

선형판별분석에서 MCMC다중대체법의 효율에 관한 연구

유희경* · 김명철**

*강원대학교 삼척캠퍼스 컴퓨터공학과 · **강원대학교 삼척캠퍼스 산업경영공학과

A Study on the efficiency of the MCMC multiple imputation In LDA

Hee Kyung Yoo* · Myung Cheol Kim**

*Dept. of computer engineering Kangwon National University

**Dept. of Industrial & Management Engineering Kangwon National University

Abstract

This thesis studies two imputation methods, the MCMC method and the EM algorithm, that take care of the problem. The performance of the two methods for the linear (or quadratic) discriminant analysis are evaluated under various types of incomplete observations. Based on simulated experiments, the effect of the imputation using the EM algorithm and the MCMC method are evaluated and compared in terms of the probability of misclassification and the RMSE. This is done for the various cases of incomplete observations. The cases are differentiated by missing rates, sample sizes, and distances between two classification groups.

The studies show that the probability of misclassification and the RMSE of the EM algorithm method is lower than the MCMC method. Therefore the imputation using the EM algorithm is more efficient than the MCMC method. And the probability of misclassification of the method that all vectors of observations with missing values are omitted from analysis is lower than the EM algorithm and the MCMC method when the samples size is small and the rate of missing values is extremely big.

Keywords : discriminant analysis, EM algorithm, MCMC

1. 서론

결측치가 포함된 관측치에 대하여 판별 분석을 실시할 경우 분류의 정확도가 떨어지기 때문에 결측치를 대체한 후에 판별분석을 실시하게 된다[1]. 결측치의 대체방법으로 관측된 자료에 대한 확률 모형과의 관련성을 이용하여 결측치를 대체하는 방법인 EM(Expectation Maximization) 알고리즘[2]과 Schafer가 제안한 MCMC(Markov Chain Monte Carlo)를 이용한 MI(multiple Imputation)방법을 적용하여 판별 분석의 분류 정확성에 미치는 영향을 비교한다[6]. 분류의 정확도 비교를 위해 두가지 방법에 대한 오분류율을 비교하고, 결측치의 추정치가 실제값과 얼마나 가까운지

살펴 보기위한 비교기준으로 RMSE (Root Mean Square Error)를 이용한다. 본 논문에서는 하나의 관찰 단위 전체가 분실된 경우인 개체 무응답은 고려하지 않고 항목 무응답(item nonresponse)인 경우의 다중대체법만을 고려한다.

본 연구는 먼저 완전 데이터인 경우의 판별 규칙을 살펴보고, 이어서 결측치가 존재할 경우 결측치 대체방법인 EM알고리즘과 MCMC를 이용한 다중 대체법을 살펴보았다. 이어서 시뮬레이션을 통하여 두가지 결측치 대체방법으로 결측치를 대체 했을 때 판별분석의 결과에 미치는 효과 비교하기 위하여 평균 오류율을 구하였다.

† 교신저자 : 유희경, 강원도 삼척시 중앙로 1번지 강원대학교 삼척캠퍼스 컴퓨터공학과

M·P: 011-701-6371, E-mail: hkyoo@kanwon.ac.kr

2009년 6월 24일 접수; 2009년 9월 4일 수정본 접수; 2009년 9월 10일 게재확정

2. 본 론

2.1 판별규칙(Classification rule)

집단 $\pi_i = 1, 2, \dots, g$ 의 확률밀도함수 $f_i(\mathbf{X})$ 는 평균벡터 $\boldsymbol{\mu}_i$, 공분산행렬 $\boldsymbol{\Sigma}_i$ 를 갖는 다변량 정규밀도 함수라고 가정할 경우 선형판별함수[3]는 다음과 같다.

$$d_i^L(\mathbf{X}) = \boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} \mathbf{X} - \frac{1}{2} \boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \ln p_i \quad (1.1)$$

여기서 p_i 는 사전 확률이다.

모집단의 공분산행렬이 동일하다고 가정할 경우 공분산행렬의 합동추정량 S_{pooled} 에 의해 선형판별 함수가 다음과 같이 추정된다.

$$\hat{d}_i^L(\mathbf{X}) = \bar{\mathbf{X}}_i' S_{pooled}^{-1} \mathbf{X} - \frac{1}{2} \bar{\mathbf{X}}_i' S_{pooled}^{-1} \bar{\mathbf{X}}_i + \ln p_i \quad (1.2)$$

여기서

$$S_{pooled} = \frac{(n_1 - 1)S_1 + 1(n_2 - 1)S_2 + \dots + (n_g - 1)S_g}{n_1 + n_2 + \dots + n_g} \quad (1.3)$$

판별규칙은 식(1.2)이 최대가 되는 집단으로 관측치 \mathbf{X} 를 분류한다.

모집단의 공분산행렬이 모두 같다고 가정 할 수 없는 경우에는 이차판별함수를 판별규칙으로 이용하며 이차 판별함수는 다음과 같다.

$$d_i^Q(\mathbf{X}) = -\frac{1}{2} \ln |\boldsymbol{\Sigma}_i| - \frac{1}{2} (\mathbf{X}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{X}_i - \boldsymbol{\mu}_i) + \ln p_i \quad (1.4)$$

표본 평균과 공분산행렬에 의해 이차판별함수는 다음과 같이 추정된다.

$$d_i^Q(\mathbf{X}) = -\frac{1}{2} \ln |S_i| - \frac{1}{2} (\mathbf{X}_i - \bar{\mathbf{X}}_i)' S_i^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}_i) + \ln p_i \quad (1.5)$$

판별규칙은 식(1.5)이 최대인 집단으로 관측치 \mathbf{X} 를 분류한다.

2.2 결측치 대체 방법

2.2.1 EM 알고리즘

EM 알고리즘은 불완전한 자료가 존재할 때, 자료가 가지고 있는 정보를 이용하여 완전한 자료를 도출한 후 최우추정치(MLE)를 구하는 반복적인 절차를 제공한다[2].

