응용통계연구 제 **21**권 2호, *2008*년, *pp. 275–290*

순차적 부분최소제곱 회귀적합에 의한 시간경로 유전자 발현 자료의 결측치 추정*

김경숙[1] 오미라[2] 백장선[3] 손영숙[4]

요 약

마이크로어레이 유전자 발현 자료는 대용량이며 또한 관측 과정이 복잡하여 결측치가 빈번하게 발생된다. 본 논문에서는 관측 시점 간에 상관성을 갖는 시간경로 유전자 발현 자료에 대한 결측치 추정을 위하여 순차적 부분최소제곱(sequential partial least squares: SPLS) 회귀적합 방법을 제안한다. 이는 순차적 기법과 부분최소제곱(partial least squares: PLS) 회귀적합 방법을 결합시킨 것이다. 세 가지의 이스트(yeast) 시간경로 자료들에 대한 몇 가지 모의실험을 통하여 제안된 결측치 추정방법의 유용성을 평가한다.

주요용어: 마이크로어레이, 시간경로 유전자 발현 자료, 결측치 추정, 부분최소제곱 회귀적합, 순차적 부분최소제곱 회귀적합.

1. 서론

마이크로어레이 실험은 수 천 혹은 수 만개의 유전자들이 동시에 관측되므로 보통 매우 크기가 큰 자료를 형성하는데, 복잡한 여러 단계의 실험 과정을 통해 얻어진 자료는 다양한 원인으로 인해 흔히 결측치를 포함하게 된다. 스폿팅(spotting) 오류나 합성의 실패 및 불충분한 해상력, 이미지 훼손, 슬라이드 상의 먼지, 긁힘 또는 지문, 의심스러운 값의 삭제, 발현 정도가 어느 수준 이하로 매우 낮은 값의 삭제, 중도절단된 자료(censored data) 등이 주요 원인들이다. 대부분의 마이크로어레이 자료들의 결측치의 개수는 전체 자료크기에 대비해서는 그리 크지 않은 수준이라 하더라도 유전자 단위로 대비해서는 매우 높은 비율로 나타난다. 이러한 현상은 결측치를 포함하는 전체 유전자들 가운데 한 개의 결측치만 있는 유전자의 비율이 상대적으로 매우 높게 발생하기 때문이다. de Brevern 등 (2004)은 자주

* 이 논문은 2005년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2005-204-C00017).

1) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 통계학과, 시간강사.
E-mail: ksook620@jnu.ac.kr
2) (500-712) 광주광역시 북구 오룡동 1, 광주과학기술원 정보통신공학과, 박사후과정.
E-mail: omr@chonnam.ac.kr
3) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 통계학과, 교수.
E-mail: jbaek@chonnam.ac.kr
4) (500-757) 교신저자. 광주광역시 북구 용봉동 300, 전남대학교 통계학과, 교수.
E-mail: ysson@chonnam.ac.kr

이용되는 일부 공공 자료(public data)에 대한 결측 유형을 정리하였는데, 한 개의 결측치를 갖는 유전자의 비율이 대부분 60%를 넘는 것으로 보고하였다.

대부분의 주요한 통계분석 기법들은 결측치가 없는 완전한(complete) 형태의 자료를 필요로 한다. 따라서 유전자 발현 프로파일 내에 결측치가 하나라도 존재하면 해당 유전자에 대한 관측 자료가 모두 통계분석에서 제외된다. 마이크로어레이 자료의 연구 초기에는 흔히 결측 자료를 무시하고 분석하거나 실험을 다시 수행하는 방법을 취해 왔으나, 재 실험을 통해 추가 자료를 얻는 방법은 매우 많은 비용과 시간이 소요되므로 현실적으로 어려움이 크다. 따라서 관측된 자료의 주어진 정보를 최대한 활용하여 자료의 유용성을 높이고, 분석 결과의 질도 향상시키기 위해서는 정확한 결측치 추정방법을 적용하여 얻어지는 추정치로 결측치를 대체하는 전처리 과정이 필요하다.

본 연구에서는 시간의 흐름에 따라 관측되는 시간경로 자료에 대한 결측치 추정을 다루고자 한다. 지금까지 연구되어 온 대부분의 기존 방법들은 관측 시점 간에 상관성을 무시하고 독립된 변수로 다루어 왔다. 한편, 특성치 분해(singular value decomposition), 주성분 분석(principal component analysis: PCA), 혹은 부분최소제곱(partial least squares: PLS) 회귀적합에 의한 결측치 추정 방법들은 독립된 성분을 추출하여 변수로 이용하므로 상관성 문제는 해결하였으나, 결측 유전자와 유사성이 매우 낮은 자료까지 모두 포함하여 추정하므로 군집의 특성이 있는 자료일 경우에는 유사성이 높은 일부 자료만 이용하는 방법에 비해 추정의 정확성이 떨어질 것으로 고려된다.

본 논문에서는 이러한 문제점들을 고려한 결측치 추정 방법으로서 순차적 부분최소제곱(sequential partial least squares: SPLS) 회귀적합 방법을 제안하고자 한다. SPLS를 수행하는 방법은 유전자를 변수로 취하는 gene-wise 방법과 관측 시점을 변수로 취하는 array-wise 방법 그리고 이 둘을 결합시키는 방법이 적용된다. 수행 절차는 먼저 결측치의 수가 적은 유전자부터 순차적으로 결측치 추정을 위한 목표(target) 유전자로 설정하고, 결측치가 없는 완전한(complete) 유전자들 가운데 이와 가장 유사한 K개의 완전 유전자를 선택한다. PLS 방법을 통해 독립된 성분들을 추출하여 회귀적합 시킨 후 추정치를 이용하여 목표 유전자를 완전한 유전자가 되게 한다. 새로운 완전 유전자는 기존의 완전 유전자들과 함께 다음 차례의 목표 유전자에 대한 결측치 추정시 이용될 수 있는 후보 자료가 된다.

원래 PLS는 chemometrics 분야에서 많이 사용되어 왔다 (Höskuldsson, 1988; Garthwaite, 1994; Abdi, 2003; Jørgensen과 Goegebeur, 2006). 마이크로어레이 자료에 대한 적용 예로서 Nguyen과 Rocke (2002)는 수천 개의 마이크로어레이 유전자 발현자료를 소수 몇 개의 성분으로 차원축소하여 판별분석하기 위해서 PLS를 적용하였다. PLS 회귀적합에 의한 결측치 추정 방법의 수식화된 설명은 Nguyen 등 (2004)에 자세히 소개되어 있다.

본 논문의 구성은 다음과 같다. 2절에서는 기존에 연구된 결측치 추정 방법들을 두 가지로 나누어 소개한다. 즉, 자료 내의 국부적(local) 유사성 구조에 기반하여 결측 유전자와 가장 유사한 K개의 완전한 자료만을 이용하는 방법과, 자료의 전체적(global) 유사성 구조에 기반하여 완전한 자료 전체를 이용하는 방법이다. 3절에서는 시간경로 자료의 특성을 고려하여 PLS에 기반한 SPLS 회귀적합에 의한 결측치 추정 방법을 제안한다. 4절에서는 SPLS 회귀적합 방법과 기존의 결측치 추정방법들에 의한 추정치의 적합수준을 비교평가

하기 위해 수치분석 한다. 이는 실제 이스트(yeast) 시간경로 유전자 발현 자료들로부터 세 가지 유형의 모의실험 자료를 각각 생성하여 수행된다. 마지막으로 5절에서 결론으로 끝맺는다.

