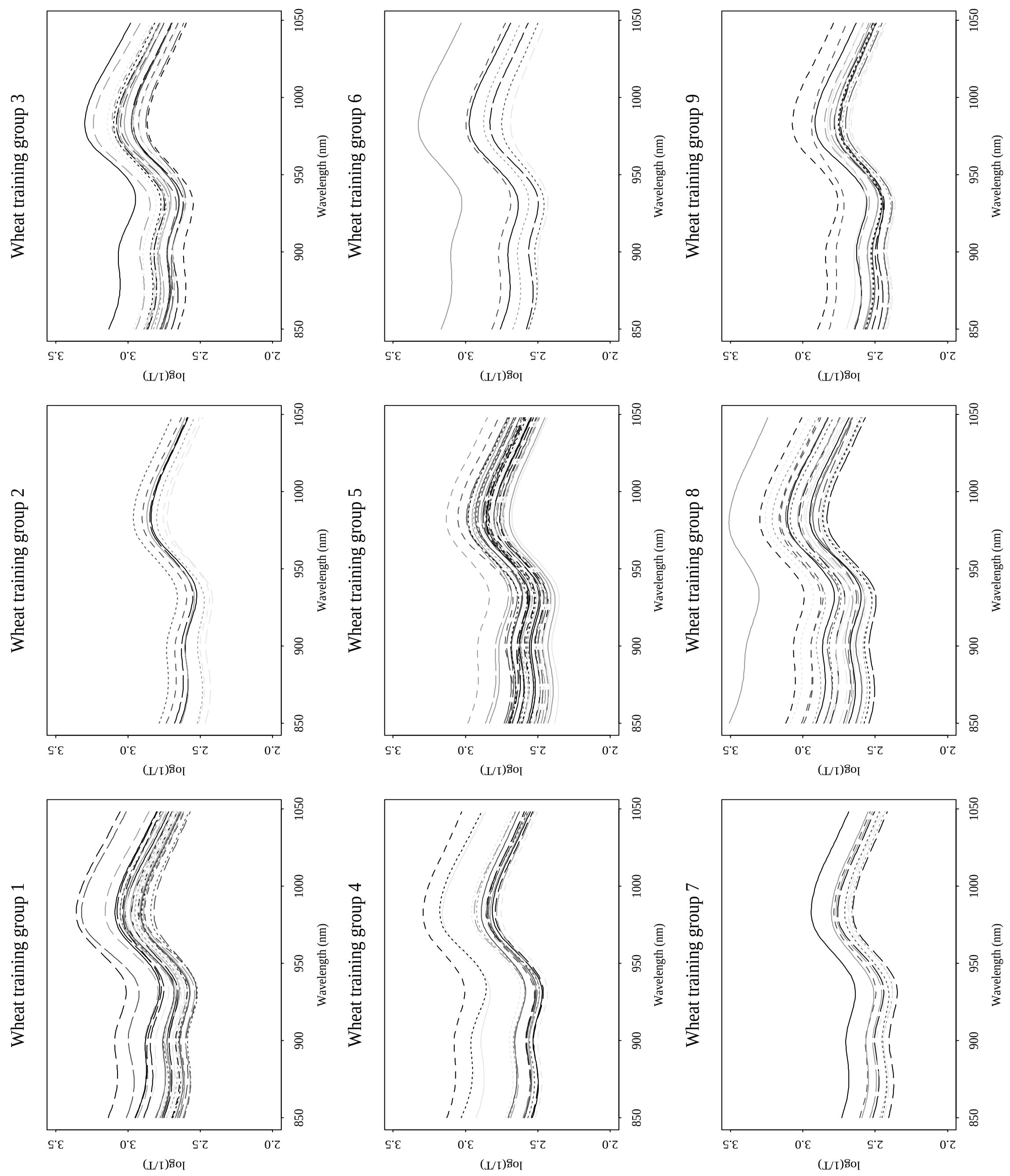

Figure 1: Transmission spectra of nine wheat varieties in the training data

plug-in classifier) as the performance criterion. Thus the overall scheme was to fit the model to the training set, use the tuning set to select a value of k, and then, having fixed the value of k, assess the performance of both plug-in and predictive versions on the validation set.

