

## Classification of Microarray Gene Expression Data by MultiBlock Dimension Reduction<sup>1)</sup>

Mira Oh<sup>2)</sup>, Seo Young Kim<sup>3)</sup>, Kyungsook Kim<sup>4)</sup>,  
Jangsun Baek<sup>5)</sup> and Young Sook Son<sup>6)</sup>

### Abstract

In this paper, we applied the multiblock dimension reduction methods to the classification of tumor based on microarray gene expressions data. This procedure involves clustering selected genes, multiblock dimension reduction and classification using linear discrimination analysis and quadratic discrimination analysis.

*Keywords* : Principal component analysis; partial least square method; linear discrimination analysis; quadratic discrimination analysis.

### 1. 서론

최근 DNA 마이크로어레이(microarray) 기술은 대량의 유전자 발현 자료를 동시에 측정하는 것을 가능하게 하였다. 마이크로어레이 자료 분석은 종양에 대한 복잡한 생물학적 이해를 가능하게 함으로써 생물학이나 의학 분야에서 암을 진단하고 예측하는데 유용하게 사용되고 있다(Golub 등, 1999; Dudoit 등, 2000; Ben-Dor 등, 2000). 특정 암의 진단에 있어서 기존의 조직검사만으로는 암의 진단이 쉽지 않기 때문에 마이크로어레이 유전자 발현자료에서 암표본과 정상표본의 정확한 분류는 암의 진단과 치료에 매우 도움을 줄 것이다. 표본 분류에 있어서 성공적인 분류를 위한 첫 번째 단계는 차원축소(dimension reduction)라 할 수 있다(Ben-Dor 등, 2000; Weston 등, 2001). 특히, 유전자 자료의 경우 표본을 분류하는데 사용되는 유전자(변수 혹은 차원)

---

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

2) Graduate student, Department of Statistics, Chonnam National University, Kwangju, 500-757, Korea.

3) BK21 post doctor, Department of Statistics, Chonnam National University, Kwangju, 500-757, Korea.

4) Graduate student, Department of Statistics, Chonnam National University, Kwangju, 500-757, Korea.

5) Professor, Department of Statistics, Chonnam National University, Kwangju, 500-757, Korea.

6) Professor, Department of Statistics, Chonnam National University, Kwangju, 500-757, Korea.

수는 수백에서 수천, 수만 개에 이른다. 따라서, 분류에 유의한 유전자 후보를 선택하고 선택된 유전자에 대한 차원축소는 분류분석에 앞서 선행되어야 할 매우 중요한 과정이다. 본 연구에서는 이미 chemometrics 등에서 사용되었던 다중 블록 차원축소법을 마이크로어레이 유전자 자료에 적용하여 다중 블록 차원축소 후 표본 분류를 수행하고자 한다.

마이크로어레이 유전자 발현자료의 분류를 위한 다양한 분류 방법들이 제안되었다. Golub 등 (1999)은 weighted voting 방법에 의해 분류하였고, Alon 등 (1999)은 군집에 기초로 한 deterministic-annealing algorithm에 의해 분류하였다. Furey 등 (2000)은 Support Vector Machine (SVM)방법을 사용하였고, Nguyen 등 (2002)은 주성분 분석(Principal Component Analysis: PCA)과 부분 최소제곱법(Partial Least Squares: PLS)을 이용한 차원축소에 의해 로지스틱 판별(Logistic Discrimination: LD)과 이차 판별분석(Quadratic Discriminant Analysis: QDA) 방법을 사용하였다.

이러한 선행연구들은 분석에 사용한 유전자 자료 전체에 대한 차원축소, 즉, 전체 차원축소(global dimension reduction)에 기초한 분류 결과들을 제시하였다. 그러나, 유전자 자료는 서로 유사한 유전자 기능을 갖는 부그룹(subgroup)으로 나누어질 수 있기 때문에 각 부그룹별로 차원축소를 수행하는 것이 정보의 관련성 측면에서 더 효과적이라 여겨진다. 차원축소 방법 중 다중 블록(MultiBlock: MB) PCA(MBPCA), 다중 블록 PLS (MBPLS)는 Wold 등(1987a,b)이 제안하였고, 이 방법은 주로 화학실험에서 얻은 데이터의 정확한 성분예측을 하기 위해 적절한 모델을 적용하여 중요한 화학적 정보를 얻는 chemometrics 기술 분석 등에 응용되어 사용되고 있다. 본 논문에서는 이러한 다중 블록 차원축소법을 마이크로어레이 유전자 자료에 적용해 보았다. 먼저 t-test 방법에 의해 선택된 분류에 유의한 유전자자료에 대해 군집화를 수행하여 얻어지는 여러 개의 군집들을 다중 블록이라 하자. 그리고 각 블록별로 PCA 혹은 PLS에 의해 각각 차원축소를 수행하는 MBPCA 혹은 MBPLS 같은 다중 블록 차원축소를 수행한다. 차원축소된 자료에 대해 선형 판별분석(Linear Discriminant Analysis: LDA)과 QDA에 의해 분류분석을 수행한다.