EM 알고리즘은 결측치의 기대치를 구하여 완전 자료의 조건부 기대값을 구하는 E(Expectation)단계와 결측치가 모두 대체 되었다는 가정 하에서 최우추정치를 구하는 M(Maximization)단계로 구성되어 있으며 기본

절차는 다음과 같다.

E 단계에서는 주어진 관측 데이터와 현재 추정된 모수로 결측치의 조건부 기대값을 구하고, 결측치에 대하여 이 기대값으로 대체한다.

개체수 n개 이고 변수 k개인 결측값을 포함한 자료 행렬 (\mathbf{Y}) 의 $\Theta^{(t)}$ 는 t 번째 반복 시점에서의 추정치 이고, $\ln f(\mathbf{Y} | \Theta)$ 는 완전한 자료의 대수우도, $f(\mathbf{Y}_{miss} | \mathbf{Y}_{obs}, \Theta)$ 는 관측치가 주어졌을 때 결측치의 밀도함수를 나타낸다. 이때 (\mathbf{Y}) 를 다음의 조건부 기대값에 의해 대체해 넣는다.

$$I_{imp} = E(\ln f(\mathbf{Y} | \Theta) | \mathbf{Y}_{obs}, \Theta^{(t)}) \quad (2.1)$$

$$= \int \ln(\Theta | \mathbf{Y}) f(\mathbf{Y}_{mis} | \mathbf{Y}_{obs}, \Theta)$$

$$\hat{y}_{ij}^{(i)} = \begin{cases} y_{ij} & , y_{ij} \text{가 관측된 경우, } i = 1, \dots, n, \\ I_{imp} & , y_{ij} \text{가 결측된 경우, } j = 1, \dots, k \end{cases}$$

M 단계에서는 완전자료에서 얻은 충분통계량을 이용하여 (t+1)번째의 모수 추정치를 구한다.

E 단계 와 M 단계를 모수 Θ 가 한 점으로 수렴 할 때까지 반복하게 된다

2.2.2 Markov Chain Monte Carlo(MCMC) 방법을 이용한 다중대체법

Rubin은 관측치가 결측치에 대해 간접적인 정보를 제공한다는 가정 하에서 예측된 분포에서 추출된 값으로 결측치를 대체하여 완전한 데이터 집합을 m개 만들고 추정된 모수와 표준오차를 결합하여 모형을 만드는 방법론을 개발 하였으며[5], Schafer는 MCMC를 이용한 다중대체법을 제안하였다[7].

다중대체법에서는 결측치를 대체하여 완전한 자료를 만들 때 이용할 알고리즘과 대체 반복수 그리고 각 점 추정치들의 결합방법을 고려해야 한다.

본 논문에서 이용할 MCMC방법의 한 방법인 자료증대(Data Argumentation)방법은 다음과 같다.

$\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{miss})$ 는 변수 K개이고 개체수 n개인 자료,

$\mathbf{Y}_{obs,i}$ 를 i번째 개체에서 관측된 변수,

$\mathbf{Y}_{miss,i}$ 를 i번째 개체에서 결측된 변수,

$\Theta^{(t)}$ 를 t번째 반복 시점에서의 모수라고 하면,

$\mathbf{Y}_{obs,i}$ 를 조건으로 하는 $\mathbf{Y}_{miss,i}$ 의 조건부 분포로부터 $\mathbf{Y}_{miss,i}$ 를 대체하는 I(imputation) 단계와 모수의 사전 정보와 완전해진 자료가 주어졌을 때 사후 모

집단의 모수를 찾는 P(Posterior) 단계로 구성된다.

[I-단계] $Y_{mis}^{(t+1)} \sim P(Y_{miss}, Y_{obs}, \Theta^{(t)})$

[P-단계] $\Theta^{(t+1)} \sim P(\Theta, Y_{obs}, Y_{mis}^{(t+1)})$

$(Y_{miss}^{(1)}, \Theta^{(1)}), (Y_{miss}^{(2)}, \Theta^{(2)}), \dots$

이렇게 충분한 반복을 통해 안정된 분포 $P(Y_{miss}, \Theta | Y_{obs})$ 로 수렴된다면 그 분포로부터 결측치를 대체할 값을 얻게 된다. 대체 결과를 평가하기 위해서 다른 난수 발생자와 다른 초기 모수 추정값을 갖는 시작값을 이용하여 과정을 반복한다.

Rubin은 결측치 비율과 대체 회수에 따른 측정의 효율성이 $(1 + (\gamma/m))^{-1}$ 로 근사함을 보였으며, γ 가 결측치 비율, m 이 대체 횟수일 때, 측정의 효율성은 다음과 같다[5].

<표 1> 다중대체법의 효율성

m	γ				
	10%	30%	50%	70%	90%
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.9772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

결측치가 대체된 m 개의 완전한 자료를 결합하여 모수를 추정하는 방법은 다음과 같다.

Q 는 m 개의 데이터에서 모수의 값, \hat{Q} 는 완전해진 자료의 점추정치, U 는 집단내 분산, B 는 집단간 분산, T 는 전체분산이다[5].

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^m \hat{Q} \tag{2.2}$$

$$\bar{U} = \frac{1}{m} \sum_{i=1}^m \hat{U} \tag{2.3}$$

$$B = \frac{1}{m-1} \sum_{i=1}^m (\hat{Q} - \bar{Q})^2 \tag{2.4}$$

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B \tag{2.5}$$

3. 시뮬레이션 비교연구

EM대체법과 MC대체법이 결측치를 포함하는 자료의 판별분석 결과에 미치는 영향을 알아보기 위해 두 그룹 선형판별분석에서 시뮬레이션 비교연구를 실시하였다.

비교연구에 이용된 자료는 동일한 공분산 행렬을 갖는 다변량 정규분포를 따르는 두 개의 그룹인 $\pi_1 \sim N_2(\mu_1, \Sigma)$ 와 $\pi_2 \sim N_2(\mu_2, \Sigma)$ 으로부터 난수를 발생시켜 얻은 자료이며 각 분포의 모수는 다음과 같다.

$$\mu_1 = [0, 0], \quad \mu_2 = [a, a], \quad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

자료의 수가 $n = 30, 50, 100, 200$ 인 경우에 대해 $\mu_1 = [1, 1]', [2, 2]', [3, 3]'$ 이고, $\rho = 0.2, 0.4, 0.6, 0.8$ 인 각각의 자료를 생성한다.

