The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.3.389
26(3), 389–399

Screening and Clustering for Time-course Yeast Microarray Gene Expression Data using Gaussian Process Regression

Jaehee Kim[a,1] · Taehoun Kim[b]

[a]Department of Statistics, Duksung Women's University
[b]Department of PrePharmMed, Duksung Women's University

(Received December 28, 2012; Revised April 16, 2013; Accepted April 26, 2013)

Abstract

This article introduces Gaussian process regression and shows its application with time-course microarray gene expression data. Gene screening for yeast cell cycle microarray expression data is accomplished with a ratio of log marginal likelihood that uses Gaussian process regression with a squared exponential covariance kernel function. Gaussian process regression fitting with each gene is done and shown with the nine top ranking genes. With the screened data the Gaussian model-based clustering is done and its silhouette values are calculated for cluster validity.

Keywords: Differentially expressed, Gaussian process regression, log marginal likelihood, squared exponential covariance function, time-course microarray expression data, yeast.

1. 서론

시간에 따른 유전자 발현 데이터에 대한 초기 연구를 살펴보면 각 시점에 대해 독립적으로 보고 분석한 경우가 대부분이다. 그러나 이렇게 시간에 관련성이 있는 실험 조건에서 얻은 데이터에 대해 정적(static)으로 분석하는 것은 내재하는 역동적인 시스템에 대한 추론을 제한하게 된다. 유전자조절과정(gene regulation process)을 보여주는 실험은 시계열 실험이 많으며 시간점에 대한 비규칙적인 표본(irregularly sampled)일 경우, 시점에 대해 반복 수가 다른 경우 등 여러 가지 상황이 발생한다. 또한 실험 상황은 완벽한 조절하에 있기 어려우므로 생물학적 상황에서 발생하는 데이터는 잡음(noise)과 섞이게 되어 측정된다. 시계열 유전자 발현 데이터 분석은 전사요인(transcription factor)에 대한 식별, 유전자조절네트웍(gene regulatory network) 복원 등 시스템 생물학(system biology)의 문제 해결에 중요한 역할을 한다.

유전자 데이터의 경우 그 수가 매우 크기 때문에 분석시 시간과 비용 등의 어려움이 많고 분석 후에 결과가 좋지 않은 경우도 발생하게 된다. 이러한 문제를 해결하기 위해 변수변환을 이용하여 데이터의 차

This research was supported by Basic Science Research Program through the National Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF 2012-0001879).
[1]Corresponding author: Professor, Department of Statistics, Duksung Women's University, 419 Samyang-ro 144 Gil 33, Dobong-Gu, Seoul 132-714, Korea. E-mail: jaehee@duksung.ac.kr

원수를 줄이거나 선별(screening)과정을 통해 유전자를 선별하여 선별된 유전자만을 가지고 분석을 하는 방법이 있다. 이와 같은 유전자 선별방법에 대한 연구로 Hero 등 (2004)은 다양한 기준을 가진 DNA 마이크로어레이 실험을 통해 얻은 유전자들 중 다르게 발현되는 유전자를 확인하는 통계적 방법에 대해 연구하였다. Eckel 등 (2004)은 마이크로어레이 유전자 발현 데이터에 대해 데이터의 차원수를 줄이는 효율적인 방법에 대해 연구하였다. Ma (2006)는 8개의 다양한 통계량에 기초한 관리된 유전자 선별법에 대해 일치성에 대해 조사하였다. 최근 Kalaitzis와 Lawrence (2011)는 시계열 유전자 발현 데이터에 대해 가우스 과정회귀(Gaussian process regression; GPR) 기법과 로그 주변우도함수(log marginal likelihood; LML)를 이용하여 유전자 순위를 주는 예를 보여주었다.

Serban과 Wasserman (2005)은 비모수적 추정과 변수 변환을 통한 군집화에 대한 방법을 연구하였는데 유전자 선별과정대신 nondifferential cluster 를 포함한 군집분석을 제안하였다. Serban과 Wasserman (2005)은 비모수적 추정과 변수 변환을 통한 군집화에 대한 방법을 연구하였는데 유전자 선별과정대신 nondifferential cluster 를 포함한 군집분석을 제안하였다. Törönen 등 (1999)은 자기 조직화 지도(SOM)방법을 이용하여 군집분석을 하였다. Zhang 등 (2003)은 이산 푸리에 변환(discrete Fourier transform)을 이용한 군집분석 결과를 보여주었다. Kim과 Kim (2008)은 미분 푸리에 계수를 사용한 군집분석을 제안하였다. 유전자를 선별한 후 데이터의 형태와 특성에 따라 이러한 군집분석 방법들을 적용할 수 있다.

본 연구에서는 가우스 과정회귀 기법을 소개하고 Kalaitzis와 Lawrence (2011) 기법을 활용하여 비활동성(inactive, quiet) 유전자 선별한 후 차별화된(differentially expressed) 유전자들을 정리하여 그 다음 단계의 분석에서의 효율성을 높이는데 기여하고자 한다. 또한 실제 데이터 분석으로는 효모데이터에 대한 적용예를 보여준다. 본 논문의 구성은 다음과 같다. 2장에서는 시계열 유전자 발현 데이터에 대한 GPR 모형과 모형 적합에 관해 설명하고 3장에서는 효모 유전자 발현 데이터에 대한 분석예를 보여주고 마지막으로 4장에서 간단한 결론을 내린다.

