The Korean Communications in Statistics
Vol. 13 No. 1, 2006, pp. 101-111

A Comparative Study of Microarray Data with Survival Times Based on Several Missing Mechanism[1)]

Jeeyun Kim[2)], Jinsoo Hwang[3)] and Seong Sun Kim[4)]

Abstract

One of the most widely used method of handling missingness in microarray data is the kNN(k Nearest Neighborhood) method. Recently Li and Gui (2004) suggested, so called PCR(Partial Cox Regression) method which deals with censored survival times and microarray data efficiently via kNN imputation method. In this article, we try to show that the way to treat missingness eventually affects the further statistical analysis.

Keywords : microarray; missingness; PCR; imputation.

1. 서론

마이크로어레이 자료는 통상적으로 변수(유전자)의 수가 관측값(표본)의 수보다 훨씬 많다는 것은 널리 알려진 사실이다. 실제 실험을 통하여 일차적으로 얻게 되는 마이크로어레이 자료는 결측치를 포함하는 경우가 많게 된다. 결측치 발생의 원인은 여러 가지가 있겠지만 대체로 이미지의 깨짐, 슬라이드에서 먼지나 긁힘 등으로 인하여 발생하는 것으로 알려져 있다.

본 연구에서는 중도 절단된 생존시간의 자료를 포함하는 마이크로어레이 자료에서 결측치를 처리하는 여러 방법을 소개하고 각 처리 방법에 따라서 PCR 생존회귀모형을 적용하고 이 모형의 타당성을 평가하는 방법 중의 하나인 위험집단간의 생존함수의 차이 검정을 실제 자료에 적용하여 비교하고자 한다.

결측치 처리 방법으로는 몇 개의 주요한 eigengene 들을 사용하는 SVD방법과 k개의 유사한 유전자들의 가중평균값으로 추정하는 kNN방법, 결측치가 많은 경우에 대치된 유전자를 재사용하는 SkNN(Sequential kNN)방법, 확률적인 주성분모형(Probabilistic PCA, Bishop (1999))에서 베이지안 방법으로 모수를 추론하는

---

1) This work was supported by INHA UNIVERSITY Research Grant. (INHA-31571)

2) Researcher, Institute for Basic Science at Inha University.
Correspondence : jeeyun@stat.inha.ac.kr

3) Professor, Department of Statistics, Inha University, Incheon 402-751, Korea.

4) Research Scientist. Division of Epidemic Intelligence Service, KCDC.

BPCA(Bayesian PCA)방법, 상관계수 또는 L_2-norm을 이용하여 가까운 유전자 일부를 선택하여 회귀모형을 이용하는 LLS(Local Least Square)방법, 대치의 불확실성 평가할 수 있는 다중대치(Multiple Imputation, MI)방법, 그리고 가장 간편한 방법으로 결측치가 속한 열이나 행에서 랜덤하게 선택하여 대치하는 Random (or Hot-deck)방법 등이 있다.

위의 여러 가지 결측치 처리 방법에 따라서 결측치를 대치한 후에 변수 수(p)가 실험 수(n) 보다 훨씬 많은($n << p$) 생존회귀모형을 적합하기 위하여 사용하는 방법으로는 Park *et al.* (2002)의 연구를 포함한 여러 연구가 있으나 본 연구에서는 부분최소제곱회귀(Partial Least Square Regression, PLSR)의 아이디어를 Cox 회귀모형에 접목시켜 Li and Gui (2004)가 제안한 Partial Cox Regression(PCR)방법을 사용하였다. PCR 방법의 개선된 형태로는 Efron *et al.* (2004)이 제안한 Stage-wise regression 인 LARS(Least Angle Regression)을 이용하여 Tibshirani (1997)가 제안한 Lasso를 마이크로어레이 자료에 사용할 수 있도록 변형한 Gui and Li (2004)의 LARS-Lasso 방법과 그 방법의 계산상의 문제를 발전시킨 Segal (2005)의 residual finesse방법 등이 있다.

