The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.4.569
26(4), 569–579

Bayesian Multiple Change-Point Estimation for Single Quantum Dot Luminescence Intensity Data

Jaehee Kim[a,1] · Hahkjoon Kim[b]

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

(Received February 15, 2013; Revised June 15, 2013; Accepted July 12, 2013)

Abstract

In the field of single-molecule spectroscopy, it is essential to analyze luminescence Intensity changes that result from a single molecule. With the CdSe/ZnS core-shell structured quantum dot photon emission data Bayesian multiple change-point estimation is done with the gamma prior for Poisson parameters and truncated Poisson distribution for the number of change-points.

Keywords: Multiple change-points, single-molecule spectroscopy, Bayesian, truncated Poisson distribution, Poisson count data, SAMC algorithm.

1. 서론

1998년 Brown 박사가 화학정보학(chemoinformatics)이란 단어를 정의한 후 화학정보학은 화학적 문제를 풀기위해 통계적 방법을 활용하고자 하는 목적으로 정보의 소스(source)를 데이터화하고 데이터로부터 통계적 분석을 통해 정보를 추출하고 화학적 지식을 얻고자하는 학문분야이다 (Chen, 2006). 화학반응의 메커니즘을 알려줄 수 있는 데이터를 얻은 후 통계적 분석을 통해 정보를 얻는 연구는 화학반응 연구 뿐만아니라 약 개발(drug discovery)에 매우 중요한 역할을 하게되었다.

Gasteiger과 Jochum (1979) 그룹은 Partial Equalization of Orbital Electronegativities(PEOE)에 의해 유기분자의 부분전하량을 계산하는 등 물리화학적(physicochemical) 효과를 측정하는 경험적 방법을 개발하였고 이와 같은 과정에서 발생하는 대량의 데이터에 대해 프로그램 패키지(program package)인 PETRA(Parameter Estimation for the Treatment of Reactivity Applications)를 개발하였다. NIH(National Institute of Health, 미국국립보건원)는 화학 구조 데이터베이스(chemical structure database) PubChem을 운영하며 작은 분자(molecule)에 대한 생물학적 활성(biological activity)에 대한 정보를 제공하고있다. 2006년 이후에는 3차원 단백질 구조에 대한 정보 데이터도 제공하며 연구자들에게 다량의 화학정보데이터를 제공한다. 이러한 데이터 분석을 위해 통계적 분석도구는 필수적이며 화학정보학은 새로운 연구분야로 인식되고 빠르게 발전하고 있으며 지속적 발전이 필요한 분야로 볼 수 있다.

[1]Corresponding author: Professor, Department of Statistics, Duksung Women's University, Samyang-ro 144 Gil 33, Dobong-Gu, Seoul 132-714, Korea. E-mail: jaehee@duksung.ac.kr

분자 수준에서 단백질과 같은 거대 분자의 기능을 이해하는 것은 신약 개발이나 질병 치료와 같은 의학 분야에 매우 중요하다. 거대 분자의 기능은 그 구조와 밀접하게 연관되어 있는데, 형광 분광법에 기초한 방법들은 거대 분자에 결합된 형광 물질로부터 발생한 형광이 주변 환경에 영향을 받아 변하는 정도를 통해 거대 분자의 구조 변화를 추정하는 방법이다 (Royer, 2006). 예를 들어 거대 분자의 구조가 변할 것으로 예측되는 부분에 결합된 형광물질이 형광을 흡수하는 부분과 가까워지면 형광의 세기가 감소하고 멀어지면 형광의 세기가 증가하는데 이를 통해 거대 분자의 구조가 어떻게 변하는지 알 수 있다. 단분자 분광학은 거대 분자 하나에 결합된 형광 변화를 관찰하는 실험으로 앙상블 평균(ensemble averaging)에 기초한 기존의 벌크(bulk) 실험이 제공할 수 없는 개별 분자의 중요한 특성에 관한 정보를 제공한다 (Lu 등, 1998; Zhuang 등, 2000). 예를 들면, 용액 내에 있는 거대 분자들의 구조가 불균일할 때 단분자 분광학을 통해 거대분자들의 분포를 알 수 있어 거대 분자의 성질을 이해하는 데 매우 유용하다 (Brasselet 등, 2000). 또한, 분자 하나만 관찰하므로 하나의 거대 분자 구조가 변하는 과정을 실시간으로 관찰할 수 있어 단백질의 접힘, 모터 단백질의 이동 메커니즘 등 분자 수준에서의 생명 현상을 연구하는 데 많이 응용되고 있다 (Michalet 등, 2006; Yildiz 등, 2003). 이러한 장점에도 불구하고 단분자 분광학을 통한 거대 분자의 구조 변화는 분자 하나에서 발생하는 매우 약한 형광 신호에 의존하기 때문에 바탕 신호로부터 형광 신호를 분리하고, 그 정보를 분석하여 결론을 도출하는데 데이터의 통계적 방법이 매우 중요하다 (Mandel과 Wolf, 1995).