2. Time-course 유전자 발현 데이터에 대한 Gaussian Process Regression

가우스 과정(Gaussian process; GP)는 시계열 유전자 발현 데이터에 포함되어있는 실제 신호와 잡음을 양적으로 표현하는 데 적절한 방법으로 활용할 수 있으며 차별화된 유전자 프로화일(differential gene profile)에 대해 순위(rank)를 줄 수 있다. GP는 다변량정규분포를 특정 분포함수들 집합에 대한 정규분포의 개념으로 확장한 것이다. 여기서 특정 분포함수들 집합이란 공분산 함수(covariance function) 또는 커널함수(kernel function)에 의해 즉 데이터 간의 유사성 측도(similarity metric)에 의해 정의되는 집합이다.

GPR에서 표준적인 공분산 함수인 제곱 지수 공분산함수(squared exponential covariance function)를 갖는 GP를 고려하면 유전자 프로화일에 내재하는 실제 신호(true signal)가 스무드 함수(smooth function) 즉 무한 미분 가능 함수를 갖는다고 가정할 수 있다. 이러한 특성 때문에 모형에 대한 강한 가정 없이 내재하는 신호를 표현하는데 GP를 활용할 수 있으나 시간에 따른 집합체 데이터에 대해 가짜 패턴(spurious pattern; false positive)을 찾을 수도 있음을 주의해야한다.

2.1. Gaussian Process Regression 모형

시계열 유전자 발현데이터 $\{Y_{iu}\}$에 대해 다음의 모형을 고려하고자 한다. i번째 곡선에 있는 u번째 데이터를 Y_{iu}라고 하면

$$Y_{iu} = f_i(x_{iu}) + \epsilon_{iu}, \quad i = 1, 2, \ldots, n, \; u = 1, 2, \ldots, m \tag{2.1}$$

와 같은 모형을 갖는다. 여기서 x_{iu}는 i번째 곡선의 u번째 time-point이고 오차항 ϵ_{iu}는 평균 $E(\epsilon_{iu}) = 0$과 분산 $\mathrm{Var}(\epsilon_{iu}) = \sigma^2$을 갖는다. 유전자 발현 데이터인 Y_{iu}는 시간점 x_{iu}에서 i번째 데이터의 로그값이다. 곡선 f_i가 스무드 함수집합에 속한다고 가정한다.

시계열 유전자 발현 데이터에 대해 GP를 적용하는 아이디어는 주어진 유전자발현 관측값에 대한 궤적 추정(trajectory estimation)을 1 차원의 함수에 대한 보간(interpolation) 문제로 보는 것이다. 각 관측값은 백색 가우시안 잡음(white Gaussian-distributed noise)을 갖는다고 가정하게 되므로 예측에 대한 계산식을 쉽게 구할 수 있다.

입력변수(input variable) x_i와 출력변수(output variable) 또는 목표변수(target variable) y_i로 표현되는 유전자 발현데이터셋 $D = \{x_i, y_i\}$, $i = 1, \ldots, n$을 고려한다. 표기편의상 한 개 유전자 관련 데이터에 대한 선형회귀모형으로 정의된 GPR 모형에 대해 살펴보고자한다.

$$y = f(x) + \epsilon, \quad f(x) = \phi' \mathbf{w}, \tag{2.2}$$

예를 들어, $\phi = (1, x, x^2)$의 경우에는 2차 곡선으로의 함수이다. i번째 유전자의 m개 시간점에 따른 유전자발현 측정값 $\mathbf{Y} = \{y_u\}$, $u = 1, \ldots, m$은 백색잡음을 포함하며 시간 입력변수는 특성공간(feature space) $\mathbf{\Phi} = \phi' \mathbf{w}$에 매핑(mapping)된다. 여기서 백색잡음은 서로 독립이며 $\epsilon \sim N(0, \sigma_n^2)$를 따른다고 가정한다.

관측값에 대한 확률모형은 정규분포를 따르므로 $\mathbf{y} = \mathbf{y}_i = (y_{i1}, \ldots, y_{im})'$의 우도함수는

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}, \mathbf{w}) &= \prod_{u=1}^{m} p(y_u|x_i, \mathbf{w}) \\ &= \prod_{u=1}^{m} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(y_i - \phi_i' \mathbf{w})^2}{2\sigma_n^2}\right) \\ &= \frac{1}{(2\pi)^{\frac{m}{2}} |\mathbf{\Sigma}_n|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{\Phi}\mathbf{w})' \mathbf{\Sigma}_n^{-1} (\mathbf{y} - \mathbf{\Phi}\mathbf{w})\right) = N\left(\mathbf{\Phi}\mathbf{w}, \mathbf{\Sigma}_n\right), \end{aligned} \tag{2.3}$$

여기서 $\mathbf{\Sigma}_n = \sigma_n^2 \mathbf{I}$이다.

