An Overview of Model Based Reject Inference for Credit Scoring

A.J. Feelders
Utrecht University
Institute for Information and Computing Sciences
PO Box 80089, 3508 TB Utrecht
The Netherlands
e-mail: ad@cs.uu.nl

Abstract

Reject inference is the process of estimating the risk of defaulting for loan applicants that are rejected under the current acceptance policy. In this survey article we show how the problem of reject inference can be viewed as one of statistical inference with incomplete data. We use a well known classification of missing data mechanisms into *ignorable* and *nonignorable* to organize the discussion of different approaches to reject inference that have been proposed in the literature.

Keywords: Credit risk management, reject inference, incomplete data, sample selection

1 Introduction

Learning from nonrandom samples is a problem that is of considerable importance to data mining in general, and to its application 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 after some time 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 nonrandom sample from the population of interest. Construction of a new decision rule based on accepted applicants only may therefore lead to incorrect results. 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.

It is fairly evident that the problem sketched here is not unique to credit scoring, but occurs in many different variations whenever some form of selection is performed and one or more additional characteristics are observed only for objects that were selected into the sample. Thus the methods discussed here are in fact relevant to such diverse problems as insurance policy acceptance, personnel selection and medical diagnosis.

We give an overview of *model based* reject inference methods that have been proposed in the literature. We do not discuss methods that require the collection of supplemental information concerning the rejected applicants. It should be clear however that if reliable information can be obtained somehow, this can lead to an improvement of the new scoring model. On the other hand, it may be impossible or very costly to obtain such information. Therefore the model based methods discussed here remain highly relevant.

In section 2 we formulate the reject inference problem as one of learning with missing data. We use the classification of missing data mechanisms into *ignorable* and *nonignorable* to organize the material. In section 3 we discuss two reject inference methods that are applicable if the missing data mechanism is ignorable: function estimation and density estimation respectively. In section 4 we discuss a a bivariate probit model that has been applied in the nonignorable case. Finally, we draw a number of conclusions and indicate possible directions for further research.

2 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 creditor, and the *outcome mechanism* that determines the response (good or bad loan) of the applicant. We also refer to selection as the *missing-data mechanism*, since it determines for which applicants the outcome is observed. In credit scoring, the primary objective is to model the outcome mechanism. The creditor is interested in using historical data to learn an updated rule that can be used to make acceptance decisions for new applicants.

We start by introducing some useful notation. We assume some vector of variables $\mathbf{x} = (x_1, \ldots, x_k)$ is completely observed for each applicant. It contains the information that is filled in on the loan application form, typically supplemented with information concerning the credit history of the applicant that is obtained from a central credit bureau.

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 [LR87], we distinguish between the cases discussed in section 2.1, 2.2, and 2.3.

2.1 Missing completely at random

The class label y is missing completely at random (MCAR) if the probability that y is observed (i.e. $a = 1$: the loan is accepted) does not depend on the value of y, nor on the value of $\mathbf{x}$, i.e.

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

This situation applies when applications are accepted at random, e.g. 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 its use [Hsi78]. Most credit institutions have a somewhat more sophisticated acceptance policy. In any case, if MCAR applies there really isn't a reject inference problem in the first place. Analysis of the accepted applicants (complete-case analysis) will give reliable results.

2.2 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.

$$P(a = 1 \mid \mathbf{x}, y) = P(a = 1 \mid \mathbf{x}). \tag{2}$$

This situation frequently occurs in practice, since many creditors nowadays use a formal selection model. In that case, y is observed only if some function g of variables occurring in $\mathbf{x}$ exceeds a threshold value, say $g(\mathbf{x}) \geq c$, where c is some constant, usually called the cut-off value.

Note that it follows from 2 that

$$P(y = 1 \mid \mathbf{x}, a = 1) = P(y = 1 \mid \mathbf{x}, a = 0) = P(y = 1 \mid \mathbf{x}), \tag{3}$$

i.e. at any fixed value $\mathbf{x}$, the distribution of the observed y is the same as the distribution of the missing y. In section 3.1 we will see that this is an important property following from the MAR assumption.

