Journal of Korean
Data & Information Science Society
2002, Vol. 13, No.2 pp. 105 113

Application of NORM to the Multiple Imputation for Multivariate Missing Data.

Hyun-Jeong Kim[1], Sung-Ho Moon[2], Jae-Kyoung Shin[3]

Abstract

The statistical analysis of incomplete data sometimes requires handling of incomplete observations. Towards this end, each case with some missing values generally should be deleted, namely, resulting in only use of non-missing cases. EM algorithm(Dempster et al., 1977) which involves prediction and estimation steps is a general method among others. In this article, we use the free software NORM developed for multiple imputation, which uses DA(Data Augmentation) algorithm in its imputation, and evaluate its efficiency through a numerical example.

Keywords : Hot-deck, EM-algorithm, NORM, Multiple Imputation

1. 서 론

결측값이란 실험이나 조사에서 어떠한 이유에 의해 관측되어야 할 값을 얻지 못할 경우를 말하며, 이런 현상은 조사자료에 대한 다변량분석을 행할 때 종종 접하게 된다. 일반적으로 자료분석에서 다루어지는 데이터는 모든 원소들이 관측된 경우이다. 그러나 결측값을 갖는 자료의 경우 어떻게 처리하는가에 대해서 많은 방법들이 연구되어 왔다.

결측값이 존재하는 경우의 자료분석을 위해서 몇 가지 편의적인 방법으로서, i) 관측된 값만을 이용하는 방법(결측율이 높을 때는 비효율적이다), ii) 관측된 값에서 샘플링하여 결측값에 대치하는 방법(유효성이 과소 평가된다), iii) 평균값 또는 조건부기대값으로 결측값을 대치하는 방법이 있다. 또한 방법 iii)의 경우 조건부기대값 이외의 경우는 결측값이 변량값과 완전 무상관(MCAR, missing completely at random)일 때는 가능하지만, 결측값이 변량값과 상관이 있고 결측 되어 있는 값에는 무상관인 경우(MAR, missing at random)는 적절하지 않다. 두 경우 모두 분산·공분산 값을

1 부산광역시 사상구 괘법동 산1-1번지 신라대학교 교양학부 전임강사
E-mail: semikim@silla.ac.kr
2 부산광역시 남구 우암동 산55-1번지 부산외국어대학교 통계학과 부교수
3 경남 창원시 사림동 9번지 창원대학교 통계학과 부교수

과소평가 한다는 사실은 잘 알려져 있다(Little and Rubin, 1987).

위의 방법들은 결측값을 포함한 자료를 분석할 때 자주 이용되는 방법들이지만 결측값 자체에 대한 정보를 얻을 수 없다는 단점도 있다. 이에 대한 개선된 방법으로서 Little and Rubin(1987)과 Johnson and Wichern(1992)은 관측값에 근거해서 iv) EM 알고리즘(Dempster et. al., 1977)을 이용한 최우추정법(Orchard and Woodbury, 1972)을 권장하고 있으며, 가장 일반적인 방법으로서는 회귀직선모형에 의한 대치법도 있다. 그러나 EM 알고리즘을 이용한 최우추정법과 회귀직선 모형에 의한 대치법들은 결측값에 대해 하나의 값만으로 대치(일대치, single imputation)하므로 편의적일 수도 있으며, 결측데이터의 결측율이 큰 경우 정확성의 정도도 문제가 될 수 있다.

본 연구에서는 위에서 지적한 i) iii)방법과 iv)의 방법의 문제점들을 개선하여 정확성의 정도를 높이는 방법으로서 결측값에 대해 여러 개의 값으로 대치하는 다대치(multiple imputation)의 유효성을 보인다. 또한 i) iii) 에서의 다대치의 결과보다 정도가 높은 보완적 방법으로서 iv)를 이용하여 결측값에 대한 초기값을 얻은 후 Markov Chain Monte Carlo(MCMC) 기술의 한 종류인 Data Augmentation(DA) 알고리즘에 의해 다대치를 생산하는 무료 소프트웨어 NORM(Schafer, 1997a, Chapter 5)의 구성을 소개하고 이를 적용한 실례를 보인다.