모수 $\mathbf{w}$에 정규 사전정보 $\mathbf{w} \sim N(\mathbf{0}, \mathbf{\Sigma}_\mathbf{w} = \sigma_w^2 \mathbf{I})$를 주어 베이지안 선형회귀(Bayesian linear regression)를 고려하며 우도함수 $\times$ 사전확률분포함수를 적분하여 모수에 대한 주변확률을 구하면 역시 정규분포를 얻게 된다. $\mathbf{y}$의 기대값과 공분산을 구해보면

$$\begin{aligned} E[\mathbf{y}] &= E[\mathbf{\Phi}\mathbf{w}] + E[\epsilon] = 0 \\ \mathrm{Cov}[\mathbf{y}] &= \sigma_w^2 \mathbf{\Phi}\mathbf{\Phi}' + \sigma_n^2 \mathbf{I} = \mathbf{K}_f + \sigma_n^2 \mathbf{I} = \mathbf{K}_\mathbf{y}. \end{aligned} \tag{2.4}$$

$\mathbf{y}$의 사후주변확률을 구해보면

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \int p(\mathbf{y}|\mathbf{x}, \mathbf{w}) p(\mathbf{w}) d\mathbf{w} \\ &= \frac{1}{(2\pi)^{\frac{m}{2}} |\mathbf{K}_y|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}' \mathbf{K}_\mathbf{y}^{-1} \mathbf{y}\right) = N(\mathbf{y}; \mathbf{0}, \mathbf{K}_\mathbf{y}). \end{aligned} \tag{2.5}$$

식 (2.5)를 이용하여 주변확률을 계산할 수 있으며 여러 모형에 대한 비교가 가능하여 모형에 대한 순위를 줄 수 있다.

2.2. Squared Exponential 공분산함수

GP는 특성공간상에서 확률과정(stochastic process)이며 정규확률분포의 일반화이다. 특정한 입력값 x에서의 함수값 $f(x)$를 가진 긴 벡터를 함수로 간주하고 가능한 함수들에 대한 함수를 가정한다. $p(y(x_1), \ldots, y(x_m))$ 즉 함수 $y(x)$의 확률분포가 정규확률분포를 따른다. 공분산행렬에서 잡음항은 $\mathbf{K_y}$가 완전순위(full rank)가 되도록 해주며 잠재함수(latent function) f에 대한 GP 사전정보를 다음과 같이 쓸 수 있다.

$$p(\mathbf{f}|\mathbf{x}) = \frac{1}{(2\pi)^{\frac{m}{2}}|\mathbf{K}_f|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{f}-\mathbf{m})'\mathbf{K}_\mathbf{f}^{-1}(\mathbf{f}-\mathbf{m})\right). \tag{2.6}$$

또는 $\mathbf{f}|\mathbf{x} \sim \mathrm{GP}(\mathbf{f}; m(\mathbf{x}), \mathbf{K}_f(x_u, x_v))$로 표기하며 평균함수(mean function) $m(\mathbf{x})$와 공분산함수 $\mathbf{K}_f(x_u, x_v)$을 가진 GP를 의미한다. GP에서 공분산함수는 중요한 요소가 되는데 이는 연구하고자 하는 함수에 대한 가정에 대한 정보를 담기 때문이다. 즉 가까운 x에 대해 비슷한 반응값 y를 갖게 되는 기본적인 가정을 포함하기 때문이다. 이 논문에서는 제곱 지수공분산함수(squared exponential(SE) covariance function)

$$k_{SE}(t) = \exp\left(-\frac{t^2}{2l^2}\right) \tag{2.7}$$

을 고려하기로 하며 여기서 모수 l은 길이척도(length-scale)이다. SE 공분산함수는 무한히 미분가능(infinitely differentiable)하므로 SE 공분산함수를 가진 GP는 모든 차수의 평균제곱 미분을 구할 수 있어 매우 스무드하게 되어 많이 사용되는 공분산함수이다. 그 외에도 Matern 공분산함수 $k_{Matern}(t)$, rational quadratic 공분산함수 $k_{RQ}(t)$ 등 여러 형태의 공분산함수들이 있으며 Rasmussen과 Williams (2005)의 4장을 참고하기 바란다.

$$k_{Matern}(t) = \left(1 + \frac{\sqrt{3}t}{l}\right)\exp\left(-\frac{\sqrt{3}t}{l}\right)$$
$$k_{RQ}(t) = \left(1 + \frac{t^2}{2\alpha l^2}\right)^{-\alpha}.$$

공분산 커널 함수

$$\mathrm{Cov}(f(x_u), f(x_v)) = K(x_u, x_v)$$

는 출력값에 대한 공분산이 입력값의 함수로 표현되는 것이다. l^2은 (x_u, x_v)에 대해 그 이후로는 서로 독립(uncorrelated)이 되는 거리를 나타낸다. 일변량인 경우 SE kernel은

$$K_y(x_u, x_v) = \sigma_f^2 \exp\left(-\frac{(x_u - x_v)^2}{2l^2}\right) + \sigma_n^2 \delta_{uv}, \tag{2.8}$$

여기서 δ_{uv}는 크로네커 델타 함수(Kronecker delta function)로 $u = v$이면 $\delta_{uv} = 1$이다. 길이척도 l^2은 입력값에 따라 함수 f의 변동량 조절 모수로 l^2이 작을수록 빠르게 변하며 l^2이 클수록 코사인(cosine) 함수와 같은 흐름을 보인다.