2. 여러 가지 결측치 추정 방법

다음에 소개되는 결측치 추정 방법들은 완전 자료 전체를 이용하는 방법과 그 중 일부만을 이용하는 방법으로 나뉘어져 있다.

2.1. 국부적(local) 자료 이용 방법

국부적 유사성이 존재하는 자료인 경우에는 전체 자료를 이용하는 것 보다 유사성이 강한 일부 자료만을 이용하여 결측치를 추정하는 방법이 적합 수준을 높여 줄 것이다. 이는 군집(cluster) 기반의 결측치 추정 방법으로 볼 수 있다.

(1) 0 대체법 및 행(열) 평균 대체법

0 대체법은 가장 간단한 방법으로서 결측치를 0으로 대체하는 방법이며, 행 (또는 열) 평균 대체법은 결측치를 추정하고자 하는 유전자 (또는 관측 시점) 내에서 결측치를 제외한 관측값들의 평균으로 대체하는 방법이다. 이는 다소 직관적인 방법들로서 자료 내의 상관구조를 전혀 이용하지 않으므로 결측치 추정능력이 상당히 떨어진다. 독립적으로 사용되기 보다는 주로 반복기법을 사용하는 다른 추정 방법에서 초기 추정값으로 사용된다.

(2) KNN(K-nearest neighbor) 방법

Troyanskaya 등 (2001)이 제안한 방법으로서 gene-wise 방법을 적용한다. 결측치를 추정하고자 하는 목표 유전자를 기준으로 하여 결측인 시점과 결측이 아닌 시점으로 자료를 분리하고 결측이 아닌 시점에 대해 목표 유전자와 완전 유전자 자료 간에 유사성을 측정한다. 가장 유사한 K개의 완전 자료만을 추출하여 유사성의 정도에 따라 가중치, $w_i = D_i^{-1} / \sum_{i=1}^{K} D_i^{-1}$, $(i = 1, 2, \ldots, K)$를 완전 유전자의 관측치에 부여하여 가중평균 내어 결측치를 추정한다. 여기서 D_i는 결측 유전자와 완전 유전자 간의 유크리드 거리를 나타낸다. 이때 유사성 측도로서 유클리드 거리 외에도 피어슨 상관계수 혹은 분산 최소화 방법을 사용하였으나, 유클리드 거리가 가장 적절한 측도인 것으로 결론지었다.

(3) SKNN(sequential KNN) 방법

Kim 등 (2004)이 제안한 방법으로서 KNN 방법을 기본으로 하며 결측치가 적은 유전자부터 순차적으로 추정해 나간다. 결측치가 모두 추정된 유전자는 완전자료 형태가 되고 다음 차례의 결측 유전자를 추정하기 위한 K개의 유사한 자료를 선택할 때 후보 자료가 된다. 결측률이 증가함에 따라 추정의 적합 수준은 KNN 기법 보다 더 좋으며 로버스트(robust)한 결과를 보였으나, 전반적으로 약 30% 이상의 높은 결측 수준에서 더 뛰어난 것으로 보고하였다. SKNN은 자료의 결측률이 상당히 높거나 array(표본)의 수가 많은 경우에 유용한 방

법이다.

(4) LS(least square) 방법

Bø 등 (2004)에 의한 방법으로서 gene-wise 방법과 array-wise 방법을 모두 고려하였다. 결측 유전자와 유사한 유전자를 K개 선택하여 최소제곱법을 통해 결측치를 추정하며, 유사성 측도로는 피어슨 상관계수의 절대값을 이용하였다. gene-wise 방법은 결측 유전자를 종속변수로 하고 유사한 유전자를 설명변수로 하여 K개의 단순회귀모형을 적합시키고, 얻어지는 K개의 추정치를 가중평균 하여 결측치를 추정하였다. 이때 가중치는 $w_i = \{r_i^2/(1-r_i^2+\epsilon)\}^2$, $(i = 1, 2, \ldots, K)$으로 주었고, 여기서 r_i^2은 i번째 유사한 유전자와의 단순회귀모형을 적합시켜 얻은 결정계수이고 ϵ은 10^{-6}을 사용하였다. array-wise 방법은 gene-wise 방법의 추정치를 초기값으로 하여 다중회귀모형을 적합시켰다. 또한 두 방법에 의한 추정치를 가중평균 내는 방법도 고려하였다.

(5) LLS(local least square) 방법

Kim 등 (2005)이 제안한 방법으로서 결측 유전자와 유사한 K개 유전자를 선택하고 다중회귀모형을 적합시킨 후 결측치 추정을 하였다. 유사성 측도로는 유클리드 거리 및 피어슨 상관계수를 이용하였다. gene-wise 방법과 array-wise 방법을 고려하였는데, 두 방법에 의한 추정 결과는 동일하게 얻어짐을 이론적으로 증명하였다.

2.2. 전체적(global) 자료 이용 방법

전체적인 자료 이용 방법은 완전 자료 전체에 대한 공분산 구조를 사용하므로 모든 완전 자료를 결측치 추정에 이용하는 방법이다.

(1) SVD(singular value decomposition) 방법

Troyanskaya 등 (2001)이 제안한 방법으로서 SVD에 기반한 이론적 배경은 Hastie 등 (1999)에 자세히 보고되었다. 완전 유전자 자료들을 gene-wise 방법으로 특성치 분해(SVD)하여 서로 독립인 성분을 추출하는데, 이 성분들은 자료 내의 모든 유전자들의 선형결합으로 구성된다. 결측치 추정 절차는 먼저, SVD 수행시 완전한 자료형태를 필요로 하므로 1차적인 결측 추정치로서 행 평균 대체값을 적용한다. SVD를 통해 설명력이 높은 K개의 고유유전자를 추출하고, 이들과 결측 유전자 간에 회귀모형을 적합시킨 후, 추정된 회귀계수를 이용하여 2차적으로 결측값을 추정한다. 최종 추정치는 EM 알고리즘을 통해 기준 임계치에 도달할 때까지 반복하여 산출하였다.

(2) BPCA(Bayesian principal component analysis) 방법

Oba 등 (2003)이 제안한 방법으로서 주성분 회귀분석(PCA regression)을 통해 얻어진 주축 벡터 (주성분)와 선형 계수 (인자점수: factor scores)를 이용하여 베이지안 방법으로 결측치를 추정한다. BPCA 추정은 다음의 세 단계로 구성된다. 먼저 주성분 회귀모형을

적합시키고, 다음으로 베이지안 모수 추정 단계를 거쳐, 마지막으로 EM과 유사한 방법의 VB(variational Bayes) 알고리즘을 적용하여 반복수행에 의해 최종 결측치를 추정하였다.

(3) PLS(partial least squares) 방법

Nguyen 등 (2004)이 제안한 방법으로서 PLS 방법을 통해 전체 유전자를 대상으로 몇 개의 유의한 주성분들을 추출하여 차원축소 후 회귀모형을 적합시켜 결측치를 추정하였다.

3. 순차적 부분최소제곱 회귀적합에 의한 결측치 추정 방법

시간경로 유전자 발현 자료는 P개의 유전자를 시간의 흐름에 따라 연속적으로 T개의 시점에서 관측함으로써 얻어지는 자료이며 다음과 같이 표현할 수 있다.

$$X_{P\times T} = \{x_{i,t_j}\}, \quad i = 1, 2, \ldots, P, \quad j = 1, 2, \ldots, T,$$

여기서 x_{i,t_j}는 i번째 유전자에 대해 시점 t_j에서 관측된 자료를 의미한다.

