Microarray 실험에서 Time course data에 관한 결측치 추정방법들에 대한 비교

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

서 원 열

Microarray 실험에서 Time course data에 관한 결측치 추정방법들에 대한 비교

지도 김 동 기 교수

이 논문을 석사 학위논문으로 제출함

2005년 8월 일

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

서 원 열

서원열의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2005년 8월 일

감사의 글

많은 생각과 방황으로 자리를 잡지못하던 몇 년간의 석사기간이, 여러 사람들의 도움으로 후회와 기쁨을 안은채 이제 종지부를 찍게되었습니다. 앞으로의 과정에서 밑거름이 될 수 있는 하나의 좋은 경험으로 남는 그런 과정이기도 합니다.

학문을 정진함에 있어, 낯설음과 고민으로 방황을 할때, 끊임없는 지도로 방향을 제시해 주신 김동기 선생님께 감사 드립니다. 인자한 얼굴로 새로운 학문인 유전통계학에 관심을 가지게 해주신 임길섭 선생님께도 깊은 감사를 드립니다. 그리고, 부족한 저에게 질책과 관심으로 논문에 많은 도움을 주신 변혜란 선생님께 깊은 감사를 드립니다.

그리고, 학교생활을 하면서 , 온갖 투정을 받아주시고 따뜻한 위안처이자, 버팀목이 되주시는 어머니 , 아버지께 너무 받기만해 죄송한 마음과 감사하다는 말씀을 드리고 싶습니다. 언제나 자상하면서 언제나 힘이되주는 형과 누나, 그리고 형수님도 고맙다는 마음 전하고 싶습니다.

조교생활을 하면서, 힘들때 많은 것을 알려주시고 고민을 들어주시는 송기준 박사님, 우직하면서 싫은표정 안하고, 많은 것을 도와줬던 성민이형, 대학동창이자 아직까지 같은 길을 걸어오며, 고민도 들어주고 힘들때 함께 해주며 도와주던 무영이, 힘들때마다 다가와 술한잔 기울이며 힘이 되주던 봉섭이, 언제나 웃음을 잃지않고 항상 밝게 살아가는 찬미씨, 듣기편한 사투리로 마지막 학기동안 자기일같이 도와주던 미영씨, 착한 미소와 목소리로 편안하게 만들어주고 열심히 할려는 모습이 보기좋은 신영이, 외부생활을 마치고 돌아와 한학기동안 지내왔던, 재미있는 말로 연구실 분위기를 즐겁게 해주는 소연이와 민진이, 싫은소리없이 열심히 배우고 일하는 모습이 보기좋은 성은이, 새로들어와 자기일을 잘해내고 앞으로 같이 지낼 은희와 혜진이한테도 고맙다는 말을 해주고 싶습니다.

그리고, 고등학교 동창이자 같은동네 살면서 힘들때 기쁠때 함께해주던 종만이, 이제는 모임까지 만들어, 평생 힘이되주며 살아갈 친구들인 상은,창희,한주,재일,대현,형선 에게도 참좋은 인연, 기쁘고 감사하게 생각합니다. 마지막으로, 학교생활 힘들때마다, 언제나 찾아와 내게 힘이 되주고, 언제나 내곁에 있어주며 참고맙다는 말밖에 할 수 없는 ,이쁜 은주한테도 곁에있어 참 행복하다는 말을 하고 싶습니다. 저도 언제나 이들에게 힘이되주는 사람이 될 수있도록 노력하고, 이모든 사람들과 언제나 함께 할수 있기를 소망합니다.

2005년 8월

서원열 올림

차 례

그림차례 iii
표차례 iv
국문요약 v
제1장 서론 1
1.1. 연구배경 1
1.2. 연구목적 및 내용 2
제2장 결측치 추정방법 3
2.1. Incremental method 3
2.2. SVD(Singular Value Decomposition) 3
2.3. KNN(K-nearest-neighbor) 7
2.3.1 개체간의 유사성 측도 7
2.3.2 가중 평균에 대한 방법 8
2.4. 회귀분석(regression) 9
2.4.1 회귀계수의 추정 10
2.5 EM(expectaton Maximixation) 알고리즘 11
제3장 이론에 의한 결측치 추정 과정과 순서 14
3.1. SVD 추정과정과 순서 15
3.2. KNN 결측치 추정과정과 순서 16
3.3. 회귀분석 결측치 추정과정과 순서 17
3.4. EM 추정과정과 순서 19
3.5. RMS error (Root Mean Square error) 19
제4장 실제자료에의 적용 20
4.1. 연구계획 20
4.2. 실험자료 21
4.2.1. 비율별로 제거한 결측치 자료 21

4.2.2. 결측치 추정에 의한 데이터 구조.....22
4.3. 제안한 방법의 결측치 추정결과.....23
4.3.1.SVD에 의한 결측치 추정결과.....23
4.3.2. KNN을 이용한 결측치 추정결과.....24
4.3.3. 회귀분석을 이용한 결측치 추정결과.....26
4.3.4. EM 알고리즘을 이용한 결측치 추정결과.....27
제5장 토의 및 결론.....32
참고문헌.....34
ABSTRACT.....37

그 림 차 례

그림 1. SVD 추정에 의한 결측치 비율 별 RMS error.....23
그림 2. KNN 추정에 의한 결측치 비율 별 k에따른 RMS error.....25
그림 3. 회귀분석 추정에 의한 결측치 비율 별 RMS error.....26
그림 4. EM 추정에 의한 결측치 비율 별 RMS error.....30
그림 5. 결측치 추정방법별 RMS error 비교.....31

표 차 례

표 1. 자료의 구조.....21
표 2. 각 시점별 제거된 데이터의 수.....22
표 3. SVD추정에대한 RMS error.....23
표 4. KNN추정에대한 RMS error.....25
표 5. 회귀분석을 알고리즘을 이용한 RMS error.....27
표 6. EM algorithm을 통한 각 결측치에따른 추정치.....28
표 7. EM algorithm을 통한 각 1% 결측치데이타의 covariance 추정치.....29
표 8. EM algorithm을 통한 각 5% 결측치데이타의 covariance 추정치.....29
표 9. EM algorithm을 통한 각 10% 결측치데이타의 covariance 추정치.....29
표 10. EM algorithm을 통한 각 20% 결측치데이타의 covariance 추정치.....30
표 11. EM algorithm을 통한 각 30% 결측치데이타의 covariance 추정치.....30
표 12. EM 알고리즘을 이용한 RMS error.....31

국 문 요 약

Microarray 실험에서 Time course data에 관한 결측치 추정방법들에 대한 비교

유전자칩 실험은, 여러 실험조건 아래 유전자의 표현정도가 큰 형태의 행렬 데이터로 얻어지는데 , 얻어지는 과정에서 자주 결측치가 발생하게 된다. 이런 결측 데이터를 정확도 높게 예측할수 있다면, 완전한 데이터를 요구하는 분석 기법에서 보다 정확한 결과를 얻을수가 있을것 이다. time course 실험 에서의 유전자 데이터에서 회귀 분석과, SVD, KNN, EM 알고리즘을 사용하여 어느것이 가장 적중률이 높은지를 평가하였다. 결측치가 작을경우에는 SVD 알고리즘이 가장 적중률이 높았으며, regression 방법이 가장 낮았다. SVD, regression, EM 알고리즘이 결측치 비율에 대해 적중률이 단조증가하는 양상을 보이고 KNN 방법이 적중률에 대해 차이가 많은 패턴을 보였다.

