다변량 자료분석

Choosing Starting Values
for the EM algorithm
in Multivariate Gaussian Mixture Models

김미성
통계학과

논문 소개

제목	Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models
저자	Christophe Biernacki, Gilles Celeux, Gerard Govaert
학술지	Computational Statistics & Data Analysis Vol.41 (2003), p.561-575
등재	SCIE, SCOPUS

목차

SECTION 1
도입

SECTION 2
GMM 추정

SECTION 3
실험 계획

SECTION 4
실험 결과

SECTION 5
결론

목 차

SECTION 1
도입

SECTION 2
GMM 추정

SECTION 3
실험 계획

SECTION 4
실험 결과

SECTION 5
결론

Mixture Model & EM algorithm

- EM 알고리즘은 초기값에 매우 민감하다.
- 특히 다변량 자료를 이용하는 경우 초기값의 영향이 매우 크다.

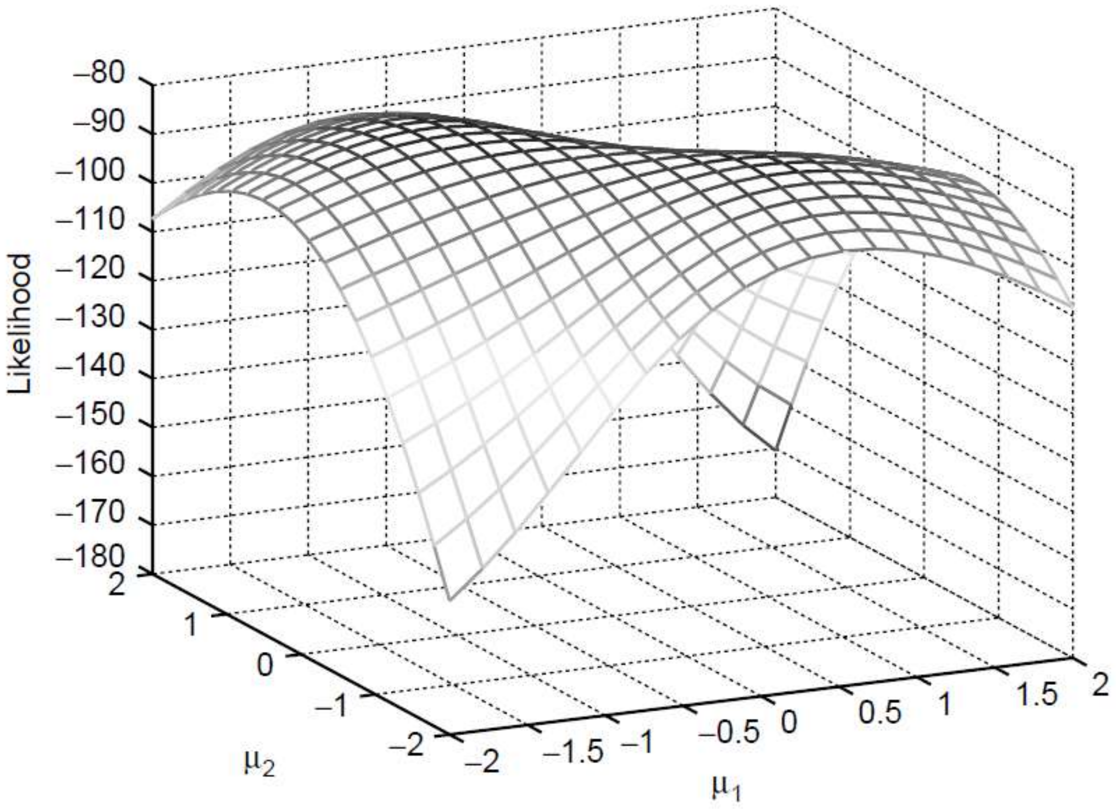


Fig. 1. A two-mode likelihood surface.