결측치 추정 문제에서는 자료를 다루는 관점에 따라 유전자 (행 단위)를 변수로 간주하고 관측 시점 (열 단위)을 관측 개체로 간주하는 gene-wise 방법과 이와 반대로 관측 시점 (열 단위)을 변수로 간주하고 유전자 (행 단위)를 관측 개체로 간주하는 array-wise 방법의 분석이 가능하다. 먼저, gene-wise 방법은 대부분의 마이크로어레이 자료가 안고 있는 변수의 수 (유전자 수)는 많고 표본의 수 (array의 수, 즉 관측 시점의 수)는 적은 문제를 고려해야 한다. 변수의 차원을 줄이기 위해 국부적 자료 이용 방법으로서 결측 유전자와 유사한 K개 완전 유전자 자료만을 선택하고, 이를 대상으로 PLS 방법으로 소수의 독립적 성분을 추출하여 변수로 이용한다. 다음으로, array-wise 방법은 시간경로 자료의 일반적 특징인 관측 시점간의 상관성 문제를 고려해야 한다. 추정 방법으로서 특히 회귀적합을 이용하는 경우 다중공선성 문제를 예상할 수 있으므로 변수간의 상관성을 제거하기 위해 PLS 방법으로 독립적 성분을 추출하여 변수로 이용한다. PLS 방법은 종속변수와 독립변수 간의 공분산을 최대로 설명하는 성분을 추출하는 방법으로서 종속변수와의 관련성을 고려하여 독립변수로부터 성분을 추출해 낸다. 한편, 이와 유사한 PCA 방법은 독립변수들만의 공분산을 고려하여 성분을 추출하므로 종속변수가 존재하는 경우 PLS 방법 보다 적합 능력이 더 떨어질 것으로 판단된다.

결측치를 추정하고자 하는 목표 유전자와 유사한 K개의 완전 유전자들을 추출하는 방법은 목표 유전자의 자료가 존재하는 시점들에서 목표 유전자 자료와 완전 유전자 자료간의 일대일 유사성(similarity) 정도를 측정하여 가장 유사한 K개의 완전 유전자들을 추출한다. 이때 유사성을 측정하는 세 가지 측도로서 유클리드 거리, 피어슨 상관계수, 피어슨 상관계수의 절대 값을 이용한다.

본 논문에서 PLS 회귀적합에 의한 결측치 추정 방법은 gene-wise 방법(PLSgene 방법)과 array-wise 방법(PLSarray 방법)을 수행한다. PLSarray 방법을 통해 결측치를 추정하는 절차를 살펴보면 먼저, 목표 유전자와 유사한 K개 완전 유전자의 자료들은 목표 유전자를 기준으로 하여 결측치 시점에 상응하는 자료와 결측치 아닌 시점에 상응하는 자료로 분리

한다. 전자를 종속 변수군, 후자를 설명 변수군으로 설정한 PLS 분석을 통해 각 변수군별로 성분들을 추출하고 이를 변수로 하여 회귀모형을 적합시킨다. 여기에 목표 유전자의 관측된 자료를 대입하여 결측치를 추정하게 된다.

SPLS 회귀적합 방법을 수행하는 절차를 살펴보면 먼저, 목표 유전자는 결측치를 갖는 유전자들 가운데 가장 적은 수의 결측치를 갖는 유전자부터 시작하여 순차적으로 하나씩 설정된다. 다음으로, 목표 유전자와 가장 유사한 K개의 완전 유전자를 선택하고, PLS 회귀적합을 수행하여 얻은 추정치를 목표 유전자의 결측치에 대체시켜 완전 유전자로 변환시킨다. 다음 차례의 목표 유전자에 대한 추정을 위해서는 기존의 완전 유전자들 뿐만 아니라 이전 단계에서 만들어진 완전 유전자도 K개의 유사한 완전 유전자들을 다시 선택하는데 후보가 되도록 한다. 이렇게 하면 결측치 수가 작은 결측 유전자의 자료가 결측치 수가 더 많은 결측 유전자와 유사성이 높은 경우 그러한 결측 유전자의 결측치 추정에 기여할 수 있는 기회를 주는 장점이 있다. K개의 유사한 완전 유전자가 새로 선택되면 PLS 회귀적합을 통해 목표 유전자를 완전 유전자로 변환시킨다. 이러한 과정을 반복하여 최후의 결측 유전자까지 모두 수행한다.

이러한 순차적인 결측치 추정과정을 진행할 때 PLS 방법을 적용하는 방법에 따라 첫째, PLSgene 방법만을 이용하는 경우에는 SPLSgene 방법이라 하고 둘째, PLSarray 방법만을 이용하는 경우에는 SPLSarray 방법이라 하며 셋째, 두 방법에 의한 추정치들을 가중평균 내어 최종 추정치를 얻는 결합방법은 SPLScombined 방법이라 명명한다. 이 때 가중치는 $w_i = R_i^2/(R_1^2 + R_2^2),\ (i = 1, 2)$과 같이 정의한다. 여기서 w_1과 w_2는 각각 SPLSgene 방법과 SPLSarray 방법에 대한 가중치이고, R_1^2과 R_2^2은 각각의 방법에 의한 회귀적합시 얻는 결정계수이다.

4. 수치 분석

본 논문에서 제안된 SPLS 회귀적합에 의한 추정치의 적합수준을 기존의 결측치 추정 방법에 의한 결과와 비교 평가한다. 고려되는 기존 방법들로는 PLS 회귀적합 방법과 최근의 결측치 추정 연구에서 빈번히 비교되는 KNN, SKNN, LLS, BPCA 등 이다. 수치 분석에 이용된 자료는 세 가지의 실제 이스트 자료들로서 첫째 자료(Y7)는 DeRisi 등 (1997)에 의해 6,400개 유전자를 7개의 시점에서 반복 측정한 자료이고, 둘째 자료(Y18)는 Spellman 등 (1998)에 의해 6,178개 유전자를 실험한 것으로서 이 가운데 alpha-factor 실험 부분에 해당하는 18개 관측 시점의 자료이다. 셋째 자료(Y24)는 Spellman 등 (1998)의 자료에서 CDC15 실험 부분의 24개 관측 시점에서 측정된 자료 가운데 세포 주기(cell cycle)에 관여하는 것으로 분류해 놓은 800개 유전자 자료이다. 세 가지 분석 자료의 특징은 Y7은 비주기성(aperiodicity)을 갖는 시간경로 자료이고, Y18은 주기성(periodicity)과 비주기성의 자료가 혼재된 형태이며, Y24는 주기성을 갖는 자료로 구성되어 있다. 하나의 유전자 내에 결측치가 너무 많이 발생한 경우에는 자료로서의 가치가 없을 뿐만 아니라 다른 결측치 추정에도 좋지 않은 영향을 미치므로, 실제 분석에서는 1/3 이상의 결측치를 포함하는 유전자는 추정 대상에서 제외하였다. 분석 자료의 구체적인 결측 유형은 표 4.1에 정리하였으며

표 4.1: 분석 자료의 결측 유형

분석 자료	관측 시점 수 (T)	유전자 수		결측 비율(%)		총 결측 유전자 수 대비 결측치 수별 비율(%)(E)							
		완전 자료 (A)	결측 자료 (B)	유전자 수 대비 (C)	총 자료 수 대비 (D)	1	2	3	4	5	6	7	8
Y7	7	6,292	64	1.01	0.25	53.1	21.9	25.0	-	-	-	-	-
Y18	18	4,489	1,567	25.88	1.93	75.1	17.9	4.1	1.6	0.7	0.5	-	-
Y24	24	632	112	15.05	1.35	50.8	17.0	7.6	8.4	3.4	7.6	1.7	3.4

* $C = 100 \times B/(A+B)$, $D = 100 \times$ (총 결측치 수)$/\{(A+B) \times T\}$,
$E = 100 \times$ (해당 결측치 수를 갖는 유전자 수)$/B$

서론에 언급하였듯이 본 실험 자료에서도 전체 결측 유전자 가운데 1개의 결측치를 포함하는 유전자 비율이 절반 이상으로 나타난다.

