The Korean Journal of Applied Statistics (2011)
24*(2), 425–433*

DOI: 10.5351/KJAS.2011.24.2.425

Reject Inference of Incomplete Data Using a Normal Mixture Model

Juwon Song[1]

[1]Department of Statistics, Korea University

(Received February 2011; accepted March 2011)

Abstract

Reject inference in credit scoring is a statistical approach to adjust for nonrandom sample bias due to rejected applicants. Function estimation approaches are based on the assumption that rejected applicants are not necessary to be included in the estimation, when the missing data mechanism is missing at random. On the other hand, the density estimation approach by using mixture models indicates that reject inference should include rejected applicants in the model. When mixture models are chosen for reject inference, it is often assumed that data follow a normal distribution. If data include missing values, an application of the normal mixture model to fully observed cases may cause another sample bias due to missing values. We extend reject inference by a multivariate normal mixture model to handle incomplete characteristic variables. A simulation study shows that inclusion of incomplete characteristic variables outperforms the function estimation approaches.

Keywords: Reject inference, mixture models, incomplete data, EM algorithm.

1. Introduction

When a customer applies for a loan, a financial institution makes a decision to accept or reject it. They should accept a loan for applicants who can pay back the loan on time and reject it for applicants who cannot. The purpose of the credit scoring system is to correctly classify all loan applicants as “good” or “bad” loaners based on the applicants’ characteristics. To develop a credit scoring system, it is necessary to have credit status from all applicants. Once the loan is approved, the financial institution has a chance to observe the applicant’s credit status. However, for applicants who were rejected for the loan, the true credit status remains unknown. Since the credit status of rejected applicants is unknown, information about rejected applicants remains as incomplete.

It is known that inference based on nonrandom sample can provide biased estimates for the population parameters (Hand, 1998; Jacobson and Roszbach, 2000). Since the characteristics of accepted applicants would be different to the characteristics of rejected applicants, accepted applicants are

This research was supported by Korea University Grant No. K0822181.

[1]Associate Professor, Department of Statistics, Korea University, Seoul 136-701, Korea.
E-mail: jsong@korea.ac.kr

not a random sample from the whole applicant population. Reject inference in credit scoring is a statistical approach to adjust for the bias of the nonrandom sample that does not have information from rejected applicants.

Reject inference has been handled by various approaches such as the augmentation method (Hsai 1978), the linear discriminant analysis (Hand and Henley, 1994), the probit model (Boyes *et al.*, 1989), the logistic model (Joans, 1993), and statistical models with parameters reflecting nonrandom sample bias (Copas and Li, 1997). Hand and Henley (1997) provides general review of statistical classification methods. Choi (2008) applies the normal mixture model in semi-supervised learning to handle reject inference.

Feelders (1999) considers reject inference as a missing data problem. Since credit statuses of rejected applicants are not observed, their unobserved credit status can be considered as missing, and analysis techniques for missing data are applied. He compares two estimation approaches, function estimation and density estimation (Friedman, 1997). In the function estimation, the conditional distribution of the credit status given characteristics is considered. If credit status is measured as "good" or "bad," a popular choice is the logistic regression model for the function estimation approach. The density estimation models a mixture of the probability densities of the characteristics for each credit status. A popular choice of the model is a mixture of the multivariate normal distributions for density estimation. In the simulation, Feelders (1999) shows that function estimation based on the logistic regression leads to higher misclassification rates compared to density estimation based on normal mixtures.

While most researches in reject inferences consider completely observed characteristic variables, the density estimation approach can be extended to handle missing characteristic variables. In that case, data contain two types of missing values: (1) credit statuses are missing for rejected participants, and (2) characteristic variables may be missing for some applicants. In function estimation, the number of complete observations would be reduced by either rejected applicants or incomplete characteristics of accepted applicants. On the other hand, density estimation should directly handle missing values to model a mixture distribution of "good" and "bad" credit status. In reject inference, group membership of the mixture distribution is missing for rejected applicants, while it is observed for accepted applicants. Moreover, if characteristic variables are missing for some accepted applicants, they should be appropriately included in the estimation.