유전자와 생존시간과의 연관 모형으로는 Cox 회귀모형이 아닌 가속 수명모형이나 일반 선형모형(예 Buckley-James 모형) 등 여러 가지가 있을 수 있으나 앞서 언급한 것처럼 본 연구에서는 여러 가지 결측치 처리 방법에 따른 추후 분석의 민감도를 조사하는 것에 초점을 맞추었기 때문에 생존회귀분석방법으로는 가장 보편적인 Cox 회귀모형에 의거하여 PCR 방법만을 사용하여 비교 연구 하였다. 실험 자료는 Rosenwald *et al.* (2002)가 사용했던 동일한 DLBCL(Diffuse Large B-cell Lymphoma) 자료를 여러 결측치 처리 방법에 따라 PCR 모형에 적용하여 위험집단들 사이의 유의확률을 사용하여 비교하였다. 제 2절에서는 사용되는 여러 결측치 처리 방법들에 대하여 간단하게 설명을 하였고 제 3절에서는 차원축소를 통한 생존회귀분석에 대하여 살펴보았고 제 4절에서는 실제 자료에 적용한 결과를 보여주며 제 5절에서는 결론과 향후 발전방향을 논의한다.

2. 결측치 처리방법

(1) SVD(Singular Value Decomposition)대치

Hastie *et al.* (1999)에서 제안한 방법으로 유전자 수가 p이고 실험 수가 n인(보통 $n << p$) 유전자 발현 행렬을 $A_{p\times n}$이라 할 때 이 행렬을 다음과 같이 분해하는 것을 SVD라고 한다.

$$A_{p\times n} = U_{p\times p}\Sigma_{p\times n}V^T_{n\times n}$$

여기서 행렬 V^T는 서로 직교하는 고유유전자(eigengene)벡터를 포함하고 있으며 고유공간(eigenspace)에서의 기여도는 대각 행렬과 영 행렬의 결합인 Σ에서의 고유치(eigenvalue)로 정해진다. 고유치의 크기 순서로 k개의 유의한 고유유전자를 선택한 후 이를 이용하여 i번째 유전자의 j번째 실험의 결측치를 회귀대치(regression

imputation)법을 이용하여 반복적으로 대치한다. SVD 방법을 사용하려면 먼저 행렬에 결측치가 없어야 하므로 처음 사용할 때는 행의 평균값으로 대치를 하여 진행한다. 즉, 초기값으로 행의 평균을 사용한 후 EM방법을 사용하여 반복적으로 회귀대치법을 수행하여 행렬의 변화가 미리 정한 허용한계(보통 0.01)이내에서 반복을 멈춘다.

(2) kNN(k Nearest Neighborhood)대치

Troyanskaya *et al.* (2001)에서 제안한 방법으로서 결측치를 포함하는 유전자와 가까운 발현 패턴을 갖는 유전자 k개를 선택한다. 유전자간의 거리는 상관계수, 유클리드 거리 등을 사용하는데 통상적으로 유클리드 거리를 사용한다. 이렇게 선택된 k개의 유전자를 이용하여 이들의 가중평균(보통 거리의 역수)으로 결측치를 대치하는 방법이다.

(3) SkNN(Sequential kNN)대치

Kim *et al.* (2004)에서 제안한 방법으로서 유전자 발현 행렬을 결측치가 없는 완전 행렬과 결측치를 포함하는 불완전 행렬로 나누고 불완전 행렬에서 결측치 포함비율이 가장 낮은 유전자를 결측치가 없는 완전한 유전자 집단에서 kNN 방법으로 대치를 하여 대치된 유전자를 완전한 유전자 집단으로 옮기고 축차적으로 다음으로 결측치가 적은 유전자를 kNN 방법으로 대치를 해나가는 방법이다. 이 방법은 유전자 발현 자료에 결측치 포함비율이 많은 경우에 유용한 방법이라고 할 수 있다.

(4) BPCA(Bayesian Principal Component Analysis)대치

Oba *et al.* (2003)에서 제안한 BPCA는 기본적으로 주성분회귀에서 얻어진 적은 수의 주축들(principle axis)과 인자점수(factor scores)를 이용하여 결측값을 대치한다. BPCA의 과정을 단계적으로 나누어보면 3단계로 이루어져 있다. 첫 단계는 주성분회귀 과정이며 두 번째 단계는 주성분에 대한 인자점수와 오차항 및 모수들에 대한 베이지안 모형을 세우고 마지막 단계는 EM과 유사한 방법으로 모수들을 추정하는 단계이다. 각 과정을 간단하게 살펴보면 다음과 같다.

