

# Choosing the Tuning Constant by Laplace Approximation

SungMahn Ahn<sup>1,a</sup>, Suhn Beom Kwon<sup>a</sup>

<sup>a</sup>College of Business Administration, Kookmin University

---

## Abstract

Evidence framework enables us to determine the tuning constant in a penalized likelihood formula. We apply the framework to the estimating parameters of normal mixtures. Evidence, which is a solely data-dependent measure, can be evaluated by Laplace approximation. According to a synthetic data simulation, we found that the proper values of the tuning constant can be systematically obtained.

Keywords: Penalized likelihood, tuning constant, evidence, Laplace approximation.

---

## 1. 서론

확률분포 추정을 위하여 벌점가능도를 최대화하는 방법은 Good (1971)이 처음으로 제안하였다. Good이 제안한 벌점가능도는 다음과 같은 형태이다.

$$\sum_{i=1}^n \log f(x_i) - \lambda J(f). \quad (1.1)$$

식 (1.1)에서 첫 번째 항은 로그우도함수이며, 두 번째 항은 벌점항으로서 추정되는 함수가 특정한 성질을 가지면 로그우도함수의 값이 작아지도록 만드는 역할을 한다. Good과 Gaskins (1971)는 추정된 함수의 모양이 부드럽게 되도록 벌점항(roughness penalty)을 부여하였다. 벌점항은 다른 목적을 위해 사용될 수도 있는데, Ahn과 Baik (2011)은 정규혼합분포의 추정에서 필요 없는 성분을 제거하기 위한 목적으로 사용하여 성과가 있음을 보였다.

Good과 Gaskins의 경우에  $\lambda$ 의 값을 크게 하면 추정된 함수의 모양이 더 부드럽게 되도록 할 수 있다. 마찬가지로 Ahn과 Baik의 경우에는  $\lambda$ 를 크게 하면 더 많은 성분을 없앨 수 있다. 이와 같이  $\lambda$ 의 값에 따라 추정된 확률분포는 달라지는데, 그렇다면 어떤 값의  $\lambda$ 를 사용하는 것이 올바른지에 대한 논의가 필요할 것이다. Eggermont와 LaRiccia (2001)에 따르면 조절상수를 선택하기 위하여 전통적으로 사용되는 방법은 추정된 분포와 실제 분포의 차이( $L_1$  혹은  $L_2$ 의 관점에서)를 최소화하는 원칙을 사용한다. 그러나 이러한 방법은 실제 분포를 알아야 한다는 문제가 있으며, 또 그렇게 하여 찾은 조절상수가 다른 데이터에도 사용된다는 보장이 없으므로 현실적으로 적용하기가 어렵다.

한편 MacKay (1992)는 베이시안의 관점에서 조절상수 혹은 모형상수를 결정하는 방법이 있음을 설명하였다. MacKay의 방법은 앞에서의 전통적인 방법과는 달리 증거체계(evidence framework)를 바탕으로 하여 순전히 데이터의 관점에서 조절상수를 결정하게 한다. 이는 ‘가장 좋은 모형은 데이터를 가장 잘 설명하는 모형’이라는 모형선택(model selection)의 관점에서 조절상수를 선택하는 방법이다.

---

<sup>1</sup> Corresponding author: Associate Professor, College of Business Administration, Kookmin University, 861-1, Jeongneung-dong, Seongbuk-gu, Seoul 136-702, Korea. E-mail: [sahn@kookmin.ac.kr](mailto:sahn@kookmin.ac.kr)

그러므로 실제 분포를 알 필요가 없으며, 또 데이터에 따라 적절한 조절상수의 값을 찾을 수 있다는 장점이 있다.

본 논문에서는 MacKay가 설명한 방법을 Ahn과 Baik이 정규혼합분포의 추정을 위하여 사용한 별점가능도에서 조절상수를 결정하는 데에 적용해 볼 것이다. 이를 위하여 2절에서는 증거체계와 증거(evidence)를 계산하는 데 사용되는 라플라스 근사(Laplace approximation)를 설명하고, 3절에서는 Ahn과 Baik의 모형과 그 모형에서의 별점가능도에서 증거를 계산하는 과정을 설명한다. 그리고 4절에서는 모의실험을 설명하고 5절에서 전체 내용을 요약하도록 한다.