시뮬레이션 데이터는 SAS IML program을 사용하여 생성하였다.

EM대체법이 최우추정법을 사용하는 것이므로 동등한 조건 하에서 두 방법을 비교하기 위해 MC대체법에서는 객관적 사전분포(objective prior)를 사용한다. 시뮬레이션에 의해 얻은 완전한 자료를 사용하여 얻은 랜덤결측(MAR : Missing At Random)형태의 자료(결측률 = 5%, 10%, 25%, 50%)에 EM대체법과 MC대체법을 적용하였다. MC대체법에는 제프리(Jeffreys)의 사전분포인 $\pi(\mu_j, \Sigma) \propto |\Sigma|^{-(p+1)/2}, j=1,2$ 을 사용하여 얻은 MCMC를 사용하였다. 그리고 MC대체법에서 다중대체회수는 5회로 하고, P단계의 초기값은 EM알고리즘에서 구한 최우추정값으로 하였다.

두 가지 방법으로 자료의 결측치를 대체 후 완전해진 자료를 식(1.2)의 일차판별함수를 이용하여 판별분석을 시행한 후 EM대체법과 MC대체법이 판별분석 결과에 미치는 효과를 비교하였다. Lachenbruch가 제안한 Cross-validation 추정법으로 오분류율을 추정하였고[4], 위 과정을 $m=100$ 번 반복 시행하여 추정한 평균 오분류율을 <부록 표 2>과 <부록 표 3>와 같다.

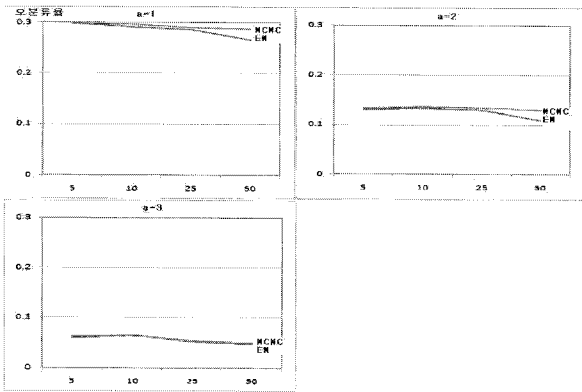
<부록 표 2>과 <부록 표 3>에서 두 집단사이의 거리가 멀어질수록 분류의 정확도는 높아지며, 특히 자료수가 적고 두 집단사이의 거리가 먼 $a=3, n=30$ 의 경우 모든 관측치가 정분류된다. 또한 자료수 n 이 클수록 그리고 ρ 가 증가함에 따라 오분류율이 증가하는 추세를 보인다.

결측치를 포함한 자료의 판별분석에서, 두 집단사이의 상관계수, 평균차이(거리), 그리고 자료수와 결측률을 모두 고려하였을 때, 판별의 오분류율은 EM대체법이 MC대체법 보다 대체적으로 낮게 나타났다.

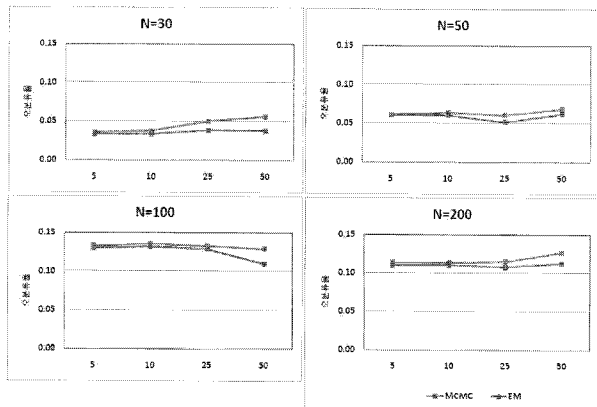
<부록 표 4>에 결측치가 모두 제거된 완전한 자료의 오분류율을 제시하였다. <부록 표 4>에서 자료수가 작은 $n=30, 50$ 의 경우 극단적으로 결측률이 높은 50%의 결측율에서는 <부록 표 2>의 MC대체법이나 EM대체법보다 결측치를 모두 제거하고 완전한 자료만을 이용하는 경우의 오분류율이 더 낮다. 그리고 $n=30$ 이고 $a=1, 3$ 인 경우 대부분 정분류되어 오분류율이 0이고, $a=2$ 인 경우도 0.15로 아주 낮은 오분류율을 나타내고 있다.

<부록 표 4>의 자료수가 $n=100, 200$ 인 경우에는 결측치를 제거한 경우의 오분류율이 몇몇 경우를 제외하고는 <부록 표 3>의 MC대체법이나 EM대체법보다 오분류율이 더 크다.

<그림 1>은 $n=100, \rho=0.6$ 인 경우 두 집단의 거리별 결측률에 따른 오분류율을 비교한 것이다. <그림 1>을 보면 두 집단의 거리가 멀어질수록 분류의 정확도가 증가하여 오분류율이 감소하며, 집단 사이의 거리에 따른 두 가지 대체방법을 비교하면 EM대체법에 따른 오분류율이 MC대체법 보다 대부분 다소 낮게 나타났다.



<그림 1> 집단 거리별 결측률에 따른 오분류율 비교($n=100, \rho=0.6$)



<그림 2> 자료수별 결측률에 따른 오분류율 비교($a=2, \rho=0.6$)

<그림 2>는 $a=2, \rho=0.6$ 인 경우 자료수별 결

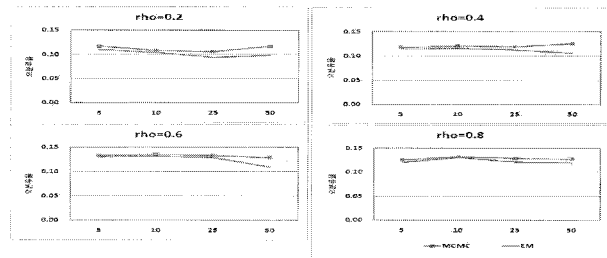
측률에 따른 오분류율 비교한 것이다. <그림 2>를 보면 자료수가 $n=30, 50, 100, 200$ 개로 증가함에 따라 오분류율이 대체적으로 증가하는 추세를 보이며, 자료수별로 두 가지 대체방법이 판별분석 결과의 정확성을 나타내는 오분류율을 비교한 결과 EM대체법에 따른 오분류율이 MC대체법의 오분류율보다 낮게 나타났다.