본 논문에서 논의된 다중 블록 차원축소법은 전립선암 자료(Singh 등, 2002)와 유방암 자료(van 't Veer 등, 2002)에 적용하여 다중 블록 차원축소에 의한 분류 결과의 정확성을 살펴본다. 2절에서는 다중 블록 차원축소에 의한 분류절차를 소개한다. 3절에서는 실제 두 가지 암 자료에 대한 분석 결과를 설명하고, 마지막으로 4절에서는 결론 및 논의사항을 언급한다.

## 2. 다중 블록 차원축소에 의한 분류절차

마이크로어레이 기술에 의한 대부분의 유전자 발현 자료는 표본의 수( $N$ )에 비해 유전자(변수, 차원)의 수( $p$ )가 훨씬 많은 경우( $N \ll p$ )에 해당된다. 일반적으로 통계적인 분류 분석에서 차원이 표본의 수보다 훨씬 많은 경우는 판별정보를 전혀 제공하지 못할 뿐만 아니라, 분석에서 필요한 행렬 계산의 문제를 야기 시킨다(Antoniadis 등, 2003). 그러므로 높은  $p$  차원 유전자 수를 훨씬 낮은  $M$  차원 유전자수( $M \ll p$ )로 축

소하는 것이 바람직하다.  $N$ 개의 표본들과  $p$ 개의 유전자들로 구성되어 있는 유전자 발현자료를  $N \times p$  자료 행렬  $X_{N \times p}$ 로 표시하자. 그리고,  $N$ 명의 환자에 대한 암의 여부를 나타내는 반응 값  $Y$ 는  $N \times 1$  벡터로 표시한다.

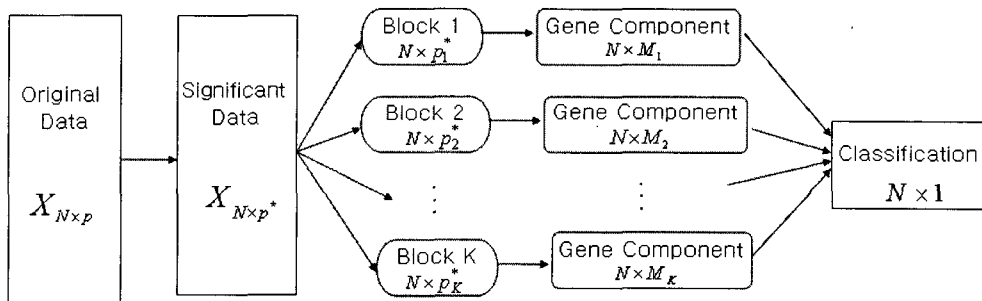
일반적으로 수 천, 수 만 개의 유전자가 모두 분류에 사용될 필요는 없다. 따라서, 차원축소를 하기 전에, 먼저 분류에 의미 있는 영향을 주는 유의한 유전자를 선택한다. 분석에서 사용되는 암 자료들은 이진 분류(0의 경우 정상 표본, 1의 경우 암 표본)에 해당되기 때문에 사용이 간단한 다음과 같은  $t$ -통계량을 사용하여 유전자를 선택한다.

$$t_i = \frac{m_{0i} - m_{1i}}{\sqrt{s_{0i}^2/n_0 + s_{1i}^2/n_1}}, \quad i = 1, 2, \dots, p,$$

여기서  $n_0(n_1)$ 는 각 분류 그룹 0(1)의 표본 수이고,  $m_{0i}(m_{1i})$ 와  $s_{0i}^2(s_{1i}^2)$ 는  $i$ 번째 유전자에 대한 각 분류 그룹 0(1)의 평균과 분산을 나타낸다. 각 유전자의  $t$ -통계치들에서, 분류에 영향을 주는 유전자로서 선택된  $p^*$ 개의 유전자는 양의  $t$ 값 중 상위  $p^*/2$ 개 및 음의  $t$ 값 중 하위  $p^*/2$ 개에 해당하는 유전자들이다. 이제 분류에 사용할  $p^*$ 개의 유전자로 구성된 자료를  $N \times p^*$ 자료 행렬인  $X_{N \times p^*}$ 로 표시한다.

