
Credit Scoring and Reject Inference With Mixture Models

A.J. Feelders

Tilburg University, The Netherlands

ABSTRACT Reject inference is the process of estimating the risk of defaulting for loan applicants that are rejected under the current acceptance policy. We propose a new reject inference method based on mixture modeling, that allows the meaningful inclusion of the rejects in the estimation process. We describe how such a model can be estimated using the EM-algorithm. An experimental study shows that inclusion of the rejects can lead to a substantial improvement of the resulting classification rule. Copyright © 2000 John Wiley & Sons, Ltd.

INTRODUCTION

Learning from non-random samples is a problem that is of crucial importance to data analysis in general, and to credit scoring in particular. In credit scoring, loan applicants are either rejected or accepted depending on characteristics of the applicant such as age, income and marital status. Repayment behaviour of the accepted applicants is observed by the creditor, usually leading to a classification as either a good or bad (defaulted) loan. As repayment behaviour of rejects is for obvious reasons not observed, complete data is available only for accepted applicants. Since the creditor does not accept applicants at random, this constitutes a non-random sample from the population of interest. Construction of a classification rule based on accepted applicants only, may therefore lead to biased estimates. In particular, one should be careful in using such a rule to assess the default risk of rejected applicants. This is, in a nutshell, what is called the reject

inference problem in the credit scoring literature.

In the next section we formulate the reject inference problem as a problem of learning with missing data. In the third section we discuss two approaches to reject inference when data are missing at random: function estimation and density estimation respectively. In the fourth section we show how density based models can be estimated from partially classified data with the EM-algorithm. Then we perform an experimental comparison of function estimation and density estimation for reject inference. Finally, we draw a number of conclusions and indicate possible directions for further research.

REJECT INFERENCE AS A MISSING DATA PROBLEM

In order to structure the following discussion, we distinguish between the *selection mechanism* that determines whether an applicant is rejected or accepted by the bank, and the *outcome mechanism* that determines the response (good or bad loan) of the applicants. We also refer to selection as the *missing-data mechanism*, since it determines for which cases the outcome is

* Correspondence to: A.J. Feelders Department of Economics and CENTER for Economic Research, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands
E-mail: A.J. Feelders@kub.nl

observed. In credit scoring, the primary objective is to model the outcome mechanism.

We assume some vector of variables $\mathbf{x} = (x_1, \dots, x_k)$ is completely observed for each applicant, and the class label y is observed for the accepted applicants, but missing for the rejected applicants. Without loss of generality we assume $y \in \{0, 1\}$, with the convention that a bad loan is labeled 0, and a good loan is labeled 1. Furthermore, we define an auxiliary variable a , with $a = 1$ if the applicant is accepted and $a = 0$ if the applicant is rejected. Note that y is observed if $a = 1$ and missing if $a = 0$. Following the classification used in Little and Rubin (1987), we distinguish between the following situations.

Missing Completely at Random

The class label is missing completely at random if acceptance is independent of both \mathbf{x} and y , i.e.

$$P(a = 1 \mid \mathbf{x}, y) = P(a = 1)$$

This situation applies when applications are accepted at random, e.g. by simply accepting all applications up to a certain number or by tossing a coin. This way of buying experience has been used to a certain extent by credit institutions, although there are obvious economic factors that constrain the use of this method (Hsia, 1978). Most credit institutions have a somewhat more sophisticated acceptance policy.

Missing at Random

The class label is missing at random (MAR) if acceptance depends on \mathbf{x} but conditional on \mathbf{x} does not depend on y , i.e.

$$P(a = 1 \mid \mathbf{x}, y) = P(a = 1 \mid \mathbf{x})$$

This situation frequently occurs in practice, since many credit institutions nowadays use a formal selection model. In that case, y is observed only if some function of variables occurring in \mathbf{x} exceeds a threshold value, say $f(\mathbf{x}_s) \geq c$, where $\mathbf{x}_s \subseteq \mathbf{x}$. Note that in case of MAR we also have

$$P(y = 1 \mid \mathbf{x}, a = 1) = P(y = 1 \mid \mathbf{x}, a = 0) = P(y = 1 \mid \mathbf{x})$$

Copyright © 2000 John Wiley & Sons, Ltd.

i.e. at any particular \mathbf{x} , the distribution of the observed y is the same as the distribution of the missing y . Later we will see that this is an important property following from the MAR assumption. The definition of MAR provides a minimal condition on which valid statistical analysis can be performed without modeling the underlying missing data mechanism.