단분자 분광학 데이터에 대한 변화점 추정은 균일한 부집단으로 나눌 수 있고 거대분자의 구조변화를 파악할 수 있어 매우 관심을 받고 있는 문제이다. 최근 진행된 연구를 살펴보면, Watkins와 Yang (2005)은 EM 군집화(expectation-maximization clustering)와 Bayesian information criterion(BIC)를 통한 일반화우도비검정(generalized likelihood ratio test)를 이용해서 영역정하기(binning)나 임계값정하기(thresholding)으로부터 발생할 수 있는 인위적인 제한 없이 단분자 분광학 발광세기 데이터에 대해 변화점을 추정했다. Jansen (2007)은 포아송 분포를 이용하여 발광세기 데이터에 대해 연속 웨이브릿(wavelet) 변환을 이용하여 변화점을 추정하였다.

본 연구에서는 단분자 분광학 데이터에 대한 다중변화점 분석을 통해 다중변화점을 추정과 더불어 분자의 구조에 대한 화학적 정보를 얻고자한다. 2장에서는 실제 분석에 사용된 단분자 분광학 데이터에 대해 설명하고 통계적 분석의 필요성을 설명하고 3장에서 포아송 데이터에 대한 베이지안 다중변화점 모형을 소개하고 4장에서는 이러한 데이터에 대한 다중변화점 추정결과와 화학적 분석 결과를 보여준다.

2. 단분자 분광학 발광 세기 데이터

이 논문에서 사용한 분광학 데이터는 다음과 같다. 지름이 약 10nm인 CdSe/ZnS core-shell 구조를 갖는 양자점(QD605, Invitrogen Corporation)에 파장이 532nm인 CW(continuous wave) 레이저를 조사 시킨 후, 발생한 발광의 시간에 따른 변화를 관찰하였다. 이를 위하여 유리 슬라이드 표면을 1% biotin이 결합된 PEG(polyethylene glycol, Laysanbio)으로 코팅한 후, biotin과 강하게 결합하는 streptavidin(Sigma)을 슬라이드 표면에 부착시켰다. 실험에 쓰인 양자점에는 약 3개의 biotin이 붙어 있는데, 매우 낮은 농도(∼ 20pM)의 양자점을 1% biotin이 결합된 PEG으로 코팅된 슬라이드 위에 가하면 그 슬라이드 표면에 양자점이 붙게 되고 piezo stage(Tritor SG 101, piezosystem jena)로 표면을 스캔 하여 양자점의 위치를 발견하면 하나의 양자점에 레이저를 조사시켜 발광 변화를 얻었다. 실험에서 샘플로부터 발광 데이터를 최대한 얻기 위해 60 × TIRFM oil immersion objective(NA 1.45, Olympus)를 사용하였으며 광자 검출기는 검출 효율이 높은 avalanche photodiode(Perkin Elmer, AQRH-13)를 사용하였다. 발광에서 나온 각 광자가 검출기에 도달한 시간을 기록하였고, 이 도달 시간에 대해 10ms 단위로 히스토그램을 그리면 Figure 2.1(a)와 같은 시간에 대한 강도 변화 그래프가 얻어지며 이