다음으로, 분류에 유의한 유전자 자료  $X_{N \times p^*}$ 에 대하여 다양한 군집화 방법이 적용되어 유사한 유전자 기능을 갖는 유전자 군집(블록)으로 형성될 수 있다.

일단 블록화가 이루어진 다음, 각 블록마다 PCA와 PLS 방법에 의해 각각 차원축소를 수행한다. 본 연구에서 사용된 다중 블록 차원축소는 선택된  $p^*$ 개의 유전자를  $K$ 개의 블록으로 블록화 한 다음 각 블록  $i$ 에서 PCA나 PLS 방법을 이용하여  $M_i$ 개의 차원으로 다중 블록 차원축소를 수행하는 방법으로서 <그림 1>에 설명되어 있다. <그림 1>에서  $p^* = p_1^* + p_2^* + \dots + p_K^*$ ,  $p_i^*$ 는 블록  $i$ 에 포함된 유전자의 수,  $K$ 는 블록의 수, 그리고  $M_i$ 는 블록  $i$ 에서 축소된 차원의 수이다.



<그림 1> 다중 블록 차원축소를 이용한 분류

차원축소 방법 중, PCA는 분산이 최대가 되도록 유전자자료들을 선형 변환하는 방법으로서, 이때 유전자들의 선형결합을 유전자성분(gene component)이라고 하자. PCA는 유전자 성분들을 찾을 때 반응 값  $Y$ 를 고려하지 않는 반면, PLS는 반응 값  $Y$

와 관련성이 높은 유전자 성분을 찾는다. 마이크로어레이 유전자자료에 대한 PCA와 PLS에 의한 전체 차원 축소법의 적용은 Nguyen 등 (2002)에 다양한 암자료에 대한 수치분석에 의해 주어진다. MB 차원축소의 방법인 MBPCA와 MBPLS는 각각의 블록 내에서 PCA 방법과 PLS 방법을 적용하는 국소 차원 축소(local dimension reduction)법이다.

다중 블록 차원축소에 의해서 최종적으로 차원축소된  $\sum_{i=1}^K M_i$  개의 유전자 성분을 새로운 변수(차원)로 하여 LDA와 QDA에 의한 분류 분석을 수행한다. LDA와 QDA에 의한 분류방법은 Johnson과 Wichern (1998)을 참고하기 바란다.

### 3. 암 유전자 자료분석

본 연구에서는 전체 차원축소 방법인 PCA와 PLS 그리고, 다중 블록 차원축소 방법인 MBPCA와 MBPLS의 비교를 위하여 전립선암과 유방암 자료를 사용하여 분석하였다. 모든 유전자 발현 값들의 평균이 0, 표준편차가 1이 되도록 표준화시키는 전처리 과정을 거친 자료들로부터  $p^* = 100$ 개의 유전자를 분류에 유의한 유전자로 선택하여 사용하였고, 블록화를 위하여 계층적군집방법(hierarchical clustering method)으로서 Ward법, 유사도(similarity)로서 피어슨의 상관계수를 사용하였다. 본 연구에서 선택된 유전자의 수를 고려하여 블록의 수는  $K = 2, 3$ 으로 변화시켜 사용하였다. 또한 유전자 성분의 수를 편의상  $\sum_{i=1}^K M_i = M$ 으로 놓고  $M = 2, 3, 4, 5, 6$ 으로 변화시켜 비교하였다.

