요인 분석을 이용한 유의한 유전자 추출

Finding significant genes using factor analysis

Jeong-Wha Lee, Hyeseon Lee, Hae-Sang Park, Chi-Hyuck Jun
{bls83;hyelee;shoo359;chjun}@postech.ac.kr
San 31, Hyoja-dong, Namgu, Pohang, Gyungbuk, Korea

Abstract

Clustering for gene expression data without filtering out noise genes may be distorted or derived inappropriate inference. Identifying significant genes and deleting noise before major analysis is necessary for meaningful discovery from genes expression pattern. We proposed a new method of finding significant genes using factor analysis which is done on transposed data matrix. We construct significance score that is sum of factor loadings for declared significant number of factor, and set threshold through replication. Our proposed method works well for simulated time-course data for finding significant genes even though variance level gets larger.

1.Introduction

유전자 데이터분석에서 실험디자인조합이 적은 경우는 유의성 검정 자체가 의미있는 추론을 제공할 수 있다. 그러나 실험과정상 노이즈로 보여지는 발현수준에 의해 유의성 검정이 왜곡될 수도 있고, 군집분석도 정확히 이루어질 수 없는 경우는 일반적으로 유전자 데이터 분석이 어려운 이유이기도 하다. 수만 개의 유전자에 대해 실험한 경우 혹은 시간대별로 여러 가지를 실험한 경우는 노이즈를 제거하지 않고 군집을 나누는 것은 swamping 혹은 masking효과에 의해 군집자체의 신뢰성이 떨어질 수 있다.

1차적으로 적절한 방법으로 노이즈를 제거하는 것, 즉 어느 수준 이상의 유의성을 갖는 유전자들을 추출하는 것은 2차적 분석에 있어서 핵심적인 기술이다.

요인분석은 몇 가지 핵심적인 요인을 찾아내어 설명해가거나 어떤 변수들이 주요요인에 높은 상관관계를 갖는지 분석하기 위해 사용되는 통계기법이다. 심리학이나 사회학 등 인지행동 분석에 많이 사용되는 기법이다. 일반적인 요인분석은 (n×p) 데이터행렬의 공분산 혹은 상관행렬에 대해 이루어진다. 본 연구에서는 유전자들 중에서 유의한 변화를 보이는 유전자들 집단을 추출하기 위한 것이므로 (p×n) 행렬의 공분산행렬에 대해 요인분석을 적용하고[1], 유의성에 대한 기각치를 결정하기 위해 유의점수(Significance Score)를 정의하고, 그에 대한 임계치를 정한다. 시간별 혹은 처치별로 유의하게 변화하는 유전자들을 추출하여 노이즈로 판단되는 유전자를 제거함으로써 보다 정확한 2차적 분석, 군집분석 등을 가능하게 하는 것이 본 연구의 목적이다.

2장에서는 요인분석을 이용한 노이즈성 유전자를 제거하는 방법을 제안하고, 3장에서는 시뮬레이션 데이터를 이용하여 2장에서 제안한 방법의 유용성을 평가한다.

2.Proposed method

2.1 요인분석

n개의 객체가 p개의 변수를 가지고 있다고 하자. y_{ij}를 i번째 객체의 j번째 변수라고 하자$(i=1,...,n,\ j=1,...,p)$. Y는 $n\times p$ 행렬로써 전체의 데이터를 각 변수들의 평균을 0으로 변환하여 변수들 간의 서로 다른 평균에 따른 영향을 줄인 행렬이다. Y의 전치행렬을 X라 하면, $X=(x_1,x_2,...,x_n)$로 나타낼 수 있다. 이때$x_i=(x_{i1},x_{i2},...,x_{pi})^T$이다.