Hunt and Jorgensen (2003) suggest a mixture model for data with missing characteristic variables. In their setting, group memberships are unknown for all observations which are subject to missing. On the other hand, in reject inference, group memberships are known for accepted applicants and unknown for rejected applicants. In this article, we extend reject inference to include missing characteristics of the normal mixtures by extending the approach by Hunt and Jorgensen (2003).

A simulation study was designed to evaluate the performance of the suggested reject inference approach using a normal mixture model when data include missing characteristics. The misclassification rates using mixture distributions were compared with the ones based on function estimation. Both results were compared with the misclassification rates when data do not include any missing values in characteristic variables. The simulation indicated that density estimation based on mixture models outperformed the function estimation using the logistic model, when characteristic variables contain missing values in characteristic variables.

In Section 2, we propose the parameter estimation method in reject inference using a normal mixture, when characteristic variables include missing values. In Section 3, a simulation study is conducted to examine the performance of the suggested method. In Section 4, the suggested approach is applied to a German credit data. Section 5 contains discussion and consideration of future research directions. Through this article, it is assumed that missing data mechanism is missing at random(MAR) (Little and Rubin, 2002) for credit status and missing completely at random(MCAR) or MAR for characteristic variables.

2. Reject Inference of Incomplete Data Using a Normal Mixture Models

Let's denote the credit status by Y and the characteristic variables by $X = (x_1, x_2, \ldots, x_p)$. Y is expressed as 0 for the "bad" credit and 1 for the "good" credit status. It is assumed that observations are independent each other and a vector of the characteristic variables, X, follows a multivariate normal distribution.

2.1. Reject inference of complete data

Suppose that among n applicants, m cases are "good" loaners and n - m cases are "bad" loaners. If all applicant's credit status is known, the likelihood function can be expressed as

$$L(\theta) = \prod_{i=1}^{m} \{\pi_1 \cdot \phi(x_i | \mu_1, \Sigma_1)\} \prod_{i=m+1}^{n} \{\pi_0 \cdot \phi(x_i | \mu_0, \Sigma_0)\},$$

where $\theta = (\pi, \mu_0, \Sigma_0, \mu_1, \Sigma_1)$, π_1 is the probability that an applicant has a "good" credit status, $\pi_0 = 1 - \pi_1$, x_i indicates the vector of the characteristic variables for the applicant i, μ_g and Σ_g are the mean vector and variance-covariance matrix of characteristic variables for group g, when $g = 0$ for "bad" credit and 1 for "good" credit status, and

$$\phi(x_i | \mu_g, \Sigma_g) = \frac{1}{\sqrt{2\pi}} |\Sigma_g|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_i - \mu_g)' \Sigma_g^{-1} (x_i - \mu_g)\right\}.$$

In reject inference, credit statuses are unknown for rejected participants and it is not possible to classify them as either "good" or "bad" credit group. Instead of finding parameters directly maximizing this likelihood, the likelihood can be re-expressed as

$$L(\theta) = \prod_{i=1}^{n} \prod_{g=0}^{1} \{z_{ig} \cdot \pi_j \cdot \phi(x_i | \mu_g, \Sigma_g)\},$$

where $z_{ig} = 0$ if $y_i \neq g$ and 1 if $y_i = g$.

Since the value of z_{ig} is unknown, the EM algorithm can be applied to maximize this likelihood and find the parameter θ as follows (Feelders, 1999):

E-step:

$$E(z_{ig} | x, \mu_g, \Sigma_g) = \begin{cases} 1, & \text{if } \ y_i = g, \\ 0, & \text{if } \ y_i \neq g, \\ \dfrac{\pi_g \phi(x_i | \mu_g, \Sigma_g)}{\sum_{g=0}^{1} \pi_g \phi(x_i | \mu_g, \Sigma_g)}, & \text{if } \ y_i \text{ is missing.} \end{cases}$$