## 2. 증거와 라플라스 근사

MacKay에 따르면, 특정 모형( $H_i$ )이 사실이라고 가정하고 데이터( $D$ )가 주어졌을 때 모형의 모수( $w$ )를 추정하는 경우에 베이즈 규칙에 의하여 모수  $w$ 에 대한 사후확률은 다음과 같다.

$$P(w|D, H_i) = \frac{P(D|w, H_i)P(w|H_i)}{P(D|H_i)}. \quad (2.1)$$

식 (2.1)의 의미는 다음과 같다.

$$\text{사후확률} = \frac{\text{우도} \times \text{사전분포}}{\text{증거}}.$$

정규화 상수인  $P(D|H_i)$ 는  $w$ 를 추정하는 단계에서는 종종 무시되지만, 모형선택의 관점에서는 매우 중요하며, 그것을  $H_i$ 에 대한 증거(evidence)라고 부른다. 이제 데이터가 주어졌을 때 어떠한 모형이 그 데이터에 가장 적합한 것인지를 추론하는 과정을 보자. 각 모형의 사후확률은 다음과 같다.

$$P(H_i|D) \propto P(D|H_i)P(H_i). \quad (2.2)$$

데이터와 관련한 항인  $P(D|H_i)$ 는  $H_i$ 에 대한 증거이며, 이것은 식 (2.1)에서 정규화 상수에 해당하는 것이었다. 다른 모형에 대한 사전 증거( $P(H_i)$ )가 서로 다르지 않다면, 모형에 대한 선호도는 증거를 계산하여 판단할 수 있을 것이다. 증거는 다음의 적분으로 정의되는데, 모수  $w$ 를 적분해서 없애므로 증거를 한계우도(marginal likelihood)라고도 부른다.

$$P(D|H_i) = \int P(D|w, H_i)P(w|H_i)dw. \quad (2.3)$$

식 (2.3)의 적분은 일반적으로 가능하지 않으므로 근사법을 사용하여 계산하며, 그 중에서 많이 사용하는 방법이 라플라스 근사(가우스 근사라고도 부른다)이다. Chickering과 Heckerman (1997)이 사용한 라플라스 근사를 지금까지 사용한 기호로 요약해 보겠다. 식 (2.3)의 피적분함수에 로그를 취한 것을  $g(w)$ 라고 하면 다음과 같다.

$$g(w) \equiv \log [P(D|w, H_i)P(w|H_i)]. \quad (2.4)$$

또 식 (2.4)을 최대화하는  $w$ 의 값을  $\tilde{w}$ 라고 하고,  $g(w)$ 를  $\tilde{w}$ 주위에서 테일러 전개하면 다음과 같다.

$$g(w) \approx g(\tilde{w}) - \frac{1}{2} (w - \tilde{w})^T A (w - \tilde{w}).$$

단,  $A = -\nabla\nabla g(w)|_{w=\tilde{w}}$ . 그러면 식 (2.3)의 피적분함수는 다음과 같다.

$$P(D|w, H_i)P(w|H_i) = \exp [g(w)] \approx \exp \left[ g(\tilde{w}) - \frac{1}{2} (w - \tilde{w})^T A (w - \tilde{w}) \right].$$

그러므로 증거는 다음과 같이 가우스 함수의 특성을 사용하여 근사치로 계산할 수 있다.

$$\begin{aligned}
 P(D|H_i) &= \int P(D|w, H_i)P(w|H_i)dw \\
 &\approx \int \exp\left[g(\tilde{w}) - \frac{1}{2}(w - \tilde{w})^T A(w - \tilde{w})\right]dw \\
 &= \exp[g(\tilde{w})] \int \exp\left[-\frac{1}{2}(w - \tilde{w})^T A(w - \tilde{w})\right]dw \\
 &= \exp[g(\tilde{w})] \cdot \frac{(2\pi)^{\frac{d}{2}}}{|A|^{\frac{1}{2}}} \int \frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}} \exp\left[-\frac{1}{2}(w - \tilde{w})^T A(w - \tilde{w})\right]dw \\
 &= \exp[g(\tilde{w})] \cdot \frac{(2\pi)^{\frac{d}{2}}}{|A|^{\frac{1}{2}}}.
 \end{aligned}$$

위의 근사식에 다시 로그를 취하면 다음과 같다.