결측치의 비율이 증가함에 따라, EM 알고리즘이 다른 방법들에 비해, 적중률이 높아 지는 것을 알 수가 있었다.

핵심되는말 : 결측치추정, KNN, 회귀분석, EM algorithm, SVD, time-course data, Incremental 방법

제1장 서론

1.1 연구배경

cDNA 유전자칩(microarray) 기술은 다양한 실험아래 수천종의 유전자(gene)의 표현 정도(expression level)를 알수 있게 해준다. 이에 따른 분석방법들도 다양하게 개발되었는데 ,이로인해 여러조건에 따른 유전자의 발현도와 유사성을 규정짓는 분석이 가능해졌다.

일반적으로 데이터의 유사성에 따라 군집화(clustering)하는 군집분석[1] 과 군집화한 상태에서의 분석기술들[2][3] 분석외의 여러 알고리즘 등이 사용되어 왔다.

이런 유전자칩 실험은, 여러 실험조건 아래 유전자의 표현정도가 큰 형태의 행렬 데이터로 얻어지는데 , 얻어지는 과정에서 자주 결측치가 발생하게 된다. 원인은 불충분한 해상도, 이미지 회손, 또는 슬라이드 상의 먼지나 긁힘등이 이에 해당되고, 그 외에도 이미지의 수치변환 과정에서 기기적 오류도 이에 포함된다. 이런 결과로 인한 데이터는 제거되고 그후 분석에서 제외된다. 이런 결측 데이터를 정확도 높게 예측할수 있다면, 완전한 데이터를 요구하는 분석 기법에서 보다 정확한 결과를 얻을수가 있을것 이다.

현재 많은 결측치 추정이론들이 많이 나와있는데, 분산분석에서의 최소자승법을이용한 방법[4] , 우도비를 이용한방법[5], 판별분석 이론을 이용한 방법[6], SVD (singular value decomposition)를 이용한 방법[7] 등이 있다.

하지만, 실험조건이 반복실험인 time course data일 경우에는 시간에 따른 연관성을 고려해 추정이론을 고려해야 할 것이다. 이 논문에서는 현재로서 가장많이 쓰이는 KNN , SVD , regression 이론과 확장된 이론인 EM 이론에 기반한 결측치 추정법을 시간에 따른 pattern을 고려하여 결측치를 추정하는방법을 소개하고 실질적 데이터에 적용해본다.

1.2 연구목적 및 내용

본 논문의 목적은 네가지 결측치 추정 이론을 유전자칩 실험에서 현재 많이 실행되고 있는, time course 실험 에서의 유전자 데이터에서 어느것이 가장 적중률이 높은지를 비교 평가하는 것이다. 점근적 증가 혹은 감소하는 repeated data의 경우 기존에 single imputation 혹은 독립실험에 대한 결측치 추정방법과는 달리, time course data 에 대한 연관성을 고려하는 각 이론들을 써서 추정을 하여야 한다. 이에 대해 각 이론의 상세한 설명을 소개하고, 이론에 해당되는 프로그래밍에 대한 상세한 순서를 제공한다. 또한, s-plus 함수를 만듬으로서 결측치 추정을 쉽게 수행하기 위함이다. 결측치 추정 방법으로 time course 에 대한 회귀분석과, SVD , EM 알고리즘을 사용하였다.

2장에서는 알고리즘에 대한 내용을 설명하고, 3장에서는 실질적인 결측치 추정 순서를 기술하였다. 그리고 4장에서는 실질적인 time point를 고려해 결측치 추정을 한후, 비교 토론을 하고자 한다.

제2장 결측치 추정방법

2.1 Incremental method

time point 에 대한 비례적 척도로서, Incremental method[8] 를 이용하는데, gene 의 개수 i=1,...,n , time을 j=1,...,k 라고 했을때, time point j 와 time point j+1의 평균의 증가분 D_{ij} 라 하고 이를 계산한다.

Y_{ij}^{*} 를 관찰된 값 , $Y_{ij}^{'}$ 를 missing인 경우의 값 이라고 하면, $D_{ij}^{*}, D_{ij}^{'}$ 를 관찰된 경우와 missing일때의 증가분이 된다. 이 증가분을 이용하여 결측치 값을 대체 할 수있다.

$$\widehat{Y_{i,j+1}} = Y_{ij} + D_{j}^{*} \quad \text{for j=1,...,k-1} \tag{2.1}$$

2.2 SVD (Singular Value Decomposition)

주성분 분석에 이용되는 이론으로서, 먼저 주성분 분석과의 관계를 따져보면 주성분 분석은 데이터의 집합을 정규직교(orthonormal) 벡터들의 선형합으로 나타나는데. Nxp 행렬 X가 차원이 L인 n개의 데이터를 표현한다고 할때, 데이터 $x \in X$ 는 다음과 같은 근사함수를 통해 그 차원을 축소할 수 있다.

$f(x, U) = m + U(U^{T}x)$

여기서 $f(x, U)$는 벡터값을 가지는 함수이고, m은 데이터 x의 평균, 그리고 행렬 U는 열이 정규직교 (d x r) 행렬이다. 사상 z=Ux는 x의 감소 공간 차원으로의 정사영인 것이다.

이러한 벡터 V와 정사영 벡터 z는 SVD를 통해 계산할수 있는데, 이때 Nxp 데이터 행렬 X는 $X = UDV^T$로 나누어질수있는데, microarray자료로서 내용을 설명하면, SVD(singular value decomposition)[9][10]의 이론을 사용함으로서, 모든 유전자를 고유유전자(eigengene)를 행렬형태로 나타낼수 있다. 이를 근사적으로 선형조합을 하여, 서로 직교하는 표현양식(expression patterns)을 집합의 기본 함수로 얻을수있다. 이 함수를 적용하여, 결측치 자료를 추정하는 방법이다.

주성분을 구하는데 이용되며, 계산하기 어려운 다차원자료를 차원을 축약하여, 보다 쉽게 성분을 이용 할 수가있다.

SVD는 N-유전자(gene) x p-배열(array) 으로 구성되는 표현양식을 L-고유배열 x L-고유유전자로 선형적으로 재구성된다. 이는 , $A = UDV^T$ 로 분해되어지며[11], A는 계수(rank)가 p이며, Nxp 표현양식인 행렬을 말하며, U는 (NxL) 행렬 , D는 L-고유유전자(eigengene) x L-고유배열(eigenarray)로 이루어진 대각행렬이고 , V^T 는 LxL 고유유전자와 배열로 이루어진 정방행렬이다. σ_i^2 는 대칭행렬 AA^T의 고유값이 되며, V^T의 행은 AA^T의 고유벡터가된다. 각 원소들의 성질을 보면,

1. $D = diag\{\sigma_1, \sigma_2, \ldots, \sigma_L\}$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r = \ldots = \sigma_L = 0, L = \min(N, p)$

   σ_i 는 singular value 로 불리며, σ_i^2 는 i번째 배열에서의 고유값이된다.

2. $UU^T = VV^T = I$, 각 고유유전자는 모든 다른 고유유전자와 연관성이없다.

3. 고유유전자와 고유배열은 유일값을 갖는다.