① 주성분회귀

특정한 유전자의 발현벡터를 $y_{n\times 1}$라 할 때 주성분회귀모형은

$$y = \sum_{l=1}^{k} x_l w_l + \epsilon \quad (k << p)$$

와 같이 표현된다. 여기서 w_l은 주축벡터이며 x_l은 인자점수로서 p개의 유전자 벡터들의 분산-공분산 행렬의 고유벡터$(u_1, \cdots, u_n)$와 고유근$(\lambda_1, \cdots, \lambda_n)$으로 다음과 같이 표기된다.

$$w_l = \sqrt{\lambda_l}\, u_l \ , \ x_l = (w_l / \sqrt{\lambda_l})^T y \, .$$

관측된 자료들에서 최소제곱법으로 위의 주성분회귀모형에서 계수역할을 하는 인자점수 $x = (x_1, \cdots, x_k)$를 추정하고 추정된 인자점수를 이용하여 결측치를 다음과 같이 추정한다.

$$y^{\text{miss}} = W^{\text{miss}} x .$$

여기서 W^{miss}는 주축벡터들 중에서 결측치에 해당하는 부분만을 포함하는 부분 행렬이다.

② Bayesian estimation

주성분회귀모형에서 인자점수 x_l와 오차항 ϵ에 대한 정규분포의 확률모형은 확률적인 주성분분석이라는 이름으로 Bishop (1999)에 의하여 제안되었고 이 모형에서 모수들을 베이지안 방법으로 추론하는 것이다. 즉 전체 자료집합 $\boldsymbol{Y}=\{y\}$가 주어졌을 때 모형의 모든 모수를 나타내는 $\boldsymbol{\theta}$와 인자점수 $\boldsymbol{X}$의 사후분포는 다음과 같이 표현된다.

$$p(\theta, \boldsymbol{X} | \boldsymbol{Y}) \propto p(\boldsymbol{Y}, \boldsymbol{X} | \theta) p(\theta)$$

모수들의 집합인 θ는 인자점수와 오차항의 정규분포의 모수들과 주축벡터 W를 포함한다.

BPCA의 주요한 부분 중 하나는 주축벡터를 나타내는 W의 계층적 사전분포로서 이는 특정한 주축의 크기가 작으면 자동적으로 불필요하다고 판단되어 없어지게 만들어 준다. 따라서 이를 자동적정성결정(Automatic Relevance Determination, ARD) 사전분포라고 부른다.

③ EM-like repetitive algorithm

만일 우리가 모수의 참값 $\boldsymbol{\theta}_{\text{true}}$ 을 알고 있다면 결측치들의 사후분포는

$$q(\boldsymbol{Y}^{\text{miss}}) = p(\boldsymbol{Y}^{\text{miss}} \mid \boldsymbol{Y}^{\text{obs}}, \theta_{\text{true}})$$

로 주어지나 모르기 때문에 모수의 사후분포 $q(\theta)$를 이용하면 결측치들의 사후분포가

$$q(\boldsymbol{Y}^{\text{miss}}) = \int \mathrm{d}\,\theta q(\theta) p(\boldsymbol{Y}^{\text{miss}} \mid \boldsymbol{Y}^{\text{obs}}, \theta)$$

로 주어진다. 따라서 모수와 결측치 자체를 추정하는 EM 알고리즘과 유사한 VB 알고리즘(variational Bayes)을 사용하여 반복적인 방법으로 사후분포 $q(\theta)$와 $q(\boldsymbol{Y}^{\text{miss}})$를 추정하여 최종적으로 다음 식을 이용하여 결측치를 추정하게 된다.

$$\hat{\boldsymbol{Y}}^{\text{miss}} = \int \boldsymbol{Y}^{\text{miss}} q(\boldsymbol{Y}^{\text{miss}}) \mathrm{d}\boldsymbol{Y}^{\text{miss}}$$

(5) LS(Least Squares)대치

Bo *et al.* (2004)에서 제안한 방법으로서 최소제곱법에 근거한 결측치 처리 방법으로는 LSimpute_gene과 LSimpute_array 그리고 이 둘을 조합한 LSimpute_combined 방법이 있다. LSimpute_gene 방법은 결측치가 있는 유전자와 상관계수의 절대값이 큰 순서로 k개의 유전자를 선택하여 각 k개의 유전자에 대하여 각각 단순선형회귀모형을 이용하여 k개의 결측치의 예측값을 구하여 이들의 가중평균값으로 대치를 하는 방법이다. 보통 가중치는 상관계수의 절대값에 비례한다. LSimpute_array 방법은, 예를 들어 m개의 결측치와 $n-m$개의 완전값을 포함하는 유전자에서 다중회귀모형을 이용하여 결측치를 추정한다. 다중회귀모형을 사용하려면 모든 자료가 완전자료이어야 하므로 LSimpute_gene방법을 이용하여 초기값으로 결측치를 추정한 다음에 사용한다. LSimpute_combined 방법은 LSimpute_gene 과 LSimpute_array에서 얻어진 값들의 가중평균값으로 대치하는 방법이다.