Missing not at Random

The class label is missing not at random (MNAR) when acceptance still depends on y , even when we condition on \mathbf{x} , i.e.

$$P(a = 1 \mid \mathbf{x}, y) \neq P(a = 1 \mid \mathbf{x})$$

This typically occurs when acceptance is partly based on characteristics that are not recorded in \mathbf{x} , for example the 'general impression' that the loan officer has of the applicant. It may also occur when a formal selection model is used, but is frequently 'overruled' by a loan officer on the basis of characteristics not recorded in \mathbf{x} . If these other (unobserved) characteristics have an additional influence on y , then

$$P(y = 1 \mid \mathbf{x}, a = 1) \neq P(y = 1 \mid \mathbf{x}, a = 0)$$

i.e. at any particular \mathbf{x} , the distribution of the observed y differs from the distribution of the missing y . In the case of MNAR, valid statistical inference cannot be performed without modeling the underlying missing data mechanism.

REJECT INFERENCE UNDER MAR

Henceforth, we assume that the acceptance/rejection decision depends only on the observed attributes of the applicant, recorded in the feature vector $\mathbf{x} = (x_1, \dots, x_k)$. In other words, we assume that the class label y is missing at random, for possible approaches to the MNAR case we refer the reader to Boyes, Hoffman and Low (1989) and Greene (1992). We are interested in modeling the outcome mechanism, i.e. the dependence of the probability of a good loan on feature vector \mathbf{x} . The question we now consider is whether we

Int. J. Intell. Sys. Acc. Fin. Mgmt. **9**, 1–8 (2000)

should include the rejects in estimating the outcome mechanism, and if so how we can do this. Before we discuss this issue it is convenient to first introduce some basic notation.

The goal of a classification procedure is to predict the class label given a set of features $\mathbf{x} = (x_1, \dots, x_k)$ measured on the same object (applicant). At a particular point \mathbf{x} the value of y is not uniquely determined. It can assume both its values with respective probabilities that depend on the location of the point \mathbf{x} in the feature space. We write

$$P(y = 1 | \mathbf{x}) = 1 - P(y = 0 | \mathbf{x}) = f(\mathbf{x})$$

Here $f(\mathbf{x})$ is a single-valued deterministic function that at every point \mathbf{x} specifies the probability that $y = 1$. We assume the goal of a classification procedure is to produce an estimate $\hat{f}(\mathbf{x})$ of $f(\mathbf{x})$ at every point in the feature space.

There are two basic approaches to producing such an estimate, sometimes called *function estimation* and *density estimation* respectively (Friedman, 1997). We give a short description of the two approaches because they have quite different implications for handling reject inference (Hand and Henley, 1993).

Function Estimation

In the function estimation setting one only models the *conditional* distribution of y given \mathbf{x} . For binary classification problems we may write in general

$$p(y | \mathbf{x}) \sim B(1, f(\mathbf{x}))$$

i.e. y is a Bernoulli random variable with 'probability of success' $f(\mathbf{x})$, and variance $\sigma_y^2(\mathbf{x}) = f(\mathbf{x})(1-f(\mathbf{x}))$. The most popular technique that uses this approach is logistic regression, where

$$f(\mathbf{x}) = \Lambda\left(\alpha + \sum_{h=1}^k \beta_h x_h\right) = \left(1 + e^{-\left(\alpha + \sum_{h=1}^k \beta_h x_h\right)}\right)^{-1}$$

where $\Lambda(\cdot)$ denotes the logistic cumulative distribution function. The goal is to obtain an estimate $\hat{f}(\mathbf{x}|T)$ using some training set T .

It is important to notice that no assumptions are made concerning the probability distri-

bution of \mathbf{x} . Under the MAR assumptions, at any particular point \mathbf{x} , the distribution of the observed y is the same as the distribution of the missing y (see above). Clearly then, using a function estimation technique on just the accepted loans yields unbiased estimates of $P(y = 1 | \mathbf{x})$.