M-step: For $g = 0, 1$, and $j, j' = 1, \ldots, p$,

$$\begin{aligned}
\pi_1 &= \sum_{i=1}^{n} \frac{z_{ig}}{n}, \\
\mu_{gj} &= \frac{E\left(z_{ig} x_{ij} \left| x_i, \mu_g, \Sigma_g \right.\right)}{\sum\limits_{j=1}^{n} z_{ij}}, \\
\Sigma_{gjj'} &= \frac{E\left(z_{ig} x_{ij} x_{ij'} \left| x_i, \mu_g, \Sigma_g \right.\right)}{\sum\limits_{i=1}^{n} z_{ig}} - \mu_{gj}\mu_{gj'}.
\end{aligned}$$

2.2. Reject inference of incomplete data

When the characteristic variables include missing values, observations are reorganized by missing data patterns (Little and Rubin, 2002) and let's assume that there exist S different missing data patterns. If applicant's credit statuses are all known, the observed data likelihood function can be expressed as

$$L(\theta) = \prod_{s=1}^{S} \left[\prod_{i=1}^{m} \prod_{i \in I(s)} \left\{ \pi_1 \cdot \phi\left(x_{obs,i} \left| \mu_1^{(s)}, \Sigma_1^{(s)} \right. \right) \right\} \prod_{i=m+1}^{n} \prod_{i \in I(s)} \left\{ \pi_0 \cdot \phi\left(x_{obs,i} \left| \mu_0^{(s)}, \Sigma_0^{(s)} \right. \right) \right\} \right],$$

where $\theta = (\pi, \mu_0, \Sigma_0, \mu_1, \Sigma_1)$, π_1 is the probability that an applicant has a "good" credit status, $\pi_0 = 1 - \pi_1$, $x_{obs,i}$ indicates the vector of the observed variables for the applicant i, $i \in I(s)$ indicates that observation i belongs to the missing data pattern s, $\mu_g^{(s)}$ and $\Sigma_g^{(s)}$ are the mean vector and variance-covariance matrix corresponding to variables that are observed in pattern s for group g, and $\phi(x_{obs,i}|\mu_g^{(s)}, \Sigma_g^{(s)}) = 1/\sqrt{2\pi}|\Sigma_g^{(s)}|^{-1/2} \exp\{-1/2(x_{obs,i} - \mu_g^{(s)})' {\Sigma_g^{(s)}}^{-1} (x_{obs,i} - \mu_g^{(s)})\}$.

In reject inference, credit statuses are unknown for rejected participants, and it is not possible to classify them as either the "good" credit group or the "bad" one. Moreover, characteristic variables are not fully observed, making this likelihood hard to maximize.

If it is assumed that missing values do not exist in both credit status and characteristic variables, the complete-data likelihood is expressed as

$$L(\theta) = \prod_{i=1}^{n} \prod_{g=0}^{1} \left\{ z_{ig} \cdot \pi_j \cdot \phi\left(x_i \left| \mu_g, \Sigma_g \right. \right) \right\},$$

where $z_{ig} = 0$ if $y_i \neq g$ and 1 if $y_i = g$.

Since the value of z_{ig} as well as $x_{mis,i}$ are unknown, the EM algorithm can be applied to maximize the complete-data likelihood to estimate the parameters θ :

E-step:

$$E\left(z_{ig} \left| x_{obs,i}, \mu_g, \Sigma_g \right.\right) = \begin{cases} 1, & \text{if } \ y_i = g, \\ 0, & \text{if } \ y_i \neq g, \\ \dfrac{\pi_g \phi\left(x_{obs,i} \left| \mu_g, \Sigma_g \right.\right)}{\sum\limits_{g=0}^{1} \pi_g \phi\left(x_{obs,i} \left| \mu_g, \Sigma_g \right.\right)}, & \text{if } \ y_i \text{ is missing.} \end{cases}$$

For $i = 1, 2, \ldots, n,\ g = 0, 1$ and $j = 1, \ldots, p$,

$$E\left(z_{ig}x_{ij}\left|x_{obs,i}, \mu_g, \Sigma_g\right.\right) = \begin{cases} z_{ig}x_{ij}, & \text{if } j \in O(s), \\ z_{ig}x^*_{ij}, & \text{if } j \in M(s), \end{cases}$$

where $x^*_{ij} = E(x_{ij}\left|x_{obs,i}, \mu_g, \Sigma_g\right.)$, $O(s)$ indicates a set of observed characteristic variables for applicant i, and $M(s)$ indicates a set of missing characteristic variables for applicant i.