결측치 추정에 관한 연구에서 통상적으로 수행하는 모의실험의 첫 단계로서 분석 자료의 준비는 전체 관측 자료 중에서 결측치가 전혀 없는 유전자 자료만 모은 완전 자료들을 대상으로 하여 가상의 결측치를 임의 생성한다. 따라서 모의실험에서는 결측치의 참값을 알 수 있으므로 여러 가지 결측치 추정 방법의 비교가 가능하게 된다.

모의실험을 위해 각 자료는 세 가지 유형으로 재구성 한다. 유형 1(type 1)은 완전 자료 전체를 대상으로 하여 원(raw) 자료의 결측 형태 및 결측률을 동일하게 유지시키는 방법으로서 Y7, Y18, Y24 자료의 결측률은 각각 0.25%, 1.93%, 1.35%로 설정하였다. 유형 2(type 2)는 완전 자료 전체를 대상으로 하되 결측률은 원 자료 보다 더 크면서 20% 이하인 경우를 고려하고, 그에 따라 결측 형태도 원 자료와 다소 달라진다. 원 자료의 결측치 수준을 고려하여 결측률은 각각 Y7 자료는 0.5%, 1%, 5%, 10%, Y18 자료와 Y24 자료는 5%, 10%, 15%, 20%로 설정하였다. 기존의 논문에서는 주로 유형 1과 유형 2 방법을 적용하였으나, 이는 1회의 실험 결과만을 이용하므로 생성된 자료에 의해 영향을 받을 것으로 판단된다. 따라서 본 논문에서는 반복 실험하는 유형 3를 추가로 고려한다. 유형 3(type 3)은 유형 2에서 설정한 결측률과 같게 설정하고 동일한 과정을 100번씩 반복한다. 이때 Y7 자료와 Y18 자료는 자료의 크기가 너무 커서 계산시간을 단축시키기 위해 완전 자료 전체로부터 500개 자료를 임의 추출하여 사용하고, Y24 자료는 비교적 작은 크기의 자료이므로 완전 자료 전체를 사용한다.

본 논문에서는 제안된 SPLS 회귀적합 방법과 기존 방법들 중 PLS 회귀적합 방법 및 KNN, SKNN, BPCA, LLS 방법들의 적합수준 및 특성을 비교 평가한다. 결측 시키기 전 원래의 값(참값)에 대한 추정치의 적합 수준을 측정하기 위한 측도로는 다음과 같은 정규화된 제곱근평균제곱오차(normalized root mean square error: NRMSE)를 사용한다.

$$\text{NRMSE} = \frac{\sqrt{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2/N}}{std(y)},$$

여기서 N은 결측치의 총 개수이며, $\hat{y}_i$는 i번째 결측치에 대한 추정치이고, y_i는 i번째 결측

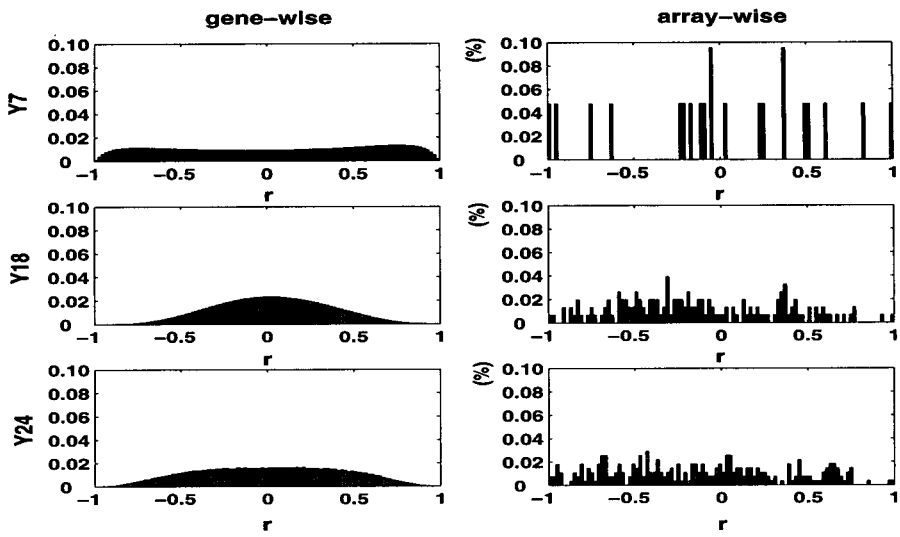

그림 4.1: 분석 자료별 유전자 간 및 관측 시점 간의 상관성 비교

치에 상응하는 원자료의 참값을 의미한다. $std(y)$는 자료 $\{y_1, y_2, \ldots, y_N\}$의 표준편차이다.

결측치를 추정하기에 앞서 자료의 유사성 구조를 살펴본 결과, 그림 4.1에서 gene-wise는 유전자간의 피어슨 상관계수 분포를 나타내고, array-wise는 관측 시점간의 피어슨 상관계수 분포를 나타낸다. 상관계수의 값이 ± 1 근처에 많이 나타날수록 국부적 상관성(유사성)이 크다고 볼 수 있다. 유전자간(gene-wise) 상관성과 관측 시점간(array-wise) 상관성을 비교해 보면, Y7 자료는 유전자간 상관성이 높게 보이며, Y18 자료와 Y24 자료는 관측 시점간 상관성이 상대적으로 더 높게 보인다. 한편 유전자간의 국부적 상관성은 Y7, Y24, Y18 자료 순으로 더 높게 나타나 있다.

다음으로는 결측 유전자와 유사한 완전한 유전자 K개를 선택하기 위해 1차적으로 후보 값 $K = 5, 10, 15, 20, ..., 500$을 설정하고, 각 경우에 대해 결측치를 추정한 후 NRMSE를 비교하여 이 가운데 최소값에 대응되는 K를 1차로 선택한다. 2차 과정으로서 1차에서 선택된 값을 중심으로 ± 4 범위 내에서 K를 다시 설정하고, 1차와 동일한 과정을 통해 선택된 K를 최종 값으로 결정한다. 예로써 Y18 자료의 실험 유형 1 자료를 대상으로 유클리드 거리를 유사성 측도로 하여 KNN 방법 및 SPLSarray 방법을 적용하여 K의 추정 과정을 그림 4.2에 제시한다. KNN 방법에서는 1차에서 $K = 10$이 선택되고, 2차에서 $K = 11$이 최종으로 선택된다. SPLSarray 방법에서는 1차에서 $K = 250$이 선택되고, 2차에서 $K = 249$가 최종으로 선택된다.