Furthermore we observe that the rejects do not provide any information concerning $P(y = 1 | \mathbf{x})$, and so it is useless to include them in the estimation process. This is quite clear if we consider the contribution of the different observations to the likelihood function. Under the usual assumption that observations are independent, the likelihood L of n observations is simply $L = \prod_{j=1}^n L_j$, with

$$L_j = \begin{cases} P(y = i | \mathbf{x}_j) & \text{if } y_j = i \ (i = 0, 1) \\ \sum_{i=0}^1 P(y = i | \mathbf{x}_j) & \text{if } y_j \text{ is missing} \end{cases}$$

Clearly, if y_j is missing it contributes a factor 1 to the likelihood leaving it unchanged. Thus including the rejects results in the same likelihood as ignoring them altogether.

Density Estimation

An alternative paradigm for estimating $f(\mathbf{x})$ in the classification setting is based on density estimation. Here Bayes' theorem

$$f(\mathbf{x}) = \frac{\pi_1 p_1(\mathbf{x})}{\pi_0 p_0(\mathbf{x}) + \pi_1 p_1(\mathbf{x})} \quad (1)$$

is applied where $p_i(\mathbf{x}) = p(\mathbf{x}|y = i)$ are the class conditional probability density functions and $\pi_i = P(y = i)$ are the unconditional ('prior') probabilities of each class. The training data are partitioned into subsets $T = \{T_0, T_1\}$ with the same class label. The data in each subset are separately used to estimate the class-conditional densities $\hat{p}_i(\mathbf{x}|T_i)$, and prior probabilities $\hat{\pi}_i$. These estimates are plugged into (1) to obtain an estimate $\hat{f}(\mathbf{x}|T)$. Examples of this approach are linear and quadratic discriminant analysis.

Now let T^A denote the training data of the accepted loans. Because the sampling fraction depends on \mathbf{x} , $\hat{p}_i(\mathbf{x}|T_i^A)$ is distorted, and if the probability of a bad loan depends (as we hope) on \mathbf{x} then $\hat{\pi}_i|T^A$ is biased as well. In general the distribution of the bads will move 'towards'

the distribution of the goods, if the selection rule is any good. In Figure 1 the distortion resulting from truncation is illustrated for a normally distributed variable. Suppose the goods have a higher mean for x than the bads, and that we reject an applicant if x is smaller than some cutoff value. In this particular example $p_1(x)$ is hardly affected by the selection mechanism since very few goods are rejected. The distribution $p_0(x)$ is severely affected, however, and has shifted towards $p_1(x)$. The selection results in an overestimate of the mean and an underestimate of the variance of x for the bad loans.

Figure 1 also suggests that one would underestimate the probability of a good loan at x , but this also depends on the estimated prior probabilities $\hat{\pi}_0$ and $\hat{\pi}_1$. Since one would tend to underestimate the prior probability of a bad loan (again, if the selection rule is any good) these estimation errors seem to work in opposite directions. We can't make any general statements however about the properties of the density based estimator, except that it tends to be biased.

There are, however, ways to avoid this bias, by including the rejected applicants into the estimation process. This idea is pursued further in the next section.

DENSITY ESTIMATION VIA MIXTURE DISTRIBUTIONS

Mixture distributions (Everitt and Hand, 1981; McLachlan and Basford, 1988) are distributions which can be expressed as 'weighted averages' of a number of component distributions.

In general, a finite mixture can be written as

$$p(x) = \sum_{i=1}^c \pi_i p_i(x; \theta_i)$$

where c is the number of components, π_i the mixing proportions and θ_i the component parameter vectors. Henceforth we assume that the number of components equals the relevant number of classes, so each component models a class-conditional distribution. All observations are assumed to be drawn from the two-component mixture

$$p(x) = \pi_0 p_0(x; \theta_0) + \pi_1 p_1(x; \theta_1)$$

where we observe the component from which an observation was drawn for the accepted loans but not for the rejected loans. We consider the contribution to the likelihood of cases with y observed (component known) and y missing (component unknown) respectively