<그림 3>은 $a=2, n=100$ 인 경우 ρ 별 결측률에 따라 오분류율 비교한 것이다. <그림 3>에서 두 집단의 상관관계에 두 가지 대체방법이 판별분석 결과의 정확성에 미치는 영향을 비교하면 EM대체법에 따른 오분류율이 MC대체법의 오분류율보다 더 작다.

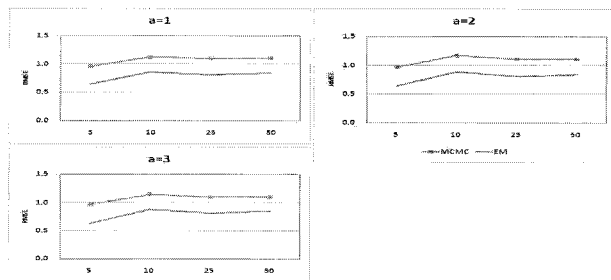
MAR 결측 상황에서의 일반적인 랜덤 패턴의 결측치를 갖는 자료를 EM대체법과 MC대체법을 이용하여 결측치를 대체 한 후 완전한 자료를 만들어 판별분석을 시행한 결과의 RMSE값은 <부록 표 5>에 제시하였다.

<부록 표 5>의 결과를 살펴보면 두 집단 사이의 상관성, 거리, 그리고 자료의 표본수와 결측률을 모두 고려하였을 때 MC대체법이 EM대체법보다 RMSE값이 크게 나타났다. 결측치를 포함한 자료의 판별분석에서 EM대체법결과 추정된 결측치의 값이 MC대체법 결과 추정된 값보다 실제값과 더 가깝다는 것을 알 수 있다.

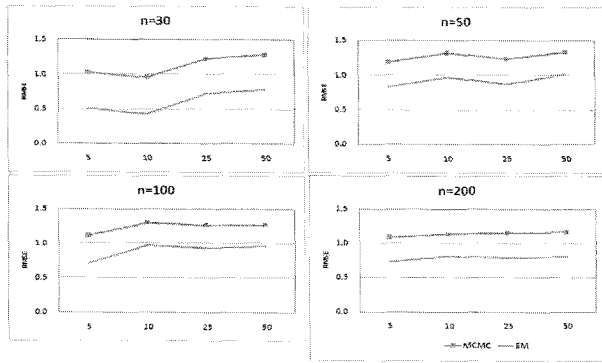
<그림 4>는 $n=100, \rho=0.6$ 인 경우 두 집단 거리별 결측률에 따른 RMSE값을 비교한 것이다. <그림 4>에서 두 집단의 거리에 따른 두 가지 대체방법을 비교하면 MC대체법의 RMSE값이 EM대체법의 RMSE값보다 더 크다.



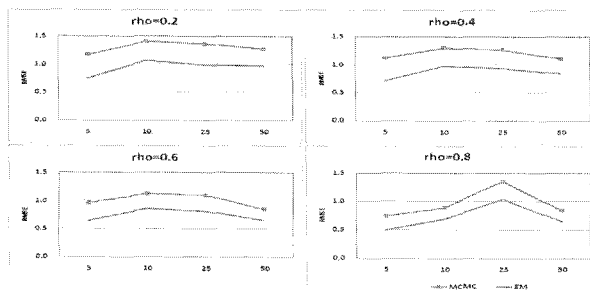
<그림 3> ρ 별 결측률에 따른 오분류율 비교 ($a=2, n=100$)



<그림 4> 집단 거리별 결측률에 따른 RMSE비교 ($n=100, \rho=0.6$)



<그림 5> 자료수별 결측률에 따른 RMSE비교
($a = 1, \rho = 0.4$)



<그림 6> ρ 별 결측률에 따른 RMSE비교
($a = 1, n = 100$)

<그림 5>는 $a = 1, \rho = 0.4$ 인 경우 자료수별 결측률에 따른 RMSE를 비교한 것이다. <그림 5>에서 자료수에 따른 두 가지 대체방법을 비교하면 MC대체법의 RMSE값이 EM대체법의 RMSE값보다 더 크다.

<그림 6>은 $a = 1, n = 100$ 의 경우 ρ 별 결측률에 따른 RMSE를 비교한 것이다. <그림 6>에서 $\rho = 0.2, 0.4, 0.6, 0.8$ 로 증가함에 따른 RMSE를 비교하면 두 집단의 상관관계가 커질수록 RMSE값은 감소하는 추세를 보인다. 또한 두 집단의 상관관계에 따른 두 가지 대체방법을 비교하면 MC대체법의 RMSE값이 EM대체법의 RMSE값보다 더 크다.

4. 결론

본 연구를 통해 결측치 처리 방법이 판별분석의 분류 정확도에 미치는 영향을 조사해 보고자 하였다. 시뮬레이션을 통하여 일반적인 결측 패턴의 MAR결측 상황에서의 EM대체법과 MC대체법에 대해 RMSE와 오분류율의 기대값을 비교하여 각 처리 방법이 판별분석의 분류 정확도에 미치는 영향을 비교해보았다. 시뮬레이션 과정에서 자료수와 결측률 그리고 판별대상 집단의 상관성과 거리를 고려하였다.

시뮬레이션 비교 결과 두 집단 사이의 거리가 멀어 질수록 분류의 정확도가 높았으며, 자료수가 많을수록 ρ 가 증가함에 따라 오분류율이 증가하는 추세를 보였다. 또한 두 집단 사이의 상관성, 거리, 그리고 표본수와 결측률을 모두 고려하였을 때, 결측치를 포함한 자료의 판별분석에서 EM대체법이 MC대체법보다 오분류율이 대체로 작게 나타났다. 그리고 결측치의 추정치가 실제값과 얼마나 가까운지 비교하기 위해 RMSE를 살펴본 결과 두 집단 사이의 상관성, 거리, 그리고 표본수와 결측률을 모두 고려하였을 때 MC대체법이 EM대체법보다 RMSE가 크게 나타났다. 즉, 일반적인 결측 패턴의 MAR 매커니즘 하에서 EM대체법이 MC대체법보다 판별분석의 분류 정확도가 더 높음을 알 수 있다.