(6) LLS(Local LS) 대치

Kim *et al.* (2005)에서 제안한 LLS 대치는 크게 두 가지 단계로 나눌 수 있다. 첫째는 L_2-norm이나 상관계수를 이용하여 k개의 유전자 그룹을 정하는 것이다. 두 번째는 회귀모형으로 추정하는 단계이다. k개의 유전자를 선택하는 과정은 기존의 다른 방법과 동일한 과정이다. 유전자의 결측치가 한 개 있는 경우를 가정하면 k개의 가까운 유전자벡터들에서 행렬 $A \in R^{k \times (n-1)}$와 두 벡터 $b \in R^{k \times 1}$와 $w \in R^{(n-1) \times 1}$을 정한다. 이 후에 최소제곱법($\min_x ||A^T x - w||_2$)으로 x를 구한 후 결측치 $\alpha = b^T x$를 구한다. 결측치가 둘 이상인 경우에도 유사하게 확장하면 된다. 실험벡터를 이용한 LLS 방법은 최소제곱을 $\min_y \|Ay - b\|_2$로 하여 y를 구한 후 결측치를 $\alpha = w^T y$와 같이 구한다. 유전자를 이용한 최소제곱법이나 실험을 이용한 최소제곱법의 결측치 해법은 아래의 식에서 처럼 동일한 결과를 준다.

$$b^T x = b^T (A^T)^+ w = (w^T A^+ b)^T = w^T y.$$

단, A^+ 는 A 의 pseudoinverse 이다.

기본적으로 LS 방법이나 LLS 방법은 회귀분석을 이용하여 대치를 하는 방법이다. LS 방법에서는 k개의 근사유전자 각각의 단순회귀를 통한 예측치의 가중평균이며 LLS는 다중회귀를 통하여 결측치를 대치한다.

(7) MI(Multiple Imputation)

Rubin (1977)에 의하여 처음 제안된 다중대치방법은 결측치를 하나의 값으로만 대치하는 것이 아니라 여러 개의 값으로 대치를 하여 대치시마다 달라지는 변동을 측정할 수 있도록 하여 대치의 불확실성을 측정할 수 있다. 예측분포 $p(Y_{\text{miss}})$에서 D개의 값을 반복적으로 추출하여 추정하고자 하는 모수 θ에 대한 완전자료 D개의 추정치와 분산을 $\hat{\theta}_d, W_d, d = 1, \cdots, D$라 하면 통합된 점 추정량은 $\bar{\theta}_D = \sum_{d=1}^{D} \hat{\theta}_d / D$와 같고 대치-내(within-imputation)평균 분산은 $\overline{W}_D = \sum_{d=1}^{D} W_d / D$가 되고 대치-간(between-imputation) 분산은 $B_D = \sum_{d=1}^{D} (\hat{\theta}_d - \bar{\theta}_D)^2 / (D-1)$가 되며 전체의 분산은 두 분산으로부터 $T_D = \overline{W}_D + (D+1) B_D / D$가 된다. 대부분의 단일대치는 B_D부분을 무시하여 전체분산을 과소 추정하는 경향이 있으나 다중대치는 결측치의 올바른 분산을 구해준다.

다중대치의 적용은 R 패키지에 포함되어 있는 MICE routine을 사용하였으나 일반적으로 R에서 제공되는 다중대치 방법은 마이크로어레이 자료와 같은 커다란 행렬의 자료에 부적당하므로 실제 프로그램이 작동하게 하려면 여러 가지 측면에서 계산이 가능하도록 프로그램을 수정해야한다.

3. 차원축소를 통한 생존회귀분석

생존시간을 포함하는 마이크로어레이 자료의 분석 방법은 중도 절단 자료의 처리문제와 변수의 수가 자료의 수보다 많기 때문에 발생하는 문제의 처리 방법에 따라서 여러 방향으로 연구가 가능하다. 본 연구에서는 Li and Gui (2004)가 제안한 PCR 방법에 대하여 간략하게 소개하고자 한다.