$$\begin{aligned}
 \log P(D|H_i) &\approx \log [g(\tilde{w})] + \frac{d}{2} \log(2\pi) - \frac{1}{2} \log |A| \\
 &= \log P(D|\tilde{w}, H_i) + \log P(\tilde{w}|H_i) + \frac{d}{2} \log(2\pi) - \frac{1}{2} \log |A|. \tag{2.5}
 \end{aligned}$$

식 (2.5)가 증거의 로그값을 라플라스 근사를 통하여 계산하는 식이다. 그런데 이 값을 계산하기 위해서는  $A$ (즉,  $g(w)$ 의  $\tilde{w}$ 에서의 헤시안의 부호를 바꾼 것)을 계산해야 하며, 이것을 근사계산하는 방법은 다음절에서 설명하도록 한다.

### 3. 별점가능도와 모형상수의 결정

Ahn과 Baik (2011)은 정규혼합모형에서 성분의 수를 결정하기 위하여 별점가능도를 사용하였는데, 그 모형에서 로그 별점가능도는 다음과 같다.

$$\sum_{i=1}^n \sum_{j=1}^g z_{ij} \log \pi_j \Phi[x_i; \mu_j, \Sigma_j] + \lambda n \sum_{j=1}^g (\alpha - 1) \log \pi_j. \tag{3.1}$$

식 (3.1)의 첫 번째 항이 정규혼합모형에서 잠재변수( $z_{ij}$ )를 포함한 로그 우도함수이며, 두 번째 항이 별점항에 해당한다. 여기서의 별점항은 혼합비율( $\pi_j$ )에 대한 것이며  $g$ 개의 혼합비율 중에서 임의의 혼합비율이 0에 가까워지면 별점항의 값이 커지도록 구성되었다. Ahn과 Baik은 식 (3.1)을 최대화함으로써 불필요한 혼합비율을 0으로 하여 모형에서 탈락시키도록 하였다. 식 (3.1)에서 조율상수는  $\lambda$ 이며 이 값이 커지면 더 많은 혼합비율을 0으로 만드는 효과가 있다. 한편 식 (3.1)에서 또 하나의 상수  $\alpha$ 가 있는데 이것은 0에서 1사이의 값으로 그 역할이  $\lambda$ 와 비슷하며 그 값이 0에 가까울수록 별점항의 값이 커진다. 그러므로  $\alpha$ 도 조율상수라고 보고 그 값을 증거체계를 사용하여 결정할 수도 있을 것이다. 그러나  $\lambda$ 만으로도 별점항의 목적을 달성할 수 있으므로, 본 연구에서는  $\alpha$ 를 0.01로 주어진 것으로 하였다.

이제 식 (3.1)을 증거의 로그값에 해당하는 식 (2.5)에 대입을 해보자. 식 (3.1)은 식 (2.4)에서 정의한  $g(w)$ 와 같은 의미이므로 식 (3.1)을 이용한 증거의 로그값은 다음과 같다.

$$\log(\text{증거}) = \log P(D|H_i)$$

$$\begin{aligned}
&= \log P(D|\tilde{w}, H_i) + \log P(\tilde{w}|H_i) + \frac{d}{2} \log(2\pi) - \frac{1}{2} \log |A| \\
&= \sum_{i=1}^n \sum_{j=1}^g \hat{z}_{ij} \log \tilde{\pi}_j \Phi[x_i; \tilde{\mu}_j, \tilde{\Sigma}_j] + \lambda n \sum_{j=1}^g (\alpha - 1) \log \tilde{\pi}_j + \frac{d}{2} \log(2\pi) - \frac{1}{2} \log |A|. \quad (3.2)
\end{aligned}$$

식 (3.2)에서  $d$ 는 추정된 모수의 수이고, 행렬  $A$ 는 다음을 의미한다.

$$A = -\nabla\nabla \left\{ \sum_{i=1}^n \sum_{j=1}^g z_{ij} \log \pi_j \Phi[x_i; \mu_j, \Sigma_j] + \lambda n \sum_{j=1}^g (\alpha - 1) \log \pi_j \right\}_{|w=\tilde{w}}.$$