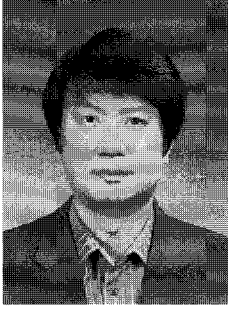
또한 결측치를 제거하고 완전한 자료만을 이용하는 방법과 MC대체법 및 EM대체법의 비교 결과 자료수가 적고 결측률이 극단적으로 큰 경우에는 결측치를 제거하고 완전한 자료값을 이용하는 방법의 오분류율이 MC대체법과 EM대체법에 비해 작았다. 반면에 자료수가 큰 경우일수록 결측치를 제거하고 완전한 자료값을 이용하는 방법의 오분류율보다 MC대체법과 EM대체법의 오분류율이 대체로 더 작았다. 그러므로 자료수와 결측률을 고려하여 분석의 목적에 맞게 결측치 대체 여부를 결정해야 할 것이다.

5. 참고 문헌

- [1] Chan, L.S., Dunn, O.J., "The treatment of missing values in discriminant analysis", Journal of the American Statistical Association, 67 (1972) : 473-477.
- [2] Dempster A.P., Laird N.M., Rubin D.B., "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society . Series B, 39 (1977) : 1-38
- [3] Johnson R.A., Wichern D.W., Applied multivariate statistical analysis, Prentice Hall.(2007)
- [4] Lachenbruch,P.A. and Mickey,M.A., "Estimation of Error Rates in Discriminant Analysis", Technometrics, 10 (1968) : 1-10
- [5] Rubin, D.B., Multiple imputation for nonresponse in surveys, New York, Wiley. (1987)
- [6] Schafer, J.L., Analysis of incomplete multivariate data, Chapman and Hall, London. (1997)
- [7] Schafer, J.L., "Multiple Imputation : A primer, Statistical Method in Medical Research", 8 (1999) : 3-15.

저 자 소 개

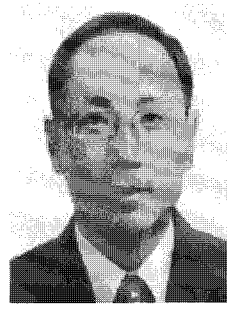
유 희 경



동국대학교에서 이학박사학위를 취득하고 현재 강원대학교 공학대학 컴퓨터공학과 교수로 재직 중이며, 주요관심분야는 데이터 마이닝, 웹마이닝, 컴퓨터시뮬레이션, 컴퓨터보안

주소 : 강원도 삼척시 중앙로 1번지 강원대학교 삼척캠퍼스 컴퓨터공학과

김 명 철



동국대학교에서 이학박사학위를 취득하고 현재 강원대학교 공학대학 산업경영 공학과 교수로 재직 중이며, 주요 관심분야는 실험계획법, 신뢰성 공학, 서비스 경영, 데이터 품질관리, 데이터 분석