2.3 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.

$$P(a = 1 \mid \mathbf{x}, y) \neq P(a = 1 \mid \mathbf{x}). \tag{4}$$

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 sometimes overruled by a loan officer on the basis of characteristics that are not recorded in $\mathbf{x}$. If these other (unobserved) charateristics have an additional influence on y, then

$$P(y = 1 \mid \mathbf{x}, a = 1) \neq P(y = 1 \mid \mathbf{x}, a = 0), \tag{5}$$

i.e. at any particular $\mathbf{x}$, the distribution of the observed y differs from the distribution of the missing y.

2.4 Ignorable and nonignorable missing data mechanisms

The missing data mechanism is said to be *ignorable* if

1. The MAR condition applies.

2. The parameters of the missing data mechanism are unrelated to those of the outcome mechanism.

Since the second condition is almost always satisfied, we may treat MAR and ignorability as equivalent conditions for all practical purposes. The missing data mechanism is called ignorable, because there is no need to include it in the model in case we are only interested in the outcome mechanism.

If MAR does not apply, the missing data mechanism is called *nonignorable*. In that case the missing data mechanism must be included in the model to get good estimates of the parameters of the outcome mechanism.

In section 3 and 4 we discuss methods that are applicable in the ignorable and nonignorable case respectively.

3 Reject inference with ignorable missing data

In this section we assume that the acceptance/rejection decision depends only on the observed attributes of the applicant, recorded in the feature vector $\mathbf{x} = (x_1, \ldots, x_k)$. In other words, we assume that the class label y is missing at random.

We are interested in modeling the outcome mechanism, i.e. the dependence of the probability of a good loan on feature vector $\mathbf{x}$. 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 $y = 1$. 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 [Fri97]. We give a short description of the two approaches because, as noted by Hand and Henley [HH93, HH94], they have quite different implications for handling reject inference.

3.1 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

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

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

$$f(\mathbf{x}) = \Lambda(\mathbf{x}\alpha) = (1 + e^{-(\mathbf{x}\alpha)})^{-1},$$

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

It is important to notice that no assumptions are made concerning the probability distribution of $\mathbf{x}$. Under the MAR assumption, at any particular point $\mathbf{x}$, the distribution of the observed y is the same as the distribution of the missing y (see section 2.2, equation 3). Clearly then, using a function estimation technique on just the accepted loans (complete case analysis) 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

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

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. Including the rejects in an iterative reclassification procedure as proposed in [Joa93] therefore seems less appropriate in this case. If the missing data mechanism is ignorable, the rejects do not provide any information and if it is nonignorable the model discussed in section 4 is more appropriate.

3.2 Density estimation

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

$$f(\mathbf{x}) = \frac{\pi_1 p_1(\mathbf{x})}{\pi_0 p_0(\mathbf{x}) + \pi_1 p_1(\mathbf{x})} \tag{6}$$

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 partioned 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 (6) to obtain an estimate $\hat{f}(\mathbf{x}|T)$. Examples of this approach are linear and quadratic discriminant analysis [McL92].

Now let $T^A = \{T_0^A, T_1^A\}$ 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 [Ave81].

To illustrate these effects, we consider a simple example where selection is based on a single variable x. Suppose that $p_0(x)$ is $N(\mu, \sigma^2)$ and we accept all applicants with $x > b$, then

$$E[x|x > b] = \mu + \sigma\lambda(\alpha)$$

where $\alpha = (b - \mu)/\sigma$ and

$$\lambda(\alpha) = \frac{\phi(\alpha)}{1 - \Phi(\alpha)},$$

is called the inverse Mills ratio or hazard function for the distribution [Gre93]. In this expression $\phi(\cdot)$ denotes the standard normal density function, and $\Phi(\cdot)$ the standard normal cdf. For the variance of the truncated variable we get

$$\text{Var}[x|x > b] = \sigma^2(1 - \delta(\alpha))$$

where $\delta(\alpha) = \lambda(\alpha)(\lambda(\alpha) - \alpha)$.

As an illustrative example, suppose $p_0(x) = N(2, 1)$ and $p_1(x) = N(6, 1)$, $\pi_0 = \pi_1 = 1/2$ and suppose an applicant is accepted if $x > 3$ (i.e. b=3). Then $E[x_0|x_0 > 3] \approx 3.53$ and $E[x_1|x_1 > 3] \approx 6.00$. Likewise, $\text{Var}[x_0|x_0 > 3] \approx 0.2$ and $\text{Var}[x_1|x_1 > 3] \approx 0.99$. We observe that the distribution of the bads is extremely distorted by the truncation: the mean has increased from 2 to 3.53, whereas the variance has decreased from 1 to 0.2. On the other hand, the distribution of the goods is hardly affected since only a small proportion of the goods is rejected. Furthermore $\pi_0|x > 3 \approx 0.14$ and $\pi_1|x > 3 \approx 0.86$, i.e. the proportion of good loans in the population is of course overestimated if the selection mechanism is any good.