전립선암 자료(Singh 등, 2002)는 12,600개의 유전자와 136개의 표본으로 구성되었다. 이 중 102개의 학습 자료는 52개는 암 조직이고, 50개는 정상 조직에 해당된다. 그리고, 34개의 시험 자료는 25개는 암 조직이고, 9개는 정상 조직으로 구성되어 있다. 유방암 자료(van 't Veer 등, 2002)는 24,188개의 유전자와 97명의 표본으로 구성되어 있다. 이 중 78명의 학습 자료는 5년 사이에 전이(metastasis)를 발견한 유방암 환자(5년 전이)는 34명, 5년 사이에 무병(disease-free)으로 남아있는 유방암 환자(5년 무병)는 44명이고, 19명의 시험 자료는 5년 전이 12명, 5년 무병은 7명으로 구성되어 있다. 분석에 사용된 두 가지 암 자료는 다음의 웹사이트에서 얻을 수 있다.

(<http://www.esat.kuleuven.ac.be/~npochet/Bioinformatics/>).

<표 1>과 <표 2>는 전체 차원축소와 다중 블록 차원축소에 의한 전립선암 및 유방암 시험자료에 대한 오분류율을 각각 나타낸다. 분류관점에서는 최소 오분류율을 주는 차원축소법이 최선이기 때문에 최소 오분류율을 기준으로 두 방법을 비교하려 한다. 각 표에서 그림자 표시된 오분류율은 전체 차원축소법 대 다중 블록 차원축소 방법을 비교하여 유전자 성분 수별로 보다 더 작은 오분류율을 나타낸 경우에 해당한다. <표 1>의 전립선암자료의 경우에 최소 오분류율을 보여주는 전체 차원축소법의 경우 PCA 성분 수 2를 사용하여 차원축소 후 QDA에 의한 판별을 한 경우로서 오분

류를 0.088을 보여준다. 반면에 다중 블록 차원축소법의 경우 2개의 블록으로 군집화하여 첫 번째 및 두 번째 군집에 대하여 각각 2개 및 1개의 PLS성분으로 차원축소 후 QDA방법을 사용한 경우가 오분류율은 0.029로서 전체 차원축소법보다 더 정확한 결과를 준다. 각 유전자 성분 수 기준으로 볼 때, 성분수 2의 경우에 전체 차원축소법에 의한 분류결과가 0.088의 오분류율을, 성분수 3,4,5,6의 경우 다중블록 차원축소법에 의한 분류결과가 각각 0.029, 0.059, 0.088, 0.059의 보다 작은 오분류율을 준다. <표 1>의 전립선암자료에서 다중 블록 차원축소법의 경우는 3개 블록 보다는 2개 블록으로의 차원축소법이 보다 작은 오분류율을 주는 반면, <표 2>의 유방암자료에서는 2개 블록 혹은 3개 블록의 많은 성분구성에서 최소 오분류율 0.053의 결과를 준다. 한편, 전체 차원축소법의 경우에 최소 오분류율은 0.211이다. <그림 2>와 <그림 3>은 각 자료의 전체자료 및 각 블록자료에서 유전자들 간의 상관계수들의 분포를 나타낸 상자그림이다. 두 자료 모두 군집화 전의 전체 자료에 비해 모두 블록자료내의 상관계수가 더 크음을 알 수 있고, 서로 유사한 유전자 기능을 갖는 유전자들의 블록으로 나누어 차원축소를 할 근거를 준다.

일반적으로 PCA 차원축소 후의 분류결과보다 PLS 차원축소 후의 분류결과가 더 좋다고 알려져 있다. 그것은 PLS 차원축소법이 유전자자료(예측변수)와 반응변수간의 공분산을 요약하는 성분을 새로운 유전자 성분(예측변수)으로 만들기 때문이다. 그러나 PCA가 PLS에 경쟁이 될 수 있을 때는 이미 반응 값을 잘 예측할 수 있는 유전자들을 선별하여 PCA를 적용한 경우이다(Nguyen 등 (2002)). 본 연구에서는 이미 수만 개의 유전자 중 정상 및 암세포를 잘 구별해주는 100개의 유전자(전립선암 자료의 경우 원 유전자 수의 0.8%, 유방암자료의 경우 0.4%에 해당)를 선별하여 PCA 및 PLS를 적용하였다.