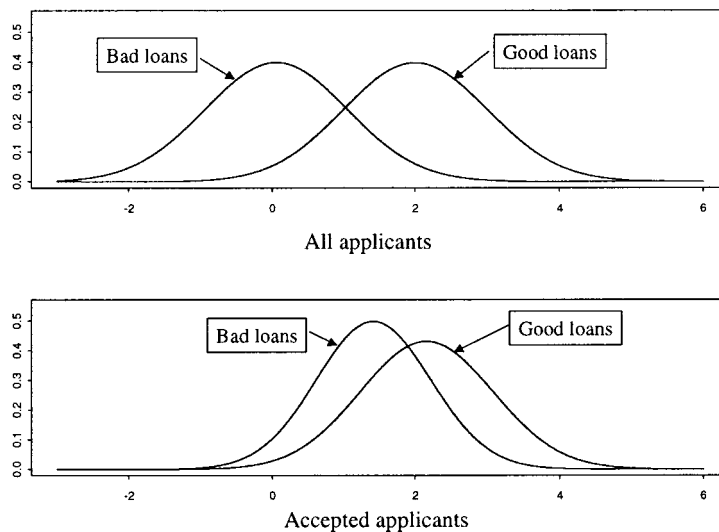


Figure 1 Distribution of x for goods and bads before and after selection.

$$L_j = \begin{cases} \pi_i p_i(\mathbf{x}_j) & \text{if } y_j = i \ (i = 0,1) \\ p(\mathbf{x}_j) = \sum_{i=0}^1 \pi_i p_i(\mathbf{x}_j) & \text{if } y_j \text{ is missing} \end{cases}$$

If there are m rejected loans and n accepted loans, the observed-data likelihood may be written

$$L_{obs}(\Psi) = \prod_{j=1}^m \left\{ \sum_{i=0}^1 \pi_i p_i(\mathbf{x}_j; \theta_i) \right\} \prod_{j=m+1}^{m+n} \left\{ \sum_{i=0}^1 z_{ij} \pi_i p_i(\mathbf{x}_j; \theta_i) \right\}$$

where $\Psi = (\pi', \theta)'$ denotes the vector of all unknown parameters, and z_{ij} equals 1 if observation j has class-label i , and zero otherwise.

For computational convenience one often considers the loglikelihood $\mathcal{L}_{obs} = \log L_{obs}$

$$\begin{aligned} \mathcal{L}_{obs}(\Psi) &= \sum_{j=1}^m \log \left\{ \sum_{i=0}^1 \pi_i p_i(\mathbf{x}_j; \theta_i) \right\} \\ &+ \sum_{j=m+1}^{m+n} \sum_{i=0}^1 z_{ij} \log(\pi_i p_i(\mathbf{x}_j; \theta_i)) \end{aligned}$$

In general this tends to be a rather complicated function of Ψ , and finding maximum likelihood estimates may require special computational algorithms. We use EM for this purpose.

The EM Algorithm

EM (Dempster, Laird and Rubin, 1977) is a general method for performing maximum likelihood estimation with incomplete data. The computational scheme consists of the alternated application of an Expectation step and a Maximization step; hence the name EM.

In the E-step, the expected value of the complete-data loglikelihood is calculated, by integrating over the possible values of the missing data under its distribution given the current parameter estimate $\theta^{(t)}$ and the observed data. In the M-step we choose the value of $\theta^{(t+1)}$ that maximizes the log-likelihood in the last E-step. It can be shown that under mild conditions the sequence $\theta^{(0)}, \theta^{(1)}, \dots$ converges to a maximum likelihood estimate of the observed data likelihood.

We illustrate the EM-algorithm with a particularly simple example that does not require EM for its solution. This allows us to discuss

the computational steps of EM without being distracted by technical detail. Consider a sequence of four independent coin tosses with the following outcome (1,1,0,?), where we use 1 to denote that heads has come up, and 0 for tails. The question mark for the fourth toss indicates that its outcome was not observed for some reason. The parameter of interest is the probability of heads, which we denote by π . We partition the complete data y into the observed part and the missing part, i.e. $y = (y_{obs}, y_{mis})$. The probability of the observed data is obtained from the probability of the complete data by summing out the missing data, i.e.

$$\begin{aligned} P(y_{obs} | \pi) &= \sum_{y_{mis}} P(y | \pi) = P((1,1,0,0) | \pi) + P((1,1,0,1) | \pi) \\ &= \pi^3(1-\pi) + \pi^2(1-\pi)^2 = \pi^2(1-\pi)\{\pi + (1-\pi)\} = \pi^2(1-\pi) \end{aligned}$$

since $\pi + (1-\pi) = 1$. As was to be expected, the observed data likelihood reduces to the likelihood obtained by ignoring the fourth toss altogether. Hence the maximum likelihood estimate is simply the fraction of heads observed, i.e. $\hat{\pi} = 2/3$.