2.3. Gaussian Process Regression 추정

표본이 없는 시간점(time-point)에서의 유전자발현추정값을 구하는 방법은 다음과 같다. 즉 알려진 표본 입력값 $\mathbf{x}$에서의 함수값 $\mathbf{f}$ 정보를 이용해 새로운 입력값 x_*에서의 함수값 f_*를 구하고자하며 결

합분포 $p(f_*, \mathbf{f})$는 정규분포이므로 조건부분포 $p(f_*|\mathbf{f})$ 역시 정규분포가 된다. $\mathbf{y} \sim N(\mathbf{0}, \mathrm{Cov}(\mathbf{y}) = N(\mathbf{0}, \mathbf{K}_f(\mathbf{x}, \mathbf{x}) + \sigma_n^2 \mathbf{I})$ 베이즈 규칙을 적용하면

$$p(f_*|\mathbf{y}) = \frac{p(f_*, \mathbf{y})}{p(\mathbf{y})} \tag{2.9}$$

으로 계산한다. 새로운 시간점 입력값 x_*와 u번째 알려진 x_u에서 $m_u = m(x_u)$, $(k_*)_u = K_f(x_u, x_*)$로 표기한다.

x_*에서 새로운 벡터 k_*는 공분산행렬 $\mathbf{K}_C$에 열과 행이 연결해 더해져 다음과 같다:

$$\mathbf{K}_{C+1} = \begin{pmatrix} \mathbf{K}_C & \mathbf{k}_* \\ \mathbf{k}_* & \kappa \end{pmatrix}. \tag{2.10}$$

평균이 $\mathbf{0}$ 벡터인 경우와 식 (2.10)을 고려하여 $(\mathbf{y}, f_*)$의 결합분포를 구하면

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_f(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I} & \mathbf{K}_f(\mathbf{X}, \mathbf{X}_*) \\ \mathbf{K}_f(\mathbf{X}_*, \mathbf{X}) & \mathbf{K}_f(\mathbf{X}_*, \mathbf{X}_*) \end{pmatrix}\right) \tag{2.11}$$

을 얻는다.

GPR 모형에서 길이척도 l, 신호 분산(signal variance) σ_f^2, 잡음 분산(noise variance) σ_n^2가 변하며 이러한 모수 (l, σ_f, σ_n)에 대한 사전 분포로 감마분포를 고려한다. 저잡음(low noise)과 빠르게 변하는 신호(quickly-varying signal)인 경우는 신호(signal)에 대한 백색잡음과정으로 볼 수 있고 고잡음(high noise)과 느리게 변하는 신호(slowly-varying signal)인 경우는 백색잡음을 포함한 일정한 신호(constant signal)로 볼 수 있다.

2.4. 사후분포와 로그 주변우도함수

사후분포의 평균과 공분산을 얻기 위해 정규분포에 대한 계산식을 이용하여 구하면 GPR에 대한 추정식을 얻을 수 있다.

$$p(f_*|\mathbf{y}) \sim N\left(f_*; m_*, \mathrm{Var}(f_*)\right), \tag{2.12}$$

여기서 $K_f = \mathbf{K}_f(\mathbf{x}, \mathbf{x})$이며

$$m_* = \mathbf{k}_* \left(\mathbf{X}_f + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{y},$$
$$\mathrm{Var}(f_*) = k\left(\mathbf{x}_*, \mathbf{x}_*\right) - \mathbf{k}_*' \left(K_f + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{k}_*.$$

이러한 식은 여러 개 시간점에서의 함수값을 추정할 때 확장 적용된다.

SE 공분산함수인 경우 GP에 대한 로그 주변우도함수를 최적화하는 모수(hyperparameter)를 찾을 수 있으며 특히 SE 공분산함수일 경우에는 해석이 용이한 모수값으로 얻게 된다.

식 (2.6)에서 주어진 사전분포 하에서 주변우도함수를 계산해보자.

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{x}) p(\mathbf{f}|\mathbf{x}) d\mathbf{f}. \tag{2.13}$$

로그우도함수 LML을 구해보면

$$\ln p(\mathbf{y}|\mathbf{x}, \theta) = -\frac{1}{2}\mathbf{y}'\mathbf{K}_\mathbf{y}^{-1}\mathbf{y} - \frac{1}{2}\ln|\mathbf{K}_y| - \frac{m}{2}\ln 2\pi, \tag{2.14}$$

여기서 $\mathbf{K_y} = \mathbf{K}_f + \sigma_n^2\mathbf{I}$이다.

Kalaitzis와 Lawrence (2011)에 따라 다음과 같이 LML 비(LML ratio)를 구해 유전자 프로화일에 순위를 줄 수 있다.

$$\ln\left(\frac{p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}_1)}{p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}_0)}\right), \tag{2.15}$$

여기서 θ_1은 대립가설인 경우를 의미하고 $\boldsymbol{\theta}_0 = (l = \infty, \sigma_f^2 = 0, \sigma_n^2)$은 귀무가설 즉 차별화되지 않은 발현인 경우를 의미한다. 각 가설 상황 하에서 LML을 구한 후 비(ratio)를 구한다. 최적값을 구하기 위해 모수 $\boldsymbol{\theta}$에 대해 편미분하여 최적화를 찾는 방법으로 척도 공액 경사도(scaled conjugate gradients)를 이용하여 LML을 최대로 하는 값과 이때의 길이척도값을 얻는다.