물론 NORM 이외에 다대치용 소프트웨어로 개발된 것으로서 로그선형모형 가정하의 범주형 자료용인 CAT(Schafer, 1997a, Chapter 7-8), 일반위치모형(general location model) 가정하의 연속형·범주형 자료를 위한 MIX(Schafer, 1997a, Chapter 9)와 다변량 선형혼합효과모형(multivariate linear mixed-effects model) 가정하의 패널자료나 집락자료에 대한 PAN(Schafer, 1997b)등이 있으나, 본 논문에서는 다변량정규분포 가정하의 연속형 자료에 대한 다대치 문제를 NORM을 이용해서 다루고자 한다.

2장에서는 NORM에 관한 소개와 사용방법 등에 관한 설명을 다루며, 3장의 수치 예에서는 주어진 결측 데이터를 1, 2장에서 언급한 기존의 대치 방법에 의한 결과와 NORM을 이용해서 얻은 결과를 비교, 검토한다.

2. 결측값 대치를 위한 NORM의 소개

NORM에서의 주요 과정은 1) 평균, 분산, 공분산(또는 상관)의 충분 통계량을 구하기 위한 EM 알고리즘, 2) 결측값의 다대치를 생성하기 위한 DA 과정과 추가적으로 3) 정규분포가 아닌 변수들의 대치를 위한 데이터(변환)의 사전·사후 대치과정, 4) DA의 수렴을 검증, 확인하기 위한 플롯(plot), 5) 불확실한 결측 자료를 병합하여 전체(overall) 추정값과 표준오차를 생성하기 위하여 점추정에 대한 Rubin(1987)의 규칙을 이용한 다대치 데이터분석의 결과를 결합하고, 다모수 추론에 이용할 수 있다.

NORM은 자료의 관측된 부분으로부터 얻어진 정보를 이용하여 결측된 부분을 $m(>1)$번 시뮬레이션 한다. 이때 완전데이터(complete data : 그림 1 참조)와 동일한 데이터세트를 m개 생성한다. 각 m개의 데이터세트는 표준 완전데이터(standard complete data : 그림 6 참조)로 분석된다. m개 세트의 결과는 결측데이터의 불확실성과 결합된 추정값과 표준오차의 한 집합을 생산하기 위하여 점추정에 대한 Rubin의 규칙들과 다모수 추론에 대한 확장된 규칙을 이용하여 결합된다. 대부분의 경우 적은 수(일반적으로 5번 내지 10번)만큼의 대치일지라도 좋은 결과를 얻을 수 있다.

NORM은 Markov Chain Monte Carlo (MCMC) 기술의 특별한 종류인 DA라고 하는 알고리즘에 의해 다대치를 생성한다.

2.1 NORM의 구성

NORM은 크게 Data, EM algorithm, Data Augmentation, Impute from parameters의 4단계로 구성되어 있으며 자세한 사항은 다음과 같다.

(1) Data

NORM을 실행하기 위하여 먼저 file 메뉴의 New session을 이용하여 데이터를 import(불러오기)한다. 데이터는 텍스트(또는 ASCII) 형식으로 파일명은 「*.dat」로 인식한다. 변수와 변수는 공백으로 구분되며 개체의 구분은 개행 또는 Enter 키로 구분된다. 결측값은 임의의 코드를 주면 되는데, 이때 관측값과 구분되는 코드를 주어야 한다. Data file은 변수 개수와 개체 개수 그리고 결측값의 코드를 나타낸다. 예를 들어, 그림 1에서는 첫 번째와 두 번째 변수에 결측값이 있으며 -9로서 결측값 코드를 나타내고 있다. Variables에서는 변수명 정의와 변수 변환 등을 할 수 있다. 예를 들어, 변수 변환을 할 때에는 Transformation항의 none부분을 더블 클릭하여 활성화되는 창에서 원하는 변수 변환을 할 수 있다. 또한 대치를 원하지 않는 변수에 대해서도 *.imp 항목의 해당 변수에 더블 클릭을 하면 된다(default일 경우는 *.imp로 대치된다).