For purposes of illustration we consider how we would arrive at this estimate using the EM computational scheme. In the E-step we form the expected complete-data loglikelihood based on the current estimate $\pi^{(t)}$,

$$\begin{aligned} Q(\pi | \pi^{(t)}) &= \pi^{(t)}(3 \ln \pi + \ln(1-\pi)) \\ &+ (1-\pi^{(t)})(2 \ln \pi + 2 \ln(1-\pi)) \end{aligned}$$

In the M-step we maximize Q with respect to π to obtain $\pi^{(t+1)}$. In this particularly simple case one may obtain a closed-form solution for the iterates: $\pi^{(t+1)} = 1/2 + 1/4 \pi^{(t)}$. Thus if we make an initial guess $\pi^{(0)} = 0.25$, we obtain the sequence 0.2500, 0.5625, 0.6406, 0.6602, 0.6650, ..., which converges to 2/3.

It is interesting to note that we may rewrite Q to get

$$Q(\pi | \pi^{(t)}) = (2 + \pi^{(t)}) \ln \pi + (1 + (1 - \pi^{(t)})) \ln(1 - \pi)$$

which shows that in this case we also may obtain the expected loglikelihood by 'filling in' the expected value of the missing data (based on the current estimate $\pi^{(t)}$) and forming the appropriate likelihood from the resulting complete data.

EM for Partially Classified Data

In this section we discuss how we can use the computational scheme of EM to find maximum likelihood estimates of Ψ in the observed-data loglikelihood \mathcal{L}_{obs} . As in the coin-tossing example, the general strategy is based on optimizing the complete-data loglikelihood

$$Q(\Psi | \Psi^{(t)}) = \sum_{j=1}^{m+n} \sum_{i=1}^c z_{ij}^{(t)} \log(\pi_i p_i(\mathbf{x}_j; \theta_i))$$

by repeated application of the E-step and M-step until convergence of the parameter estimates. In the first E-step, one uses some initial estimate $\Psi^{(0)}$, to calculate the expectation of the complete-data loglikelihood. This is done by calculating the posterior probabilities of group membership for the unclassified cases, and entering these as values of $z_{ij}^{(0)}$ in the complete-data loglikelihood. In the M-step, the algorithm chooses $\Psi^{(t+1)}$ that maximizes the complete-data loglikelihood that was formed in the last E-step. The E and M steps are alternated repeatedly until convergence. It has been shown that, under very weak conditions, this algorithm will yield a local maximum of the likelihood \mathcal{L}_{obs} of the observed data. For a more detailed and rigorous account of the application of EM to this problem, the reader is referred to (McLachlan, 1992, pp 39–43).

EXPERIMENTAL COMPARISON OF FUNCTION ESTIMATION AND DENSITY ESTIMATION

In this section we perform an experimental comparison of the function estimation and density estimation approach to reject inference in order to gain some insight as to whether and when inclusion of the rejects leads to a better classification rule. We are especially interested in the relative performance in the reject region.

In order to make a meaningful comparison we consider the following experiment. We assume the class-conditional distribution of the feature vector $\mathbf{x} = (x_1, x_2)$ is bivariate normal with $\mu_0 \neq \mu_1$ and $\Sigma_0 \neq \Sigma_1$. Under these assumptions the optimal allocation rule is a quadratic

function of \mathbf{x} , the well-known quadratic discriminant function:

$$\log \left(\frac{P(y=1)}{P(y=0)} \right) = \log \left(\frac{\pi_1}{\pi_0} \right) - \frac{1}{2} \{ \delta_1(\mathbf{x}) - \delta_0(\mathbf{x}) \} - \frac{1}{2} \left\{ \log \left(\frac{|\Sigma_1|}{|\Sigma_0|} \right) \right\}$$

where

$$\delta_i(\mathbf{x}) = (\mathbf{x} - \mu_i)' \Sigma_i^{-1} (\mathbf{x} - \mu_i)$$

is the squared Mahalanobis distance between \mathbf{x} and μ_i with respect to Σ_i ($i=0,1$). We assign \mathbf{x} to class 0 if $\log(P(y=1)/P(y=0)) \leq 0$ and to class 1 otherwise.