행렬D에서 계수가 r이라고 하면 , SVD분해는 ,

$$\widehat{X}_J = U_J D_J V_J^T \text{ , } (J \leq p) \qquad (2.2)$$

로 이루어지며, 여기에서,

$$X = \beta V^T + \varepsilon$$

에대한 선형모형을 구하기 위해서는, 최소자승법 으로 문제를 풀수가 있다.

$$\min_{\beta} ||x - V_j\beta||^2 = \min_{\beta} \Sigma_{i=1}^{p} (x_i - \Sigma_{j=1}^{J} v_{ij}\beta_j)^2 \quad (2.3)$$

이에대한 해는 ,

$$\hat{\beta} = (V_j^T V_j)^{-1} V_j^T x \quad (2.4)$$

로 나타낼수있으며, V_j^T 가 정방행렬이라면, $\hat{\beta} = V_j^T x$ 가 될 수있다. 따라서, 각 행에 대한 회귀계수를 이용하여 공식(1.1) 에 적용하여 결측치 값을 추정 할 수 있다. 결측치가 있는 Xm 데이터에서 역시 이 방법을 이용하여 각 원소들을 구할 수가 있는데, 결측 데이터에서의 β는 ,

$$\min_{\beta} \Sigma_{non-missing} (x_i - \Sigma_{j=1}^{J} v_{ij}\beta_j)^2 \quad (2.5)$$

를 이용하여 구할 수 있다. 여기에서, 고유값의 개수를 정할때는 표현양식의 비율 (fraction of eigenexpression) P_i와 "Shannon entropy"[12] 를 이용할수 있다.

$$P_i = \sigma_i^2 / \Sigma_{k=1}^{L} \sigma_k^2 \quad (2.6)$$

Shannon entropy를 d라고 하면,

$$0 \leq d = \frac{-1}{log(L)} \Sigma_{k=1}^{L} p_k log(p_k) \leq 1 \quad (2.7)$$

$d \fallingdotseq 0$ 이라면 한 개의 고유유전자로 표현이 가능하다는것이고, $d \fallingdotseq 1$ 이라면 모든 고유유전자를 이용해야한다는 의미가 된다.

2.3 KNN(K-nearest-neighbor)

이 이론은 다른 결측치 추정방법과는 달리 특별한 유전자에 대한 영향을 받지 않는다. 계산이 간편하면서도 적중률이 어느정도 높기 때문에 많이 이용되고 있으며, 이방법에 대한 이론[13][14] 은, 우선 결측치를 가진 유전자와 가장 표현이 유사한 유전자 K개를 찾아서 그에 따른 가중치를 부여하여 평균값을 계산하여, 추정하는 이론이다. 유사성을 측정하는 방법은, 개체간의 거리를 이용한다. 거리를 계산하여 가장 유사성이 높은 K개의 유전자 집합을 얻을수 있다.

2.3.1 개체간의 유사성 측도

Minkiwski 거리

$$d_{ij} = d(x_i, x_j) = [\sum_{k=1}^{p} |X_{ik} - X_{jk}|^m]^{1/m}$$

Mahalanobis 거리

$$d_{ij} = d(X_i, X_j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$$

유클리드 거리

$$d_{ij} = d(X_i, X_j) = \sqrt{(X_i - X_j)' (X_i - X_j)} \quad (2.8)$$

이중 mahalanobis 거리는 유클리드에 공분산 행렬을 나누어 보정한 거리이다.

2.3.2 가중 평균에 대한 방법

가중 평균에 대한방법은 거리의 분포에 따라 계산되어지는데, 거리의 분포계수는,

$$\Delta_i = \frac{d_{ij}}{\sum_{x_i \in N_k(x)} d_{ij}} \tag{2.9}$$

로 나타날 수 있다. 구할수있으며, 이에따른 결측치 추정값은,

$$\widehat{Y(x)} = \sum_{x_i \in N_k(x)} \Delta_i y_i \tag{2.10}$$

으로 추정할수있다.

2.4 회귀분석(regression)

자연의 많은 현상들에 있어서 목적변수(결과적인현상)는 많은 관계된 입력변수들에 의해 결정되어진다. 이처럼 하나의 변수가 다른 많은 변수들에 의해 어떻게 설명 혹은 예측되는지를 알아보기 위해 적절한 함수식으로 표현하여 자료 분석을 하는 통계적 방법, 즉 회귀분석은 목적변수를 표현할수 있는 함수를 찾기위한 모델링의 하나의 기법이다[19].

반응 변수 **y** 와 설명변수 x사이에 다음과 같은 관계와 **y**의 기대값을 $\hat{y}$라 할 때,

$$\hat{y}=\beta_0+\beta_1 x \tag{2.11}$$

라고 표현할수 있다. 이와 관련해 독립변수가 여러개일 경우 다중회귀분석을 이용하게 되는데, 다중회귀분석의 기본가정을 살펴보면,

1. 독립변수, 종속변수가 모두 등비이고 변수간의 관계와 설명, 예측 (중요한 인자가 명목이나, 서열이어도 Dummy 변수화 하여 분석할수 있다.)

2. 기본가정으로 데이터가 선형성, 정규성 , 등분산성을 만족해야한다. 잔차는 정규성, 등분산성을 만족한다.

3. 가설은 $\beta_1=\beta_2=\ldots=\beta_n=0$ (선형회귀모형이 존재하지 않는다.) 대립가설은 적어도 하나의 회귀계수는 0 이 아니다. (선형회귀모형이 존재한다.)로 나타나며, 다중회귀모형은,

$$\hat{y}=\beta_0+\beta_1 x_1+\ldots+\beta_n x_n+\varepsilon_i \tag{2.12}$$

와 같이 나타난다. 다중회귀분석은 단순회귀분석의 확장형으로, 다중회귀분석은 다중공선성이라는 개념이 하나 추가되고 , 나머지의 개념은 단순회귀분석과 동일하다.

2.4.1 회귀계수의 추정

β_0와 β_1는 알지못하는 모수로서 β_0와 β_1의 추정치는 잔차의 제곱합이 최소의 값을 가지도록 하여 모수를 추정하는 최소제곱법[14]을 이용하여 추정한다.

$$\widehat{\beta_0} = \bar{y} - \beta_1 \bar{x}\ ,\quad \widehat{\beta} = COV(x, y) / Var(x) \tag{2.13}$$

2.5 EM (Expectation Maximization) 알고리즘

EM 알고리즘은 기본적으로 기대값을 구하는과정과 거기에따른 모수를 극대화 시키는 두가지 과정을 거치게 되는데, 주변 사후 분포, $p(\Theta|y)$ 의 모수를 찾는데 유용하다. EM이면의 기본적 개념[15][16]은 결측 자료(missing data), Ψ 라 하는 모수 벡터를 확대시키는 것이다. 그래서 결과를 얻기 어려운 $p(\Theta|y)$ 를 직접 최대화하기 보다는, 확대 사후 분포 $p(\Theta,\Psi|y)$ 를 다룬다. Ψ에 대해 이용 가능한 몇가지 경우가 있지만, 결정은 주관적으로 이루어진다. 반복과정은 다음을 따른다.

1) E-단계 - 주어진 관측치와 $p(\Theta|y)$의 현재의 모수를 추측하고 ln $p(\Theta,\Psi|y)$의 기대값을 계산한다.
2) M-단계 - 기대 로그 사후 분포인 (1)단계에서 얻은 Θ 의 함수를 Θ에 대해 최대화 시킨다. (1)단계 에서 Θ를 최신값으로 하여 수렴에 도달할때까지 이과정을 반복한다.