Mixture Model & EM algorithm

- EM 알고리즘은 초기값에 매우 민감하다.
- 특히 다변량 자료를 이용하는 경우 초기값의 영향이 매우 크다.
- Mixture model 추정에서 likelihood의 local maximum을 선택할 위험이 있다.

  ex) cluster의 개수를 잘못 추정할 수 있음

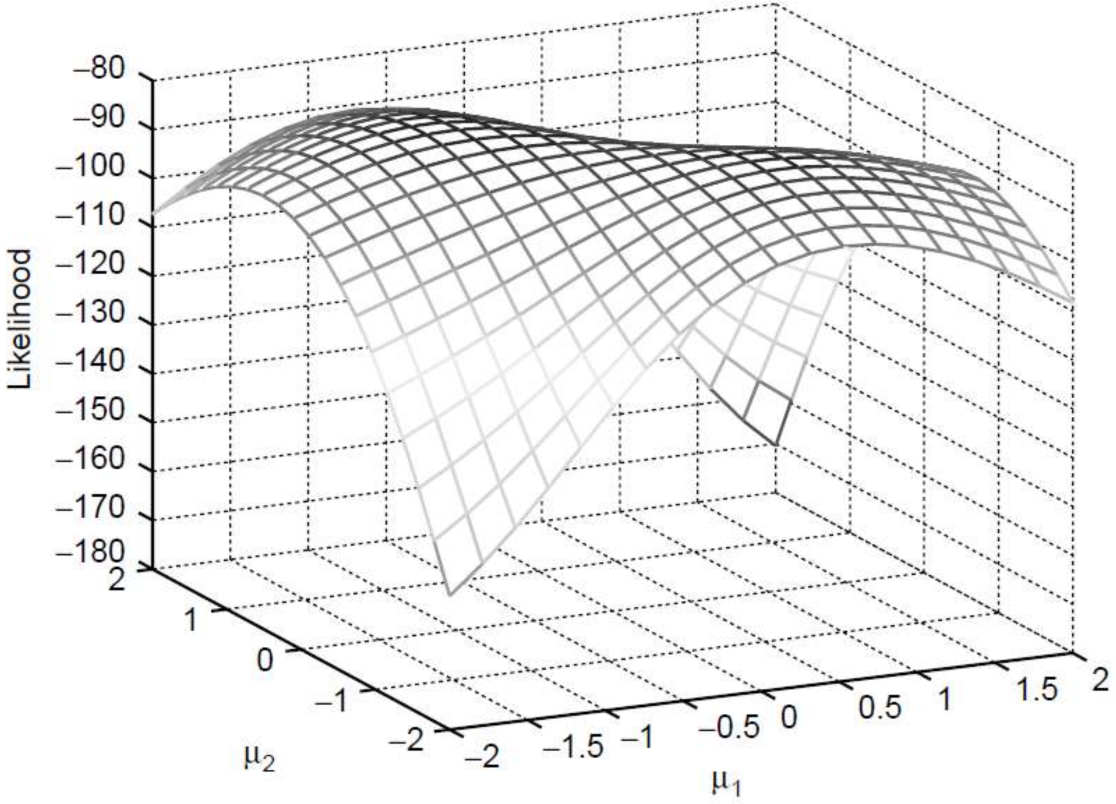


Fig. 1. A two-mode likelihood surface.

Mixture Model & EM algorithm

- EM 알고리즘은 초기값에 매우 민감하다.
- 특히 다변량 자료를 이용하는 경우 초기값의 영향이 매우 크다.
- Mixture model 추정에서 likelihood의 local maximum을 선택할 위험이 있다.
  ex) cluster의 개수를 잘못 추정할 수 있음