3. 데이터 분석

효모 데이터에 대해 GPR 활용 유전자 선별 작업을 하기 전에 우선 모의실험을 수행하여 선별 결과에 대한 정보를 얻고자한다.

3.1. 모의실험

모의실험에 사용한 함수는 다음과 같으며 함수 f_1은 차별화되지않는 실현값을 갖는 데이터이다. f_1은 300개, 다양한 변화 패턴을 갖는 나머지 5개 함수는 가각 40개씩으로 총 500개의 유전자에 대해 $m = 20$ 시점에서 데이터를 발생시킨 후 GPR 적합을 구하였다. 여기서 $t_i = i/m$이다.

$$\begin{aligned}
f_1 &= 0,\\
f_2 &= 2t_i\\
f_3 &= \min\left\{\left(\frac{5-5t_i}{2}\right), \left(\frac{5t_i-2}{3}\right)^2 + \sin\left(\frac{5\pi t_i}{2}\right)\right\},\\
f_4 &= -2t_i + \sin\left(\frac{5\pi t_i}{2}\right),\\
f_5 &= \cos(2\pi t_i),\\
f_6 &= 2|t_i - 0.3|.
\end{aligned}$$

이러한 함수에 정규분포 $N(0,\sigma^2)$를 따르는 오차항이 더해져 실험데이터가 구성되며 300번 반복 실험을 하였다. Table 3.1에서는 결정점(cut-off point)으로 LML ratio가 1.6과 1.5인 경우 σ가 0.5, 1.0인 경우에 대해 차별화된 유전자를 선별한 결과를 보여준다. 선별된 퍼센트(screened percentage)는 실제 차별화된 유전자 대비 GPR 적합 후 차별화된 유전자 비율로 $\sigma = 0.5$의 경우 초과선별 경향을 보여준다. 민감도(sensitivity), 특이도(specificity), FDR(false discovery rate), FNR(false nondiscovery rate) 값을 보면 분산에 따라 변화가 있음을 알 수 있으며 데이터의 특성에 따라 결정값도 구해야한다. 함수의 진동폭 등 함수 형태와 오차항의 분산정도에 따라 선별 결과가 달라질 수 있으므로 분석하고자하는 데이터에 대하 탐색적 연구가 필요하다.

3.2. 효모 데이터에 대한 Gaussian process regression과 유전자 선별

효모는 단세포로서 세포분열 단계를 잘 나타내주는 생물재료로 인식되고 있으며 Spellman 등 (1998)이 데이터를 제공한 이후로 이러한 효모 데이터에 대한 분석은 세계 각국의 생물정보학(bioinformatics)

Table 3.1. Simulation results with GPR screening with 500 genes and $m = 20$ time points in 300 repetitions

LML ratio	σ	screened percentage(%)	sensitivity	specificity	FDR	FNR
1.5	0.5	110.8	0.928	0.881	0.121	0.072
1.6		105.6	0.906	0.910	0.105	0.094
1.5	1.0	81.2	0.602	0.867	0.145	0.398
1.6		75.3	0.558	0.876	0.132	0.441

Table 3.2. Functions of top 9 genes based on GPR LML ratio

Rank	Probe ID	Gene Name	Gene Function
1	3073	YLR452C (SST2)	GTPase-activating protein for Gpa1p, 알파인자 페로몬 감지 조절
2	2261	YJR004C (SAG1)	Alpha-agglutinin of alpha-cells, Agglutination 조절
3	2664	YLL024C (SSA2)	ATP binding protein, 단백질 접합과 액포로의 수송에 관여
4	3238	YMR058W (FET3)	Ferro-O2-oxidoreductase, 철 흡수와 구리에 대한 내성 반응 조절
5	387	YCR098C (GIT1)	Plasma membrane permease, inositol과 인의 대사조절에 관여
6	2524	YKL177W	Dubious open reading frame, STE3 유사 유전자
7	1342	YGL089C (MF(Alpha2))	Mating pheromone alpha-factor, mating type 조절인자
8	2874	YLR216C (CPR6)	Cyclophilin, 펩티드 결합의 cis-trans isomerization 촉매
9	3848	YOR135C (IRC14)	Dubious open reading frame, YOR136W 유사 유전자