행렬  $A$ 는 다차원이 될수록 그리고 성분의 수가 늘어날수록 커지므로 일반적으로 계산이 매우 어렵다. 그러므로 본 논문에서는 큰 표본을 가정하고 Raftery (1995)가 사용한 방법으로 근사계산을 하기로 한다. Raftery에 따르면  $|A| \approx n^d |i|$ 이다.  $i$ 는 단일 관측치에 대한 피셔정보행렬의 기대값을 의미한다. 이 근사식을 식 (3.2)에 대입한 뒤, 피셔정보행렬의 기대값의 행렬식은  $O(1)$ 에 해당하므로 생략하고 정리하면 다음과 같다.

$$\sum_{i=1}^n \sum_{j=1}^g \hat{z}_{ij} \log \tilde{\pi}_j \Phi[x_i; \tilde{\mu}_j, \tilde{\Sigma}_j] + \lambda n \sum_{j=1}^g (\alpha - 1) \log \tilde{\pi}_j + \frac{d}{2} \log(2\pi) - \frac{d}{2} \log(n). \quad (3.3)$$

이제 주어진 데이터에 대하여 서로 다른  $\lambda$ 를 사용하여 모형을 추정된 뒤에 식 (3.3)을 사용하여 각 모형에 대한 증거의 로그값을 계산할 수 있다. 이 값을 가장 크게 하는  $\lambda$ 가 적절한 값이며 그때의 모형이 주어진 데이터에 가장 잘 적합 되는 모형이라 할 수 있다.

#### 4. 모의 실험

앞 절에서 설명한 방법으로 최적의 조율상수를 결정하는 것을 보이기 위하여 두 개의 분포에 대하여 모의실험을 해 보았다.

##### 4.1. 일변량 정규혼합분포

이 분포는 Marron과 Wand (1992)에 제시된 분포 중의 하나로서 ‘Skewed unimodal’이라는 이름을 가지고 있으며 다음과 같이 정의된다.

$$\frac{1}{5}N(0, 1) + \frac{1}{5}N\left(\frac{1}{2}, \left(\frac{2}{3}\right)^2\right) + \frac{3}{5}N\left(\frac{13}{12}, \left(\frac{5}{9}\right)^2\right).$$

분포의 모양과 성분의 구조를 그리면 Figure 1과 같다.

Figure 1에서 보듯이 분포는 왼쪽으로 늘어져 있으며 혼합비율이 큰 성분 한 개와 작은 성분 두개로 이루어져 있다. 분포는 단봉분포이며 그림으로 볼 때 성분은 잘 분리되어 있지 않다. 이 분포로부터 크기가 3000인 표본을 생성하여 서로 다른  $\lambda$ 의 값에 대하여 별점가능도를 최대화하여 분포를 추정하고 각각의 경우에 식 (3.3)의 값을 계산한 결과가 Figure 2이다.

Figure 2에서  $x$ 축은 조율상수( $\lambda$ )의 값이며 값은 0.00001에서 0.001까지 증가하고 있으며,  $y$ 축은 우도 혹은 증거의 값으로서 그림에서 실선은 증거를 점선은 우도를 나타내고 있다. Figure 2에서 보듯이 조율상수의 값이 증가함에 따라 우도는 전반적으로 감소하고 있지만 증거의 값은 증가와 감소를 반복하고 있으며 특정한  $\lambda(0.00027)$ 의 값에서 최대값을 가진다. Figure 2에서 증거를 나타내는 선이 급격하