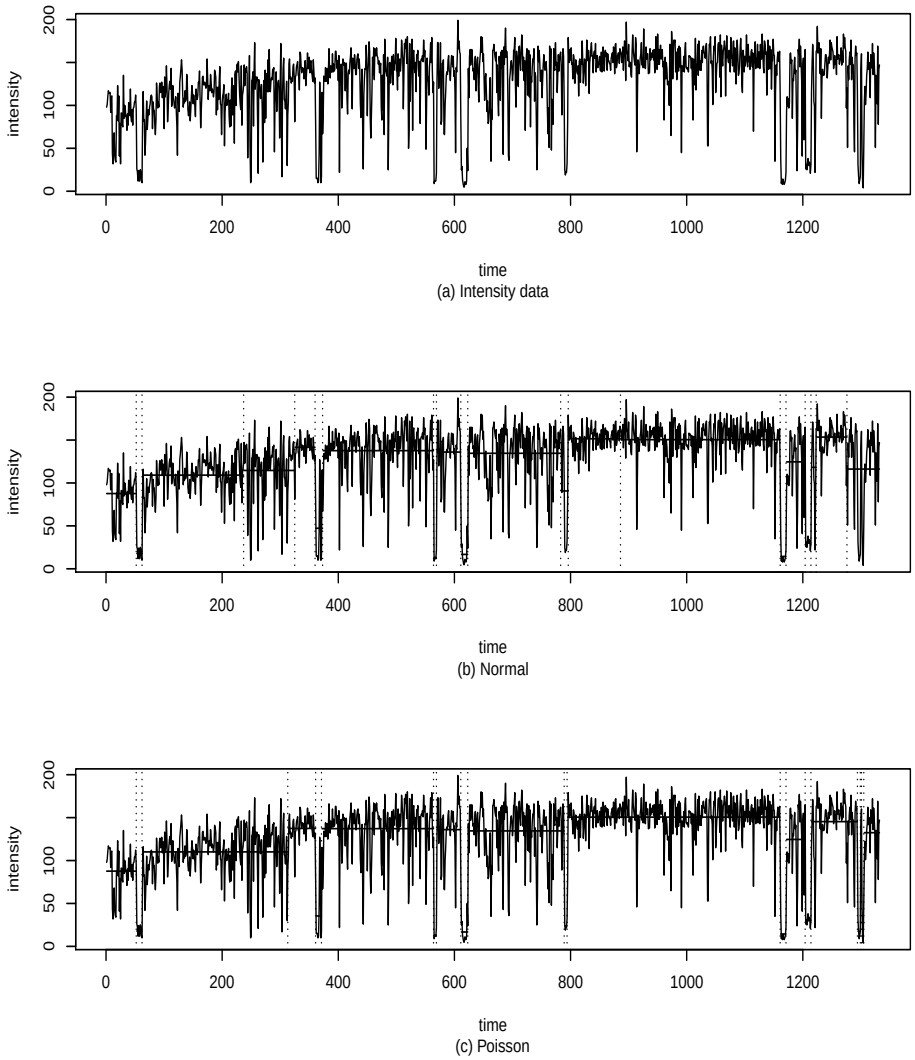

Figure 2.1. Intensity data and change-point estimation

그림에서 볼 수 있듯이 일반적인 양자점은 에너지를 흡수하여 발광이 발생하다 어느 순간에 발광이 멈추는 깜박거림 현상이 나타나는데 그 지속 시간은 수 마이크로초에서 수십 밀리초까지 다양하다. 이와 같이 발광의 깜박거림 세기와 지속 시간은 랜덤하게 변하여 직관적으로 발광 세기의 변화를 추정하는 것은 거의 불가능하다 (Nirmal 등, 1996; Kuno 등, 2000). 따라서 발광 세기의 유의미한 변화는 통계적 접근에 의해서만 가능하다.

3. Poisson count data에 대한 다중변화점 추정

데이터가 동일한 분포를 따른다고 볼 수 없을 때 변화점 문제가 발생하며 통계적 분석을 위해 변화점을 추정한 후 각 부집단에 대한 추론을 해야한다. 단분자 분광학 발광세기 데이터는 포아송분포를 따르는

개수(count) 데이터이므로 포아송분포 가정이 적절하다. 그러나 짧은 시간에 다량의 측정값이 얻어지므로 일정구간을 정해 평균값으로 데이터를 활용하는 경우에는 정규분포를 활용할 수도 있다.

Raftery와 Akman (1986)은 포아송 분포의 경우 한 개 변화점이 있을 경우 improper prior를 주고 사후분포를 유도하여 베이지안 변화점을 추정법을 제안하였고 Jarrett (1979)은 석탄 광산 사고 데이터에 대해 포아송 분포의 다중변화점을 MCMC 기법을 이용하여 추정하였다. Ensign과 Pande (2010)은 단분자 궤도를 나타내는 포아송분포를 따르는 데이터에 대해 Bayes factor를 이용하여 다중변화점을 추정하였다. Kim과 Cheon (2010)은 변화점의 개수에 대해 절단포아송 분포를 가정하고 포아송분포와 정규분포에 대한 베이지안 다중변화점 추정법을 제안했고 Cheon과 Kim (2010)은 다변량의 경우 다중변화점 추정방법을 제안하였으며 본 연구에서 이러한 방법을 활용하고자한다.

발생 순서에 따른 서로 독립인 관측값들을 $\mathbf{Z} = (z_1, z_2, \ldots, z_n)$이라하자. k개 변화점(change-point) $\{c_1, \ldots, c_k\}$를 $0 = c_0 < c_1 < c_2 < \cdots < c_{k+1} = n$라 할 때 변화점을 통해 지시변수집합(index set) $\{1, 2, \ldots, n\}$을 각 부집단으로 나눌 수 있으며 이렇게 형성된 부집단내의 관측값들은 부집단별로 동일한 분포를 따르게 된다. 그러므로 변화점은 이분벡터(binary vector) $\mathbf{x} = (x_1, x_2, \ldots, x_n)$로 표현할 수 있고 $x_{c_1} = x_{c_2} = \cdots = x_{c_k} = 1$을 만족한다. k개 변화점을 가진 모형은 다음과 같다:

$$z_i \sim f_r(\cdot|\phi_r), \quad c_{r-1} < i \le c_r, \tag{3.1}$$

여기서 $r = 1, 2, \ldots, k+1$이고 f_r는 분포에 대한 모수에 의존하는 확률밀도함수이고 $\phi_r \in \Phi$이다.

변화점 개수에 대한 사전분포로는 $n^* = n/2$ 절단 포아송(truncated Poisson) 분포

$$\pi(k) = \frac{\lambda^k}{\sum_{j=0}^{n^*} \frac{\lambda^j}{j!}} \frac{(n^* - k)!}{n^*!} \tag{3.2}$$

를 가정한다.