How do all these distortions influence the estimated probability of a good loan? On the basis of the true distributions we would compute the probability of a good loan at $x = 4$ to be 0.5; using a normal model but based on the mean and variance after truncation we compute a probability of 0.39. In figure 1 we show the computed probability of a good loan at different values of x on the basis of the true distribution (solid line) and the normal distribution with mean and variance after truncation (dashed line). We can see that the truncated version is way off in the reject region ($x < 3$) whereas it is reasonably close in the accept region ($x > 3$). It is however very hard to draw any general conclusions concerning the bias caused by truncation since this of course critically depends on the parameters of the true distributions as well as the selection rule.

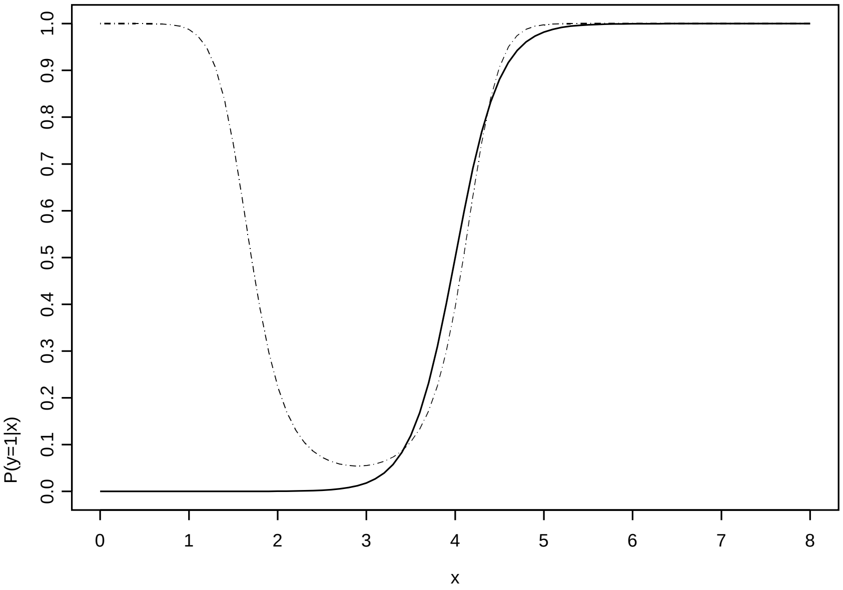

Figure 1: $P(y = 1|x)$ for true distribution (solid line) and truncated distribution (dashed line).

There are however ways to avoid this bias, by including the rejected applicants into the estimation process. A straightforward way to do this is by using a mixture distribution formulation of the problem. Mixture distributions [EH81, MB88] are distributions which can be expressed as "weighted averages" of a number of component distributions.

In general, a finite mixture can be written as

$$p(\mathbf{x}) = \sum_{i=1}^{c} \pi_i p_i(\mathbf{x}; \theta_i) \tag{7}$$

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. For the credit scoring problem, all observations are assumed to be drawn from the two-component mixture

$$p(\mathbf{x}) = \pi_0 p_0(\mathbf{x}; \theta_0) + \pi_1 p_1(\mathbf{x}; \theta_1), \tag{8}$$

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

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

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

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

where $\mathbf{\Psi} = (\pi', \theta')'$ 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 $\mathcal{L}_{obs} = \log L_{obs}$

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

In general this tends to be a rather complicated function of $\mathbf{\Psi}$, and finding maximum likelihood estimates may require special computational algorithms.

One can use the Expectation-Maximization (EM) algorithm [DLR77] for this purpsose. The general strategy is based on optimizing the complete-data loglikelihood

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

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 $\mathbf{\Psi^{(0)}}$, to calculate the expectation of the complete-data loglikelihood. For the problem under consideration, this can be 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 $\mathbf{\Psi^{(t+1)}}$ that maximizes the complete-data loglikelihood that was formed in the last E-step. In case of normal components one can find closed-form solutions for the M-step [MB88]. 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 [McL92], pages 39–43.