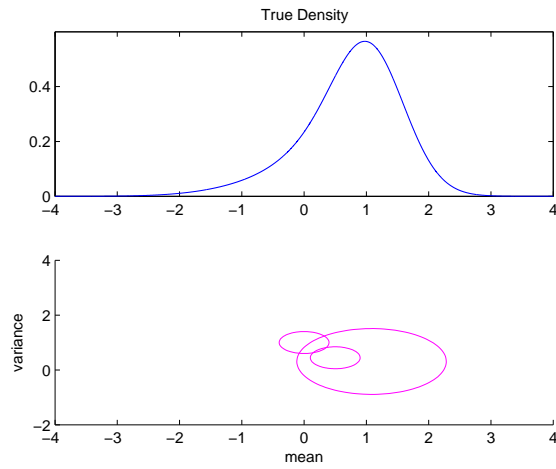


Figure 1: True distribution

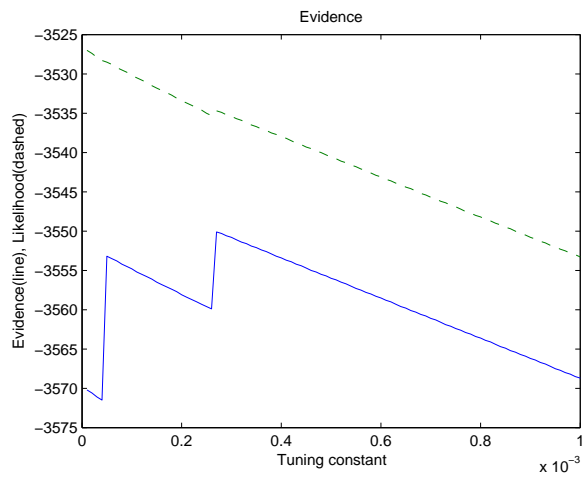


Figure 2: Evidence and likelihood

계 상승하는 곳이 두 곳이 있는데, 그 이유는 해당하는 조율상수의 값에서 추정된 분포의 성분의 수가 감소하기 때문이다. 이 경우는  $\lambda = 0.00005$ 에서 성분의 수가 5에서 3으로 감소하고,  $\lambda = 0.00027$ 에서 성분의 수가 3에서 2으로 감소한다.

증거의 값이 상승하는  $\lambda$ 의 주위에서 우도와 증거의 값, 그리고 추정된 분포의 성분의 수를 나타낸 것이 Table 1이다. 지금까지의 논의에 따르면 증거를 최대화하는 조율상수의 값이 최적의 값이므로, 그 값( $\lambda = 0.00027$ )을 사용하여 추정된 모형이 최적모형이라고 할 수 있을 것이다. Table 1에서 볼 수 있듯이 최적모형은 성분이 두개인 모형으로 그 모형을 나타낸 것이 Figure 3이다.

그런데, 증거를 최대화하는 조율상수의 값으로 추정된 모형은 실제 모형의 성분의 수와 일치하지 않는다. Figure 2와 Table 1을 보면 추정된 모형 중에 성분의 수가 3개인 경우도 다수가 있음을 알 수 있고, 그 경우에 증거를 최대화하는 모형은  $\lambda = 0.00005$ 일 때이다. 만약 우리가 성분이 3개가 있는 모

Table 1: Selected models according to various tuning constants

tuning constant	evidence	likelihood	number of components
0.00004	-3571.5	-3528.3	5
0.00005	-3553.2	-3528.5	3
⋮	⋮	⋮	⋮
0.00026	-3559.9	-3535.3	3
0.00027	-3550.1	-3534.7	2
0.00028	-3550.3	-3534.9	2

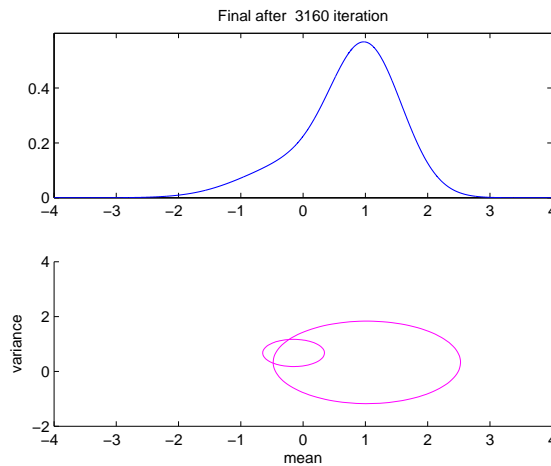


Figure 3: Model selected by maximizing evidence

형이 옳다는 정보가 있었다면  $\lambda = 0.00005$ 를 사용하여 추정된 모형을 선택하는 것이 옳을 것이다. 그렇지만 사전에 그런 정보가 없는 경우에는 실제 분포가 몇 개의 성분을 가지고 있는지 알 수 없으므로  $\lambda = 0.00027$ 을 사용하여 얻은 모형을 선택하는 것이 옳을 것이다. 불행하게도 이 경우는 실제 분포와 성분의 수가 같지 않지만 ‘Skewed unimodal’ 분포와 같이 성분이 잘 분리되지 않는 분포에 대하여 우리가 내릴 수 있는 최선의 결과이다. 그럼에도 불구하고, Figure 1과 Figure 3을 비교했을 때 두 분포의 성분의 수는 다르지만 전체적인 모양은 크게 다르지 않음을 알 수 있다. 이것은 비슷한 모형이 있을 때 단순한 모형을 선호한다는 Occam’s razor의 원칙을 따른다는 점에서 긍정적이다.