For $i = 1, 2, \ldots, n,\ g = 0, 1$ and $j, j' = 1, \ldots, p$,

$$E\left(z_{ig}x_{ij}x_{ijj'}\left|x_{obs,i}, \mu_g, \Sigma_g\right.\right) = \begin{cases} x_{ij}x_{ij'}, & \text{if } j, j' \in O(s), \\ z_{ig}x^*_{ij}x_{ij'}, & \text{if } j \in M(s) \text{ and } j' \in O(s), \\ a_{jj'} + z_{ig}x^*_{ij}x^*_{ijj'}, & \text{if } j, j' \in M(s), \end{cases}$$

where $a_{jj'} = \mathrm{Cov}(x_{ij}x_{ij'}\left|x_{obs,i}, \mu_g, \Sigma_g\right.)$.

<u>M-step</u>: For $g = 0, 1$, and $j, j' = 1, \ldots, p$,

$$\pi_1 = \sum_{i=1}^{n} \frac{z_{ig}}{n},$$

$$\mu_{gj} = \frac{E\left(z_{ig}x_{ij}\left|x_{obs,i}, \mu_g, \Sigma_g\right.\right)}{\sum\limits_{j=1}^{n} z_{ij}},$$

where μ_{gj} is the j^{th} element of μ_g.

$$\Sigma_{gjj'} = \frac{E\left(z_{ig}x_{ij}x_{ij'}\left|x_{obs,i}, \mu_g, \Sigma_g\right.\right)}{\sum\limits_{i=1}^{n} z_{ig}} - \mu_{gj}\mu_{gj'},$$

where $\Sigma_{gjj'}$ is the $(j,\, j')^{th}$ element of Σ_g.

E-step and M-step can be easily implemented by using Sweep operator (Schafer, 1997).

3. Simulation

To evaluate the performance of the suggested estimation method, we conducted a simulation. It was assumed that the credit status Y was measured as "good" or "bad", and two characteristic variables, $X = (x_1, x_2)$, were continuously measured. As done by Feelders (1999), n hypothetical applicants were generated from a mixture distribution,

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

where $\pi_0 = \pi_1 = 0.5$, $\mu_0 = \binom{0}{0}$, $\mu_1 = \binom{1.5}{1.5}$, $\Sigma_0 = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}$ and $\Sigma_1 = \begin{pmatrix} 2.0 & 1.6 \\ 1.6 & 2.0 \end{pmatrix}$.

Suppose that applicants were rejected if x_1+ x_2< c, where c was adjusted to attain the chosen rejection rates. We considered $n = 150$ and 300 for the training data set, and the misclassification rates were calculated using 10,000 test samples. Rejection rates were considered as either 10% or 30%. Missing values of characteristic variables were assumed to follow three missing data mechanisms. In the first missing data mechanism, MCAR, a chosen percentage of randomly selected x_2 values were considered as missing. In the second missing data mechanism, MAR1, the chosen percentage of x_2 values corresponding to the largest x_1 values were considered as missing. In the third missing data