#### 4. 맺음말

본 연구에서는 화학실험의 chemometrics 기술분석법 등에서 주로 사용되는 차원축소 방법인 MBPCA 와 MBPLS를 두 가지의 암 자료에 적용하여 차원축소 후 LDA와 QDA에 의한 분류분석을 수행하여 보았다. 그리고, 전체차원축소 방법인 PCA와 PLS를 적용하여 분류분석한 경우와 비교한 결과 MBPCA, MBPLS에 의한 차원축소법에 의한 분류결과에 최소오분류율이 속해 있음을 알 수 있었다. 본 논문연구를 위하여 많은 마이크로어레이자료에 대한 자료분석을 하였는데, 최소 오분류율 기준에서 전체블록차원축소법이 다중 블록 차원축소법보다 더 좋은 자료들도 있었고 본 논문의 두 가지 자료처럼 다중 블록 차원축소법이 전체 블록 차원축소법보다 더 좋은 자료들도 있었다. 본 논문에서 의도하는 바는 다중 블록 차원축소법이 더 좋은 경우도 있다는 것을 보이는 것이다. 본 논문에서 제안된 분류의 결과는 대상이 되는 유전자 발현자료, 전처리방법, 분류에 유의한 유전자의 선택방법, 군집분석방법, 사용된 유사도, 판별방법, 블록의 수, 축약된 유전자 성분의 수 등의 선택에 따라 달라질 수 있으며, 향후 연구에서는 여러 방법이나 축도를 달리 했을 때 분류결과를 비교분석하고 특히 수많은 마이크로어레이 유전자 발현자료들의 분류분석을 통하여 다중 블록차원축소법이

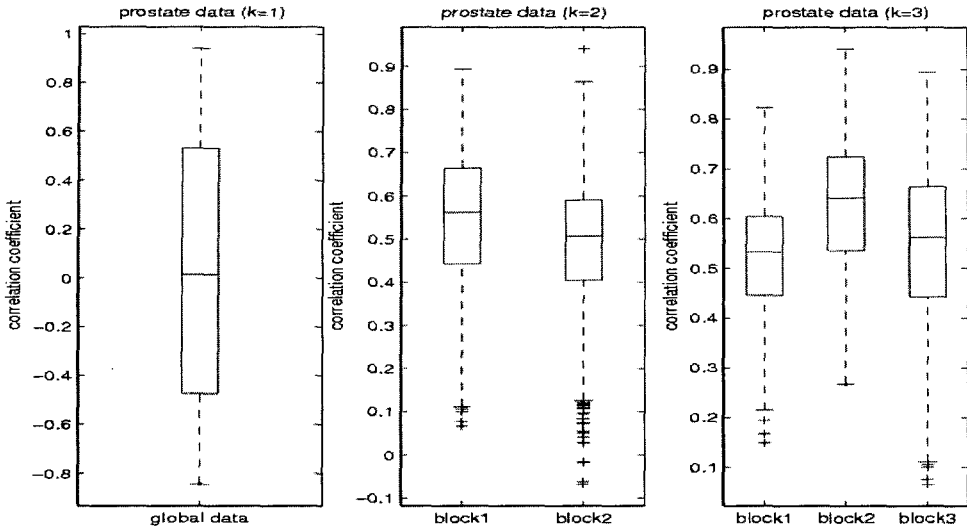
전체차원축소법보다 더 정확한 분류결과를 주는 유전자 발현자료들의 성질을 규명해야 할 것이다..