요인분석에 사용되는 요인의 수는 보통 고유값이 1보다 큰 개수를 사용하거나[2], random matrix theory에 의해서 Y의 데이터를 각 행별로 무작위로 섞은 후의 행렬의 최대 고유치보다 큰 Y의 고유치의 개수로 정한다[3]. 여기서는 후자를 이용한다. 이때 사용되는 요인의 수를 m이라 하면 X는 (1)과 같은 식으로 나타낼 수 있다.

$$X^T=\Lambda F+E \qquad (1)$$

이때 $F=(f_1,f_2,...,f_m)^T$로 $m\times p$행렬로 요인을 나타내며, $E=(\epsilon_1,\epsilon_2,...,\epsilon_n)^T$로 오차를 나낸다. Λ는 $m\times n$행렬로 요인적재량이다.

$$\Lambda=\begin{bmatrix}\lambda_{11}\ \lambda_{12}\ ...\ \lambda_{1m}\\ \lambda_{21}\ \lambda_{22}\ ...\ \lambda_{2m}\\ \vdots\ \ \vdots\ \ \ \ \vdots\\ \lambda_{n1}\ \lambda_{n2}\ ...\ \lambda_{nm}\end{bmatrix}$$

요인적재량은 관측유전자와 요인간의 상관관계를 의미한다. 따라서 각 x_i에 대한 요인 적재량 $\lambda_i=[\lambda_{i1},\lambda_{i2},...,\lambda_{im}]$은 x_i를 중요 요인으로 설명할 수 있는 값이다.

2.2 유의한 유전자 판별

요인분석 결과인 Λ를 이용하여 노이즈 데이터를 판별하고자한다. 여기서는 significance score(SC)를 정의하고 사용하고자 한다. SC의 정의는 식(2)와 같다.

$$SC_i=\sum_{j=1}^{m}\lambda_{ij}^2,\ i=1,...m \qquad (2)$$

λ_{ij}의 절대값이 클수록 중요 요인으로 설명력이 높다고 이야기 할 수 있으므로 SI의 값이 클수록 그 객체는 패턴을 지닌다고 이야기 할 수 있다. 그렇다면 유의한 유전자를 판별하는 기준을 어떻게 정할 것인가가 중요하다. 이를 위해서는 다음과 같은 방법을 사용한다.

i번째 객체의 최대값과 최소값을 인자로 가지는 균등분포에 따라 새로운 i번째 객체를 생성한다. 이렇게 새롭게 만들어진 데이터를 이용하여 새롭게 요인분석을 한다. 이때 만들어진 새로운 요인적재량을 Λ^k라 하자. 이때의 SC를 $SC_i^{(k)}$라 하면, 새롭게 만들어진 K개의 데이터 집합들의 SC 평균을 식(3)으로 표현할 수 있다.

$$\nu_i=\frac{1}{K}\sum_{k=1}^{K}SC_i^{(k)} \qquad (3)$$

ν는 $n\times 1$행렬로써 각 ν_i는 i번째 객체들의 유의성을 판별하는 기준값이 된다. 즉 $SC_i>\nu_i$이라면 i번째 객체는 유의한 객체로 판명을 하며 $SC_i\le\nu_i$이라면 i번째 객체는 노이즈로 판명을 한다.

3. Numerical experiments

3.1 Data sets

시뮬레이션을 위해서 식(4)를 이용하였다[4]. 노이즈가 아닌 데이터는 5개의 군집을 이루고 있다. 즉 식(4)에 사용된 k는 5이다. 변수의 개수는 10개이다.

$$D(i,j)=\delta_j+\lambda_j(\alpha_i+\beta_j\phi(i,j)) \qquad (4)$$
$$\text{for } j=1,...,n\ \ j=1,...,10$$

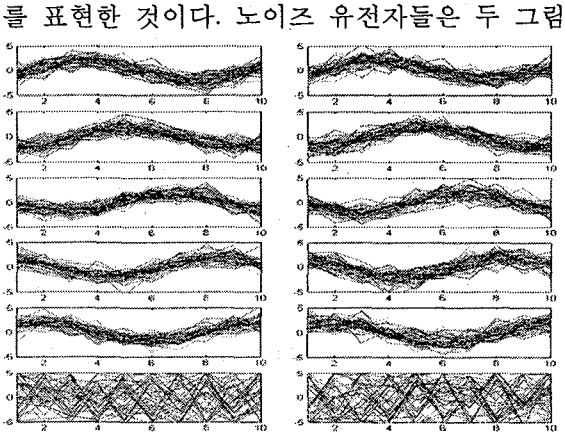
이때 $\phi(i,j)=\sin(\frac{2\pi\times j}{8}-\frac{2\pi\times k}{5})$이다. 각 변수들에 대한 분포는 식(5)와 같다.