PCR은 PLS(Partial Least Squares)의 아이디어를 이용하여 위험률과 특정 유전자 (X_j)와의 관계식을 유도한 것이다. $X_1, \cdots, X_p$ 는 유전자를 나타내고, $x_j = \{x_{1j}, x_{2j}, \cdots, x_{nj}\}^T$ 는 j번째 유전자의 관측치를 나타낸다. 샘플을 사람으로 할 때 $(t_i, \delta_i, x_{i1}, x_{i2}, \cdots, x_{ip})$는 n개의 샘플 중에 i번째 사람의 자료를 나타낸다. 여기서 t_i는 생존시간, δ_i는 센서링 지시기를 나타낸다. PCR은 서로 직교하는 유전자의 선형결합$(T_1, T_2, \cdots, T_k)$을 축차적으로 찾아서 이들을 이용하여 Cox의 비례위험모형을 적합시킨다. 즉,

$$\begin{aligned}\lambda(t) &= \lambda_0(t)\exp(\beta_1 T_1 + \beta_2 T_2 + \cdots + \beta_k T_k) \\ &= \lambda_0(t)\exp[f(X)],\end{aligned}$$

여기서 $\lambda_0(t)$는 기저 위험함수이고, 각 선형결합 T_i는 $X = \{X_1, \cdots, X_p\}$의 선형결합이므로 T_i들의 선형결합 역시 X의 함수형태 $f(X)$로 표현할 수 있다. $T_i (i = 1, \cdots, k)$ 를 구하는 과정을 간단히 설명하면 다음과 같다. 먼저 유전자별로 자료의 평균을 0으로 만든다. 즉, $V_{1j} = X_j - \bar{x}_{.j}$ 단, $\bar{x}_{.j} = \sum_{i=1}^{n} x_{ij}/n$ 이다. 표준화된 $V_{1j} (j = 1, \cdots, p)$ 에 대하여 Cox 회귀모형을 적합 시킨다. $\lambda(t) = \lambda_0(t)\exp(\beta_{1j} V_{1j})$. 이 때 추정된 회귀계수를 이용하여 유전자들의 선형결합(T_1)을 구한다.

$$T_1 = \sum_{j=1}^{p} w_{1j}\hat{\beta}_{1j} V_{1j} \quad , \quad \sum w_{1j} = 1 \ .$$

여기서 w_{ij} 는 PCR 가중치로 V_{ij} 의 분산에 비례한다. 다음은 V_{1j}를 T_1에 회귀한 잔차 V_{2j}를 구하고 (T_1, V_{2j})를 이용하여 Cox 회귀모형을 적합시킨다. 즉,

$$V_{2j} = V_{1j} - \frac{V_{1j}^T T_1}{T_1^T T_1} T_1 \ ,$$

$$\lambda(t) = \lambda_0(t)\exp(\beta_1 T_1 + \beta_{2j} V_{2j}) \ .$$

추정된 계수를 이용하여 유전자들의 선형결합(T_2)을 구한다. $T_2 = \sum_{j=1}^{p} w_{2j}\hat{\beta}_{2j} V_{2j}$. 같은 방법으로 축차적으로 $T_3, \cdots, T_k$ 를 구한다.

$$V_{(i+1)j} = V_{ij} - \frac{V_{ij}^T T_i}{T_i^T T_i} T_i \quad ,$$

$$T_{i+1} = \sum_{j=1}^{P} w_{(i+1)j}\hat{\beta}_{(i+1)j} V_{(i+1)j} \ .$$

이처럼 구해진 구성요소 $T_1, T_2, \cdots$를 PCR 구성요소라 부르며 이러한 k개의 구성요소를 이용하여 Cox 모형의 위험점수 $f(X)$를 적당한 계수 β_j^*에 대하여 원래 변수인 X로 표현한다.

$$f(X) = \sum_{j=1}^{p} \beta_j^* V_{1j} = \sum_{j=1}^{p} \beta_j^* (X_j - \bar{x}_{.j}) \quad .$$

또 하나의 방법으로 Gui and Li (2004)가 제시한 발전적인 방법은 Tibshirani (1997)가 제시한 Lasso 방법과 Efron *et al.* (2004)이 제안한 LARS 방법을 융합한 LARS-Lasso 방법이 있다. 유전자들 간의 높은 상관성과 자료의 수보다 많은 변수의 문제를 극복하기 위하여 일반적으로 penalized 가능도함수를 사용하게 된다. 이러한 penalized 가능도함수의 관점에서 본다면 Li and Luan (2003)은 L_2-penalty를 Cox 모형에 사용하였고, Gui and Li (2004)는 LARS-Lasso에서 L_1-penalty를 사용하였다. 최근에는 L_1과 L_2 penalty를 결합한 모형으로는 Zou and Hastie (2005)가 제안한 Elastic-Net방법이 제시되었다. 이 외에도 연관된 연구로는 Segal (2005)이 Li and Gui (2004) 방법에서 계산상의 복잡함과 가법적인 Cox 모형에서도 사용할 수 있는 "residual finesse" 방법을 제안하였다. 본 연구는 결측치 처리방법의 민감성을 보여주는 것이 주된 목적이므로 여러 결측치 처리 방법에 따른 생존회귀모형은 PCR 방법에 국한하여 비교하였다.