<표 1> 전립선암 자료: 136개 표본 중 34개 시험자료의 오분류률

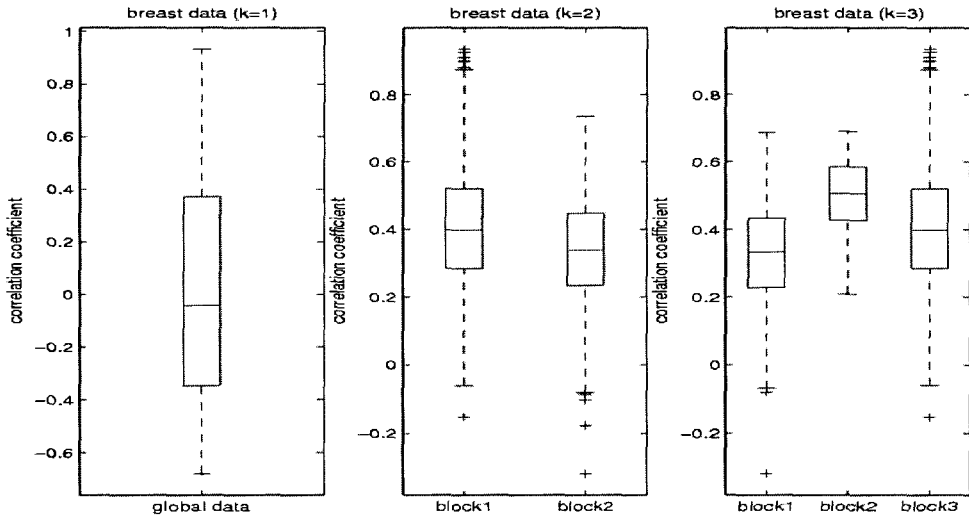
성 분 수	전체 차원축소법				블 록 수 ( <i>K</i> )	성 분 수 ( <i>M</i> )	<i>M<sub>i</sub></i> 구 성	다중블럭 차원축소법				
	시험자료							시험자료				
	LDA		QDA					LDA		QDA		
	PCA	PLS	PCA	PLS				PCA	PLS	PCA	PLS	
2	0.147	0.235	<b>0.088</b>	0.235	2	2	(1, 1)	0.147	0.147	0.147	0.147	
3	0.147	0.147	0.147	0.147		3	(1, 2)	0.147	0.235	0.147	0.265	
						(2, 1)	0.176	0.118	0.118	<b>0.029</b>		
4	0.147	0.118	0.176	0.176		4	(1, 3)	0.147	0.176	0.147	0.206	
						(2, 2)	0.147	0.235	0.147	0.235		
						(3, 1)	0.206	0.147	0.176	<b>0.059</b>		
5	0.147	0.118	0.176	0.147		5	(1, 4)	<b>0.088</b>	0.206	0.176	0.206	
						(2, 3)	0.147	0.176	0.147	0.235		
						(3, 2)	0.147	0.176	0.176	0.206		
						(4, 1)	0.176	0.206	0.176	0.235		
6	0.176	0.178	0.147	0.176		6	(1, 5)	0.118	0.235	0.206	0.235	
						(2, 4)	0.147	0.147	0.206	0.206		
						(3, 3)	0.176	0.147	0.147	0.147		
						(4, 2)	0.147	0.206	0.147	0.265		
						(5, 1)	0.088	0.147	<b>0.059</b>	0.235		
						3	3	(1, 1, 1)	0.147	0.147	0.176	0.176
							4	(1, 1, 2)	0.147	0.088	0.147	0.088
								(1, 2, 1)	0.147	0.176	0.147	0.206
					(2, 1, 1)		0.147	0.265	0.147	0.265		
					5		(1, 1, 3)	0.147	0.147	0.176	<b>0.088</b>	
							(1, 2, 2)	0.206	0.147	0.176	0.118	
							(1, 3, 1)	0.147	0.176	0.118	0.176	
							(2, 1, 2)	0.147	0.265	0.176	0.147	
							(2, 2, 1)	0.147	0.265	0.147	0.265	
					(3, 1, 1)		0.147	0.147	0.206	0.235		
					6		(1, 1, 4)	0.176	0.206	0.147	0.294	
							(1, 2, 3)	0.176	0.147	0.176	0.118	
							(1, 3, 2)	0.176	0.088	0.147	0.118	
							(1, 4, 1)	0.118	0.176	0.118	0.235	
							(2, 1, 3)	0.147	0.147	0.176	0.118	
							(2, 3, 1)	0.147	0.235	0.147	0.265	
							(3, 1, 2)	0.176	0.206	0.176	0.147	
					(3, 2, 1)		0.147	0.176	0.206	0.265		
(4, 1, 1)	0.147	0.147	0.206	0.206								