3.3 Function estimation or density estimation?

We have discussed two methods that are applicable in the case of ignorable missing data. Which, if any, is to be preferred?

The attractive property of the function estimation approach is that we can use a standard method of analysis, e.g. logistic regression, using just the accepted cases. The downside is that this is not fully efficient, since not all information can be used: the information on the rejected applicants is ignored.

A density based approach allows the use of information available in the rejects, but requires more complicated computational techniques. Furthermore one has to specify an appropriate probability model for the component distributions. As remarked by Hand [HH93], 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 [Sch97], which allows for the occurrence of discrete variables but is still based on normality for the continuous part. Feelders, Chang and McLachlan [FCM98] discuss a method for modelling non-normal distributions, based on modelling the *class-conditional* distributions as mixtures as well. They show how the evidently non-normal distributions of a number of financial ratios can be modelled as mixtures of two normal components. Unfortunately this resulted in only a marginal improvement in classification accuracy.

Feelders [Fee99] performs a small simulation study to compare the function estimation and density estimation approach to reject inference. The data were drawn from two normal distributions, one for each class, with different means and covariance matrices. Hence the optimal classification rule is a quadratic function. Then selection was performed by taking a linear combination of the variables, and rejecting all cases with a score below a certain cutoff level for this linear combination. The experiments indicate that for moderate sample size the predictive performance of the density based approach is better, especially in the reject region. As the sample size gets larger, the bias component of prediction error becomes dominant over the variance component. Since the correct specification was used for both models, the difference in predictive performance disappeared as the sample size increased.

4 Reject inference with nonignorable missing data

In this section we review an approach that has been proposed in the literature for cases where the missing-data (selection) mechanism is nonignorable. Recall that this situation is characterized by

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

The selection mechanism may be nonignorable when not all the relevant decision variables are recorded in the dataset, and the variables that are not recorded do have an additional (i.e. additional to the variables that *are* recorded) influence on the outcome.

The model that has been used in the literature for this case is a bivariate probit model with sample selection [BHL89, Gre98, Gre92, JR98]. It consists of two equations, one for the selection mechanism (i.e. the accept/reject decision) and one for the outcome (good/bad loan):

$$\begin{aligned} a_i^* &= \mathbf{x}_i\alpha + \varepsilon_i & (9) \\ y_i^* &= \mathbf{x}_i\beta + \upsilon_i \quad \text{for } i = 1, 2, \ldots, n & (10) \end{aligned}$$

In these equations a_i^* and y_i^* are unobserved numeric variables.

The binary variable a_i takes value 1 if the loan was accepted and 0 if the application was rejected:

$$a_i = \begin{cases} 0 \text{ if loan rejected } (a_i^* < 0) \\ 1 \text{ if loan accepted } (a_i^* \geq 0) \end{cases}$$

Likewise, the binary variable y_i takes the value 0 if the loan is classified as bad, and the value 1 otherwise:

$$y_i = \begin{cases} 0 \text{ if bad loan } (y_i^* < 0) \\ 1 \text{ if good loan } (y_i^* \geq 0) \end{cases}$$

Furthermore, y_i is only observed if $a_i = 1$.

The disturbances are assumed to be bivariate normally distributed

$$\begin{pmatrix} \varepsilon_i \\ \upsilon_i \end{pmatrix} \sim N(\mu, \Sigma) \qquad \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

According to this model, there are three types of observations, rejected loans, accepted bad loans and accepted good loans with respective probabilities:

$$\begin{aligned} a = 0 &: P(a = 0) = 1 - \Phi(\mathbf{x}\alpha) \\ a = 1, y = 0 &: P(a = 1, y = 0) = \Phi(\mathbf{x}\alpha) - \Phi_2(\mathbf{x}\alpha, \mathbf{x}\beta; \rho) \\ a = 1, y = 1 &: P(a = 1, y = 1) = \Phi_2(\mathbf{x}\alpha, \mathbf{x}\beta; \rho) & (11) \end{aligned}$$

where $\Phi(\cdot)$ represents the univariate standard normal cdf and $\Phi_2(\cdot, \cdot; \rho)$ the bivariate standard normal cdf with correlation coefficient ρ.