4. 실제자료의 적용

Rosenwald *et al.* (2002)에서 사용한 Diffuse Large B-cell lymphoma(DLBCL) 자료는 환자 240명이며 유전자는 7399개가 있는데 이중 결측치를 포함하는 유전자수는 6925개가 된다. 이 자료를 보면 결측치를 거의 대부분이 포함하고 있음을 알 수 있다. 환자들의 생존시간은 0-21.8년이고 조사 종료 후까지 생존자는 102명(중도절단 자료), 조사기간 중 사망자는 138명이다. 먼저 주어진 DLBCL 자료를 kNN, SkNN, BPCA, LLS, MI, Random방법으로 결측치를 처리한 후 훈련 집합(160명)과 검사 집합(80명)으로 나누어 훈련 집합에서 모형을 세우고 검사 집합에서 적합하여 위험점수를 계산하였다. 위험점수를 양수와 음수 집단으로 나누어 고위험군과 저위험군으로 나누어 두 군간의 차이를 알아보기 위하여 군별 Kaplan-Meier curve 와 log-rank test를 하였다. <표 1>은 결측치를 여러 가지 방법들에 의해 처리한 후 훈련 집합을 PCR 모형에 적합 시켰을 때의 각 구성요소의 유의확률 값을 보여주고 있다.

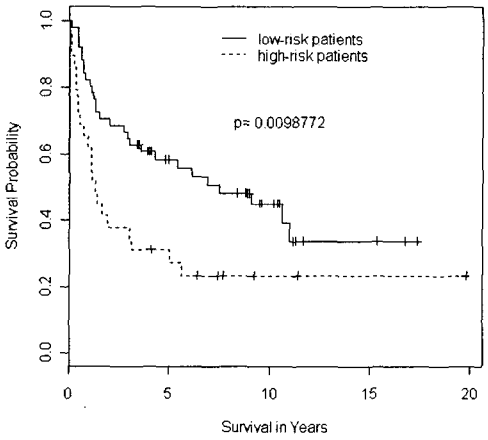
kNN 방법을 사용한 것은 첫 번째부터 일곱 번째 구성요소까지가 유의함을 보이고 있고, SkNN, BPCA, LLS은 첫 번째부터 여섯 번째, 그리고 여덟 번째 구성요소가 유의하고, MI는 첫 번째부터 여섯 번째 까지, Random은 첫 번째, 두 번째, 여섯 번째 그리고 아홉 번째가 유의함을 보여주고 있다. 대체적으로 7 개 정도의 구성요소가 모형에 유의하게 나오고 있음을 알 수 있다. Li and Gui (2004)에서는 PCR을 하기 전에 주성분분석을 하는 PC-PCR방법을 추가로 제안하였고 그 모형의 경우에는 유의한 요소가 3 개로서 PCR보다 적게 나오고 있음을 보고하고 있다.

<표 1> 결측치 처리방법에 따른 구성요소들의 유의확률 값

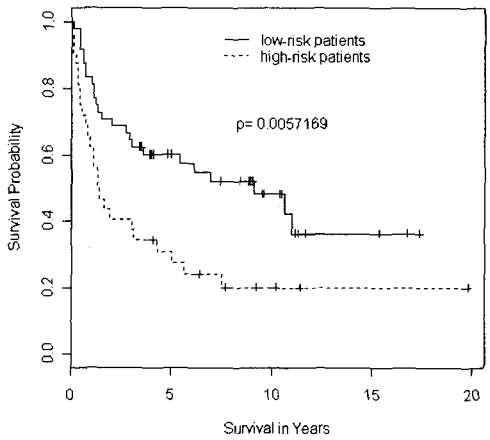
구성 요소	결측치 방법					
	kNN	SkNN	BPCA	LLS	MI	Random
1	7.64E-13	6.38E-13	1.09E-12	9.80E-13	4.45E-13	4.73E-14
2	3.50E-09	7.70E-09	1.32E-08	1.00E-08	4.50E-09	7.37E-06
3	9.19E-10	4.77E-09	1.70E-08	5.59E-09	4.96E-10	5.38E-01
4	1.79E-05	2.02E-05	4.08E-05	2.85E-05	2.15E-05	3.54E-01
5	1.55E-03	7.71E-04	7.89E-04	1.48E-03	1.09E-03	7.54E-02
6	4.18E-03	1.47E-03	1.13E-03	1.20E-03	2.94E-03	3.59E-02
7	3.05E-02	7.35E-02	1.28E-01	8.92E-02	5.35E-02	6.50E-02
8	7.83E-02	2.46E-02	1.21E-02	1.76E-02	1.04E-01	9.83E-01
9	2.98E-01	1.69E-01	9.09E-02	2.15E-01	2.92E-01	1.30E-02
10	2.51E-01	2.77E-01	9.52E-02	2.06E-01	3.64E-01	1.17E-01