<표 2> 유방암 자료: 97개 표본 중 19개 시험자료의 오분류율

전체 차원축소법					다중블럭 차원축소법						
성분수	시험자료				블록 수 (K)	성분 수 (M)	$M_i$ 의 구성	시험자료			
	LDA		QDA					LDA		QDA	
	PCA	PLS	PCA	PLS				PCA	PLS	PCA	PLS
2	0.211	0.263	0.263	0.263	2	2	(1, 1)	0.211	0.263	0.211	0.211
3	0.211	0.263	0.263	0.316		3	(1, 2)	0.105	0.211	0.158	0.105
							(2, 1)	0.158	0.263	0.105	0.316
4	0.263	0.368	0.211	0.368		4	(1, 3)	0.053	0.263	0.158	0.263
							(2, 2)	0.105	0.211	0.105	0.263
							(3, 1)	0.158	0.211	0.105	0.158
5	0.316	0.316	0.211	0.316		5	(1, 4)	0.053	0.211	0.211	0.211
							(2, 3)	0.053	0.263	0.053	0.263
							(3, 2)	0.158	0.211	0.105	0.158
							(4, 1)	0.158	0.211	0.105	0.158
6	0.316	0.263	0.211	0.421		6	(1, 5)	0.053	0.263	0.158	0.211
							(2, 4)	0.053	0.211	0.105	0.211
							(3, 3)	0.053	0.211	0.053	0.158
							(4, 2)	0.158	0.211	0.105	0.158
							(5, 1)	0.158	0.211	0.158	0.211
					3	3	(1, 1, 1)	0.053	0.211	0.053	0.105
							(1, 1, 2)	0.053	0.211	0.053	0.158
						4	(1, 2, 1)	0.053	0.211	0.053	0.053
							(2, 1, 1)	0.053	0.316	0.105	0.158
						5	(1, 1, 3)	0.105	0.211	0.105	0.211
							(1, 2, 2)	0.053	0.263	0.053	0.158
							(1, 3, 1)	0.158	0.211	0.158	0.263
							(2, 1, 2)	0.053	0.316	0.053	0.211
							(2, 2, 1)	0.053	0.263	0.105	0.263
						6	(3, 1, 1)	0.053	0.158	0.105	0.158
							(1, 1, 4)	0.105	0.211	0.105	0.211
							(1, 2, 3)	0.105	0.211	0.105	0.211
							(1, 3, 2)	0.105	0.263	0.105	0.105
							(1, 4, 1)	0.211	0.158	0.211	0.263
							(2, 1, 3)	0.105	0.211	0.105	0.211
							(2, 3, 1)	0.211	0.263	0.105	0.211
							(3, 1, 2)	0.053	0.211	0.053	0.211
							(3, 2, 1)	0.053	0.211	0.105	0.158
(4, 1, 1)	0.053	0.211	0.158	0.211							



<그림 2> 전립선암자료의 전체 및 각 블록별 상관계수의 분포



<그림 3> 유방암자료의 전체 및 각 블록별 상관계수의 분포

참고문헌

[1] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Mack, S.Y.D. and Levine, J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences, USA*. Vol. 96. 6745-6750.



- [2] Antoniadis, A., Lamber-Lacroix, S. and Leblance, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*. Vol. 19. 563-570.
- [3] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z. (2000). Tissue Classification with Gene Expression Profiles. *Journal of Computational Biology*. Vol. 7. 559-584.
- [4] Dudoit, S., Fridlyand, J. and Speed, T.P. (2000). Comparison of discrimination methods for the classification of tumors using gene expression data. *Technical Report 576*. Department of Statistics, University of California, Berkeley.
- [5] Fraley, C. and Raftery, A.E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*. Vol. 94. 611-631.
- [6] Furey, T.S., Cristianini, N., Duffy, N., Shummer, B.M. and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*. Vol. 16. 906-914.
- [7] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. and Lander, E. (1999). Molecular Classification of Cancer: Class discovery and class prediction by gene expression monitoring, *Science*. Vol. 286. 531-537.
- [8] Johnson, R.A. and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis(4th ed.)*. Prentice Hall.
- [9] Nguyen, D.V. and Rocke, D.M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*. Vol. 18. 39-50.
- [10] Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R. and Sellers, W.R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*. Vol. 1. 203-209.
- [11] van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. Vol. 405. 530-536.
- [12] Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T. and Vapnik, V. (2001). Feature selection for SVMs. *Advances in Neural Information*

*Processing Systems.* Vol. 13. 668-674.

- [13] Wold, S., Geladt, P., Esbensen, K. and Ohman, J. (1987a). Multiway Principal Components and PLS Analysis. *Journal of Chemometrics*. Vol. 1. 41-56.
- [14] Wold, S., Hellberg, S., Lundstedt, T., Sjostrom, M. and Wold, H. (1987b). PLS modeling with latent variable in two or more dimensions. *Proc. Symp. on PLS Model Building: Theory and Application*. Frankfurt am Main. also Tech. Rep., Department of Organic Chemistry, Umea University.

[Received June 2006, Accepted September 2006]