포아송분포 Poi(ϕ_r)를 따르는 포아송 확률밀도함수 f_r이라 하자. 모형 $\boldsymbol{\eta}^{(k)} = (\phi_1, \ldots, \phi_{k+1})$에 대한 로그우도함수(log-likelihood function)는

$$\log L\left(\mathbf{Z}|\boldsymbol{\eta}^{(k)}\right) = \sum_{i=1}^{k+1} \left\{ \left(\sum_{j=c_{i-1}+1}^{c_i} z_j \right) \log \phi_i - (c_i - c_{i-1})\phi_i - \sum_{j=c_{i-1}+1}^{c_i} z_j! \right\}. \tag{3.3}$$

ϕ_i에 대한 공액사전분포로 다음의 감마분포 $G(\gamma, \delta)$를 고려하면

$$\log P\left(\boldsymbol{\eta}^{(k)}\right) = a_k + \sum_{i=1}^{k+1} \{(\gamma - 1)\log \phi_i - \phi_i \delta\}, \tag{3.4}$$

여기서 $a_k = (k+1)(\gamma \log \delta - \log \Gamma(\gamma)) + \log(n^* - k)! + k \log \lambda$.

사후분포를 $\phi_1, \ldots, \phi_{k+1}$에 대해 적분하여 변화점모형에 대한 사후분포를 구하면 다음과 같다:

$$\log P\left(\mathbf{x}^{(k)}|\mathbf{Z}\right) = a_k + \sum_{i=1}^{k+1} \left\{ \log \Gamma\left(\gamma + \sum_{j=c_{i-1}+1}^{c_i} z_j \right) - \left(\sum_{j=c_{i-1}+1}^{c_i} z_j + \gamma \right) \log\left(c_i - c_{i-1} + \delta\right) - \sum_{j=c_{i-1}+1}^{c_i} \log z_j! \right\}. \tag{3.5}$$

정규분포$N(\mu, \sigma^2)$를 따르는 정규 확률밀도함수 f_r에 대해서는 평균에 대해서는 improper prior와 분산 모수에 대해서는 역감마분포(inverse Gamma) IG(γ, δ)를 고려한 후 $\phi_1, \ldots, \phi_{k+1}$에 대해 적분하여 변화점모형에 대한 사후분포를 구하면

$$\log P\left(\mathbf{x}^{(k)}|\mathbf{Z}\right) = a_k + \frac{k+1}{2}\log 2\pi - \sum_{i=1}^{k+1}\left\{\frac{1}{2}\log\left(c_i - c_{i-1}\right) - \log\Gamma\left(\frac{c_i - c_{i-1} - 1}{2} + \gamma\right)\right.$$
$$\left. + \left(\frac{c_i - c_{i-1} - 1}{2} + \gamma\right)\log\left[\delta + \frac{1}{2}\sum_{j=c_{i-1}+1}^{c_i} z_j^2 - \frac{\left(\sum_{j=c_{i-1}+1}^{c_i} z_j\right)^2}{2(c_i - c_{i-1})}\right]\right\} \tag{3.6}$$

을 얻는다. 여기서 $a_k = (k+1)(\gamma\log\delta - \log\Gamma(\gamma)) + \log(n-1-k)! + k\log\lambda$이고 γ, δ, λ는 hyperparameter이다.

변화점 모형선택은 Kass와 Raftery (1995)가 언급한 바와 같이 로그 Bayes factor로 근사되는 BIC를 최소화하도록 선택한다. 여기서 BIC $= -2$(log maximized likelihood) $+$ $(\log n)$(number of parameters)로 정의되며 사후분포확률을 이용하고 모형에 포함되는 모수의 개수에 대한 벌칙항을 포함한다. 사후분포 계산에는 SAMC(Stochastic Approximation Monte Carlo) 알고리즘 (Liang 등, 2007)을 활용하여 계산한다.

다중 변화점 문제에 적용된 SAMC에 대해 간단히 설명하고자한다. 가변적인 모수 공간에서 움직이는 함수의 최대/최소값을 구하는 경우 적절한 표본추출 알고리즘이 필요한데 SAMC 알고리즘은 이러한 상황에서 적절하게 활용될 수 있다. Liang 등 (2007)은 reversible MCMC (Green, 1995)에 비해 SAMC는 국소적인 최대 최소값에 빠지지않고 전체 모수 공간에서의 최적값을 구할 수 있음을 보였다.

목표확률밀도함수(target probability density function) $p(\mathbf{x}) = c\psi(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$, 여기서 c는 모르는 상수(unknown constant)이고 $\psi(\mathbf{x})$는 표본공간(sample space) $\mathcal{X}$에서 정의된 음이 아닌 함수이다. $E_1, \ldots, E_m$는 $\mathcal{X}$의 partition이라놓고 $\theta_i = \log(\int_{E_i}\psi(\mathbf{x})d\mathbf{x})$, $i = 1, \ldots, m$. 여기서 θ_{ti}는 t번 반복 후 θ_i에 대한 추정값이다. $\boldsymbol{\theta}_t = (\theta_{t1}, \theta_{t2}, \ldots, \theta_{tm})$. $E_1, \ldots, E_m$을 우도함수를 고려하여 오름차순으로 $e^{\theta_1}, \ldots, e^{\theta_m}$의 확률을 갖도록 배열할 수 있으며 각 부집단에 대한 표본분포는 $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$, $0 < \pi_i < 1$, $\sum_{i=1}^m \pi_i = 1$이도록 한다.