The computation of the Laplace approximation may be prone to rounding errors as it involves differencing similar sized quantities. To compute $\ln \det H$ which is the most likely source of errors in (7) we summed the logs of the eigenvalues of H as given by the MATLAB routine eig. The $g = 9$ predictive probabilities as computed typically summed to 1 ± 0.02. They were then normalised to add to 1.

3.3 Results

The chosen value of k gave 36/58 correct assignments on the tuning set; not by any means perfect discrimination but in line with what was expected for this problem. Some feel for the size of k can be obtained by comparing the two parts of $H(\theta)$ in (8). Since $H(\theta)$ is an inverse variance matrix it corresponds to the familiar $X^T X + kI$ in ordinary ridge regression. The chosen value of k is the same size as the 79th largest out of the 168 eigenvalues of the first, summed, term in (8), so there is substantial shrinkage in at least half of the dimensions involved.

Using the plug-in and the predictive classifiers on the validation samples, each with the chosen value of k, gave 38/58 and 37/58 correct respectively. Only one case is classified differently, and that is the result of very small changes in two almost equal probabilities. In fact the two approximations give quite similar probabilities throughout, and other measures of success, such as Brier score, also fail to differentiate between their performances. Figure 2 shows the two sets of probabilities plotted against each other. The predictive probabilities are noticeably less extreme than the plug-in ones, but the effect is not large.

Since there is a good deal of residual uncertainty about group membership even after conditioning on Y, it is important that the membership probabilities are well calibrated. Put another way, it is bad enough not knowing which group many of the samples belong to, but it would be even worse if we thought we did know.

Table 2 is an attempt to examine this calibration. For each method the $58 \times 9 = 522$

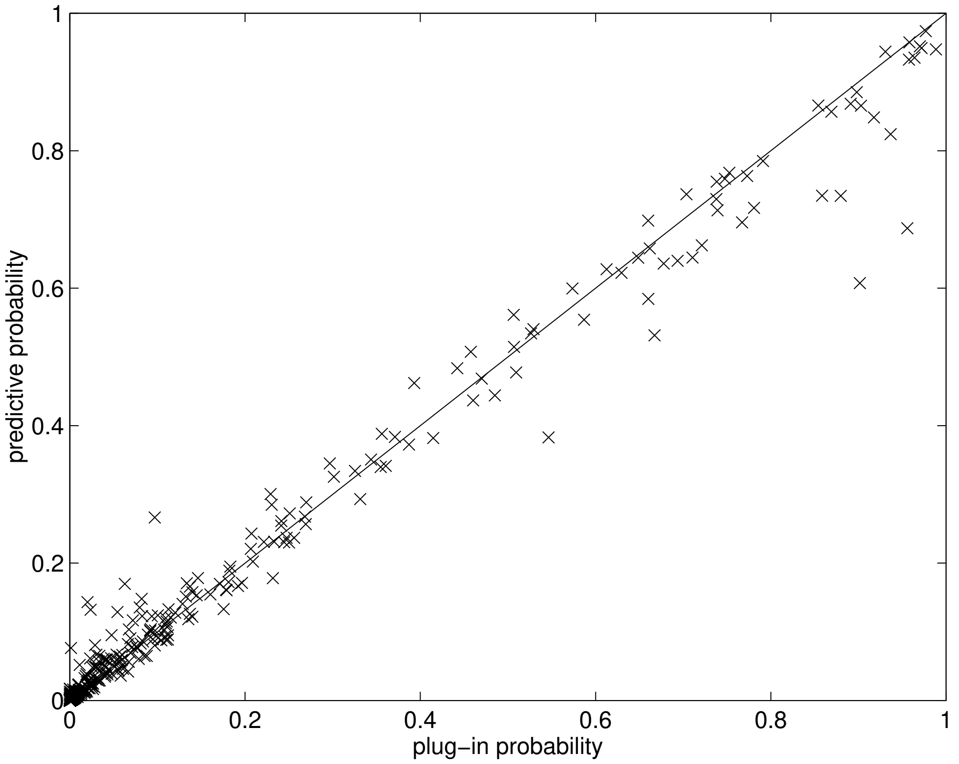

Figure 2: Comparison of plug-in and predictive group membership probabilities for validation samples