#### 4.2. 이변량 정규혼합분포

두번째로 사용한 분포는 Xu와 Jordan (1996)에서 사용한 분포로서 다음과 같이 정의된다.

$$0.345N\left(\begin{bmatrix} 2.5 \\ 1.0 \end{bmatrix}, \begin{bmatrix} 4.0 & -0.9 \\ -0.9 & 0.3 \end{bmatrix}\right) + 0.32N\left(\begin{bmatrix} -1.8 \\ 2.2 \end{bmatrix}, \begin{bmatrix} 3.50 & 0.75 \\ 0.75 & 0.30 \end{bmatrix}\right) + 0.335N\left(\begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix}, \begin{bmatrix} 2.0 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}\right).$$

이 분포는 앞의 ‘Skewed unimodal’ 분포와는 달리 성분이 상대적으로 잘 분리되어 있으며, 모양과 성분의 구조를 그리면 Figure 4와 같다.

Figure 4에서 분포는 세 개의 봉을 가지고 있고 성분은 잘 분리되어 있다. 이 분포로부터 크기가 3000인 표본을 생성하여 서로 다른  $\lambda$  값에 대하여 별점가능도를 최대화하여 분포를 추정하고 각각의 경우에 식 (3.3)의 값을 계산한 결과가 Figure 5이다.

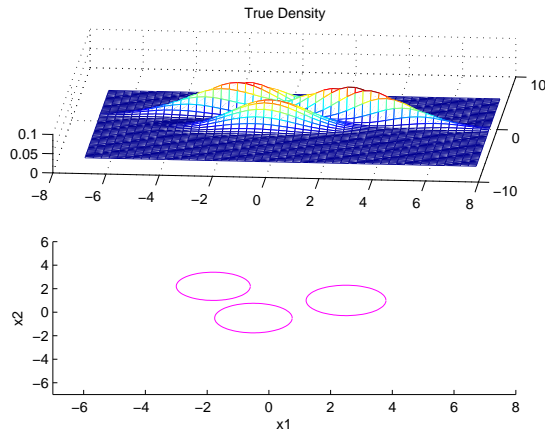


Figure 4: True distribution

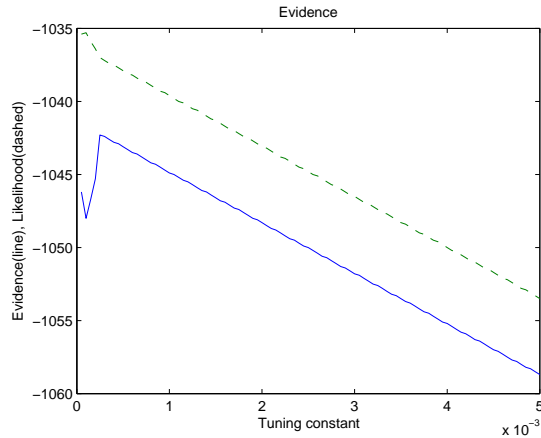


Figure 5: Evidence and likelihood

Figure 5에서 조율상수( $\lambda$ )의 값은 0.00005에서 0.005까지 증가하고 있으며, ‘Skewed unimodal’의 경우와 비슷하게 조율상수의 값이 증가함에 따라 우도는 전반적으로 감소하고 있지만 증거의 값은 증가와 감소를 반복하고 있으며 특정한  $\lambda(0.00025)$ 의 값에서 최대값을 가진다. 이 경우는  $\lambda = 0.00025$ 에서 성분의 수가 5에서 3으로 감소한다.

$\lambda = 0.00025$ 의 주위에서 우도와 증거의 값, 그리고 추정된 분포의 성분의 수를 나타낸 것이 Table 2이다. Table 2에서 볼 수 있듯이 최적모형은 성분이 세 개인 모형으로 그 모형을 나타낸 것이 Figure 6이다.

‘Skewed unimodal’의 경우와는 달리, Figure 4와 Figure 6을 보면 추정된 모형과 실제모형이 성분의 수도 같고 형태도 유사하므로 논의의 여지가 없이 잘 추정된 모형이라고 할 수 있을 것이다. 이런 결과를 낳은 가장 큰 이유는 실제 모형이 성분이 다른 성분과 잘 분리되어 있고 개별 성분도 정규분포이기 때문이다. 물론 앞의 ‘Skewed unimodal’의 경우처럼 실제 분포가 성분이 분리가 잘 되지 않는다면 추정된 분포도 실제 분포와 다르게 될 수 있음을 예상할 수 있다.