주소 : 강원도 삼척시 중앙로 1번지 강원대학교 삼척캠퍼스 산업경영공학과

별첨 부록 : 표 2, 표 3, 표 4, 표 5,

부 록

<표 2> 오분류율 비교 (n=30, 50) () 안의 값은 표준편차

a	결측률	ρ	n=30		n=50	
			MCMC	EM	MCMC	EM
1	5%	0.2	0.2356(0.0198)	0.2333(0.0000)	0.2086(0.0152)	0.2044(0.0083)
		0.4	0.2378(0.0117)	0.2333(0.0000)	0.2312(0.0154)	0.2200(0.0000)
		0.6	0.2600(0.0187)	0.2490(0.0167)	0.2429(0.0080)	0.2400(0.0000)
		0.8	0.2556(0.0206)	0.2497(0.0167)	0.2482(0.0137)	0.2400(0.0000)
	10%	0.2	0.2422(0.0295)	0.2333(0.0000)	0.2181(0.0192)	0.2026(0.0101)
		0.4	0.2422(0.0153)	0.2333(0.0000)	0.2424(0.0163)	0.2424(0.0065)
		0.6	0.2600(0.0258)	0.2497(0.0167)	0.2503(0.0132)	0.2496(0.0100)
		0.8	0.2444(0.0241)	0.2520(0.0166)	0.2601(0.0176)	0.2666(0.0163)
	25%	0.2	0.2267(0.0580)	0.1600(0.0413)	0.2373(0.0303)	0.2154(0.0173)
		0.4	0.2222(0.0430)	0.1830(0.0347)	0.2428(0.0288)	0.2292(0.0187)
		0.6	0.2422(0.0320)	0.2230(0.0331)	0.2463(0.0285)	0.2422(0.0218)
		0.8	0.2467(0.0394)	0.2087(0.0275)	0.2614(0.0269)	0.2648(0.0291)
	50%	0.2	0.1933(0.0901)	0.1053(0.0426)	0.2160(0.0546)	0.1652(0.0430)
		0.4	0.2022(0.0527)	0.1263(0.0488)	0.2258(0.0506)	0.1970(0.0396)
		0.6	0.2400(0.0607)	0.1573(0.0515)	0.2334(0.0454)	0.2180(0.0427)
		0.8	0.2111(0.0709)	0.2027(0.0441)	0.2564(0.0414)	0.2626(0.0349)
2	5%	0.2	0.0257(0.0155)	0.0260(0.0139)	0.0235(0.0088)	0.0200(0.0000)
		0.4	0.0358(0.0090)	0.0333(0.0000)	0.0425(0.0079)	0.0400(0.0000)
		0.6	0.0357(0.0102)	0.0333(0.0000)	0.0606(0.0036)	0.0600(0.0000)
		0.8	0.0576(0.0177)	0.0557(0.0158)	0.0901(0.0100)	0.0906(0.0100)
	10%	0.2	0.0230(0.0181)	0.0250(0.0145)	0.0264(0.0124)	0.0200(0.0000)
		0.4	0.0367(0.0117)	0.0333(0.0000)	0.0458(0.0106)	0.0438(0.0079)
		0.6	0.0370(0.0122)	0.0333(0.0000)	0.0632(0.0076)	0.0600(0.0000)
		0.8	0.0605(0.0168)	0.0607(0.0129)	0.0882(0.0115)	0.0836(0.0077)
	25%	0.2	0.0267(0.0303)	0.0160(0.0167)	0.0414(0.0195)	0.0286(0.0100)
		0.4	0.0466(0.0246)	0.0327(0.0047)	0.0516(0.0153)	0.0402(0.0020)
		0.6	0.0499(0.0224)	0.0383(0.0120)	0.0600(0.0188)	0.0508(0.0100)
		0.8	0.0659(0.0227)	0.0713(0.0116)	0.0918(0.0219)	0.0854(0.0117)
	50%	0.2	0.0379(0.0374)	0.0130(0.0163)	0.0357(0.0338)	0.0096(0.0194)
		0.4	0.0501(0.0314)	0.0320(0.0066)	0.0538(0.0346)	0.0448(0.0296)
		0.6	0.0557(0.0282)	0.0377(0.0113)	0.0684(0.0322)	0.0618(0.0283)
		0.8	0.0707(0.0325)	0.0623(0.0210)	0.0893(0.0330)	0.0904(0.0264)
3	5%	0.2	0.0000(0.0000)	0.0000(0.0000)	0.0096(0.0100)	0.0020(0.0060)
		0.4	0.0003(0.0030)	0.0000(0.0000)	0.0199(0.0013)	0.0200(0.0000)
		0.6	0.0001(0.0015)	0.0000(0.0000)	0.0200(0.0000)	0.0200(0.0000)
		0.8	0.0009(0.0055)	0.0000(0.0000)	0.0202(0.0022)	0.0200(0.0000)
	10%	0.2	0.0003(0.0033)	0.0000(0.0000)	0.0086(0.0101)	0.0044(0.0083)
		0.4	0.0003(0.0033)	0.0000(0.0000)	0.0192(0.0042)	0.0200(0.0000)
		0.6	0.0001(0.0021)	0.0000(0.0000)	0.0201(0.0027)	0.0200(0.0000)
		0.8	0.0018(0.0091)	0.0000(0.0000)	0.0221(0.0073)	0.0200(0.0000)
	25%	0.2	0.0023(0.0093)	0.0000(0.0000)	0.0104(0.0109)	0.0082(0.0099)
		0.4	0.0026(0.0094)	0.0000(0.0000)	0.0199(0.0071)	0.0200(0.0000)
		0.6	0.0021(0.0099)	0.0000(0.0000)	0.0209(0.0088)	0.0200(0.0000)
		0.8	0.0078(0.0171)	0.0000(0.0000)	0.0280(0.0117)	0.0236(0.0077)
	50%	0.2	0.0049(0.0136)	0.0000(0.0000)	0.0072(0.0142)	0.0000(0.0000)
		0.4	0.0035(0.0119)	0.0000(0.0000)	0.0114(0.0181)	0.0070(0.0115)
		0.6	0.0055(0.0150)	0.0007(0.0047)	0.0169(0.0203)	0.0128(0.0152)
		0.8	0.0173(0.0219)	0.0057(0.0126)	0.0225(0.0216)	0.0180(0.0210)

<표 3> 오분류율 비교(n=100, n=200)