여기에서 $p(\Theta|y)$ 에따른 공식은,

$$p(\Theta|y)=\frac{p(\Theta,\Psi|y)}{p(\Psi|\Theta,y)} \tag{2.14}$$

가되고, 여기에따른 로그사후분포는 다음과 같다.

$$\ln p(\Theta|y)=\ln p(\Theta,\Psi|y)-\ln p(\Psi|\Theta,y) \tag{2.15}$$

여기서 $\ln p(\Theta|y)$ 는 로그사후 분포 또는 Θ 의 로그우도 함수이다.

그리고 $\ln p(\Theta, \Psi|y)$ 는 EM에 관련된 문헌에서 완전자료의 로그우도 함수로 언급한다[17].

식 3.1의 양변에 $\ln p(\Psi|\Theta^{[t]}, y)$에 관한 기대값을 구한다. 여기서 $\Theta^{[t]}$ 는 t번 반복했을 경우 추정된 최빈수이다. ($\ln p(\Theta|y)$ 는 Ψ의 함수가 아니다.) 그러면 다음을 얻을수 있는데,

$$\begin{aligned}\ln p(\Theta|y) &= \int \ln p(\Theta, \Psi|y) p(\Psi|\Theta^{[t]}, y) d\Psi - \int \ln p(\Psi|\Theta, y) p(\Psi|\Theta^{[t]}, y) d\Psi \\ &= Q(\theta|\theta^{[t]}) - H(\theta|\theta^{[t]}) \end{aligned} \tag{2.16}$$

$\ln p(\Theta|y)$를 최대화 시키는 값을 찾기 위해, 식 3.3의 우변의 양 항을 이용 할 수 있는데, EM 알고리즘은 실질적으로 $H(\Theta|\Theta^{[t]})$ 항은 무시하고 첫번째 항 $Q(\Theta|\Theta^{[t]})$ 만을 풀어나간다. EM 알고리듬의 두 번째 단계는 E-단계에서 $Q(\Theta|\Theta^{[t]})$ 를 계산하고, M-단계에서 Θ 에 관해 반복적인 방법으로 $Q(\Theta|\Theta^{[t]})$를 최대화하는 과정이다. 이후에 보여지는 것처럼, 반복계산은 $\ln p(\Theta|y)$ 를 단조 증가시킨다. 즉, $\ln p(\Theta^{[t+1]}|y) > \ln p(\Theta^{[t]}|y)$ 이다.

주변 사후 분포 밀도 함수, $p(\Theta|y)$ 각 단계에서 증가함에 따라, EM 알고리즘에 의해 (거의 예외없이) 지역 최빈수에 수렴한다.

$\Theta^{[0]}, \Theta^{[1]}, \ldots, \Theta^{[t+1]}$ 로 반복이 진행된다고 할 때 반복 간에 나타나는 $\ln p(\Theta|y)$의 값의 차이가 다음과 같이 주어졌다고 하자.

$$\begin{aligned}\ln p(\Theta^{[t+1]}|y) - \ln p(\Theta^{[t]}|y) &= [Q(\Theta^{[t+1]}|\Theta^{[t]}) - Q(\Theta^{[t]}|\Theta^{[t]})] \\ &\quad - [H(\Theta^{[t+1]}|\Theta^{[t]}) - H(\Theta^{[t]}|\Theta^{[t]})] \end{aligned} \tag{2.17}$$

모든 Θ 에 대해 $H(\Theta^{[t]}|\Theta^{[t]})$와 $H(\Theta|\Theta^{[t]})$의 차이는 다음과 같다.

$$H(\Theta^{[t]}|\Theta[t]) - H(\Theta|\Theta^{[t]}) = \int f(\Psi|\Theta^{[t]}, y) f(\Psi|\Theta^{[t]}, y) d\Psi$$

$$- \int \ln f(\Psi|\Theta, y) f(\Psi|\Theta^{[t]}, y) d\Psi$$

$$= \int - \ln \left[\frac{f(\Psi|\Theta, y)}{f(\Psi|\Theta^{[t]}, y)} \right] f(\Psi|\Theta^{[t]}, y) d\Psi$$

$$= - E[\ln g(\Psi)]$$

어떤 볼록 함수의 상태를 말하는 Jensen 의 부등식, $E[f(x)] \geq f[E(x)]$ 을 적용하면, 다음과 같다.

$$H(\Theta^{[t]}|\Theta^{[t]}) - H(\Theta|\Theta^{[t]}) \geq - \ln[E(g(\Psi))]$$

$$= - \ln \int \left[\frac{f(\Psi|\Theta, y)}{f(\Psi} |\Theta^{[t]}, y) \right] f(\Psi|\Theta^{[t]}, y) d\Psi = - \ln(1) = 0$$

그러므로 $H(\Theta^{[t]}|\Theta[t]) - H(\Theta|\Theta^{[t]}) \geq 0$ 이다. EM 은 Θ 의 함수인 $Q(\Theta|\Theta^{[t]})$가 Θ 의 함수로 최대화 되도록 $\Theta^{[t+1]}$ 을 선택하므로 식 (3.4)의 첫 항은 매 반복마다 증가한다.

그리고 두 번째 항은 음수 또는 0이므로 , $\ln(\Theta^{[t+1]}|y) > \ln f(\Theta^{[t]}|y)$ 일반화된 EM 알고리듬(GEM)[15]을 이용할 수도 있는데 이것은 전체적으로 $Q(\Theta|\Theta^{[t]})$를 전역 최대화 시키는 Θ 값을 찾기 보다는 $Q(\Theta^{[t+1]}|\Theta^{[t]}) \geq Q(\Theta^{[t]}|\Theta^{[t]})$ 가 되도록 $\Theta^{[t+1]}$ 을 찾는다.

제3장 이론에 의한 결측치 추정 과정과 순서

3.1 SVD 추정과정과 순서

데이터 X에서 결측치가 없는 데이터부분은 Xc, 결측치가 하나라도 있는 부분을 Xm라고 하자. 이제 Xc의 SVD에 의한 선형관계를 알기위해 식(1.1)을 적용 하면,

$$Xc = U_J D_J V_J^T \quad , \ (J \leq p)$$

와 같이 Xc에관한 계수 p, U는 (NxL)행렬 , D는 L-고유유전자(eigengene) x L-고유배열(eigenarray)로 이루어진 대각행렬이고 , V^T 는 LxL 고유유전자와 배열로 이루어진 정방행렬표현양식을 얻을수 있다[14].

여기에서 X에 관한 V^T의 선형관계를 알기위해 식(2.2)을 적용하여 최소자승법으로 선형모형을 구할 수가 있는데, 먼저 주성분에 따른 순위(rank)결정은 고유유전자 결정계수인 식(2.4)와 (shannon entropy) 식(2,6)를 이용하여 주성분에 따른 유전자의 표현양식을 얻어 설명력에 따른 순위를 결정한 뒤, 식(2.3)을 적용. 이에 따른 Xm의 선형회귀계수를 얻을수 있다.

여기에서 회귀계수를 이용하여 Xc에 관한 결측치 값을 추정할수 있는데, 여기에서 Xm에따른 선형모형은 Xm과 유사한 형태로,

$$\min_{\beta} \sum_{i \, non-missing} (x_i - \sum_{j=1}^{J} v_{ij} \beta_j)^2$$

와 같이 결측치를 추정할수 있는데, 과정을 살펴보면, 먼저 Xc 에따른 V_j에서 결정된 순위에 따른 고유값에 따라 제거된 V_j를 V_j^*라고 하자. 그러면 Xm에

따른 $\hat{\beta}$는,

$$\hat{\beta} = (V_J^{*T} V_J^*)^{-1} V_J^{*T} x^*$$

으로 정해진다. 여기에서 문제는 결측치 때문에 ,유전자 i 번째에서의 j배열값이 결측치라면, 그에 해당되는 k고유유전자의 j번째 값이 계산이 되질 않는다. 그렇기 때문에 계수를 계산하기에 앞서 결측치값을 초기치로 각 행에 따른 평균값으로 대체하는게 필요하다.