비감소(non-decreasing) 양(positive)의 수열 $\{\tau_t\}$은 다음의 조건을 만족할 때

$$\text{(i)}\ \sum_{t=1}^{\infty}\tau_t = \infty, \qquad \text{(ii)}\ \sum_{t=1}^{\infty}\tau_t^{\zeta} < \infty, \tag{3.7}$$

여기서 $\zeta \in (1, 2)$. $\{\tau_t\}_{t\geq 0}$는 gain factor sequence로 불리며

$$\tau_t = \frac{T_0}{\max(T_0, t)}, \quad t = 0, 1, 2, \ldots,$$

여기서 미리 정한 $T_0 > 1$을 사용할 수 있다.

다중변화점 문제에 대해서 표본추출 단계는 다음과 같다.

(a) (Sampling) Metropolis-Hastings 기법으로 다음의 목표확률밀도함수를 고려하여 $\mathbf{x}_t$를 최신화한다.

$$P_{\theta_t}(\mathbf{x}) = \frac{1}{D_t}\sum_{i=1}^{m}\frac{\psi(\mathbf{x})}{e^{\theta_{ti}}}I(\mathbf{x} \in E_i), \quad t = 1, 2, \ldots, \tag{3.8}$$

여기서 D_t는 t번째 반복에서 표준화상수(normalizing constant)이고 I는 지시함수(indicator function)이다.

(a.1) proposal distribution $q(\mathbf{x}_t, \mathbf{y})$에 따라 $\mathbf{y}$를 발생시킨다.

(a.2) 다음의 비(ratio)를 계산한다.

$$G = \frac{e^{\theta_t}}{e^{\theta^*}} \frac{\psi(\mathbf{y})}{\psi(\mathbf{x}_t)} \frac{q(\mathbf{y}, \mathbf{x}_t)}{q(\mathbf{x}_t, \mathbf{y})}.$$

(a.3) $\min(1, G)$의 확률로 proposal을 받아들인다. 만약 받아들이면 $\mathbf{x}_{t+1} = \mathbf{y}$로 놓고 그렇지않은 경우에는 $\mathbf{x}_{t+1} = \mathbf{x}_t$로 놓는다.

(b) (Weight updating)

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}_t + \tau_{t+1}(\mathbf{e}_{t+1} - \boldsymbol{\pi}),$$

여기서 $\mathbf{e}_{t+1} = (e_{t+1,1}, \ldots, e_{t+1,m})$, 만약 $\mathbf{x}_t \in E_i$이면 $e_{t+1,i} = 1$, 아니면 0으로 놓는다. 만약 $\theta^* \in \Theta$이면 $\theta_{t+1} = \theta^*$; 그렇지않으면 $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}^* + \mathbf{c}^*$으로 놓는다. 여기서 $\mathbf{c}^* = (c^*, \ldots, c^*)$는 $\boldsymbol{\theta}^* + \mathbf{c}^* \in \Theta$ 조건을 만족하는 임의 벡터이다. 여기서 $\Theta = \left[-100^{100}, 10^{100}\right]^m$와 같이 고려할 수 있다.

4. 단분자 분광학 데이터에 대한 다중변화점 추정

본 연구에 사용한 단분자 분광학 발광세기 데이터는 Figure 2.1(a)에서 보여주며 이와 같이 발광의 세기가 계속 변하는 이유는 Ha (2009)를 참조하여 다음과 같이 설명할 수 있다. 양자점은 레이저에 의해 들뜬 상태가 된 후, 바닥상태로 떨어지면서 발광을 발생시키는데, 우연히 중심에 있던 전자가 표면으로 튀어 나오게 되면 표면은 음전하를 띄고 중심은 양전하를 띄게 된다. 이렇게 되면 양자점이 레이저의 에너지를 흡수하여 들뜬 상태가 되긴 하지만, 이 에너지를 발광이 아닌 열로 방출하게 되어 더 이상 빛이 나오지 않게 된다. 표면에 있던 전자가 중심으로 들어가게 되면 다시 양자점에서 발광이 발생하게 된다. 따라서 발광이 발생되다 멈추는 과정은 양자점의 표면 전하상태를 말해준다.

단백질과 같은 거대 분자의 구조 변화에 의해 발생하는 형광 변화는 일반적으로 랜덤하므로, 양자점에 의한 발광 신호의 랜덤한 변화점 추정은 거대 분자를 분석하기에 좋은 모델 시스템이다. 또한, 양자점의 움직임에 대한 변화점 연구는 용액 내에서 확산에 의한 형광 물질의 랜덤한 움직임에 의한 급격한 형광 변화를 연구하는 데에도 좋은 모델 시스템이다 (Zhang과 Yang, 2005). 형광에 의한 에너지 전이 (Förster resonance energy transfer; FRET)에 의하면 두 형광 물질 사이에 에너지 전이가 일어나면 형광의 세기가 줄어들거나 증가하는데 이는 두 형광물질 간의 거리에 의존한다 (Förster, 1948). 따라서 단백질과 같은 거대 분자내 특정 작용기간 거리나 단백질과 결합한 양자점 간의 거리와 같은 공간 정보를 얻기 위해서는 정확한 형광 변화 추정은 필수적이다 (Joo 등, 2004).