Summarize는 데이터의 요약통계량을 나타내는 파일과 제목을 지정할 수 있다. 이때

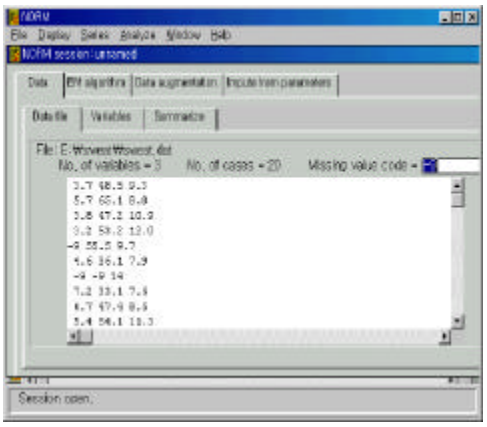

그림 1. NORM의 데이터

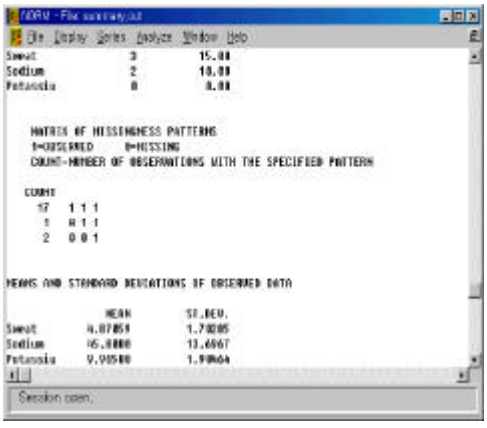

그림 2. Summary Out

파일명을 따로 지정하지 않고 Run 버튼을 눌러 실행하면 결과는 summary.out이라는 default file명으로 저장되어 그림 2와 같은 화면이 자동 표시된다. 여기서는 변수와 개체의 수, 결측율, 결측 패턴, 관측값만을 이용하여 얻은 평균, 표준편차 등을 알 수 있다.

(2) EM algorithm

결측값이 존재할 때 모수의 최우추정값 계산을 위한 일반적인 방법의 하나가 EM

알고리즘(Dempster et al., 1977)이다. EM 과정은 이 방법의 기술적인 면으로 추정(estimation)과 최대화(maximization)의 두 단계를 반복하는 계산이다. 이 화면에서는 공분산 또는 상관행렬을 선택하여 출력할 수 있으며, 모수의 초기값을 data 또는 파일로부터 지정할 수 있다. EM을 실행하면 두 개의 파일을 생성하는데 하나는 EM의 결과를 알려주는 output file(*.out)이고 다른 하나는 모수 추정 결과를 저장한 parameter file(*.prm)이다. EM 실행이 끝날 때 output file은 자동 출력되나 parameter file은 자동으로 표시되지 않는다. 이 두 파일은 메뉴의 화면 표시(Display)를 이용하여 언제든지 확인할 수 있다. Computing option을 이용하면 최우추정의 반복 횟수(Maximum iterations)와 수렴의 정도(Convergence criterion)를 조절할 수 있다. 그림 3에서 Maximum iterations를 1000회로 정하였는데, 만약 EM이 1000회까지 수렴하지 않는다면 1000회에서 멈추게 된다. 추정방법에 따라 MLE와 Posterior mode를 선택 할 수 있는데, EM에서는 default로 MLE가 선택된다. 이 과정은 결측데이터를 대치하기 위한 전 단계로서 먼저 EM을 실행하여 다음 과정의 DA에서 이용할 모수의 초기값을 제공하게 된다.

이때 EM의 수렴성은 결손 정보율(Rate of missing information: Rubin, 1987)과 관련이 있으며 결손율이 높으면 수렴성이 느리다.

(3) Data Augmentation(DA)

NORM에서 DA 알고리즘은 결측데이터에 대한 적합한 다대치를 생성하는 단계로, 데이터의 사후분포로부터 결측데이터와 모수의 임의의 값을 시뮬레이션한다. DA는 MCMC의 또 다른 종류로서, 반복적인 시뮬레이션 기술이다. DA를 실행하기 전 EM을 먼저 실행하여야만 모수에 대한 좋은 초기값을 얻을 수 있다.

DA 화면은 그림 4와 같이 Run의 실행 버튼과 Computing, Imputation, Series의 3가지 옵션 버튼이 있다. Run button을 누르면 시뮬레이션을 위하여 모수를 초기화한다. 이것을 실행하면 두 개의 파일이 생성된다(output file(*.out)은 결과를 표시하고, parameter file(*.prm)은 모수의 최종 추정값이 저장된다). 이 실행을 마치면 Output 항목에서 파일명을 따로 주지 않으면 da.out 파일이 생성되어 자동 표시되나