$\hat{\beta}$를 얻었다면 앞서 계산된 V_j^*을 이용하여 , 결측치값을 $\hat{X} = V_J^* \hat{\beta}$ 으로 대체한다. 그리고 초기치값을 $\hat{X}$값으로 대체후 과정을 반복하여 정해진 오차이내에 이를때까지 과정을 반복하여 최종 $\hat{X}$값을 얻을수 있다.

(1) 결측치값을 각 유전자의 time course 연관성에 대한 식(2,1)를 적용하여 time 에 대한 비례적 값으로 대체

(2) Xc에 대한 SVD를 계산후 선형모형을 구한후 Xm에 적용하여 결측치 대체.

(3) 과정2에서 얻어진 값에대한 $||M^i - M^{i+1}||/||M^i||$ 의 값이 정해진 오차 아래가 될 때까지 과정(2)를 반복

3.2 KNN 결측치 추정과정과 순서

우선 결측치를 가진 유전자와 가장 표현이 유사한 유전자 K개를 찾기 위해서 유사성 측도인 거리함수를 사용해야하는데, 초기치에 영향을 받지않을 유클리드 거리를 선택하는 것이 가장 적절하나 time point 간의 비례적관계를 고려하기위해 가중치를 부여한 초기치를 선택한다.

그에 따른 가중치를 부여하여 추정하는 이론이다. 유사성을 측정하는 방법은, 개체간의 거리를 이용한다. 거리를 계산하여 가장 유사성이 높은 K개의 유전자 집합을 얻을수 있다.

식(2.1)를 적용하여 유사성을 측정한 뒤 순위에 따른 순위계수 식(2.2)를 적용하여 보정결측치값을 얻기위해, 결측치값이 i번째 유전자의 j번째 실험이라면 K개의 순위에 해당하는 유전자 j번째 값들과 순위계수를 이용해 식(2.3)대입하여 결측치값을 보정할수 있다.

(1) 유사한 유전자의 개수 K를 선택하고, time course 연관성에 대한 식(2,1)를 적용하여 time 에 대한 비례적 값으로 삽입한다.

(2) 모든 유전자에 대한 유사성을 측정하고, 가장 유사한 K개를 선택.

(3) 거리에 따른 가중치를 부여한 평균값을 결측치에 삽입한다.

3.3 회귀분석 결측치 추정과정과 순서

이 방법은 회귀분석의 회귀계수를 이용하는 방법으로 유전자 i의 결측치 j값을 얻기위하여 다른 유전자의 j행에 대한 다른행의 관계를 구명하여 유전자 i에 적용하는 방법이다.

incremental method 의 D값을 이용해 다음과 같은 regression model을 구할수 있다.

$$D_{ij} = D_{.j} + A_j(Y_{ij} - Y_{.j}) + \epsilon_{ij} \tag{3.1}$$

$D_{.j}$ 는 sample 에 대한 평균값이 된다. 식 (3.1)를 사용하여 관찰된 data에 적용하게 되면,

$$D^*_{ij} = D^*_{.j} + A^*_j(Y_{ij} - Y^{**}_{.j}) + \epsilon^*_{ij} \tag{3.2}$$

$Y^{**}_{.j}$는 $\{Y_{ij} | Y_{i,j+1} =$관찰된값$\}$ 경우의 평균값으로 한다. A^*_j를 least square method 를 이용하여 추정하면 결측치에 적용하면,

$$\widehat{D_{ij}} = D^*_{.j} + A^*_j(Y_{ij} - Y_{.j}) \tag{3.3}$$

가 되고, 식(3.3)를 이용하여,

$$\widehat{Y_{i,j+1}} = Y_{ij} + \widehat{D_{ij}} \quad \text{for j=1,...,k-1} \tag{3.4}$$

결측치값을 구할 수 있다.

미지의 모수 β_0, β_1 를 추정하기 위해서 식(2.13)을 적용하여 추정할수 있으며, 이는 유전자들의 실험상의 패턴이 다르다라는 가정하에 이루어지며 한유전자의 결측치값이 많을 경우 계산이 불가능하므로 초기값으로 식(3.3)값을 초기값으로 대치한다.

(1) time course 연관성에 대한 식(3.3)를 적용하여 time 에 대한 비례적 값으로 삽입한다.

(2) 결측 유전자 i를 제외한 나머지 유전자의 D값을 이용해 나머지열의 회귀분석을 적용시킨다.

(3) 회귀계수를 이용하여 비례적 관계를 고려하여 결측치값에 적용시켜 값을 사용한다.

3.4 EM 추정과정과 순서

이 이론을 적용하기 위해서는 먼저 complete data에서의 평균과 분산값을 이용을 한다. 추정치를 초기값으로 하여 최대우도 함수를 이용한다.

EM 추정법은 초기치에 민감하기 때문에, time course 연관성 수치를 구하기 위한 식(3.3)를 적용하여, time 에 대한 비례적 값을 삽입후, 그에따른 μ 와 Σ를 초기치로 설정한다.

(1) time에 대한 비례적 값을 삽입한후 complete data에서 μ 와 covariance를 estimate한다.

(2) $E(y_{imiss}|y_{obs}, \hat{\mu}, \hat{\Sigma})$ and $Cov(y_{miss}|y_{obs}; \hat{\mu}, \hat{\Sigma}), i = 1, 2, \ldots, N$ 를 계산한다.

(3) 얻어진 새로운 μ 와 Σ 를 가지고 step 1,2를 반복한다.

3.5 RMS error (Root Mean Square error)

결측치를 추정후에 적중률에 대한 척도로서 RMS error를 사용한다[14].

$$RMSerror = \sqrt{\frac{(y - \hat{y})^2}{N}} \tag{3.5}$$

여기서 **y**는 결측치를 제거하지않은 데이터가 되고, $\hat{y}$는 결측치 보정후 데이터가 된다.

제4장 실제 자료에의 적용

4.1 연구계획

결측치 추정법인 SVD, KNN, 회귀분석, 그리고 확장된 EM알고리즘을 가지고, 실제 연구하여 수집된 자료로 각각의 추정과정과 각 이론에 의해 추정된 결과의 적중률을 비교해보고자 한다. 먼저 수집된 자료에서 기기에 의한 오류는 제거를 하고 random하게 결측치를 1%,5%,10%,20%,30%로 결측치 비율을 증가시켜가며 제거하여 ,각 알고리즘 별로 결측치를 추정하여 제거하지않은 row-data 와 비교를 하여 각각의 경우의 적중률을 비교해보고, 최종적으로 여러조건의 실험상에서의 결측치 추정에 어느 이론이 가장 좋은지를 알아보고, 각 이론별로 장단점을 알아본다.

4.2 실험자료

본 연구에서 실험에 사용한 자료는 (표1)과 같다.