발광세기 데이터는 광자의 개수 데이터이므로 포아송 분포를 가정할 수 있다. 변화점 개수에 대한 사전분포의 절단 포아송 분포에서 $\lambda = 10$로 주었으며 사후분포 계산에서 SAMC 적용시 5,000,000번 반복계산하였으며 변화점 개수 k의 범위는 $k_{\min} = 15 \leq k \leq k_{\max} = 20$로 하였다. 또한 발광세기 데이터는 개수 데이터로서는 이산형이지만 영역(bin) 안에 들어온 평균을 취한 값으로 보면 연속형 정규분포로 가정할 수 있다. 정규분포에 대한 사전분포로는 Inverse Gamma$(0.05, 0.05)$를 주었으며 포아송분포에 대한 사전분포로는 Gamma$(r = 1.0, \delta = 0.01)$로 주었다.

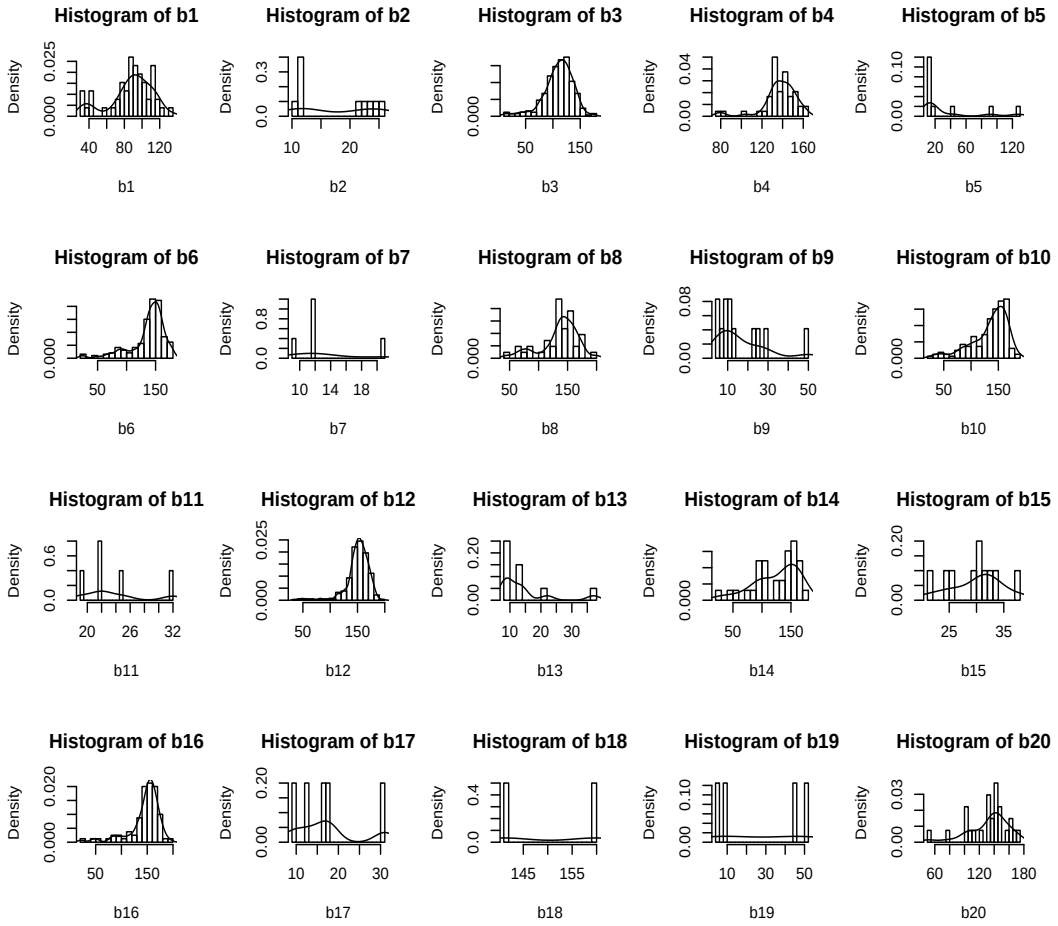

Figure 4.1. Histogram and density for intensity data in each segment

Figure 2.1(a)는 원래 데이터, (b)는 정규분포를 가정하고 추정한 결과 (c)는 포아송 분포를 가정하고 변화점을 추정한 결과이다. Figure 4.1는 포아송 분포에 대해 변화점을 추정하여 Figure 2.1(c)에서 보여주는 변화점에 의해 분리되는 부집단별 분포를 나타내는 히스토그램과 커널함수에 의한 확률분포추정선으로 각 부집단에 대한 분포가 다름을 알 수 있다. 정규분포의 경우 추정된 변화점은 19개이며 변화점 위치는 (52, 62, 237, 325, 360, 373, 564, 569, 611, 623, 783, 796, 886, 1161, 1171, 1204, 1214, 1223, 1276)이고 Table 4.1은 추정된 변화점에 의해 나누어지는 각 부집단의 평균과 분산 추정값이다. Table 3.2는 포아송 분포의 경우로 정규분포의 경우와 마찬가지로 변화점이 19개이고 변화점의 위치는 (52, 62, 313, 361, 371, 564, 569, 611, 623, 789, 794, 1161, 1171, 1204, 1214, 1294, 1299, 1301, 1305)이다. 변화점의 위치가 비슷하나 정규분포의 경우 886에 위치한 변화점은 포아송 분포의 경우에는 추정되지않았으며 1301에 위치한 변화점은 포아송 분포의 경우에만 추정되었다. 정규분포 가정하에 추정된 변화점 886은 분산모수만의 변화 때문에 추정된 것으로 생각되어 개수 데이터에 대한 적절한 통계적 변화점 모형이 다중변화점 추정에 중요함을 알 수 있다.

Table 4.1. Mean and Standard deviation of each regime as Normal distribution

regime	mean	sd
Regime 1	87.65385	25.10908
Regime 2	17.8	6.646637
Regime 3	109.1486	21.52066
Regime 4	114.5795	36.33931
Regime 5	141.5429	10.60062
Regime 6	47.23077	43.13961
Regime 7	137.6335	29.74279
Regime 8	13.2	4.549725
Regime 9	136.0238	32.96154
Regime 10	16.75	13.15727
Regime 11	134.575	33.62455
Regime 12	90.69231	61.63655
Regime 13	151.5	12.94609
Regime 14	150.4436	23.80347
Regime 15	14.4	9.008638
Regime 16	124.5455	38.85397