K개의 유의한 유전자들이 선정되면 이를 대상으로 PLS 방법에 의해 최소 1개부터 최대로는 종속 변수군의 차원 수와 독립 변수군의 차원 수 중에서 작은 수에 해당하는 개수까지 모든 가능한 주성분 수에 대해 결측치 추정을 수행한다. 이때 최소의 NRMSE에 상응하는 주성분 수를 선택한다.

이제, 선택된 K개의 완전 자료를 가지고 여러 가지 추정 방법을 적용한 후 추정 방법

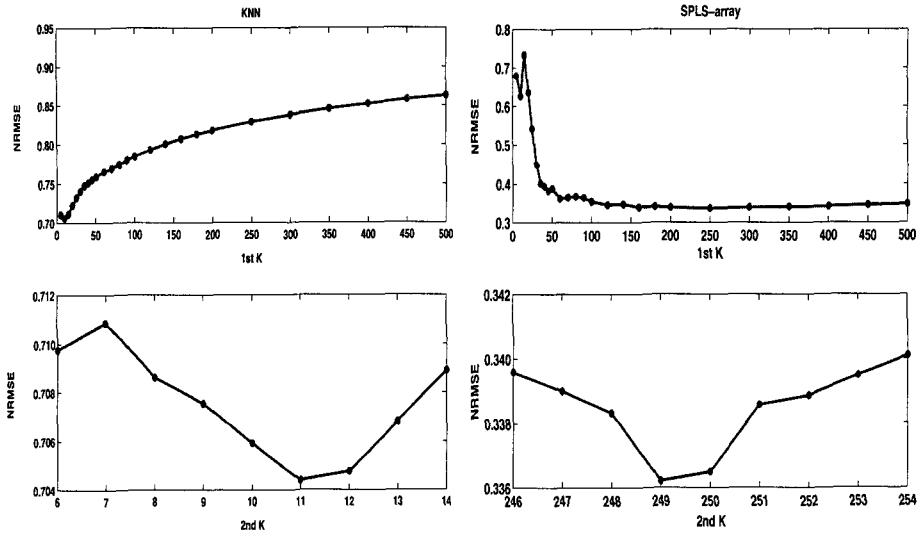

그림 4.2: 최적 K의 추정 예

들의 특징과 적합 수준을 비교 평가한다. 첫째, Y7 자료에 대한 결과를 살펴보면, 먼저 여러 가지의 SPLS 방법과 PLS 방법의 결과가 그림 4.3(a)에 있다. 원 자료의 결측 형태 및 낮은 결측률을 그대로 유지 시킨 실험 유형 1에서는 array-wise 방법이 gene-wise 방법 보다 더 좋고, 여러 방법들 중 SPLSarray와 PLSarray 방법이 거의 비슷하게 가장 좋은 결과를 보인다. 결측률이 높은 실험 유형 2와 유형 3에서는 반대로 gene-wise 방법이 array-wise 방법 보다 더 좋고, 전반적으로 PLSgene, SPLSgene 그리고 SPLScombined가 비슷하게 가장 좋은 결과를 보인다. 다음으로 최적의 SPLS 방법과 최적의 PLS 방법 그리고 그 외의 방법들에 의한 결과가 그림 4.3(b)와 표 4.2에 있다. 모든 실험에서 전반적으로 SPLS가 가장 좋고 PLS도 비등하게 좋은 수준을 보이며, 그 외 방법들은 이와 상당한 차이를 보이는데 그 중에서는 유전자간의 국부적 유사성 구조를 이용하는 LLS, KNN, SKNN이 BPCA 보다 더 좋은 경향을 보인다. 이는 비주기성의 Y7 자료가 유전자간에 강한 국부적 유사성을 갖는데 BPCA는 이러한 특성을 고려하지 않고 자료 전체를 이용하기 때문인 것으로 고려된다.

둘째, Y18 자료에 대한 결과를 살펴보면, 먼저 여러 가지의 SPLS 방법과 PLS 방법의 결과가 그림 4.4(a)에 있다. PLS의 일부 결과를 제외하고는 array-wise 방법이 gene-wise 방법보다 더 좋고, 여러 방법들 중 SPLSarray 방법이 전반적으로 가장 좋은 결과를 보인다. 이는 주기성 자료와 비주기성 자료가 혼재된 Y18 자료가 유전자간 보다 관측 시점간에 더 강한 국부적 유사성 구조를 갖기 때문인 것으로 고려된다. 다음으로, 최적의 SPLS 방법과 최적의 PLS 방법 및 그 외의 방법들에 의한 결과가 그림 4.4(b)와 표 4.3에 있다. 모든 실험 유형에서 SPLS가 가장 좋은 결과를 보이고, 결측수준이 낮을수록 PLS, LLS, BPCA는 KNN, SKNN 보다 큰 차이로 더 우수한 결과를 보인다. Y7 자료와 다르게 BPCA가 KNN과 SKNN 보다 더 좋은 결과를 보이는데, 이는 Y18 자료의 유전자간 국부적 유사성이 약한 편이기 때문인 것으로 고려된다.

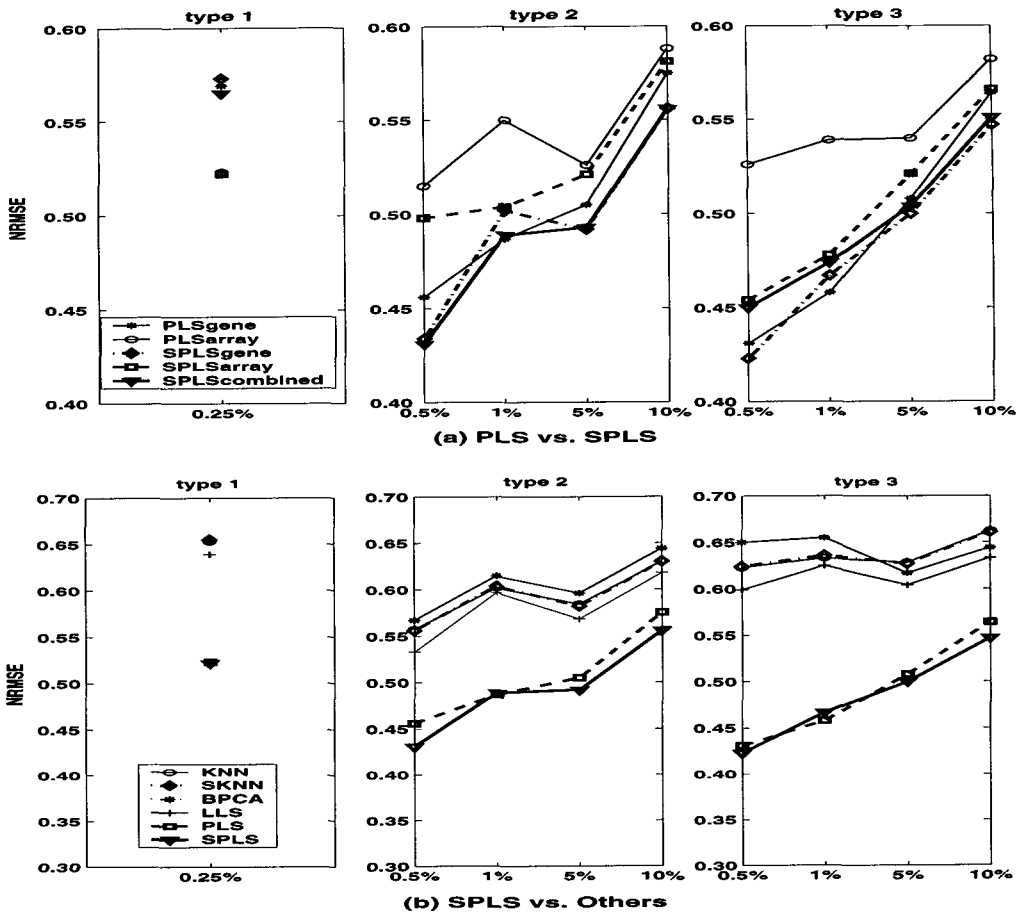

그림 4.3: Y7 자료에 대한 각 실험방법의 유형(type)별 결측치 추정 결과의 비교

표 4.2: SPLS 및 기존의 결측치 추정 방법의 비교 (자료: Y7)