표 1. 자료의 구조

자료출처	stanford 대학에서실험한 Yeast *Saccharomyces cerevisiae gene*	
자료내용	유전자칩 실험에서 Yeast *Saccharomyces cerevisiae gene*에 대해 발현추이를 10,30,50,70,80,90,100분 별로 7번 반복실험한 time course data	
자료크기	유전자 데이터 8830개	
입력변수	ch2b	GREEN에 대한 background
	ch2i	GREEN에 대한 foreground
	ch1b	RED에 대한 background
	ch1i	RED에 대한 foreground
	⋮	⋮
	ch7b	GREEN에 대한 background
	ch7i	GREEN에 대한 foreground
	ch7b	RED에 대한 background
	ch7i	RED에 대한 foreground

4.2.1 비율별로 제거한 결측치 자료

앞서 제시한 에 의해 제거 비율에 의해 비율을 증가시켜가며 제거되는 유전자의 수를 살펴보았다. 제거의 비율은 약 1% 5% 10% 20% 30% 유전자가 제거될 때까지 시행하였다.

표 2. 각 시점별 제거된 데이터의 수

time / 제거된 비율	10분	30분	50분	70분	80분	90분	100분
1%	99	79	99	84	100	66	84
5%	459	466	475	451	443	436	461
10%	894	912	841	867	864	967	902
20%	1824	1786	1774	1773	1735	1815	1693
30%	2435	2548	2468	2571	2415	2348	2848

선택된 데이터에 대해 정규화를 시행하기위해, RED 와 GREEN의 비율에 로그 형태로 변환해준, $m = log_2 \frac{GREEN_{fg} - GREEN_{bg}}{RED_{fg} - RED_{bg}}$ 형태로 변환하였다. 여기서 결측치 추정이론은 이 m값을 이용하게된다.

4.2.2 결측치 추정에 의한 데이터 구조

결측치 추정이론에 의해 데이터를 결측치 여부에 따라 구분하여야만한다. 그래서 다음과 같이 데이터 분할을 시행하였다.

Xc : 필터링에 의해 제거되는 유전자가 ch1~7까지 하나라도 제거되지않은 완전한 데이터 집합

Xm : ch1~7까지 적어도 한 row에서 하나라도 제거된 결측치를 가지고 있는 데이터 집합

4.3. 제안한 방법의 결측치 추정결과

4.3.1. SVD에 의한 결측치 추정결과

각 데이터에서 결측치가 1% ,5% ,10% ,20% ,30%인경우에 SVD추정방법으로 분석을행하여 그에따른 RMS를 구하여 보았다.

결측치값을 Incremental method를 이용해, 비례적 값으로 대체후, $||M^i - M^{i+1}||/||M^i||$가 0.000001보다 작아질때까지 Xc에대한 선형모형을 Xm에 적용시킨 결과는 아래와 같다. 예상대로 30%가 가장 큰 error를 가지고 조금씩 증가하는 패턴을 보이는 것을 알수가 있다.

표 3. SVD추정에대한 RMS error

missing data 결측비율	RMS error
1%	0.2525791
5%	0.3436622
10%	0.3622532
20%	0.4125123
30%	0.4435335

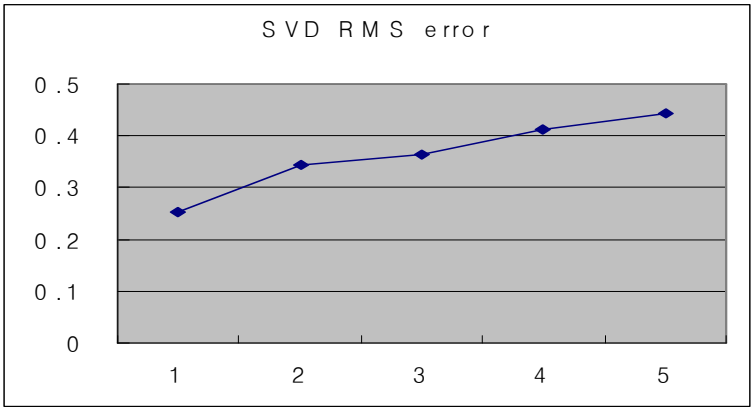

그림 1. SVD 추정에 의한 결측치 비율 별 RMS error

4.3.2. KNN을 이용한 결측치 추정결과

결측치값을 time point 에대한 비례적 값으로 대체후, 유사한 유전자의 개수인 k를 3,5,7,9,11 로 정하고 유클리드 거리를 측정하여 가장가까운 k개를 골라 유사성을 측정하여, knn imputation을 행하였다. 각 결측비율별로 k를 달리하여 RMS error를 측정한 것은 (표4)와 같다. k가 3일때에 큰변화가 없기는하지만, 대체로 k개가 5인경우가 RMS error가 줄어드는 양상을 보이고 있고, 5%일때는 제외하고는 가장 적중률이 가장 높은 것으로 나타난다. 그리고 결측비율이 30%일경우에 크게 늘어 났다.

표 4. KNN 추정에대한 RMS error

missing data 결측비율	k의 개수	RMS error
1%	3	0.3353235
	5	0.3435435
	7	0.3735353
	9	0.3946346
	11	0.4135326
5%	3	0.3447633
	5	0.3984552
	7	0.4354733
	9	0.4435338
	11	0.4425620
10%	3	0.3646312
	5	0.3335234
	7	0.3835313
	9	0.4247324
	11	0.4768547
20%	3	0.4132535
	5	0.4094363
	7	0.4032652
	9	0.4536213
	11	0.4622355
30%	3	0.4942652
	5	0.4624742
	7	0.4936421
	9	0.5843621
	11	0.5946422

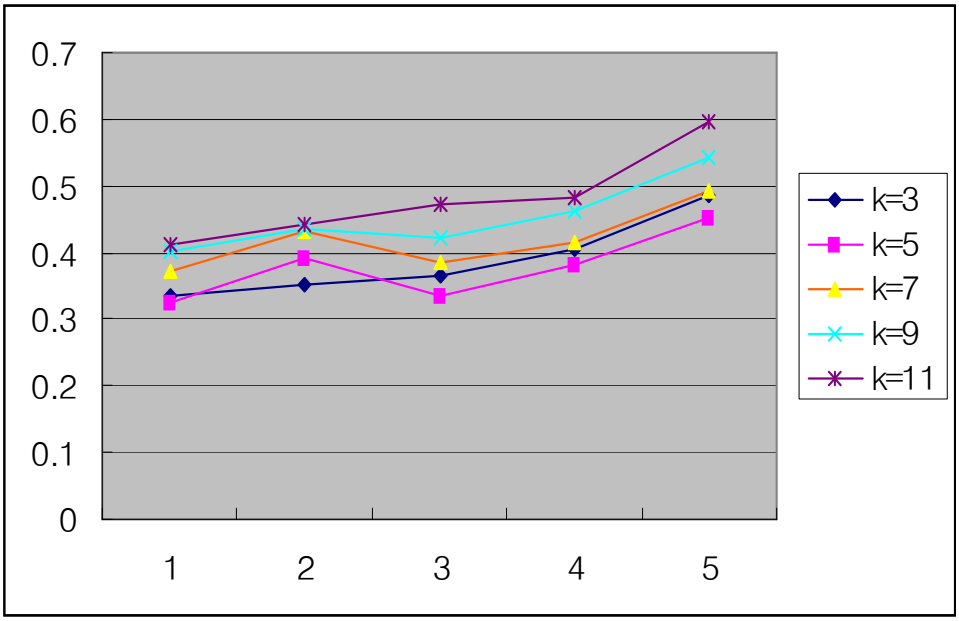

그림 2. KNN 추정에 의한 결측치 비율 별 k에따른 RMS error

4.3.3. 회귀분석을 이용한 결측치 추정결과

결측치값을 Incremental method를 이용해 비례적값으로 대체후, 비례적 값에 대한 missing row가 있는 각각 row별로 그 row를 제외한 나머지 유전자의 row의 비례적 값에 대한 회귀분석을 시행한 beta값을 산출하였다. (식3-5)을 이용하여 추정한 RMS error 는 (표5)와 같다.