Table 2: Selected models according to various tuning constants

tuning constant	evidence	likelihood	number of components
0.00005	-1046.2	-1035.4	6
0.0001	-1048	-1035.3	7
0.00015	-1046.7	-1035.9	6
0.0002	-1045.3	-1036.4	5
0.00025	-1042.3	-1037	3
0.0003	-1042.4	-1037.2	3

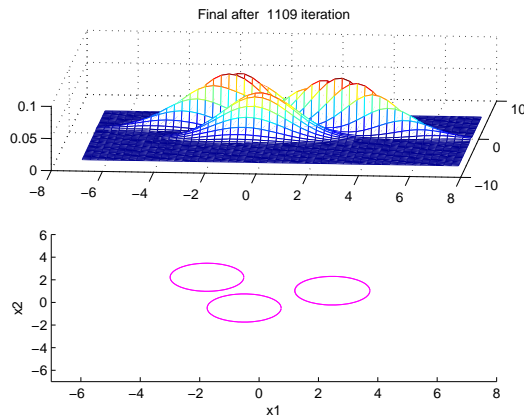


Figure 6: Model selected by maximizing evidence

## 5. 요약 및 맺음말

지금까지 증거체계(evidence framework)를 바탕으로 하여 데이터의 관점에서 조절상수를 결정하는 방법을 설명하고, 그것을 정규혼합분포의 추정을 위하여 사용한 벌점가능도에서 조절상수를 결정하는데에 적용해 보았다. 증거(evidence)를 계산하기 위하여 라플라스 근사(Laplace approximation)를 사용하였고, 3절에서 두 개의 분포에 대하여 그 방법을 적용해 본 결과 의미 있는 결과를 얻은 것으로 판단된다. 즉, 일반적으로 벌점가능도에서 사용되는 조율상수를 직관을 통해서 결정하는 것이 아니라 체계적으로 적용할 수 있음을 확인하였다. 물론 많은 조율상수의 값을 사용하여 각각의 경우에 벌점가능도를 최적화하는 과정을 거치므로 시간이 많이 소요되는 단점은 있지만, 그것은 최적의 조율상수를 찾기 위한 비용이라고 봐야 할 것이다.

Figure 2와 Figure 5에서 보듯이 증거의 곡선은 그 모양이 부드럽지 못하고 급격하게 변하는 지점이 있다. 이것은 식 (3.2)를 계산하기 위하여 헤시안행렬을 근사계산함으로써 나타난 결과이다. 만일 헤시안행렬을 직접계산하였다면 그 선이 부드럽게 될 것으로 기대한다. 그렇게 되면 성분의 수가 줄어드는 위치에서 증거의 값이 급격하게 상승하는 형태의 곡선이 아니고 그 지점 주위에서 상승하고 하락하는 모양이 나타날 것이라고 예상할 수 있다. 그렇게 되면 조금 더 정확한 조율상수의 값을 찾을 수 있을 것으로 기대한다.

## References

Ahn, S. and Baik, S. W. (2011). Estimating parameters in multivariate normal mixtures, *Korean Communications in Statistics*, **18**, 357–366.



- Chickering, D. M. and Heckerman, D. (1997). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables, *Machine Learning*, **29**, 181–212.
- Eggermont, P. P. B. and LaRiccia, V. N. (2001). *Maximum Penalized Likelihood Estimation*, Springer.
- Good, I. J. (1971). A nonparametric roughness penalty for probability densities, *Nature*, **229**, 29–30.
- Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities, *Biometrika*, **58**, 255–277.
- MacKay, D. J. C. (1992). Bayesian interpolation, *Neural Computation*, **4**, 415–447.
- Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error, *Annals of Statistics*, **20**, 712–736.
- Raftery, A. (1995). Bayesian model selection in social research. In Marsden, P. (Ed.), *Sociological Methodology*, Blackwells, Cambridge, MA.
- Xu, L. and Jordan, M. I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures, *Neural Computation*, **8**, 129–151.

2012년 3월 4일 접수; 2012년 4월 8일 수정; 2012년 4월 27일 채택