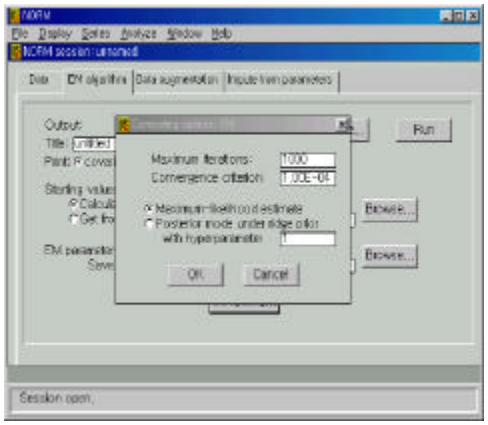

그림 3. EM 알고리즘 옵션

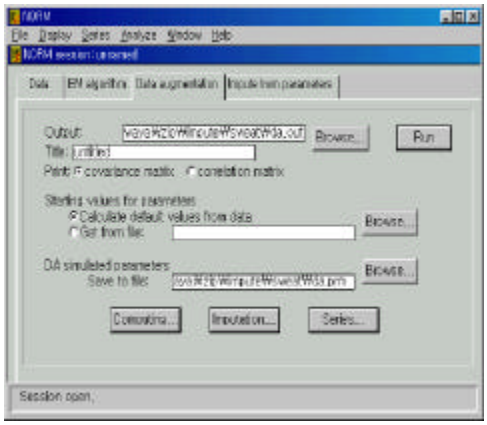

그림 4. Data Augmentation

parameter file은 자동 표시되지 않는다. DA Computing options는 반복횟수(Number

of iterations), 난수를 생성하는 초기값의 설정(Random seed), 사전분포(prior distribution)를 지정할 수 있다. Imputation option은 대치의 종류를 선택할 수 있는데, 이때 생성되는 데이터 집합들은 *.imp file로 저장되는 View imputation immediately를 지정하면 매회 생성되는 데이터를 *_n.imp file로부터 바로 확인할 수 있다.

Series 옵션은 DA의 수렴성 진단을 위해 plotting에 대한 parameter series(*.prs)의 시뮬레이션된 추정 모수의 흐름을 보여준다. 시뮬레이션되는 모수를 parameters series(*.prs)에 어떻게 저장할 것인가를 선택할 수 있다. Save all parameters at every k-th cycle : 모든 모수(평균, 분산, 공분산)가 *.prs에 저장된다. 이 파일은 각 모수에 대해 시계열과 자기상관 플롯으로 나타난다. 이 옵션의 장점은 DA의 수렴이 느리거나 이상하게 반응한다면 각 플롯은 이러한 원인을 진단하는데 도움을 준다. 그리고 이 옵션의 단점은 k의 증가가 아주 적은 값일지라도 *.prs file이 매우 커진다는 것이다. Save only worst linear function of parameters : 이 옵션은 가장 심한(격한) 선형 함수의 모수를 *.prs 파일에 저장하는 것이다. 단지 이 함수에 대한 시계열과 자기상관 플롯을 검토할 수 있다. 이 파일은 실행 후 자동 출력되지 않으므로 메뉴 series에서 *.prs 파일을 불러낸 후 다시 Series 항목의 plot을 선택하면 Worst linear function과 Autocorrelation function의 시계열이 나타난다. 이것은 Series 옵션에서 Save only worst linear function을 선택하여 얻을 수 있다.

(4) Impute from parameter sheet

Impute from parameters는 NORM의 마지막 단계로서 모수로부터의 대치를 위한 과정이다. 이 화면에서는 NORM parameter(*.prm) 파일과 NORM parameter series (*.prs) 파일로부터 얻어진 모수값에 의한 결측값을 위한 랜덤 대치를 생성한다. 이 과정을 실행하면 두 개의 파일이 생성되는데, 하나는 실행결과의 output(*.out) 파일이고, 다른 하나는 대치된 데이터집합을 포함하고 있는 대치(*.imp) 파일이다. 결과 파일에서는 결측값에 대치된 데이터를 이용하여 얻은 평균, 분산, 공분산 등의 값을 나타낸다. NORM의 많은 사용자는 DA를 실행하는 동안 적합한 다대치가 생성되기 때문에 이 과정을 사용할 필요는 없다. 그림 5에 있는 Run 버튼을 클릭하면 parameter(*.prm) file 또는 parameter series(*.prs) file로부터 얻어진 모수를 이용하여 랜덤 대치를 생성한다. Computing 버튼은 난수 생성을 위해 보존되어 있는 초기값(seed)을 설정할 수 있다.