표 5. 회귀분석을 알고리즘을 이용한 RMS error

missing data 결측비율	RMS error
1%	0.3235233
5%	0.3523363
10%	0.3657331
20%	0.3795218
30%	0.4385411

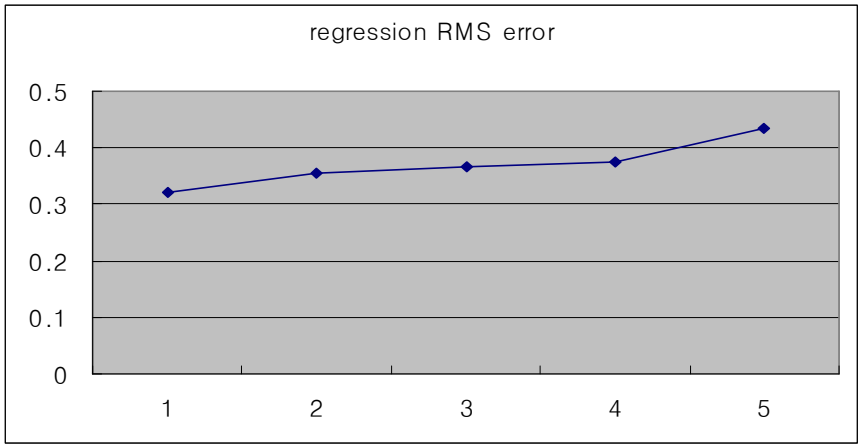

그림 3. 회귀분석 추정에 의한 결측치 비율 별 RMS error

4.3.4. EM 알고리즘을 이용한 결측치 추정결과

EM 알고리즘을 수행하기위해 time에 대한 비례적 값을 삽입한후 complete data에서 μ 와 covariance estimate 한후 M-step 에 따른 최대우도함수를 이용하여 iteration=10000 하여 나온 $E(y_{imiss}|y_{obs}, \hat{\mu}, \hat{\Sigma})$ $Cov(y_{miss}|y_{obs}; \hat{\mu}, \hat{\Sigma})$ 추정치는 (표 6)와 같다.

결과에 따른 RMS error를 구해본결과 (표 12)에서 1%~30% 로 갈수록 단조증가하는 양상을 보이는 것을 알수 있다.

표 6. EM algorithm을 통한 각 결측치에따른 μ 추정치

time / 제거된 비율	10분	30분	50분	70분
1% μ	0.00935	0.00083	0.00006	-0.00054
5% μ	0.00033	-0.00062	-0.00531	-0.00062
10% μ	0.00272	0.00305	-0.00251	-0.00135
20% μ	-0.00252	-0.00412	0.01076	0.00205
30% μ	0.00101	0.00374	-0.00183	0.00395

time / 제거된 비율	80분	90분	100분
1% μ	0.00003	0.00006	-0.00047
5% μ	0.00272	-0.00063	0.00161
10% μ	-0.00295	0.00164	-0.00105
20% μ	-0.00285	-0.00215	-0.00261
30% μ	0.00507	0.00029	-0.00076

표 7. EM algorithm을 통한 각 1% 결측치데이타의 covariance 추정치

1.00E+00	5.75E-01	4.23E-01	2.21E-01	1.26E-02	2.36E-01	2.72E-02
	1.00E-01	6.14E-01	1.68E-01	4.34E-02	6.80E-02	7.31E-02
		1.00E+00	4.05E-01	1.31E-01	3.03E-01	-2.51E-02
			0.99E-01	3.21E-01	5.55E-01	1.59E-01
				1.00E+00	1.95E-01	6.25E-01
					1.00E+00	7.12E-02
						1.00E-01

표 8. EM algorithm을 통한 각 5% 결측치데이타의 covariance 추정치

0.99637	0.595508	0.450815	0.232318	0.016367	0.277255	0.027935
	1.000118	0.604562	0.168017	0.045109	0.068758	0.075593
		1.00853	0.405454	0.133178	0.300705	-0.01968
			1.009495	0.319282	0.560379	0.16443
				0.998989	0.184176	0.631368
					0.999245	0.079673
						0.998874

표 9. EM algorithm을 통한 각 10% 결측치데이타의 covariance 추정치

1.000394	0.601351	0.454632	0.230145	0.012426	0.303515	0.017351
	0.994921	0.547425	0.262223	0.025321	0.093574	0.088129
		0.996943	0.403531	0.134831	0.311953	-0.0224
			1.009237	0.324116	0.5525	0.161824
				1.001152	0.17865	0.632215
					0.998335	0.075124
						1.00166

표 10. EM algorithm을 통한 각 20% 결측치데이타의 covariance 추정치

1.003065	0.645126	0.489314	0.281515	0.000262	0.351354	0.019353
	0.999357	0.673653	0.174476	0.031727	0.072675	0.098416
		0.992274	0.416437	0.119066	0.310705	-0.037816
			0.992224	0.216484	0.53743	0.126423
				1.008138	0.183525	0.524533
					1.004107	0.035351
						1.006723

표 11. EM algorithm을 통한 각 30% 결측치데이타의 covariance 추정치

1.007611	0.603782	0.462639	0.228081	0.012397	0.277275	0.019127
	0.98859	0.597766	0.165418	0.04469	0.061632	0.080639
		0.991608	0.407959	0.135752	0.31051	-0.01244
			1.004788	0.314489	0.541495	0.172616
				1.002196	0.176121	0.633366
					0.991879	0.084082
						0.989287

표 12. EM 알고리즘을 이용한 RMS error

missing data 결측비율	RMS error
1%	0.2903512
5%	0.3236213
10%	0.3385153
20%	0.3544361
30%	0.3953358