Table 3.1. The misclassification rate with $n = 150$

Missing data	% of	% of Rejects					
mechanism	missing	10%			30%		
for x_2	in x_2	All Data	EM	LR	All Data	EM	LR
MCAR	0%	0.244	0.248	0.256	0.243	0.262	0.298
	10%	0.243	0.249	0.272	0.244	0.262	0.313
	20%	0.244	0.252	0.289	0.244	0.263	0.335
	40%	0.243	0.257	0.326	0.244	0.275	0.374
MAR1	0%	0.244	0.246	0.257	0.244	0.253	0.297
	10%	0.244	0.250	0.260	0.244	0.258	0.299
	20%	0.244	0.256	0.275	0.244	0.273	0.307
	40%	0.244	0.281	0.329	0.244	0.315	0.342
MAR2	0%	0.244	0.246	0.257	0.244	0.257	0.300
	10%	0.244	0.247	0.258	0.244	0.256	0.301
	20%	0.244	0.250	0.264	0.244	0.257	0.304
	40%	0.244	0.246	0.257	0.244	0.257	0.300

mechanism, MAR2, the chosen percentage of x_2 values corresponding to the smallest and largest x_1's were considered as missing. In both MAR1 and MAR2, since the missingness of x_2 values depends on completely observed x_1 values, missing data mechanism is missing at random (Little and Rubin, 2002).

Based on each generated training data set, the parameters were estimated. In the function estimation, $P(Y = 1)$ was calculated from the logistic model including all quadratic terms. For test data, if $\log[P\,(Y = 1)/\{1 - P\,(Y = 1)\}] \leq 0$, applicants were classified as "bad" credit and otherwise as "good" credit. In the density estimation based on a normal mixture distribution, all parameters were estimated using the EM algorithm. Applicants were classified as "bad" or "good" credit using the quadratic discriminant function with these estimated parameters. The number of simulations was 1,000.

The results are shown in Tables 3.1 and 3.2. The column labeled as "All Data" shows the misclassification rates when none of applicants were rejected and when characteristic variables have no missing values. The column labeled as "EM" shows the misclassification rates using the suggested EM algorithm in the density estimation, when data include missing values in both credit status Y and characteristic variables X. The column labeled as "LR" shows the misclassification rates using the logistic regression in the function estimation, when data include missing values in both Y and X.

Table 3.1 shows the mean misclassification rates when $n = 150$. Under MCAR missing data mechanism and when there were no missing values in x_2 and 10% percentage of applicants were rejected, the misclassification rates for all three methods were not very different, even if the ones based on logistic regression were a little higher. When the percentage of missing values in x_2 increases, the misclassification rates increases in both reject inference based on the EM algorithm and the logistic regression, but the incremental rates were much higher in the logistic regression than the EM algorithm suggested here.

Under MCAR missing data mechanism and when there were no missing values in x_2 and 30% percentage of applicants were rejected, the misclassification rates based on the logistic regression were higher than the other two methods. When the percentage of missing values in x_2 increases, the misclassification rates increases in both reject inference based on the EM algorithm and the

Table 3.2. The misclassification rate with $n = 500$

Missing data	% of	% of Rejects					
mechanism	missing	10%			30%		
for x_2	in x_2	All Data	EM	LR	All Data	EM	LR
MCAR	0%	0.239	0.240	0.242	0.239	0.241	0.259
	10%	0.239	0.242	0.260	0.239	0.243	0.276
	20%	0.239	0.243	0.277	0.239	0.246	0.297
	40%	0.239	0.248	0.311	0.239	0.252	0.333
MAR1	0%	0.239	0.239	0.242	0.239	0.240	0.259
	10%	0.239	0.240	0.243	0.239	0.241	0.259
	20%	0.239	0.242	0.249	0.239	0.248	0.263
	40%	0.239	0.255	0.282	0.239	0.289	0.288
MAR2	0%	0.239	0.239	0.242	0.239	0.240	0.259
	10%	0.239	0.239	0.244	0.239	0.239	0.244
	20%	0.239	0.240	0.246	0.239	0.241	0.263
	40%	0.239	0.243	0.260	0.239	0.246	0.280

Table 4.1. Description of analysis variables

Variable	Type	Description
Y	Binary	Credit Status
X_1	Binary	Existing checking account has money
X_2	Numeric	Duration in month
X_3	Numeric	Credit amount
X_4	Numeric	Amount in savings account
X_5	Numeric	Duration in current employment in years
X_6	Numeric	Age
X_7	Binary	Has a telephone registered under the customer's name

logistic regression, but the incremental rates were much higher in the logistic regression than the EM algorithm. The similar trends were shown under both MAR1 and MAR2 mechanisms.