2.2 NORM의 사이트

대치된 데이터집합을 확인하기 위해서는 메뉴의 display의 imputation항을 이용하여 imp.out 파일을 열면 그림 6과 같이 그림 1에서 결측되어 있는 부분이 대치되어 있음을 알 수 있다. 이때 파일명은 default일 경우는 데이터 파일명_n.imp로 된다. n은 다대치의 대치 횟수를 의미한다. 다대치에 대한 NORM의 분석 결과는 표 3.2에 나타나듯이 유용성이 높음을 알 수 있다.

NORM은 결측값이 있는 데이터의 분석을 다루는 과학적 연구자를 위한 서비스 차원에서 만들어졌다. 마지막으로 NORM은 정규모형 하에서 불완전 다변량 데이터의 다대치용 소프트웨어로서, 다음의 URL http://www.stat.psu.edu/~jls/misoftwa.html에서 무료로 내려 받을 수 있다.

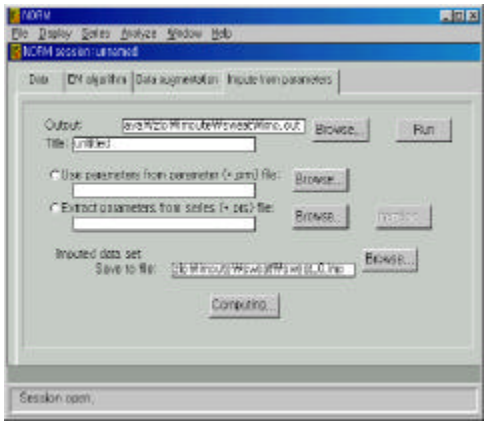

그림 5. Impute from parameters sheet

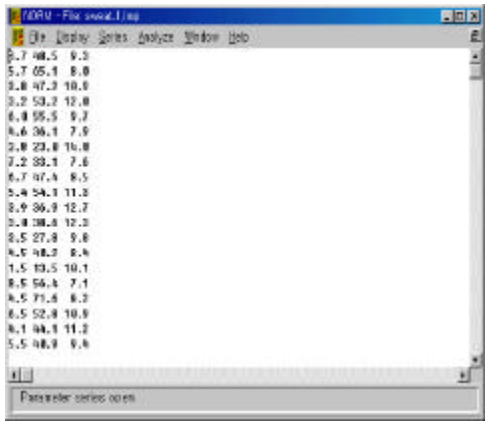

그림 6. 결측값의 대치

또한 결측자료에 대한 보다 상세한 설명은 Tanner and Wong(1987), Schafer (1997a, 1997b, 1999)와 Schafer and Olsen(1998)을 참고하기 바란다.

3. 수치 예와 토론

수치적인 예로서 여성 건강조사의 세 변수(X_1: sweat rate, X_2: sodium content, X_3: potassium content)에 대한 20명의 자료($p = 3$, $n = 20$)에 적용시켜 보았다. 이 자료는 표 3.1과 같으며 결손값에 관한 연구를 위해 세 변수의 상호관계를 고려하여 두 변수에 5개의 결손값을 만들었는데, *가 붙어있는 값이 결손값을 의미한다.

먼저 미지 분포에 대한 공분산 행렬의 최대고유값을 추정한다고 가정하자(Efron, 1994).

$$\theta = \text{maximum eigenvalue of } \hat{\Sigma}$$

여기서 $\hat{\Sigma}$는 어떤 분포로부터 유도된 공분산 행렬이다.

이와 같은 결측데이터가 있는 경우 먼저 가장 일반적인 방법으로 결측 개체를 생략하여 관측된 개체만을 이용하여 얻은 추정값은 $\hat{\theta} = 193.70$, 결측변수의 관측 자료만을 이용한 샘플링(hot-deck)으로 대치한 경우 모수 추정값은 $\hat{\theta} = 173.84$이었다. 또한 결측변수의 관측된 데이터의 평균값을 대치(mean imputation)하여 $\hat{\theta} = 170.40$을 얻었다. 관측된 개체만을 이용하여 얻은 추정값은 다른 두 방법보다 실제 추정값에 근사하였는데, 이는 결측 변수의 자료값들 중 낮거나 높은 값들이 결측되어 비교적 낮은 분산을 보이고 있기 때문이다. 실제의 경우는 이러한 결측 정보를 얻을 수 없으므로 결측데이터에 대한 대치적 방법으로는 부적합하다. 위의 세 가지 방법의 추정값들은 관측된 자료만을 이용한 경우인데 실제 추정값 $\theta = 200.46$보다 모두 낮게 추정되었다.