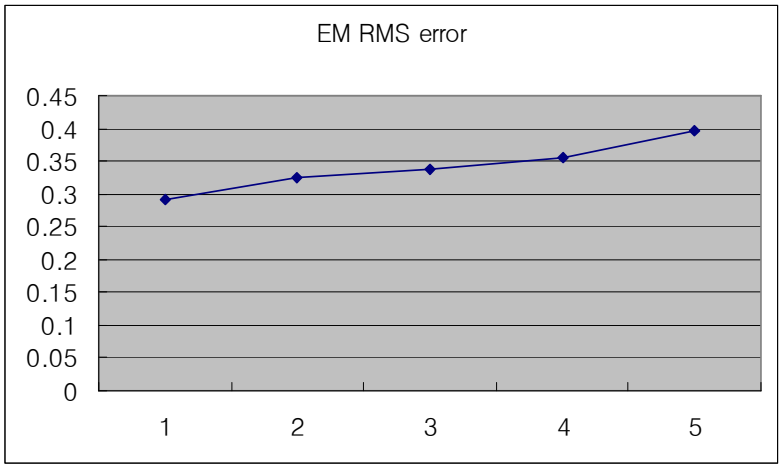

그림 4. EM 추정에 의한 결측치 비율 별 RMS error

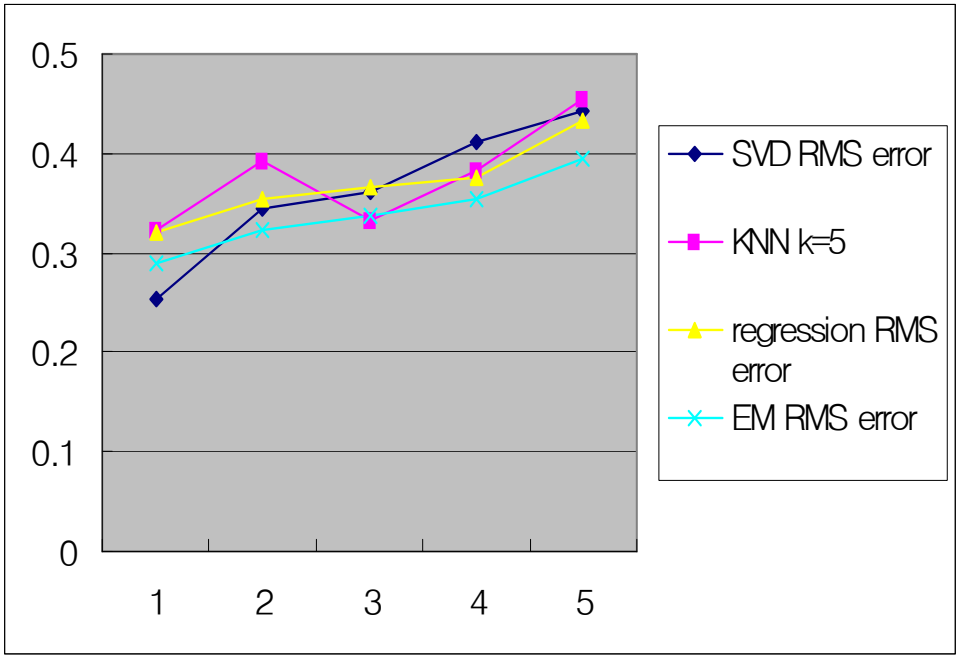

그림 5. 결측치 추정방법별 RMS error 비교

5. 토의 및 결론

지금까지 time course data 에대한, 결측치 추정방법 4가지 이론을 비교하여 보았다. 결측치가 제일작은 1%인 경우 SVD 알고리즘이 가장 적중률이 높았으며, regression 방법이 가장 낮았다. SVD, regression, EM 알고리즘이 결측치 비율에 대해 적중률이 단조증가하는 양상을 보이고 KNN 방법이 적중률에 대해 차이가 많은 패턴을 보였다.

KNN의 경우 k=3인경우가 낮은 결측치에서는 대체로 낮고 적중률 감소가 단조로운 양상을 보였으나 k=5인경우가 대부분에서 가장 RMS error 가 낮은경향을 보여 k=5인경우를 채택하였다.

k=5인경우만 따로 선택하여 다른 방법들을 비교하였는데, 이경우에 결측치가 10%인 경우에 가장 낮았으며, 대체로 1%를 제외한 나머지 비율에 대한 결측치에 대해 EM 알고리즘이 비율이 커짐에 따라 점차 적중률의 떨어짐이 가장적어 30%에서는 가장 높은 적중률을 보이고 있다. microarray gene expression 실험의 경우 실험혹은 filtering에 의한 결측치가 많은 것으로 미루어보아 EM알고리즘이 더 합리적이라고 생각이든다. 이 논문에서는 다루지 않았지만 KNN에서 결측된 gene과 유사한 gene 들을 뽑아 이들의 모수를 EM 알고리즘의 초기치에 적용하면 더욱 적중률이 높아지는 결과를 얻을수 있을거라고 생각한다. 혹은 SVD를 적목시키는 방법도 생각해볼 수 있다.

그리고, 본실험의 data는 시간에 따른 gene 발현도를 측정한 data이다. 그래서 단조로운 증가 혹은 감소하는 양상을 보이는 data의 일종이라고 할 수 있다.

따라서, 이결과가 repeated data 에 관해서만 파생되는 결과일지도 모른다는 생각을 할수 있다. 그룹이 나뉘어졌다던지 , 독립적인 실험인 경우 에서는 다른 양상을 보일수도 있다. 여러 가지 data를 이용하여 분석을 하지못한 아쉬움이 남는다.

참 고 문 헌

[1] Eisen MB, Spellman PT, and Brown PO, et al,. Clustering analysis and display of genome-wide expression patterns. *PNAS*,1998;95:14863-14868

[2] Heyer LJ, Kruglyak S, and Yooseph S. Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome*,1999;9:1106-1115

[3] Tamayo P, Slonim D, and Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps:Methods and application to hematopoietic differentiation. *PNAS*,1999;96:2907-2912

[4] Yates Y. The analysis of replicated experiments when the field results are incomplete. *Emp. J. exp.* Agric,1933;1:129-142

[5] Wilkinson GN. Estimation of missing values for the analysis of incomplete data. *Biometrics*,1958;14:257-286

[6] Loh WY, and Vanichsetakul N. Tree-structured classfication via generalized discriminant analysis (with discussion).*JASA*,1988;83:715-728

[7] Alter O, Brown PO, and Botstein D. Singular value decomposition for genome-wide expression data processing and modeling.*PNAS*,2000;97:10101-10106

[8] Khutoryansky NM. Implutation of longitudinal incomplete data in clinical trials.Novo Nordisk Pharmaceuticals, Inc.,*Princeton,NJ*,2002

[9] Alizadeh,AA, Eisen MB, and Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.*Nature*,2000;403:503 – 511

[10] Anderson TW. *An introduction to multivariate statistical analysis.* Wiley,New York,1984

[11] Johnson RA, and Wichern DW. *Applied multivariate statistical analysis*. Prentice-hall.London,1998

[12] Shanon CE (1948), A Mathematical Theory of Communication, *Bell Syst. Tech. J.*, 27, 379-423, 623-656

[13] Venables, W. N. and Ripley, B. D. *Modern Applied Statistics with S,*Fourth edition.Springer,2002.

[14] Troyanskaya O, Cantor M, and Sherlock G. Missing value estimation methods for DNA microarrays. *Bioinformatics,* 2001;17(6):520-525

[15] McLachlan GJ, and Krishnan T. *The EM algorithm and extensions*. Wiley, New York,1996

[16] Dempster AP., Laird N, and Rubin DB (1977).Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion).*Journal of the Royal Statistical Society*,1977,Series B,1977;39:1-38

[17] Dempster AP,Laird N, and Rubin DB. Maximum likelihood estimation(with discussion). *Applied Statistics*,1977;43:49-93

[18] Schafer JL. *Analysis of Incomplete multivariate data*. Chapman hall, London,1997

[19] 강명욱,김영일,안철환,이용구. *회귀분석:모형개발과 진단*. 율곡출판사,1995

ABSTRACT

Comparison of imputation method for time course data in Microarray experiment

Seo, Won Youl

Dept. of Biostatistics and Computing

The Graduate School

Yonsei University

DNA microarray experiment is got with matrix data of the form that a lot of experiment condition following gene expression level is large, but missing data is occurred in a got process. If predict can do missing data accurately, we can get the more correct results in an analysis technique to request complete data.

we used regression, SVD, KNN, and EM algorithm so that comparison evaluated what it was that accuracy was the highest. When missing data is small, the accuracy which used SVD algorithm was the highest and the accuracy which used regression was the lowest.

On missing data propotion, SVD, regression, EM algorithm showed the aspect that accrucy increased, and KNN algorithm showed a pattern to have a lot of different pattern for accuracy. EM algorithm compared to three other method according to propotion of missing value increasing and, accuracy was

high increased on missing data propotion

Key words : Imputation, KNN, regression, EM algorithm, SVD, time-course data, Incremental method