The appropriate specification for the logistic regression model in this case is to include all quadratic terms. For the bivariate case, we then get

$$\log \left(\frac{P(y=1)}{P(y=0)} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

A sample of applicants is drawn from the mixture

$$\pi_0 N(\mu_0, \Sigma_0) + \pi_1 N(\mu_1, \Sigma_1)$$

with $\pi_0 = \pi_1 = 0.5$.

$$\mu_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Sigma_0 = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}$$

and

$$\mu_1 = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix} \quad \Sigma_1 = \begin{pmatrix} 2.0 & 1.6 \\ 1.6 & 2.0 \end{pmatrix}$$

Next we apply a linear selection rule that rejects an applicant if $x_1 + x_2 < c$, where c is chosen in such a way that a preset fraction of applications is rejected. This selection rule represents the current acceptance policy of the bank. Clearly this rule is not optimal, but it tends to accept the good loans and reject the bad loans. Both models are then estimated on this training sample, i.e.

- (1) The mixture is estimated on both the accepted loans and the rejected loans (with unknown class label!). The starting values $\Psi^{(0)}$ for EM are simply computed from the accepted loans. Henceforth we denote this model by QDA^{ri}.

(2) The quadratic logistic regression function is estimated on the accepted loans only (it has no means of gaining any information from the unclassified rejects). This model is denoted by QLR^{ri} .

Since the quadratic discriminant model assumes more than quadratic logistic regression, it is to be expected that it is more efficient if its assumptions apply. This seems to give the mixture model an unfair advantage. In order to estimate the magnitude of this advantage, we also fit a quadratic discriminant function and quadratic logistic regression to the complete training data (i.e. including rejects with their class labels) and compute their error rates on the test set. This allows us to judge whether any observed difference between the two approaches to reject inference may be due to the natural advantage of discriminant analysis in this case.

Table 1 summarizes the results of an experiment with $n = 150$ applicants, and a rejection rate of 30%. The upper part of the table shows estimates of the error rates of the different approaches. All estimates are averages of 100 independent trials, and in each trial the error rate was estimated on a test set of 10,000 observations. The first column shows the error rate on the total feature space, the second column on the accept region, and the third column on the reject region.

The first two rows show the error rates of the different approaches to reject inference. Thus we read from the table that the quadratic logistic regression (QLR^{ri}) has an overall error rate of about 30.3% against 25.9% for quadratic discriminant analysis (QDA^{ri}). In the lower part of the table we have taken ratios of error rates

Table 1 Relative performance of quadratic discriminant analysis (QDA) and quadratic logistic regression (QLR), with $n = 150$ and 30% rejects

	Total	Accept	Reject
QDA^{ri}	0.259	0.262	0.250
QLR^{ri}	0.303	0.270	0.383
QDA	0.245	0.258	0.213
QLR	0.247	0.260	0.217
QDA^{ri}/QLR^{ri}	0.86	0.97	0.65
QDA/QLR	0.99	0.99	0.98

Table 2 Relative performance of quadratic discriminant analysis (QDA) and quadratic logistic regression (QLR), with $n = 150$ and 10% rejects

	Total	Accept	Reject
QDA^{ri}	0.249	0.250	0.233
QLR^{ri}	0.256	0.253	0.287
QDA	0.245	0.249	0.209
QLR	0.247	0.251	0.221
QDA^{ri}/QLR^{ri}	0.97	0.99	0.81
QDA/QLR	0.99	0.99	0.94

to allow meaningful comparison. Note that the overall difference between QLR^{ri} and QDA^{ri} is largely due to the horrible performance of the former in the reject region. Note also that on complete data (rows 3 and 4), QLR and QDA hardly differ, so the observed difference between rows 1 and 2 cannot be explained by the natural advantage of QDA.

In table 2 we used exactly the same datasets as in Table 1, but now with only 10% rejects. In accordance with intuition QLR^{ri} has a relatively better predictive performance now, since its sample size increases from 105 (0.7×150) to 135 (0.9×150). Its overall error rate has dropped from 30.3% to 25.6%. Performance in the reject region is still considerably worse than that of QDA^{ri} , however (28.7% against 23.3%).