실험	결측률	KNN	SKNN	BPCA	LLS	PLS	SPLS
유형 1	0.25%	.653	.655	.656	.639	.523 (Ar, Cr)	.522 (Ar, Cr)
유형 2	0.5%	.557	.556	.567	.533	.456 (Gn, Cr)	.431 (Gn/Cb, D)
	1%	.602	.604	.615	.597	.487 (Gn, D)	.489 (Cb, D)
	5%	.584	.582	.596	.568	.505 (Gn, D)	.492 (Gn/Cb, D)
	10%	.630	.630	.644	.618	.575 (Gn, D)	.556 (Gn/Cb, D)
유형 3	0.5%	.623(.19)	.624(.19)	.650(.24)	.599(.23)	.431(.18) (Gn, Cr)	.423(.19) (Gn, D)
	1%	.633(.12)	.636(.12)	.655(.17)	.625(.16)	.458(.14) (Gn, Cr)	.467(.14) (Gn/Cb, D)
	5%	.628(.06)	.627(.06)	.617(.08)	.604(.07)	.508(.07) (Gn, D)	.500(.07) (Gn/Cb, D)
	10%	.662(.05)	.661(.05)	.644(.06)	.633(.05)	.564(.06) (Gn, D)	.547(.06) (Gn/Cb, D)

* 유형3에서 수치는 평균(표준편차)를 의미함.
* (,): 가장 좋은 경우를 표시함.
* Ar: array-wise 방법, Gn: gene-wise 방법, Cb: 결합 방법, Gn/Cb: gene-wise 방법 또는 결합 방법.
* D: 유클리드 거리, Cr: 상관계수.

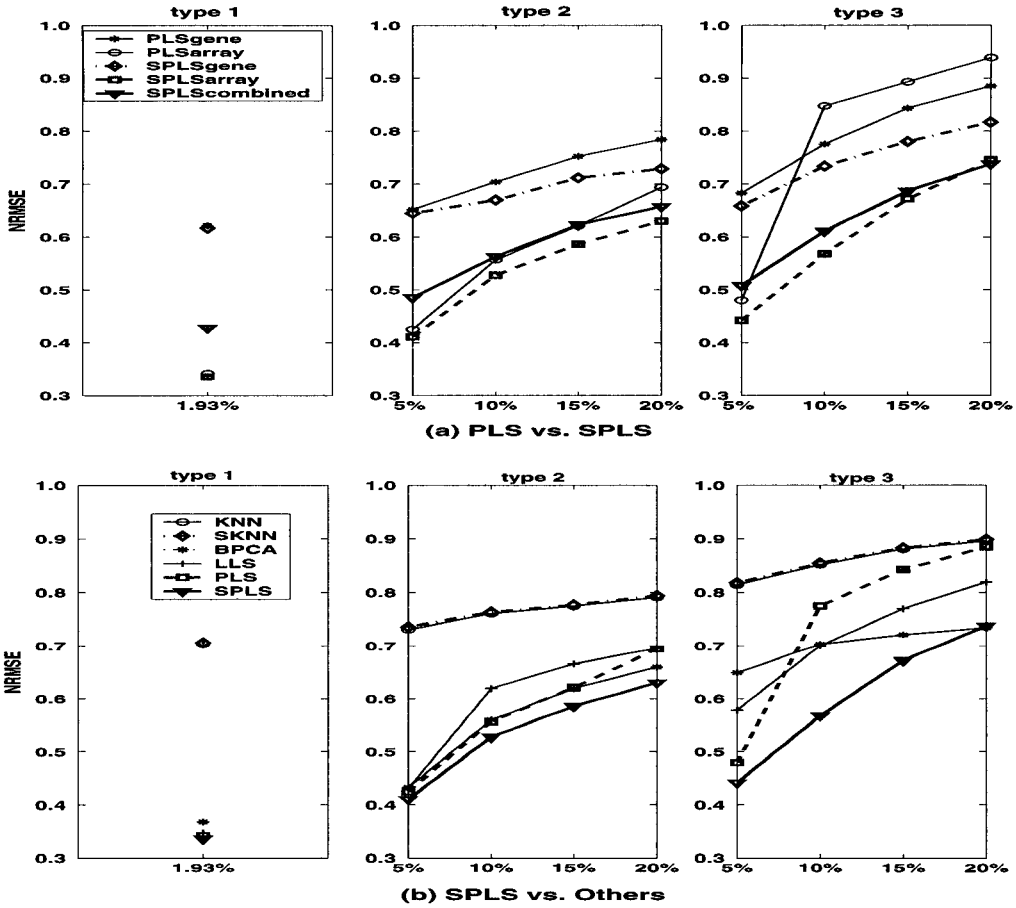

그림 4.4: Y18 자료에 대한 각 실험방법의 유형(type)별 결측치 추정 결과의 비교

표 4.3: SPLS 및 기존의 결측치 추정 방법의 비교 (자료: Y18)

실험	결측률	KNN	SKNN	BPCA	LLS	PLS	SPLS
유형 1	1.93%	.704	.706	.368	.348	.342 (Ar, Cr)	.336 (Ar, D)
유형 2	5%	.731	.736	.433	.431	.425 (Ar, D)	.411 (Ar, D)
	10%	.761	.764	.560	.619	.557 (Ar, AC)	.528 (Ar, D)
	15%	.775	.777	.619	.666	.621 (Ar, AC)	.586 (Ar, AC)
	20%	.791	.794	.660	.695	.694 (Ar, AC)	.630 (Ar, AC)
유형 3	5%	.814(.03)	.818(.03)	.649(.05)	.579(.05)	.480(.11) (Ar, AC)	.442(.05) (Ar, AC)
	10%	.852(.02)	.855(.02)	.703(.08)	.700(.04)	.775(.03) (Gn, AC)	.568(.04) (Ar, AC)
	15%	.881(.02)	.883(.02)	.721(.03)	.770(.03)	.843(.03) (Gn, AC)	.672(.04) (Ar, AC)
	20%	.896(.02)	.899(.02)	.735(.03)	.819(.03)	.885(.03) (Gn, AC)	.738(.03) (Ar, AC)

* 유형3에서 수치는 평균(표준편차)를 의미함.
* (,): 가장 좋은 경우를 표시함.
* Ar: array-wise 방법, Gn: gene-wise 방법.
* D: 유클리드 거리, Cr: 상관계수, AC: 상관계수의 절대값.