또한, 추정(estimation)과 최대화(maximization)의 두 단계를 반복하는 계산인 EM 알고리즘(Dempster et. al., 1977)을 이용한 결측 대치법의 추정값은 179.15이었다. 이 방법을 보완하여 EM 알고리즘에 의해 대치된 데이터 세트에 대한 모수를 초기화하여 NORM의 DA를 2.1절에 따라 실행하면 다음과 같다.

표 3.1. Sweat Data

	X_1	X_2	X_3
Individual	Sweat rate	Sodium	Potassium
1	3.7	48.5	9.3
2	5.7	65.1	8.0
3	3.8	47.2	10.9
4	3.2	53.2	12.0
5	3.1*	55.5	9.7
6	4.6	36.1	7.9
7	2.4*	24.8*	14
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3
11	3.9	36.9	12.7
12	4.5*	58.8*	12.3
13	3.5	27.8	9.8
14	4.5	40.2	8.4
15	1.5	13.5	10.1
16	8.5	56.4	7.1
17	4.5	71.6	8.2
18	6.5	52.8	10.9
19	4.1	44.1	11.2
20	5.5	40.9	9.4

Source : *Johnson and Wichern(1992).*

표 3.2 대치 방법에 따른 추정값

방법	$\hat{\theta}$
실제값	$\theta = 200.46$
관측값만을 이용	193.70
hot-deck	173.84
평균값 대치	170.40
EM	179.15
NORM	196.58

NORM 실행 단계

1단계(Data). 결측값 코드 "-9"를 포함한 데이터가 입력된 파일을 그림 1과 같이 연다. 다음으로 Variables 시트에서는 VAR_1을 더블 클릭하여 각 해당 변수명을 입력하고, Summarize 시트로 가서 출력될 파일명을 정하고 Run을 실행한다.

2단계(EM). 특정한 path를 설정하지 않는다면 default로 두고 Computing 버튼을 눌러 반복횟수와 추정방법을 선택한다(그림 3 참조).

3단계(DA). Computing의 분포선택, Imputation의 방법과 Series의 저장방법을 선택한 후 Run을 실행한다(그림 4 참조).

4단계(Impute). 생성된 impute 파일을 생성하기 위해 Run을 실행한 후 메뉴의 Display에서 impute file을 선택하여 생성된 *_n.imp 파일을 열면 그림 6과 같이 대치된 파일을 n개 볼 수 있다. 이 데이터 세트들의 각 추정값을 구하여 평균한 값이 $\hat{\theta}$가 된다.

이와 같이 NORM을 이용하여 모수의 최종적인 값이 생성되고 이로부터 얻은 다대치의 추정값은 196.58이었다. 이 값은 결측성분을 EM과 DA의 두 과정을 이용함으로써 실제 추정값에 훨씬 가까운 값으로 근사함을 알 수 있다.

결측값 대치에 관한 상세한 설명은 김현정 외(2001)을 참고하기 바란다.

또한, 차후의 연구과제로 1절에서 언급했던 NORM이외의 다대치용 S/W인 CAT, MIX, PAN등을 활용한 수치 예에 대한 논의도 유용할 것으로 생각된다.

참고문헌

1. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum-likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series*, B39, 1-38.
2. Efron, B. (1994). Missing Data, Imputation, and Bootstrap, *Journal of the American Statistical Association*, Vol 89, No 426, 463-479.
3. Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis 3rd Ed.*, Prentice Hall
4. Liitle, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, John Wiley and Sons.
5. Orchard, T. and Woodbury, M. A. (1972). A missing information principle: Theory and application, *Proc. 6th Berkely Symposium on Math. Statist. and Prob.* 1, 697-715.
6. Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley.
7. Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall, Chapter 3.
8. Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: a data analyst and perspective. *Multivariate*

Behavioral Research, 33, 545-571.
9. Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1), 3-15.
10. Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82, 528-550.
11. **김현정**, **문승호**, **신재경** (2001). "**결측자료의 대치에 관한 방법적 비교**", ***한국통계학회 2001년 춘계 학술발표회 논문집***, 101-104.

[**2002년 8월 접수**, **2002년 9월 채택**]