$$\begin{aligned}&\delta_j\sim N(0,\sigma^2) \qquad (5)\\ &\lambda_j\sim N(1,0.1^2)\\ &\alpha_i\sim N(0,1^2)\\ &\beta_i\sim N(2,0.3^2)\end{aligned}$$

노이즈성 유전자는 유의한 유전자의 최대값과 최소값을 인자로 가지는 균등분포를 따른다. 이를 식(6)에 나타내었다.

$$n_{kl}=Unif(\min(D(i,j),\max(D(i,j)) \qquad (6)$$

시뮬레이션 데이터는 각 군집이 40개의 유전자를 가지며 총 200개의 유의한 유전자를 가진다. 이때 σ를 0.1, 0.2, 0.3, 0.4 그리고 0.5로 변화하면서 유의한 유전자의 분산이 커질 때 제안한 방법이 어떠한 경향을 보이는지 알아보고자 한다. 또한 노이즈 유전자의 수도 40, 80 그리고 120개로 증가하면 결과에 어떤 영향을 주는지도 함께 알아보고자 한다. 시뮬레이션에 사용되는 데이터 집합은 총 15개이다. significant [그림1]은 시뮬레이션 데이터를 표현한 것이다. 노이즈 유전자들은 두 그림



[그림1]노이즈성 유전자 40, σ=0.1, 0.5인 경우

모두 40개이다. [그림1]의 오른쪽은 σ가 0.1이며 왼쪽은 σ가 0.5이다.

3.2 False Discovery Rate(FDR)

False Discovery Rate(FDR)이란 [표 1]에 나타난 값 중 $E[V/R]$을 나타내는 값으로써 유의하게 판별된 값 중 실제로는 유의하지 않은 데이터의 비율로써 절대적인 오류의 수를 나타내는 FWER (family- wise error rate)와 달리 오류의 비율을 비교할 때 많이 사용된다. [표1]에서 m은 전체 데이터 개수를 의미하며, m_0는 실제로 유의한 데이터의 수를 의미한다. T는 true positive, U는 false negative, S는 false positive 그리고 V는 true negative를 의미한다. FDR은 1종 오류의 비율을 조절하고자 할 때 많이 사용된다[5].

[표1]유의성여부 분류표

		판정		
		유의함	유의하지 않음	합
실제	유의함	T	U	m_0
	유의하지 않음	S	V	$m-m_0$
합		R	$m-R$	m

3.3 Result

제안한 방법의 평가 비교를 위해서 Storey et al.(2005)의 유의성검정 방법을 이용하여 동일한 개수를 노이즈로 설정하여 FDR을 계산하였다[6]. 분석소프트웨어는 EDGE(Extraction and analysis of Differential Gene Expression)을 사용하였고[7], 시뮬레이션 데이터가 time-course 데이터의 형태이므로 EDGE에서 natural cubic spline을 이용한 time-course data 유의성 검정값을 계산하였다.

각 데이터에 대한 FDR의 결과는 [표 2]와 같다. [표3]은 True Positive의 개수를 보여준다. 표를 살펴보면 제안한 방법의 경우 대부분의 유의한 실제 데이터를 대부분 찾아낸 것을 볼 수 있다.

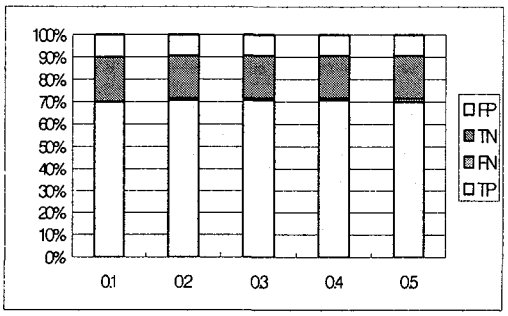
[그림2]은 잡음의 개수가 80개인 경우 제안한 방법의 결과를 나타낸 것이다. [그림3]은 EDGE의 같은 데이터에 대한 결과이다. FP는 False Positive, TN은 True Negative, FN은 False Negative 그리고 TP는 True Positive를 나타낸다. 제안한 방법이 TP와 TN의 비율이 모두 높은 것을 볼 수 있다.