실험실에서 이루어지고 있다. 세포주기(cell cycle)는 유전자적 불안정성(genomic instability) 또는 비정상적 세포분할(abnormal cell division)과 관련 있는 세포복제(cell replication), 악성과 재발성 질병(malignancy and reproductive disease) 등을 나타내는데 중요한 역할을 한다. 분석에 이용한 데이터는 효모 마이크로어레이 발현 데이터로 cdc15 동기조절방법(cdc15 synchronization)을 이용한 후 측정한 시계열 데이터로 http://genome-www.stanford.edu/cellcycle/에서 다운로드받았으며 결측값을 포함한 유전자를 제거하였다. 본 연구에서는 결측값을 제거 후 남은 4381개 유전자들이 매 7분마다의 24개 시간점에서 측정된 유전자 발현 값이다. 우선 효모 유전자 발현 데이터에 대해 GPR 을 적용하여 로그 주변우도함수를 계산하여 큰 값부터 우선순위를 준다. 순위가 1부터 9까지의 유전자 각각에 대해 GPR 적합 그림과 최적값의 길이척도값을 Table 3.2와 Figure 3.1에서 각각 보여준다. 우선순위로 차별화된 유전자 100개와 낮은 순위로부터의 100개 유전자들에 대한 시간에 따른 Figure 3.2를 보면 발현 진폭에 차이가 남을 볼 수 있다. 우선순위 100개는 기존 연구 결과들을 참조하여 임의로 선택한 숫자이며 결정점에 의해 차별화된 유전자로 선택되는 개수가 차이가 나게 된다.

데이터의 상황을 반영할 수 있는 결정점에 대한 연구가 필요하다. 본 연구에서는 LML ratio가 1.6 이상인 경우만 선택하기로 하면 367개 유전자가 선택된다. 여기서 선별되는 유전자들의 특성과 개수 등을 고려하며 결정점을 정하였다. 이와 같이 선택된 367개 유전자들에 대해 모형기반 군집화(model based clustering)를 고려하고자한다. 모형 기반 군집화는 Fraley와 Raftery (2002, 2006, 2007)에 의한 제안되고 정리되었으며 가우시안 혼합모형(Gaussian mixture model) 기반 군집화 방법이다. Serban과 Wasserman (2005), Kim과 Kim (2008) 등에서 효모 유전자 데이터에 대한 모형 기반 군집화를 활용하였다. 가우시안 모형기반 군집 결과 최적으로 4개 군집으로 각 군집은 42, 43, 117, 165개의 유전자로 형성된다. 이와 같은 군집 결과에 대한 군집의 gene ontology와 평균을 각각 Table 3.3과 Figure 3.2에서 보여주는데 군집별로 유전자 발현 곡선 모양의 차이가 나타난다. 군집 유효성 측도로 Rousseeuw (1987)가 제안한 실루엣(silhouette) 측도를 계산한 결과 실루엣 값 평균은 0.22이며 군집별 각 유전자에 대한 실루엣 값은 Figure 3.3에서 보여준다.

Lee 등 (2011)에서는 검정통계량에 5개의 푸리에 계수를 사용하였을 경우 FDR(false discovery rate)을 통해 유전자 선별작업을 수행한 결과 전체 유전자수 4381개 중에서 418개 유전자가 선별되었다. 이와 같이 유전자 선별방법에 따라 차이가 날 수 있는데 적용하는 통계량에 대한 선택, FDR이나

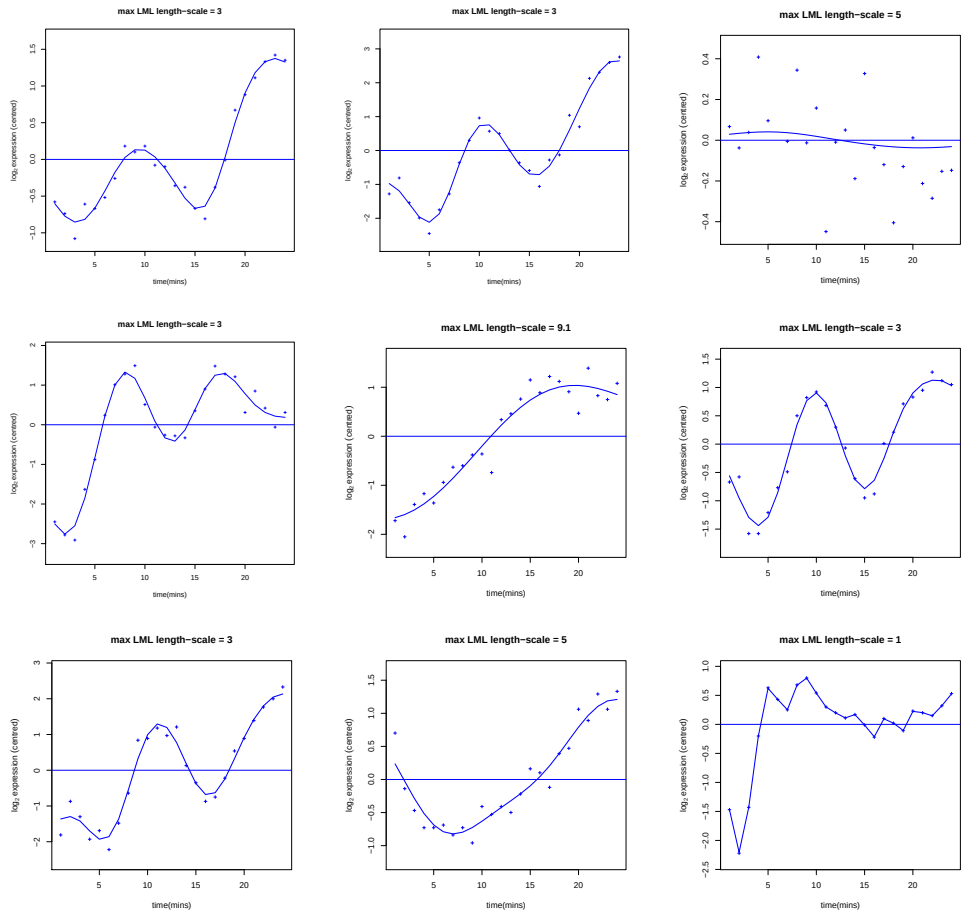
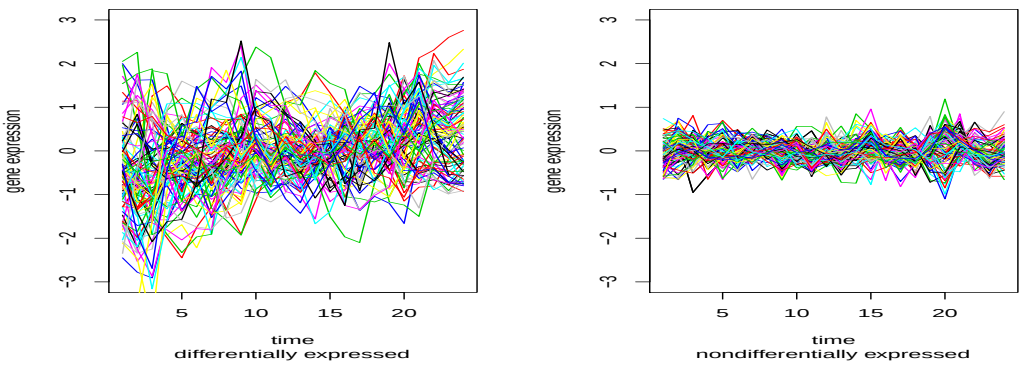