	plug-in			predictive		
τ	n	c	e	n	c	e
0.0-0.2	439	17	10.4	439	17	12.2
0.2-0.4	29	4	8.1	30	5	8.9
0.4-0.6	14	6	7.0	15	7	7.2
0.6-0.8	21	15	14.9	23	16	16.9
0.8-1.0	19	16	17.6	15	13	13.7

Table 2: Assessing the calibration of the probabilities of group membership. For each method n is the number of probabilities in the given range, c is the number and e the expected number of correct groups.

group membership probabilities have been put into bins of width 0.2. Then the number of correct groups in each bin is compared with its expected number, the sum of the probabilities. From the table, the probabilities produced by either approach do seem to be realistic. The only sign of poor calibration is that it appears that some of the smallest probabilities are rather too small, but the agreement in general is good.

4 Discussion

For this example at least the predictive probabilities are very similar to the plug-in ones. However, more examples need to be studied before any general conclusions can be reached on this front. What is very encouraging is that the group membership probabilities produced by this approach do seem to be realistic ones. Many approaches to discrimination give similar classification success rates on these data. The approach described here gives useful probabilities as well.

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 data.

References

Aitken, C. G. G. (1978) Methods of discrimination in multivariate binary data. *Compstat 1978, Proc. Computational Statistics.* Vienna: Physica-Verlag, 155–161.

Anderson, J. A. (1982). Logistic discrimination. *Handbook of Statistics* **2** (P. R. Krishnaiah and L. Kanal, eds.). Amsterdam: North- Holland, 169–191.

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

Brown, P. J. (1982). Multivariate calibration (with discussion). *J. Roy. Statist. Soc. B* **44**, 287–321.

Brown, P. J., Fearn, T, and Haque, M. S. (1999). Discrimination with many variables. *J. Amer. Statist. Assoc.*, **94**, 1320-9.

Dawid, A. P. (1976). Properties of diagnostic data distributions. *Biometrics* **32**, 647–658.

Duffy, D. E. and Santner, T. J. (1989). On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression numbers. *Comm. Statist. Theory and Methods* **18**, 959–980.

Efron, B. (1975). The efficiency of logistic regression compared to normal discrimination. *J. Amer. Statist. Assoc.* **70**, 892–898.

Evans, D. G., Scotter, C. N. G., Day, L. Z. and Hall, M. N. (1993). Determination of the authenticity of orange juice by discriminant analysis of near infrared spectra: A study of pretreatment and transformation of spectral data. *Journal of Near Infrared Spectroscopy* **1**, 33–44.

Farewell, V. T. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika* **66**, 27–32.

Kahn, M. J. and Raftery, A. E. (1996). Discharge rates of Medicare stroke patients to skilled nursing facilities: Bayesian logistic regression with unobserved heterogeneity. *J. Amer. Statist. Assoc.* **91**, 29–41.

Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Appl. Statist.* **41**, 191–201.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion).). *J. Roy. Statist. Soc. B* **34**, 1–41.

Malec, D., Sedransk, J., Moriarity, C. L. and LeClere, F. B. (1997). Small area inference for binary variables in the National Health Interview Survey. *J. Amer. Statist. Assoc.* **92**, 815–826.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* New York: Wiley.

Osborne, B. G., Fearn, T. and Hindle, P. H. (1993). *Practical NIR Spectroscopy.* Harlow: Longman.

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.