The appropriate log-likelihood function is readily derived from (11):

$$\begin{aligned} \ln L(\alpha, \beta, \rho) &= \sum_{i=1}^{n} \{(1 - a_i) \ln(1 - \Phi(\mathbf{x}_i\alpha)) \\ &+ a_i(1 - y_i) \ln(\Phi(\mathbf{x}_i\alpha) - \Phi_2(\mathbf{x}_i\alpha, \mathbf{x}_i\beta; \rho)) \\ &+ a_i y_i \ln \Phi_2(\mathbf{x}_i\alpha, \mathbf{x}_i\beta; \rho)\} \end{aligned}$$

This model is identified except for some pathological cases [MS85], and can be estimated with maximum likelihood.

The correlation coefficient of the disturbances provides the link between the two equations. If this is the correct specification (a big if), and $\rho = 0$ then we are back to

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

and MAR applies after all. On the other hand, if $\rho > 0$ then

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

i.e. at a fixed point $\mathbf{x}$, the probability of a good loan is *higher* among the accepts than among the rejects. This is what you would expect when the decision of the model is overruled by loan officers for "good reasons" that are however not recorded in $\mathbf{x}$. Finally if $\rho < 0$, then

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

i.e. at a fixed point $\mathbf{x}$, the probability of a good loan is *lower* among the accepts than among the rejects.

Somewhat surprisingly perhaps, Jacobson and Roszbach [JR98] Boyes et al. [BHL89] and Greene [Gre98, Gre92] found *negative* values for ρ of -0.9234, -0.353 and -0.1178 respectively [1]. Jacobson and Roszbach conclude that the bank involved does not appear to be minimizing the default risk. This follows not only from the negative correlation found, but also from the fact that many of the variables that make the bank approve loans are not among those that reduce the probability of default (i.e. are significant in the selection equation but not in the default equation, or appear with opposite signs). Boyes et al. [BHL89] make similar observations, and explain these findings by the hypothesis that the bank follows a lending policy where they pick out loans with higher default risk because they have higher returns due to the size of the loan. The findings of Jacobson and Roszbach contradict this hypothesis because they find that the size of the loan does not affect default risk. They come to the conclusion that the results bear evidence of a lending institution that has *attempted* to minimize default risk or maximize a simple return function, but without success.

[1] The correlations reported in [BHL89, Gre98, Gre92] are in fact positive, but the the value of y was defined the other way around

The ultimate question from a practical viewpoint is whether modeling the selection mechanism leads to a better default equation in terms of predictive accuracy. Unfortunately this question is hard to answer with real credit data because the true class label of the rejected applicants is unknown. Neither of the studies [JR98, BHL89, Gre98] attempted to answer this question. Banasik et al. [BCT02] report on a study that did compare the predictive performance of the single equation model and the bivariate probit model. They conclude that the adoption of a bivariate probit model only marginally improves predictive performance. This observation is confirmed by [AM02], but there clearly is scope for further empirical study in this direction.

5 Summary and conclusions

We have given an overview of model based methods for reject inference in credit scoring, based on the distinction between ignorable and nonignorable selection mechanisms. If the selection mechanism is ignorable, we can use function estimation (e.g. logistic regression) on accepted loans only and obtain unbiased estimates (provided of course the model assumptions are correct). Furthermore, we cannot do better than that since the rejects contain no information at all concerning the model parameters in this case. Therefore function estimation is not fully efficient.

Alternatively one may use a density estimation approach such as linear or quadratic discriminant analysis. In that 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.

If the selection mechanism is nonignorable, we have to include it in our model to obtain valid inferences. Several authors have proposed to use a bivariate probit model with sample selection in this case. Preliminary studies show however that the gains of this approach in terms of predictive performance are only marginal. Furthermore, the results tend to be rather sensitive to departures from the normality assumption.

The foregoing observations give rise to a number of interesting directions for further research. In case the selection mechanism is ignorable, the mixture modelling approach is more efficient than the function estimation approach. On the other hand, the normality assumption of the standard discriminant analysis model is not very realistic for credit scoring, where usually both numeric and categorical variables occur. Therefore, it would be interesting to look at different component distributions in the mixture model, that allow for a combination of numeric and categorical variables, and for non-normal numeric data. We have suggested the general location model as an interesting alternative.