[표2]제안한 방법과 Storey et al(2005)의 FDR 비교

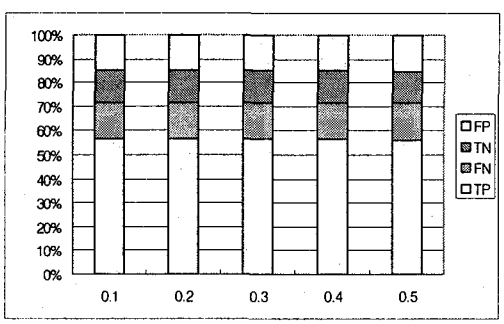
노이즈 수	σ	Proposed	Storey et al.
40	0.1	0.0619	0.1213
	0.2	0.0609	0.1228
	0.3	0.0598	0.1212
	0.4	0.0616	0.1243
	0.5	0.0621	0.1278
80	0.1	0.1190	0.2078
	0.2	0.1118	0.2069
	0.3	0.1156	0.2070
	0.4	0.1150	0.2079
	0.5	0.1172	0.2146
120	0.1	0.1677	0.2711
	0.2	0.1691	0.2772
	0.3	0.1693	0.2796
	0.4	0.1753	0.2814
	0.5	0.1728	0.2867

[표3]제안한 방법과 Storey et al.(2005)의 True Positive 비교

노이즈 수	σ	Proposed	Storey et al.
40	0.1	198.7	175.7
	0.2	198.5	175.4
	0.3	197.8	175.8
	0.4	197.9	175.1
	0.5	197.0	174.4
80	0.1	196.4	158.5
	0.2	198.5	158.6
	0.3	198.3	158.6
	0.4	197.8	158.4
	0.5	197.3	157.1
120	0.1	198.7	145.8
	0.2	198.7	144.5
	0.3	198.3	144.1
	0.4	198.0	143.7
	0.5	197.2	142.7



[그림2]노이즈성 유전자가 80개일대 σ의 변화에 따른 제안된 방법의 FN, TN, FP, TP의 변화



[그림3]노이즈성 유전자가 80개일대 σ의 변화에 따른 Storey et al.(2005)의 FN, TN, FP, TP의 변화

4.Conclusion

3장의 결과를 살펴보면 제안한 방법은 Storey et al.(2005)에 비해 정확하게 노이즈를 찾는다는 결과를 보였다.

데이터의 크기가 커지고, 노이즈 데이터가 아니더라도 분산이 커지는 경우 노이즈 데이터에 의해 영향을 많이 받을수 있다. 따라서 이러한 데이터일수록 유의한 유전자들을 미리 분리하는 것이 필요하다. 하지만 제안한 방법에 사용된 요인분석의 경우 유전자의 수가 많아지면 계산해야 할 분량이 많아지는 단점이 있다.

제안한 방법에서 유의한 유전자를 판별하는데 사용한 SC 임계치 산정방식은 노이즈 유전자를 제거하는데 적절한 것으로 평가된다.

Reference

[1] Hair, J.F., Anderson, Jr.R.E., Tatham, R.L., Black, W.C., Multivariate Data Analysis, Prentice Hall, 1995.

[2] Sharma, S., Applied Multivariate Techniques, John Wiley & Sons, Inc., 1996

[3] Brody, T. A. et al., Random-matrix physics: spectrum and strength fluctuations, Rev. Mod. Phys. vol 53, Issue 3, pp. 385–479, 1981

[4] Yeung, K. Y., Ruzzo, W. L., Principal component analysis for clustering gene expretion data, Bioinformatics, Vol. 17, No. 9, pp. 763-774, 2001

[5] Benjamini, Y, Hochberg,Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Statist. Soc. B vol. 57 Part 1, pp289-300, 1995

[6] Storey, John, Xiao, W., Leek, J. T., Tompkins, R.G., Davis, R.W., Significance analysis of time course microarray experiments, PNAS, Vol.102, No.36, pp12837-12842

[7] Leek, J. T. et al. EDGE: Extraction and analysis of Differential Gene Expression, Bioinformatics, Vol. 22, No. 4, pp. 507-508, 2006.