Figure 3.1. Top 9 genes based on LML ratio with Gaussian process regression (from top left to right)



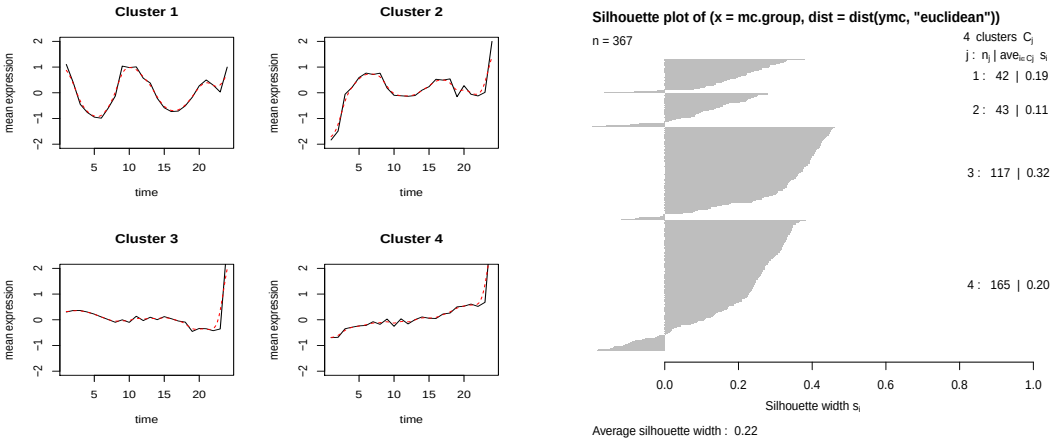
(a) Highest 100 differential genes

(b) Lowest 100 nondifferential genes based on LML ratio

Figure 3.2. Differentially vs. nondifferentially expressed genes

Table 3.3. GO genes in model based clusters

cluster	representative GO	function	ID	Gene Name
Cluster 1	GO:0006457	Protein Folding	3122833	Heat shock protein SSA2
			3129618	Heat shock protein SSA3
			3123898	Heat shock protein SSC1, mitochondrial
			3123667	ATP-dependent molecular chaperone HSC82
			3127063	Heat shock protein homolog SSE2
			3128448	Heat shock protein SSA4
			3128997	Heat shock protein SSC3, mitochondrial
			3122790	Heat shock protein 60, mitochondrial
Cluster 2	GO:0005618	Cell wall	3125902	Cell wall protein YLR194C
			3127813	Alpha-agglutinin
			3127120	Cell wall protein CWP2
			3122630	Probable glycosidase CRH1
			3127083	Protein SIM1
Cluster 3	GO:0031224	Intrinsic to memebrane	3126065	Cell membrane protein YLR413W
			3126361	Probable transporter MCH4
			3126842	Flavin carrier protein 1
			3125015	Putative uncharacterized protein YPR123C
			3124222	Phosphate metabolism protein 6
			3123111	Putative uncharacterized protein YGR160W
Cluster 4	GO:0010629	Negative regulation of gene expression	3129286	Condensin complex subunit 1
			3128999	Zinc finger protein RME1
			3128680	NAD-dependent deacetylase HST2
			3125675	Transcription factor BAF1



(a) Mean function in model based clusters (b) Silhouette values for model based clusters

Figure 3.3. Cluster profiles and silhouette values

FWER(family-wise error rate) 또는 GPR인 경우 LML 비 사용 등의 기준에 따라 차이가 날 수 있기 때문이다.