When $n = 500$, the differences between all data and estimation methods become much smaller, but the same trend was still clear. Misclassification rates of the EM algorithm were much lower than the ones using the logistic regression, even if the misclassification rates are increased with larger percentage of rejects or larger percentage of missing values in the x_2 variable.

4. Application to German Credit Data

German credit data include credit status (measured by either good or bad) and 20 characteristic variables (7 numerical and 13 categorical ones) from 1,000 cases. The 700 cases had "good" credit status and the other 300 had "bad" credit status. To evaluate the performance of the suggested reject inference method, the randomly selected 700 cases (490 good and 210 bad credits) were used to build the credit scoring model and the remaining 300 used for the cross-validation. In this application, we focused only on 7 characteristic variables, either numeric or binary, since the suggested reject inference assumes a multivariate normal distribution. Variables considered in the analysis are described in Table 4.1. Numeric variables were log transformed to follow the normal distribution.

The probability of "good" credit status was calculated by using a logistic regression model. The cases whose probabilities belong to the lowest 30% were considered as rejected cases. If cases have longer than 30 years of the duration in current employment, their duration was assumed to be missing, resulting in about 20% of missing cases in the characteristic variable, X_5. Reject inference were conducted based on the suggested method and the misclassification rates were calculated from the 300 cross-validation cases. The misclassification rate among them was 29.3%. We also include the misclassification rate based on the logistic regression and it was 30.3%, which is one percent higher than the one based on the suggested method. On the other hand, the misclassification rate of complete data (without any rejected case and missing duration) was 27%.

5. Discussion

When characteristic variables include missing values, it may cause a bias in estimation, similar to missing credit status. We extended the EM algorithm to handle both missing credit status and missing characteristics in reject inference. The simulation showed that under all three missing data mechanisms and all percentages of missing values, misclassification rates of the suggested EM algorithm were lower than the ones using the logistic regression. Even if their rates (when the percentage of missing values becomes large) were higher than the ones based on all data, these differences occur due to smaller available information from missing data. Similarly, logistic regression showed large misclassification rates, since it discarded more available information by deleting observations with missing characteristic variables.

We assume that all characteristic variables are normally distributed. This assumption is often used in the analysis due to easiness in modeling among characteristic variables using a multivariate normal distribution. However, characteristic variables may be observed as categorical variables in real data. The extension of density estimation to handle mixed types of variables would be worthy.

References

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

Choi, B. J. (2008). *Semi-Supervised learning Based on Independent Gaussian Mixture Models*, Ph.D. dissertation, Korea University, Korea.

Copas, J. B. and Li, H. G. (1997). Reject inference for non-random samples, *Journal of the Royal Statistical Society, Series B*, **20**, 55–95.

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

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

Hand, D. J. (1998). *Reject Inference in Credit Operation, In: E. Mays ed. Credit Risk Modeling: Design and Application*, American Management Association, New York, 181–190.

Hand, D. J. and Henley, W. E. (1994). *Inference about Rejected Cases in Discriminant Analysis, In: Diday, E., Lechevallier, Y., Schader, M., Bertrand, P. and Burtschy, B. eds, New Approaches in Classification and Data Analysis*, Springer, New York, 292–299.

Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review, *Journal of the Royal Statistical Society, Series A*, **160**, 523–541.

Hsai, D. C. (1978). Credit scoring and the equal credit opportunity act, *The Hastings Law Journal*, **30**, 371–448.

Hunt, L. and Jorgensen, M. (2003). Mixture model clustering for mixed data with missing information, *Computational Statistics & Data Analysis*, **41**, 429–440.

Jacobson, T. and Roszbach, K. F. (2000). *Evaluating Bank lending Policy and Consumer Credit Risk, In: Yaser S. Abu-Mostafa et al. eds. Computational Finance 1999*, the MIT Press, Cambridge, 535–548.

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

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, John Wiley, New York.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall, New York.