For the nonignorable case the bivarate probit model with sample selection has been applied in a number of studies. Because of the sensitivity of this model to departures from normality, it would be interesting to investigate semi-parametric alternatives in order to relax the normality assumption. As a general strategy, it is advisable to compare several plausible nonignorable mechanisms and analyse how much the conclusions differ between them, and how much they differ from those obtained by ignoring the missing data mechanism altogether.

From a practical viewpoint, it might be preferrable to avoid nonignorable selection mechanisms. Usually the creditor knows the rule that was used to accept credit in the past, and in the case of overrules it might be worth the effort to find out the reasons for overruling and recording them in the dataset. In any case, ignorability is a question of degree, and the more variables we include that are predictive for acceptance, the closer we get to ignorability.

Our final conclusion is that there is still much scope for further research in this area, that might benefit not only credit risk modeling, but many other data mining problems that involve similar sample selection mechanisms.

References

[AM02] D. Ash and S. Meester. Best practices in reject inferencing. Presentation at conference credit risk modeling and decisioning, Wharton FIC, University of Pennsylvania, http://fic.wharton.upenn.edu/fic/creditrisk.html, 2002.

[Ave81] R.B. Avery. Credit scoring models with discriminant analysis and truncated samples. Research Papers in Banking and Financial Economics 54, Board of Governors of the Federal Reserve System, 1981.

[BCT02] J. Banasik, J. Crook, and L. Thomas. Sample selection bias in credit scoring models. Presentation at conference credit risk modeling and decisioning, Wharton FIC, University of Pennsylvania, http://fic.wharton.upenn.edu/fic/creditrisk.html, 2002.

[BHL89] W.J. Boyes, D.L. Hoffman, and S.A. Low. An econometric analysis of the bank credit scoring problem. *Journal of Econometrics*, 40:3–14, 1989.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[EH81] B.S. Everitt and D.J. Hand. *Finite mixture distributions.* Chapman and Hall, London, 1981.

[FCM98] A.J. Feelders, S. Chang, and G.J. McLachlan. Mining in the presence of selectivity bias and its application to reject inference. In R. Agrawal, P. Stolorz, and G. Piatetsky Shapiro, editors, *Proceedings of the fourth international conference on knowledge discovery and data mining (KDD-98)*, pages 199–203. AAAI Press, 1998.

[Fee99] A.J. Feelders. Credit scoring and reject inference with mixture models. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 8:271–279, 1999.

[Fri97] J.H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.

[Gre92] W.H. Greene. A statistical model for credit scoring. Working Paper EC-92-29, Leonard N. Stern School of Business, 1992.

[Gre93] W.H. Greene. *Econometric Analysis (second edition).* Macmillan, New York, 1993.

[Gre98] W.H. Greene. Sample selection in credit-scoring models. *Japan and the World Economy*, 10:299–316, 1998.

[HH93] D.J. Hand and W.E. Henley. Can reject inference ever work? *IMA Journal of Mathematics Applied in Business and Industry*, 5(4):45–55, 1993.

[HH94] D.J. Hand and W.E. Henley. Inference about rejected cases in discriminant analysis. In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, editors, *New approaches in classification and data analysis*, pages 292–299. Springer, New York, 1994.

[Hsi78] D.C. Hsia. Credit scoring and the equal credit opportunity act. *The Hastings law journal*, 30:371–448, November 1978.

[Joa93] D.N. Joanes. Reject inference applied to logistic regression for credit scoring. *IMA Journal of Mathematics Applied in Business and Industry*, 5(4):35–43, 1993.

[JR98] T. Jacobson and K. Roszbach. Bank lending policy, credit scoring and value at risk. SSE/EFI Working Paper Series in Economics and Finance 260, Stockholm School of Economics, 1998.

[LR87] R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data.* John Wiley & Sons, New York, 1987.

[MB88] G. J. McLachlan and K. E. Basford. *Mixture models, inference and applications to clustering.* Marcel Dekker, New York, 1988.

[McL92] G. J. McLachlan. *Discriminant analysis and statistical pattern recognition.* Wiley, New York, 1992.

[MS85] C.L. Meng and P. Schmidt. On the cost of partial observability in the bivariate probit model. *International Economic Review*, 26(1):71–85, 1985.

[Sch97] J.L. Schafer. *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London, 1997.