Regime 17	30.1	4.909175
Regime 18	118.2222	47.52572
Regime 19	153.4528	19.11158
Regime 20	116.1607	51.01899

Table 4.2. Mean and Standard deviation of each regime as Poisson distribution

regime	mean	sd
Regime 1	87.65385	25.10908
Regime 2	17.8	6.646637
Regime 3	110.0438	27.41916
Regime 4	137.3958	16.7176
Regime 5	35.4	40.82809
Regime 6	137.1503	30.06388
Regime 7	13.2	4.549725
Regime 8	136.0238	32.96154
Regime 9	16.75	13.15727
Regime 10	134.5301	33.41251
Regime 11	24	4.949747
Regime 12	150.5886	21.93114
Regime 13	14.4	9.008638
Regime 14	124.5455	38.85397
Regime 15	30.1	4.909175
Regime 16	145.3375	30.61064
Regime 17	17.6	8.294577
Regime 18	150.5	13.43503
Regime 19	27.5	24.5017
Regime 20	132.4444	27.77289

5. 결론

양자점의 발광 변화는 거대 분자에서 일어나는 형광 변화와 매우 유사하기 때문에 양자점의 발광 변화를 거대 분자의 형광 변화에 적용할 수 있다. 또한, 정확한 형광 변화점 분석은 에너지 전이를 계산하는데 응용되어 거대 분자내 특정 부분간의 거리를 계산과 같은 공간 정보를 얻는데 꼭 필요하다.

양자점은 일반적인 유기발광물질과 다르게 그 강도가 강하고 수명이 길어 단분자 분광학에서 많은 주목을 받고 있다. 그러나, 표면 상태에 따라 발광이 일어나지 않는다면 발광세기 변화가 거대 분자의 구조 변화 때문인지 표면상태 때문인지 판단하기 어려워진다. 따라서, 발광 변화가 적거나 거의 없는 양자점을 만드는 것이 단분자 분광학에서 매우 중요한데, 본 연구에서와 같이 단일 양자점의 세기 변화를 측정하고 분석하면 발광 물질로서 양자점의 상태를 알 수 있게 된다.

본 연구에서는 카드뮴셀레나이드/황화아연의 중심-껍질 구조를 갖는 양자점에 대해 단분자 분광학 실험으로 얻은 발광세기 데이터에 대해 베이지안 다중변화점 모형을 적용하여 각 변화점의 위치를 추정한 후 전체 데이터를 평균과 분산이 동일한 각 부분으로 나눌 수 있음을 보였고 각 변화점은 발광한 상태가 변화하는 점으로 상태 변화에 대한 중요한 정보를 제공한다. 이와 같이 변화점의 개수가 많은 데이터에 대해 사전정보를 이용한 베이지안 접근의 다중변화점 추정법을 통해 상태변화점을 추정하고 균일한 상태로 나눌수 있으므로 본 연구에서 활용한 다중변화점 추정법은 의미있는 변화분석법으로 생각된다. 앞으로 단분자 분광학 데이터에 대한 통계적 모형 연구를 통해 화학연구에 통계 기법이 많이 활용되기를 기대한다.

References

Brasselet, S., Peterman, E. J. G., Miyawaki, A. and Moerner, W. E. (2000). Single-molecule fluorescence resonant energy transfer in calcium concentration dependent cameleon, *Journal of Physical Chemistry B*, **104**, 3676–3682.