a	결측률	ρ	n=100		n=200	
			MCMC	EM	MCMC	EM
1	5%	0.2	0.2523(0.0161)	0.2431(0.0120)	0.2632(0.0089)	0.2609(0.0069)
		0.4	0.2784(0.0117)	0.2724(0.0065)	0.2964(0.0071)	0.2958(0.0048)
		0.6	0.2974(0.0107)	0.2992(0.0072)	0.3256(0.0084)	0.3237(0.0060)
		0.8	0.3153(0.0091)	0.3206(0.0074)	0.3302(0.0045)	0.3271(0.0025)
	10%	0.2	0.2645(0.0160)	0.2546(0.0088)	0.2547(0.0123)	0.2420(0.0109)
		0.4	0.2775(0.0125)	0.2737(0.0094)	0.2902(0.0109)	0.2854(0.0085)
		0.6	0.2961(0.0131)	0.2922(0.0116)	0.3179(0.0116)	0.3161(0.0069)
		0.8	0.3072(0.0125)	0.3118(0.0104)	0.3319(0.0061)	0.3318(0.0030)
	25%	0.2	0.2588(0.0241)	0.2455(0.0192)	0.2577(0.0156)	0.2460(0.0098)
		0.4	0.2752(0.0225)	0.2653(0.0179)	0.2904(0.0161)	0.2813(0.0096)
		0.6	0.2899(0.0207)	0.2861(0.0176)	0.3174(0.0141)	0.3131(0.0092)
		0.8	0.2960(0.0168)	0.2943(0.0138)	0.3299(0.0098)	0.3308(0.0070)
	50%	0.2	0.2670(0.0338)	0.2352(0.0252)	0.2589(0.0221)	0.2336(0.0182)
		0.4	0.2747(0.0301)	0.2483(0.0231)	0.2829(0.0205)	0.2704(0.0196)
		0.6	0.2869(0.0274)	0.2659(0.0217)	0.2967(0.0179)	0.2961(0.0133)
		0.8	0.2907(0.0233)	0.2679(0.0220)	0.3158(0.0147)	0.3200(0.0115)
2	5%	0.2	0.1117(0.0060)	0.1100(0.0000)	0.0857(0.0044)	0.0827(0.0025)
		0.4	0.1178(0.0064)	0.1136(0.0048)	0.1073(0.0065)	0.1089(0.0032)
		0.6	0.1327(0.0048)	0.1300(0.0000)	0.1135(0.0057)	0.1103(0.0051)
		0.8	0.1256(0.0075)	0.1205(0.0022)	0.1397(0.0041)	0.1416(0.0024)
	10%	0.2	0.1080(0.0108)	0.1040(0.0088)	0.0808(0.0056)	0.0779(0.0033)
		0.4	0.1210(0.0103)	0.1156(0.0054)	0.1054(0.0061)	0.1050(0.0039)
		0.6	0.1355(0.0072)	0.1321(0.0041)	0.1134(0.0064)	0.1101(0.0049)
		0.8	0.1317(0.0082)	0.1304(0.0070)	0.1401(0.0061)	0.1382(0.0046)
	25%	0.2	0.1057(0.0172)	0.0941(0.0096)	0.0783(0.0087)	0.0696(0.0051)
		0.4	0.1191(0.0170)	0.1119(0.0101)	0.0979(0.0106)	0.0909(0.0062)
		0.6	0.1329(0.0151)	0.1289(0.0107)	0.1145(0.0100)	0.1074(0.0062)
		0.8	0.1283(0.0183)	0.1211(0.0164)	0.1428(0.0089)	0.1357(0.0060)
	50%	0.2	0.1171(0.0236)	0.0981(0.0150)	0.0840(0.0140)	0.0577(0.0062)
		0.4	0.1255(0.0221)	0.1057(0.0168)	0.1020(0.0143)	0.0848(0.0101)
		0.6	0.1288(0.0245)	0.1093(0.0165)	0.1265(0.0141)	0.1127(0.0090)
		0.8	0.1272(0.0185)	0.1197(0.0149)	0.1419(0.0140)	0.1295(0.0103)
3	5%	0.2	0.0307(0.0026)	0.0300(0.0000)	0.0074(0.0039)	0.0062(0.0029)
		0.4	0.0414(0.0039)	0.0400(0.0000)	0.0202(0.0011)	0.0200(0.0000)
		0.6	0.0632(0.0067)	0.0600(0.0000)	0.0298(0.0024)	0.0298(0.0011)
		0.8	0.0778(0.0050)	0.0800(0.0000)	0.0396(0.0020)	0.0400(0.0000)
	10%	0.2	0.0256(0.0062)	0.0242(0.0050)	0.0087(0.0041)	0.0059(0.0024)
		0.4	0.0409(0.0072)	0.0414(0.0035)	0.0224(0.0036)	0.0200(0.0000)
		0.6	0.0651(0.0098)	0.0661(0.0099)	0.0318(0.0034)	0.0300(0.0000)
		0.8	0.0724(0.0079)	0.0723(0.0093)	0.0409(0.0024)	0.0400(0.0000)
	25%	0.2	0.0339(0.0109)	0.0293(0.0082)	0.0105(0.0052)	0.0059(0.0025)
		0.4	0.0472(0.0118)	0.0445(0.0083)	0.0204(0.0053)	0.0156(0.0016)
		0.6	0.0549(0.0117)	0.0528(0.0088)	0.0318(0.0058)	0.0263(0.0023)
		0.8	0.0665(0.0117)	0.0680(0.0080)	0.0379(0.0051)	0.0371(0.0025)
	50%	0.2	0.0362(0.0150)	0.0346(0.0123)	0.0138(0.0061)	0.0070(0.0025)
		0.4	0.0505(0.0154)	0.0487(0.0146)	0.0208(0.0071)	0.0109(0.0023)
		0.6	0.0617(0.0146)	0.0629(0.0141)	0.0317(0.0080)	0.0239(0.0041)
		0.8	0.0654(0.0150)	0.0603(0.0126)	0.0389(0.0070)	0.0344(0.0045)

<표 4> 결측치가 모두 제거된 완전한 자료의 오분류율 비교

			완전데이터			
a	결측률	ρ	n=30	n=50	n=100	n=200
1	5%	0.2	0.0000	0.2337	0.2422	0.2685
		0.4	0.0000	0.2346	0.2844	0.3053
		0.6	0.0000	0.2563	0.3056	0.3422
		0.8	0.0000	0.2563	0.3267	0.3421
	10%	0.2	0.0000	0.2202	0.2444	0.2498
		0.4	0.0000	0.2708	0.2778	0.2942
		0.6	0.0000	0.2708	0.2889	0.3219
		0.8	0.0000	0.2917	0.3222	0.3498
	25%	0.2	0.0000	0.2753	0.2396	0.2600
		0.4	0.0000	0.2753	0.2568	0.3000
		0.6	0.0000	0.2753	0.2837	0.3467
		0.8	0.0000	0.2515	0.2837	0.3600
	50%	0.2	0.0000	0.2132	0.2386	0.2810
		0.4	0.0000	0.2132	0.2614	0.2916
		0.6	0.0000	0.2132	0.2386	0.3117
		0.8	0.0000	0.2132	0.3019	0.3507
2	5%	0.2	0.0357	0.0217	0.1157	0.0894
		0.4	0.0357	0.0426	0.1157	0.1158
		0.6	0.0357	0.0643	0.1367	0.1159
		0.8	0.0714	0.1069	0.1261	0.1475
	10%	0.2	0.0357	0.0238	0.1000	0.0833
		0.4	0.0357	0.0446	0.1111	0.1109
		0.6	0.0357	0.0685	0.1333	0.1109
		0.8	0.0742	0.0893	0.1333	0.1388
	25%	0.2	0.0000	0.0551	0.0904	0.0800
		0.4	0.0982	0.0551	0.1051	0.1000
		0.6	0.0982	0.0551	0.1173	0.1267
		0.8	0.0982	0.1101	0.1173	0.1533
	50%	0.2	0.1500	0.0000	0.1169	0.1016
		0.4	0.1500	0.0294	0.1169	0.1228
		0.6	0.1500	0.0294	0.1347	0.1228
		0.8	0.1500	0.0294	0.0942	0.1736
3	5%	0.2	0.0000	0.0000	0.0315	0.0052
		0.4	0.0000	0.0217	0.0419	0.0209
		0.6	0.0000	0.0217	0.0632	0.0315
		0.8	0.0000	0.0217	0.0842	0.0421
	10%	0.2	0.0000	0.0000	0.0222	0.0057
		0.4	0.0000	0.0238	0.0444	0.0225
		0.6	0.0000	0.0238	0.0556	0.0336
		0.8	0.0000	0.0238	0.0667	0.0445
	25%	0.2	0.0000	0.0000	0.0269	0.0067
		0.4	0.0000	0.0313	0.0391	0.0200
		0.6	0.0000	0.0313	0.0513	0.0333
		0.8	0.0000	0.0313	0.0538	0.0400
	50%	0.2	0.0000	0.0000	0.0227	0.0094
		0.4	0.0000	0.0000	0.0406	0.0295
		0.6	0.0000	0.0000	0.0406	0.0389
		0.8	0.0000	0.0000	0.0406	0.0496