Li and Gui (2004)에서는 PCR을 하기 전에 주성분분석을 하는 PC-PCR방법을 추가로 제안하였고 그 모형의 경우에는 유의한 요소가 3 개로서 PCR보다 적게 나오고 있음을 보고하고 있다.

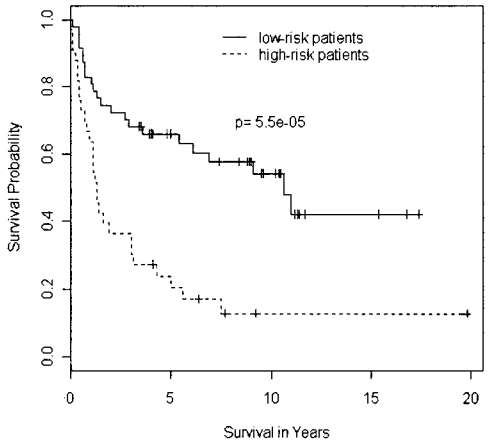
다음 그림은 각 결측치 처리 방법별로 유의한 구성요소만을 포함한 모형을 검사 집합에 적합 시켜 위험점수를 계산하여 저위험군과 고위험군으로 나누어 Kaplan-Meier curve와 log-rank test의 유의확률 값을 표시한 결과이다.