Chen, W. L. (2006). Chemoinformatics: Past, present, and future, *Journal of Chemistry Inference Model*, **46**, 2230–2255.

Cheon, S. and Kim, J. (2010). Multiple change-point detection of multivariate mean vectors with the Bayesian approach, *Computational Statistics & Data Analysis*, **54**, 406–415.

Ensign, D. and Pande, V. S. (2010). Bayesian detection of intensity changes in single molecule and molecular dynamics trajectories, *Journal of Physical Chemistry B*, **114**, 280–292.

Förster, T. (1948). Intermolecular energy migration and fluorescence, *Annuals of Physics*, **2**, 55–75.

Gasteiger, J. and Jochum, C. (1979). An algorithm for the perception of synthetically important rings, *Journal of Chemical Information and Computer Science*, **19**, 43–48.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82**, 711–732.

Ha, T. (2009). How nanocrystals lost their blink, *Nature*, **459**, 649–650.

Jansen, M. (2007). Multiscale change point analysis in Poisson count data, *Chemometrics and Intelligent Laboratory Systems*, **85**, 159–169.

Jarrett, R. (1979). A note on the intervals between coal-mining disasters, *Biometrika*, **66**, 191–193.

Joo, C., McKinney, S. A., Lilly, D. M. J. and Ha, T. (2004). Exploring rare conformational species and ionic effects in DNA Holliday junctions using single-molecule spectroscopy, *Journal of Molecular Biology*, **341**, 739–751.

Kass, R. and Raftery, A. (1995). Bayes factors, *Journal of the American Statistical Association*, **90**, 773–795

Kim, J. and Cheon, S. (2010). Bayesian multiple change-point estimation with annealing stochastic approximation Monte Carlo, *Computational Statistics*, **25**, 215–239.

Kuno, M., Fromm, D. P., Hamann, H. F., Gallagher, A. and Nesbitt, D. J. (2000). Nonexponential "blinking" kinetics of single CdSe quantum dots: A universal power law behavior, *Journal of Chemical Physics*, **112**, 3117–3120.

Liang, F., Liu, R. and Carroll, R. (2007). Stochastic approximation in Monte Carlo computation, *Journal of the American Statistical Association*, **102**, 305–320.

Lu, P. H., Xun, L. and Xie, X. (1998). Single-molecule enzymatic dynamics, *Science*, **282**, 1877–1882.

Mandel, L. and Wolf, E. (1995). *Optical Coherence and Quantum Optics*, Cambridge.

Michalet, X., Weiss, S. and Jager, M. (2006). Single-molecule fluorescence studies of protein folding and conformational dynamics, *Chemical Reviews*, **106**, 1785–1813.

Nirmal, M., Dabbousi, B. O., Bawendi, M. G., Macklin, J. J., Trautman, J. K., Harris, T. D. and Brus, L. E. (1996). Fluorescence intermittency in single cadmium selenide nanocrystals, *Nature*, **383**, 802–804.

Raftery, A. and Akman, V. (1986). Bayesian analysis of a Poisson process with a change-point, *Biometrika*, **73**, 85–89.

Royer, C. A. (2006). Probing protein folding and conformational transitions with fluorescence, *Chemical Reviews*, **106**, 1769–1784.

Watkins, L. P. and Yang, H. (2005). Detection of intensity change points in time-resolved single-molecule measurements, *Journal of Physical Chemistry B*, **109**, 617–628.

Yildiz, A., Forkey, J. N., Mckinney, S. A., Ha, T., Goldman, Y. E. and Selvin, P. R. (2003). Myosin V walks hand over hand: single fluorophore imaging with 1.5 nanometer localization, *Science*, **300**, 2061–2065.

Zhuang, X., Bartley, L. E., Babcock, H. P., Russell, R., Ha, T., Herschlag, D. and Chu, S. (2000). A single-molecule study of RNA catalysis and folding, *Science*, **288**, 2048–2051.

Zhang, K. and Yang, H. (2005). Photon-by-photon determination of emission bursts from diffusing single chromophores, *Journal of Physical Chemistry B*, **109**, 21930–21937.

단일 양자점으로부터 발생한 발광세기 변화에 대한 베이지안 다중 변화점 추정

김재희[a,1] · 김학준[b]

[a]덕성여자대학교 정보통계학과, [b]덕성여자대학교 화학과

(2013년 2월 15일 접수, 2013년 6월 15일 수정, 2013년 7월 12일 채택)

요 약

단일 분자에서 발생한 발광의 세기 변화를 분석하는 문제는 단분자 분광학에서 반드시 필요하다. 본 연구에서는 카드뮴셀레나이드/황화아연의 중심-껍질 구조를 갖는 양자점에 대한 단분자 분광학 데이터에 대해 Poisson count data로서 베이지안 접근으로 모수에 대한 공액 감마분포와 변화점 개수에 대한 절단포아송 분포로 사전분포를 주고 다중변화점을 추정하였다.

주요용어: 다중변화점, 단분자 분광학, 베이지안, 절단 포아송분포, 포아송 count data, SAMC 알고리즘.

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