To determine the effect of increasing the number of applicants, we performed the same experiment with $n = 500$. The results are summarized in Tables 3 and 4 for the 30% and 10% rejection rate respectively. Again in accordance with intuition, the disadvantage of QLR^{ri} diminishes with increasing sample size, because

Table 3 Relative performance of quadratic discriminant analysis (QDA) and quadratic logistic regression (QLR), with $n = 500$ and 30% rejects

	Total	Accept	Reject
QDA^{ri}	0.241	0.253	0.213
QLR^{ri}	0.259	0.254	0.273
QDA	0.238	0.253	0.204
QLR	0.239	0.254	0.205
QDA^{ri}/QLR^{ri}	0.93	1.00	0.78
QDA/QLR	1.00	1.00	1.00

Table 4 Relative performance of quadratic discriminant analysis (QDA) and quadratic logistic regression (QLR), with $n=500$ and 10% rejects

	Total	Accept	Reject
QDA ^{ri}	0.239	0.243	0.201
QLR ^{ri}	0.241	0.244	0.221
QDA	0.238	0.243	0.196
QLR	0.239	0.244	0.197
QDA ^{ri} /QLR ^{ri}	0.99	1.00	0.91
QDA/QLR	1.00	1.00	0.99

the variance component of error decreases, and extrapolation into the reject region becomes more reliable. Still, at 30% rejection rate the performance of QLR^{ri} is clearly worse than QDA^{ri} in the reject region (27.3% against 21.3%). At the 10% rejection rate the overall performance is already quite close, with still a slight edge for QDA^{ri} in the reject region.

SUMMARY AND CONCLUSIONS

We have discussed two different approaches to reject inference under the assumption that class labels are missing at random; one based on function estimation and the other on density estimation. Using a function estimation approach we may obtain unbiased estimates (provided the model assumptions are correct) by ignoring the rejects altogether in the estimation process. Furthermore, we cannot do better than that since the rejects contain no information at all concerning the model parameters. A popular example of the function estimation approach is logistic regression.

Alternatively one may use a density estimation approach such as linear or quadratic discriminant analysis. In this case, ignoring the rejects leads to distortion of the class-conditional densities and class prior probabilities. It is however possible to include the rejects in the estimation process by using a mixture model formulation of the problem. The parameters can then be reliably estimated with the EM-algorithm.

We have performed an experiment on artificial data to compare the predictive perform-

ance of the two approaches. From a practical viewpoint, one may prefer the function estimation approach because it allows the use of standard software and saves us the trouble of specifying a probability model for the covariates. The experiments indicate, however, that for moderate sample size the predictive performance of the density based approach is better. More elaborate experiments are required to study the behaviour when the number of components of x is more realistic, say 10 to 20, and contains discrete as well as continuous variables. As noted by Hand and Henley (1993), credit scoring problems tend to contain many discrete variables and non-normal marginal distributions. An interesting alternative might be to use the general location model (Schafer, 1997), which allows for the occurrence of discrete variables but is still based on normality for the continuous part. Further study is clearly required to judge whether mixture models are interesting for the practice of reject inference.

References

- Boyes WJ, Hoffman DL, Low SA. 1989. An econometric analysis of the bank credit scoring problem. *Journal of Econometrics* **40**: 3–14.
- Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* **39**: 1–38.
- Everitt BS, Hand DJ. 1981. *Finite Mixture Distributions*. Chapman and Hall: London.
- Friedman JH. 1997. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* **1**(1): 55–77.
- Greene WH. 1992. A statistical model for credit scoring. Working paper, Leonard N. Stern School of Business.
- Hand DJ, Henley WE. 1993. Can reject inference ever work? *IMA Journal of Mathematics Applied in Business and Industry* **5**(4): 45–55.
- Hsia DC. 1978. Credit scoring and the equal credit opportunity act. *The Hastings Law Journal* **30**: 371–448, November.
- Little RJA, Rubin DR. 1987. *Statistical Analysis with Missing Data*. Wiley: New York.
- McLachlan GJ, Basford KE. 1988. *Mixture Models, Inference and Applications to Clustering*. Marcel Dekker: New York.
- McLachlan GJ. 1992. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley: New York.
- Schafer JL. 1997. *Analysis of Incomplete Multivariate Data*. Chapman & Hall: London.