4. 결론

본 연구에서는 효모 유전자 발현 데이터에 대해 GPR을 적용하여 차별화된 유전자에 대해 로그주변우

도함수를 구하고 로그주변우도함수 비를 이용하여 유전자를 선별하였으며 선별된 유전자들에 대해 모형기반 군집을 형성하여 군집별 유전자 함수의 특성을 알아보았다. 시계열 유전자 발현 데이터에 대해 functional data의 관점으로 GPR을 적용하여 선별할 경우 유전자 발현의 전반적인 패턴이 고려되므로 각 유전자에 대한 함수를 기반으로 선별할 수 있는 장점이 있다. 그러나 결정점 기준에 따라 선택되는 유전자 개수가 달라질 수 있으므로 몇 가지 경우로 시행해보며 정해야한다. GPR을 적용하여 선별된 유전자 집합과 Lee 등 (2011)에서 시계열 반영 검정통계량과 FDR을 이용한 경우 선별된 유전자 집합이 일치하지는 않는다. 선별을 위한 통계량 계산과 선별 기준에 따라 차이가 날 수 있으므로 데이터 상황에 대한 파악이 필요하다. 앞으로 GPR을 적용하여 결정점을 구하는 문제, GPR과 군집화를 동시에 해결하는 문제, 데이터 형태를 잘 반영하는 공분산함수 탐구 문제 등 아직 해결해야할 연구 문제들이 있으며 또한 생물학적 의미를 고려한 방법 등 다각적인 연구가 필요하다.

References

Eckel, J. E., Gennings, C., Chinchilli, V. M., Burgoon, L. D. and Zacharewski, T. R. (2004). Empirical Bayes gene screening tool for time-course or dose-response microarray data, *Journal of Biopharmaceutical Statistics*, **14**, 647–670.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, **97**, 611–631.

Fraley, C. and Raftery, A. E. (2006). *MCLUST Version 3 for R: Normal mixture modeling and model-based clustering*. Technical Report No. 504.

Fraley, C. and Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering, *Journal of Classification*, **24**, 155–181.

Hero, A. O., Fleury, G., Mears, A. J. and Swaroop, A. (2004). Multicriteria gene screening for analysis of differential expression with DNA microarrays, *Journal on Applied Signal processing*, **2004**, 43–52.

Kalaitzis, A. and Lawrence, N. (2011). A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression, *BMC Bioinformatics*, **12**, 180.

Kim, J. and Kim, H. (2008). Clustering of change using Fourier coefficient, *Bioinformatics*, **24**, 184–191.

Lee, K., Kim, T. and Kim, J. (2011). Gene screening and clustering of yeast microarray gene expression data, *The Korean Journal of Applied Statistics*, **24**, 1077–1094.

Ma, S. (2006). Empirical study of supervised gene screening, *BMC Bioinformatics*, **7**, 537.

Rasmussen, C. E. and Williams, C. K. (2005). *Gaussian Processes for Machine Learning*, MIT Press

Rousseeuw, P. T. (1987). Silhouettes: Graphical aid to the interpretation and validation of cluster analysis. *Journal of Computation Applied Math*, **20**, 53–65.

Serban, N. and Wasserman, L. (2005). CATS: Clustering after transformation and smoothing, *Journal of the American Statistical Association*, **471**, 990–999.

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.

Törönen, R., Kolehmainen, M., Wong, G. and Castren, E. (1999). Analysis of gene expression data using self-organizing maps, *Federation of European Biochemical Societies*, **451**, 142–146.

Zhang, L., Zhang, A. and Ramanathan, M. (2003). Fourier harmonic approach for visualizing temporal patterns of gene expression data, *IEEE Computer Society Bioinformatics Conference*, **2**, 137–147.

효모 마이크로어레이 유전자 발현데이터에 대한 가우시안 과정 회귀를 이용한 유전자 선별 및 군집화

김재희[a,1] · 김태훈[b]

[a]덕성여자대학교 정보통계학과, [b]덕성여자대학교 프리팜메드학과

(2012년 12월 28일 접수, 2013년 4월 16일 수정, 2013년 4월 26일 채택)

요 약

본 연구에서는 가우시안 과정회귀방법을 소개하고 시계열 마이크로어레이 유전자 발현데이터에 대해 가우시안 과정회귀를 적용한 사례를 보이고자한다. 가우시안 과정회귀를 적합하여 로그 주변우도함수 비를 이용한 유전자를 선별방법에 대한 모의실험을 통해 민감도, 특이도, 위발견율 등을 계산하여 선별방법으로의 활용성을 보였다. 실제 효모 세포주기 데이터에 대해 제곱지수공분산함수를 고려한 가우시안 과정회귀를 적합하여 로그 주변우도함수 비를 이용하여 차변화된 유전자를 선별한 후, 선별된 유전자들에 대해 가우시안 모형기반 군집화를 하고 실루엣 값으로 군집유효성을 보였다.

주요용어: 가우시안 과정회귀, 로그주변우도함수, 시계열 마이크로어레이 발현데이터, 제곱지수공분산함수, 차별화 발현, 효모.

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF 2012-0001879)입니다.

[1]교신저자: (132-714) 서울시 도봉구 삼양로 144길 33, 덕성여자대학교 정보통계학과, 교수.
E-mail: jaehee@duksung.ac.kr