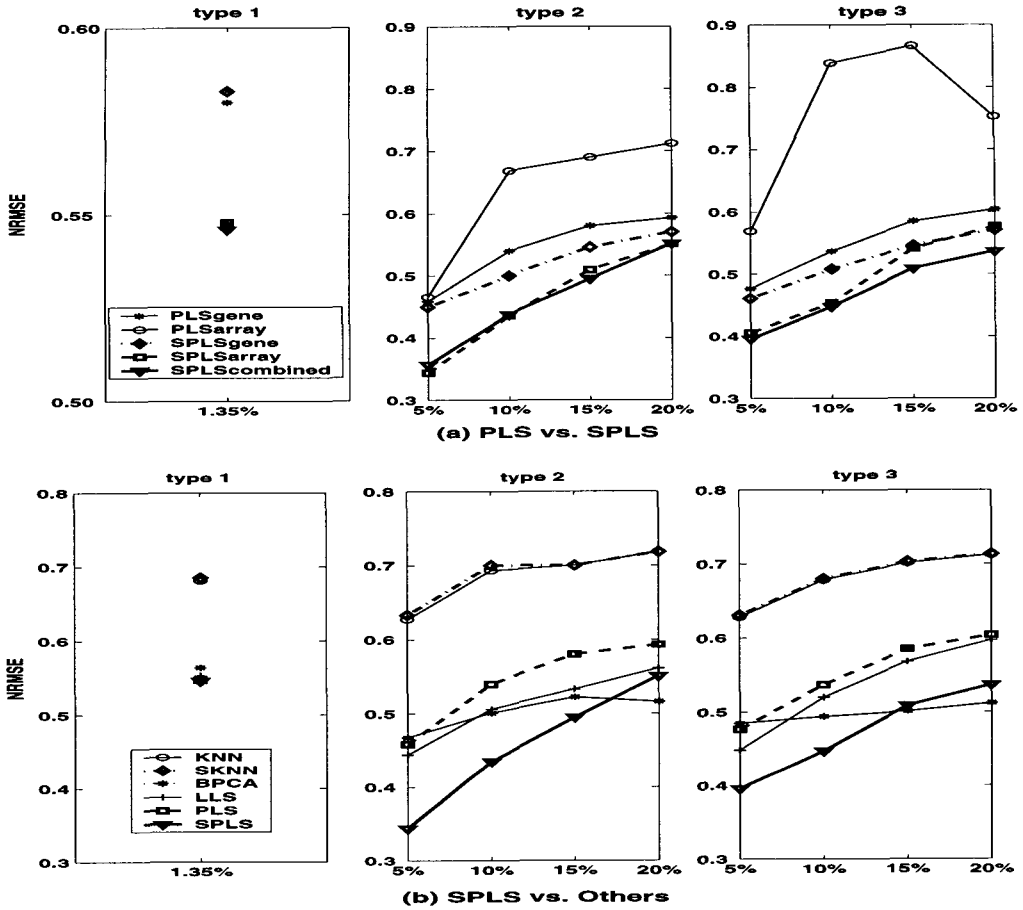

그림 4.5: Y24 자료에 대한 각 실험방법의 유형(type)별 결측치 추정 결과의 비교

표 4.4: SPLS 및 기존의 결측치 추정 방법의 비교 (자료: Y24)

실험	결측률	KNN	SKNN	BPCA	LLS	PLS	SPLS
유형 1	1.35%	.681	.685	.564	.553	.547 (Ar, AC)	.546 (Ar/Cb, AC)
유형 2	5%	.627	.633	.468	.445	.459 (Ar, AC)	.344 (Ar/Cb, AC)
	10%	.693	.700	.501	.506	.540 (Gn, D)	.436 (Ar/Cb, AC)
	15%	.700	.700	.522	.533	.580 (Gn, AC)	.495 (Ar/Cb, AC)
	20%	.718	.718	.516	.561	.593 (Gn, D)	.551 (Ar/Cb, AC)
유형 3	5%	.628(.02)	.631(.02)	.485(.03)	.449(.03)	.477(.02) (Gn, AC)	.396(.02) (Cb, AC)
	10%	.678(.02)	.680(.02)	.494(.02)	.520(.02)	.537(.02) (Gn, D)	.448(.02) (Cb, AC)
	15%	.701(.02)	.703(.02)	.501(.02)	.568(.02)	.585(.02) (Gn, D)	.509(.02) (Cb, AC)
	20%	.712(.01)	.713(.01)	.512(.01)	.597(.01)	.604(.01) (Gn, AC)	.537(.01) (Cb, AC)

* 유형3에서 수치는 평균(표준편차)를 의미함.
* (,): 가장 좋은 경우를 표시함.
* Ar: array-wise 방법, Gn: gene-wise 방법, Cb: 결합 방법, Ar/Cb: array-wise 방법 또는 결합 방법.
* D: 유클리드 거리, AC: 상관계수의 절대값.

셋째, Y24 자료에 대한 결과를 살펴보면, 먼저 여러 가지의 SPLS 방법과 PLS 방법의 결과가 그림 4.5(a)에 있다. 여러 방법들 중 SPLScombined 방법이 가장 좋은 결과를 보이고 SPLSarray 방법도 근소한 차이로 우수한 결과를 보인다. 주기성을 갖는 Y24 자료의 경우에는 관측 시점 뿐만 아니라 유전자 간에도 어느 정도의 국부적 유사성이 존재하는 것에 기인한 것으로 고려된다. 다음으로, 최적의 SPLS 방법과 최적의 PLS 방법 및 그 외의 방법들의 결과가 그림 4.5(b)와 표 4.4에 있다. 유형 1에서는 SPLS, PLS, LLS 및 BPCA가 비슷한 수준으로 KNN, SKNN 보다 더 좋은 결과를 보인다. 한편 유형 2와 유형 3에서는 BPCA가 결측률 변화에 따른 적합수준에 있어서 로버스트한 경향을 보였는데, 결과적으로 유형 2의 15% 이하 결측률과 유형 3의 10% 이하 결측률에서는 SPLS가 가장 좋으나, 각 유형에서 그 이상의 결측률에서는 BPCA가 가장 좋고 SPLS가 그 다음 순으로 나타나 있다.

5. 결론

본 논문에서는 시간경로 마이크로어레이 유전자 발현 자료에 대한 전처리로서 결측치 추정 문제를 다루었다. 관측 자료의 유용성을 높이기 위해 소수의 결측치가 있는 유전자들에 대해서는 결측치를 대신할 적절한 값을 추정하여 더 심화된 분석을 가능하게 하고자 한다. 시간경로 자료는 관측 시점 간에 상관성이 존재하므로 이를 고려하는 순차적 부분최소제곱(SPLS) 회귀적합 방법을 제안하였다. 기존의 여러 방법들과 비교 평가하기 위해 비주기성, 주기성, 혼합형의 시간경로 자료에 대해 분석하고, 자료 전체에 대한 결측 수준은 너무 높지 않은 수준인 20% 이하를 고려하였다.

결측치 추정에 대한 적합수준을 평가하기 위해 NRMSE 척도를 이용하여 SPLS 회귀적합 방법과 기존의 방법인 PLS 및 KNN, SKNN, LLS, BPCA를 비교하였다. 이 가운데 참값에 대한 추정치의 적합수준은 전반적으로 SPLS 방법이 좋은 결과를 보였고, SPLS를 기반으로 하는 여러 방법들 중에서는 자료의 국부적(local) 유사성 정도에 따라서 최적 방법이 다르게 나타났다. 즉, 유전자간 유사성은 약하고 관측 시점간 유사성이 강하면 SPLSarray 방법이, 그 반대이면 SPLSgene 방법이 그리고 두 측면 모두 어느 정도의 유사성이 존재하면 SPLScombined 방법이 최적인 것으로 나타났다. 한편, BPCA 방법은 자료의 전체적(global) 유사성 구조를 이용하는 gene-wise 방법으로서 유전자간 국부적 유사성이 상대적으로 약한 Y18 자료와 Y24 자료에 대해서는 로버스트한 성향을 보였으며, 특히 Y24 자료 중 가장 높은 결측율인 20% 수준에서는 SPLS 보다 더 좋은 결과를 보이기도 했다.