<표 5> MC대체법와 EM대체법의 RMSE 비교

a	결측률		n=30		n=50		n=100		n=200	
			MCMC	EM	MCMC	EM	MCMC	EM	MCMC	EM
1	5%	2	0.9515	0.5934	1.2341	0.8736	1.1651	0.7397	1.1784	0.7945
		4	1.0294	0.5142	1.1884	0.8287	1.1128	0.7095	1.0878	0.7282
		6	0.8509	0.3987	1.1078	0.7788	0.9577	0.6388	0.9631	0.6476
		8	0.6793	0.3654	0.8349	0.5999	0.7384	0.4906	0.6998	0.4872
	10%	2	0.9681	0.4951	1.3120	0.9452	1.3990	1.0661	1.2342	0.8655
		4	0.9556	0.4295	1.3180	0.9667	1.2990	0.9752	1.1368	0.8077
		6	0.8601	0.3340	1.2096	0.9628	1.1225	0.8595	1.0003	0.7195
		8	0.6569	0.3020	0.9281	0.7608	0.8891	0.6849	0.7414	0.5491
	25%	2	1.2110	0.7088	1.2739	0.8488	1.3560	0.9807	1.2416	0.8517
		4	1.2258	0.7247	1.2349	0.8669	1.2683	0.9324	1.1536	0.7828
		6	1.1014	0.6728	1.1436	0.8358	1.1005	0.8080	0.9805	0.6704
		8	0.8600	0.5401	0.8921	0.6345	0.8350	0.6165	0.7236	0.4989
50%	2	1.2765	0.7756	1.3889	0.9882	1.3544	1.0324	1.2551	0.8660	
	4	1.2871	0.7783	1.3385	1.0197	1.2689	0.9623	1.1662	0.7994	
	6	1.1915	0.7501	1.2229	0.9544	1.1067	0.8387	1.0152	0.6818	
	8	0.9764	0.5876	0.9166	0.7214	0.8394	0.6486	0.7495	0.5016	
2	5%	2	0.9535	0.6028	1.2670	0.9328	1.1801	0.7408	1.1603	0.7801
		4	0.9042	0.5350	1.2224	0.8288	1.0795	0.6882	1.0939	0.7299
		6	0.8537	0.3954	1.1203	0.7652	0.9654	0.6394	0.9383	0.6432
		8	0.6642	0.3631	0.8268	0.5897	0.7371	0.5030	0.7000	0.4830
	10%	2	0.9848	0.4998	1.2822	0.9105	1.3750	1.0139	1.2444	0.8692
		4	0.9664	0.4318	1.3009	0.9738	1.3239	0.9786	1.1461	0.8145
		6	0.8940	0.3357	1.2132	0.9626	1.1696	0.8886	0.9907	0.7172
		8	0.6558	0.3006	0.9490	0.7751	0.8793	0.6739	0.7508	0.5509
	25%	2	1.2528	0.7487	1.2877	0.8680	1.3419	0.9628	1.2335	0.8457
		4	1.2258	0.7495	1.2509	0.8698	1.2547	0.9121	1.1406	0.7774
		6	1.0956	0.7038	1.1482	0.8183	1.1052	0.8041	0.9806	0.6782
		8	0.8595	0.5399	0.8914	0.6296	0.8360	0.6215	0.7254	0.4994
50%	2	1.2769	0.7677	1.3767	0.9920	1.3749	1.0429	1.2559	0.8673	
	4	1.2710	0.7977	1.3500	1.0193	1.2674	0.9660	1.1713	0.8059	
	6	1.2101	0.7666	1.2385	0.9544	1.1044	0.8424	1.0150	0.6846	
	8	0.9732	0.5809	0.9192	0.7211	0.8378	0.6458	0.7495	0.4999	
3	5%	2	0.9822	0.5995	1.2111	0.8526	1.1381	0.7129	1.1775	0.7856
		4	0.9450	0.5215	1.2024	0.8243	1.0802	0.7116	1.0795	0.7350
		6	0.8775	0.3994	1.0714	0.7338	0.9695	0.6337	0.9500	0.6428
		8	0.6245	0.3692	0.8076	0.5890	0.7183	0.4968	0.6980	0.4813
	10%	2	0.9352	0.4968	1.3309	0.9608	1.4279	1.0590	1.2421	0.8654
		4	0.9658	0.4315	1.3035	0.9661	1.3029	0.9720	1.1467	0.8167
		6	0.8875	0.3285	1.2283	0.9666	1.1491	0.8802	1.0004	0.7196
		8	0.6682	0.3059	0.9440	0.7624	0.8815	0.6792	0.7496	0.5511
	25%	2	1.1919	0.6754	1.2887	0.8551	1.3541	0.9635	1.2397	0.8521
		4	1.2403	0.7421	1.2691	0.8630	1.2590	0.9211	1.1448	0.7814
		6	1.1344	0.6962	1.1716	0.8323	1.1026	0.8136	0.9862	0.6798
		8	0.8523	0.5377	0.8790	0.6345	0.8320	0.6115	0.7254	0.4992
50%	2	1.2945	0.7912	1.3804	0.9827	1.3797	1.0415	1.2660	0.8741	
	4	1.2729	0.7767	1.3719	1.0291	1.2648	0.9629	1.1725	0.8066	
	6	1.2061	0.7526	1.2341	0.9622	1.1030	0.8539	1.0128	0.6837	
	8	0.9641	0.5804	0.9292	0.7306	0.8386	0.6429	0.7534	0.4998	