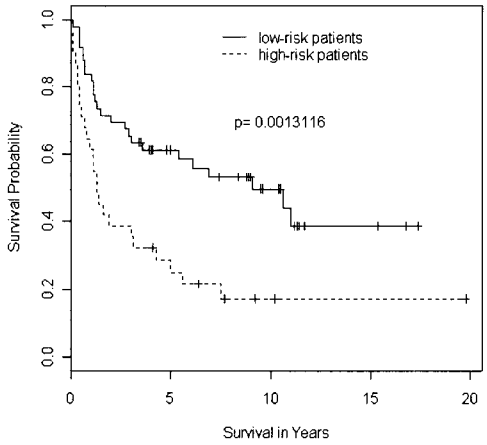
<그림 1> kNN : 구성요소 수=7



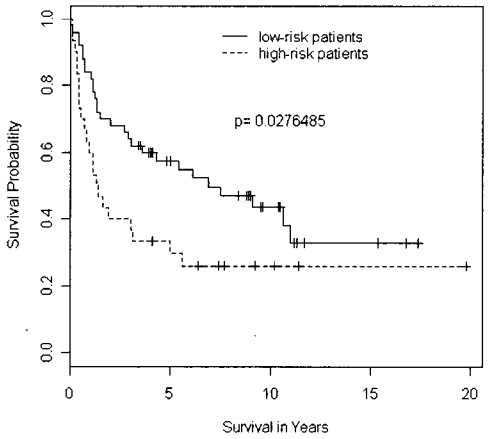
<그림 2> SkNN : 구성요소 수=7



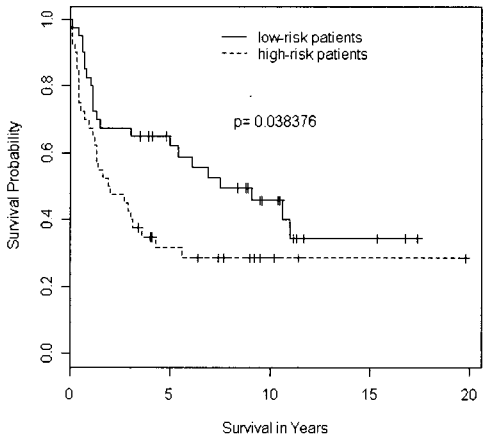
<그림 3> BPCA : 구성요소 수=7



<그림 4> LLS : 구성요소 수=7



<그림 5> MI : 구성요소 수=6



<그림 6> Random : 구성요소 수=4

5. 결론 및 토의

앞 절의 결과를 보면 결측치를 처리하는 방법에 따라서 저위험군과 고위험군간의 변별력이 차이가 남을 알 수가 있다. kNN 방법에서는 Li and Gui (2004)의 결과와 같은 유의확률 값을 얻을 수 있었고 그들의 논문에서 주장한 개선된 PC-PCR 방법(주성분분석을 한 후 PCR을 적용)에서도 두 군간의 유의확률 값은 0.0033으로 나왔다. 다른 결측치 처리 방법인 LLS 와 BPCA를 사용한 후 PCR모형을 적용하여 동일한 방법으로 저위험군과 고위험군간의 유의확률값은 0.0033보다 작은 값이 나왔다. 이 결과에 의하면 BPCA가 가장 변별력이 있게 나왔으나 이것이 BPCA방법의 우월성을

이야기한다고 할 수는 없다. 즉 결측치 처리방법에 따라서 동일한 분석방법인 PCR을 비교하려면 예측력을 구하여 좋은 예측력을 보이는 방법을 찾으면 될 것이나 현재의 결과를 가지고는 예측력을 비교할 수는 없다는 것이다. 생존회귀모형과 무관하게 결측치 처리 방법들 간의 비교에서는 BPCA 방법이나 LLS방법이 kNN방법보다는 좋은 방법이라는 연구 결과가 보고되고 있다. 본 논문에서는 생존 자료가 포함된 마이크로어레이 자료에서의 결측치를 처리하는 방법에 따라서 동일한 생존회귀분석방법을 사용하여도 유의하게 다른 결과가 나올 수 있음을 보여주고 있다. 결측치를 처리하는 방법 중에서 MI 방법은 결측치를 대치하는 불확실성을 측정할 수 있기 때문에 개념적으로 바람직한 방법이라고 여겨지고 있으나 실제 계산 결과 값은 만족스럽지 않았다.

모형의 적정성을 서로 비교하는데 있어서 일반적인 선형모형과는 달리 센서링이 있는 생존회귀모형에서는 잔차에 대한 일관성 있는 접근이 어렵기 때문에 모형의 적합도 또는 예측력을 비교하려면 ROC 곡선을 이용하여 AUC(area under curve)를 구하여 비교하는 방법을 사용하는 것도 하나의 방법이라고 생각된다.

결측치 처리 방법들만의 비교와 생존회귀모형들만의 비교도 의미가 있겠지만 이들을 결합한 결과의 비교도 의미가 있다고 생각된다. 향후 과제로는 LARS-Lasso 방법이나 L_1과 L_2 penalty의 결합형태인 Elastic-Net 방법을 포함하는 여러 가지 마이크로어레이 생존회귀모형과 결측치 처리 방법들 간의 폭 넓은 비교 연구가 필요하다.

참고문헌

[1] Bishop, C.M. (1999). Variational principal components. *In IEE Conference Publication on Artificial Neural Networks*, 509-514.

[2] Bo, T.H., Dysvik, B. and Jonassen, I. (2004). Lsimpute:accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Research*, Vol. 32, No.3 e34.

[3] Efron, B., Johnston, I., Hastie, T. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, Vol. 32, 407-499.

[4] Gui, J. and Li, H. (2004). Penalized Cox Regression Analysis in the High-Dimensional and Low-sample Size Settings with Applications to Mi-croarray Gene Expression Data. *Center for Bioinformatics & Molecular Biostatistics.*

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

[6] Kim, H., Golub, G.H. and Park, H. (2005). Missing value estimation for DNA microarray gene expression data : local least squares imputation. *Bioinformatics*, Vol. 21, 187-198.

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

[8] Li, H. and Gui, J. (2004). Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, Vol. 20, i208-i215.

[9] Li, H. and Luan, Y. (2003). Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing*, 65-76.

[10] 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*, Vol. 19, 2088-2096.

[11] Park, P.J., Tian, L. and Kohane, I.S. (2002). Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, Vol. 18, S120-S127.

[12] Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E., Fisher, R.I., Gascoyne, R.D., Muller-Hermelink, H.K., Smeland, E.B. and Staudt, L.M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine*, Vol. 346, 1937-1947.

[13] Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, Vol. 72, 538-543.

[14] Segal, M.R. (2005). Microarray gene expression data with linked survival phenotypes : Diffuse large-B-cell lymphoma revisited. *Center for Bioinformatics & Molecular Biostatistics.*

[15] Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, Vol. 16, 385-395.

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

[17] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, Vol. 67, 301-320.

[Received August 2005, Accepted December 2005]