이 외에 기존 방법들의 적합수준을 비교 평가해 보면 먼저, PLS는 자료에 따라 적합수준의 변동이 크게 나타났는데 유전자간 국부적 유사성이 강한 Y7 자료는 SPLS와 거의 비등한 결과를 보여 2순위로 나타났으나, 국부적 유사성이 다소 약한 Y18 자료와 Y24 자료에서는 KNN과 SKNN 보다는 좋지만 LLS와 BPCA 보다는 좋지 못한 결과를 보였다. 다음으로, KNN과 SKNN은 거의 모든 결과에서 비슷한 수준을 보였다. 또한, KNN, SKNN과 BPCA 방법간에는 gene-wise 방법만 적용된 점은 같으나 자료의 유사성 구조에 따라 적합수준의 우위가 다르게 나타났다. 즉, Y7 자료는 KNN과 SKNN이 BPCA 보다 더 좋은 결과를 보였고, Y18 자료와 Y24 자료는 BPCA가 KNN과 SKNN 보다 상당히 좋은 결과를 보였

다. 마지막으로, LLS와 KNN 간에는 국부적 유사성 구조를 이용하는 점은 같으나 회귀적 합을 통해 더 정확한 추정치를 얻는 LLS가 더 좋은 적합수준을 보였고, LLS와 BPCA 간에는 국부적 유사성이 약한 자료에 대해서는 BPCA가 더 좋은 적합수준을 보였다.

결측치 추정 방법들의 계산 시간은 KNN, SKNN, BPCA, LLS, PLS 그리고 SPLS 순으로 더 많이 소요된다. SPLS는 PLS를 기초로 하여 순차적으로 진행될 때 점점 더 많아지는 완전 유전자들과 결측 유전자 간의 유사도 측정 시 상당한 시간이 소요되는데 주로 기인한다. SPLS 방법은 이러한 단점은 있으나, 결측치에 대한 추정의 적합수준은 기존 방법들 보다 전반적으로 더 좋은 결과를 보였다. 또한 이는 관측 시점 간 상관성이 큰 시간경로 자료의 특성을 고려하기 위해 시도된 방법이지만, Y7 자료처럼 관측 시점간 상관성이 그리 높지 않은 시간경로 자료, 또는 시간경로 자료가 아닌 경우에 대해서도 기존 방법들보다 개선된 결과를 얻을 것으로 예상된다.

결측치 추정시 한 가지 주목할 사항은 결측치를 추정하는 단계에서 발생한 오차는 뒤따르는 다음 단계의 분석 결과에 영향을 미치게 되므로 결측치 추정의 정확성 문제는 중요하다 할 것이다. 궁극적으로 결측치에 대한 최적 추정치는 본(main) 분석에 가장 적합한 결과를 주는 값으로 결정하는 것이 합당할 것으로 본다. 예를 들면 판별분석을 목적으로 하는 경우는 판별력을 가장 높게 해주는 방향으로 결측치 추정하는 방법을 고려해야 할 것이다. 또한 추정된 결측치를 이용함에 있어서 중요한 생물학적 결론을 이끌 때에는 주의를 요할 필요가 있다.

참고문헌

Abdi, H. (2003). Partial least squares regression (PLS-regression), In *M. Lewis-Beck, A. Bryman, T. Futing (Eds): Encyclopedia for research methods for the social sciences*, Thousand Oaks (CA): Sage, 792–795.

Bø, T. H., Dysvik, B. and Jonassen, I. (2004). LSimpute: Accurate estimation of missing values in microarray data with least squares methods, *Nucleic Acids Research*, **32**, e34.

de Brevern, A. G., Hazout, S. and Malpertuy, A. (2004). Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering, *BMC Bioinformatics*, **5**, 114.

DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, **278**, 680–686.

Garthwaite, P. H. (1994). An interpretation of partial least squares, *Journal of the American Statistical Association*, **89**, 122–127.

Hastie, T., Alter, O., Sherlock, G., Eisen, M., Tibshirani, R., Bostein, D. and Brown, P. (1999). Imputation of missing values in DNA microarrays, *Technical Report Stanford University Statistics Department.*

Höskuldsson, A. (1988). PLS regression methods, *Journal of Chemometrics*, **2**, 211–228.

Jørgensen, B. and Goegebeur, Y. (2006). Module 8: Partial least squares regressions II, *STO2: Multivariate Data Analysis and Chemometrics*, http://statmaster.sdu.dk/courses/ST02.

Kim, H., Golub, G. H. and Park, H. (2005). Missing value estimation for DNA microarray gene expression data: Local least squares imputation, *Bioinformatics*, **21**, 187–198.

Kim, K. Y., Kim, B. J. and Yi, G. S. (2004). Reuse of imputed data in microarray analysis increases imputation efficiency, *BMC Bioinformatics*, **5**, 160.

Nguyen, D. V. and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics*, **18**, 39–50.

Nguyen, D., Wang, N. and Carroll, R. J. (2004). Missing value estimation for cancer microarray gene. expression data, *Journal of Data Science*, **2**, 347–370.

Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K. and Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data, *Bioinformatics*, **19**, 2088–2096.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Molecular Biology of the Cell*, **9**, 3273–3297.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Bostein, D. and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**, 520–525.

[2007년 6월 접수, 2007년 12월 채택]

Missing Values Estimation for Time Course Gene Expression Data Using the Sequential Partial Least Squares Regression Fitting*

Kyungsook Kim[1] Mira Oh[2] Jangsun Baek[3] Young Sook Son[4]

ABSTRACT

The size of microarray gene expression data is very big and its observation process is also very complex. Thus missing values are frequently occurred. In this paper we propose the sequential partial least squares(SPLS) regression fitting method to estimate missing values for time course gene expression data that has correlations among observations over time points. The SPLS method is to combine the sequential technique with the partial least squares(PLS) regression fitting method. The usefulness of method proposed is evaluated through some simulation study for three yeast time course data.

Keywords: Microarray, time course gene expression data, missing value estimation, partial least squares regression fitting, sequential partial least squares regression fitting.

* This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2005-204-C00017).

1) Instructor, Dept. of Statistics, Chonnam National University, 300 Yongbong-dong, Buk-gu, Gwangju 500-757, Korea.
E-mail: ksook620@jnu.ac.kr

2) Post-Doctor, Dept. of Information and Communications, Gwangju Institute of Science Technology, 1 Oryong-dong, Buk-gu, Gwangju 500-712, Korea.
E-mail: omr@chonnam.ac.kr

3) Professor, Dept. of Statistics, Chonnam National University, 300 Yongbong-dong, Buk-gu, Gwangju 500-757, Korea.
E-mail: jbaek@chonnam.ac.kr

4) Corresponding author. Professor, Dept. of Statistics, Chonnam National University, 300 Yongbong-dong, Buk-gu, Gwangju 500-757, Korea.
E-mail: ysson@chonnam.ac.kr