→ 제한된 반복횟수 안에서 적절한 초기값을 선택하는 몇 가지 방법을 제안 & 비교

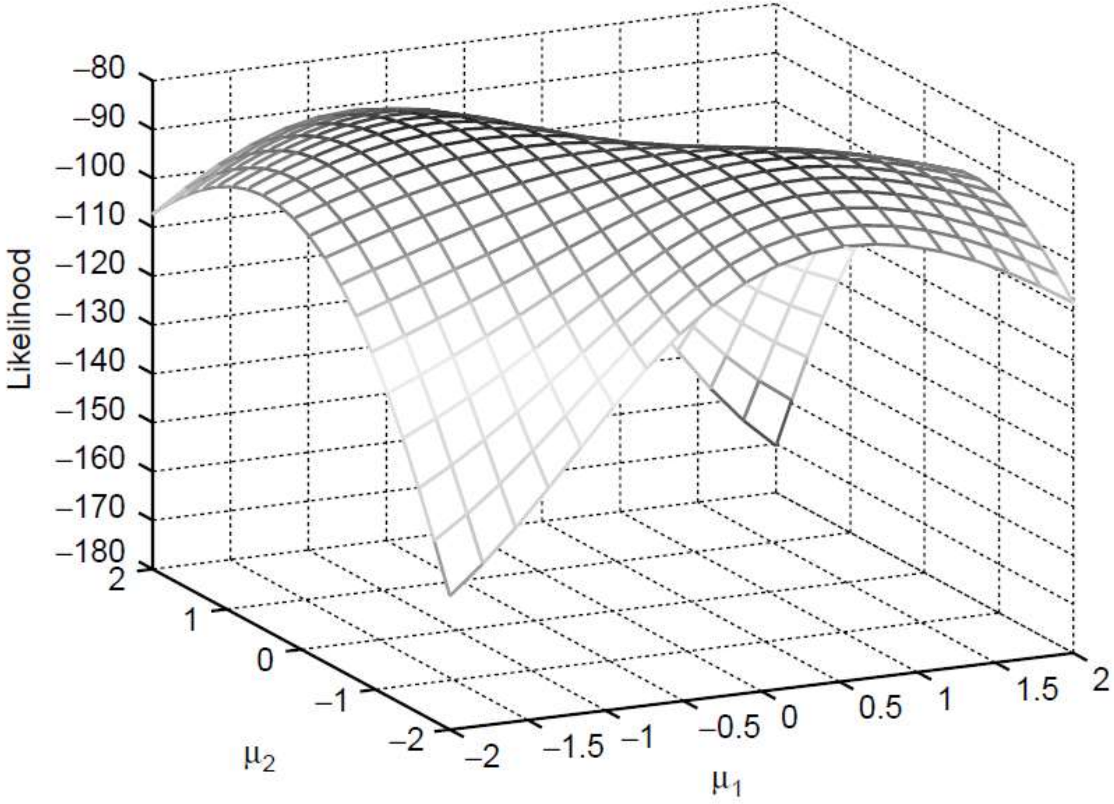


Fig. 1. A two-mode likelihood surface.

목 차

SECTION 1
도입

SECTION 2
GMM 추정

SECTION 3
실험 계획

SECTION 4
실험 결과

SECTION 5
결론

Gaussian Mixture Model

$\mathbf{x}_1, \dots, \mathbf{x}_n$ in $\mathbf{R}^d$ from a random vector with density

$$f(\mathbf{x}) = \sum_{k=1}^{K} p_k \phi(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

p_k: mixing proportions, $k = 1, \dots, K$

$\phi(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$: multivariate Normal density with mean vector $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$

ML Estimateion using the EM algorithm

$\theta = (p_1, \ldots, p_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots \boldsymbol{\Sigma}_K)$ are estimated by maximizing the log-likelihood

$$l(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \ln\left[\sum_{k=1}^{K} p_k \phi(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right]$$

ML Estimateion using the EM algorithm

$\theta = (p_1, \ldots, p_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots \boldsymbol{\Sigma}_K)$ are estimated by maximizing the log-likelihood

$$l(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \ln\left[\sum_{k=1}^{K} p_k \phi(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right]$$

Starting from an initial parameter θ^0, iterate the following E-M steps:

E step Compute $\hat{p}_k(\mathbf{x}_i)$ which are the current conditional probabilities that $\mathbf{x}_i$ from the kth cluster

M step Update the ML estimates $(\hat{p}_k, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)$ using $\hat{p}_k(\mathbf{x}_i)$ as conditional mixing weights

Methods for Choosing Starting Values

1. Short runs of EM
2. CEM (Classification EM)
3. SEM (Stochastic EM)

CEM algorithm (Classification EM)

E step Compute $\hat{p}_k(\mathbf{x}_i)$ which are the current conditional probabilities that $\mathbf{x}_i$ from the kth cluster

C step Design a partition $P = (P_1, \dots, P_K)$ of $\mathbf{x}_1, \dots, \mathbf{x}_n$ by assigning $\mathbf{x}_i$ to the cluster maximizing $\hat{p}_k(\mathbf{x}_i)$

M step Compute the ML estimates $(\hat{p}_k, \hat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k)$ using the cluster P_k as sub-sample of the kth cluster

CEM algorithm (Classification EM)

E step Compute $\hat{p}_k(\mathbf{x}_i)$ which are the current conditional probabilities that $\mathbf{x}_i$ from the kth cluster

C step Design a partition $P = (P_1, \dots, P_K)$ of $\mathbf{x}_1, \dots, \mathbf{x}_n$ by assigning $\mathbf{x}_i$ to the cluster maximizing $\hat{p}_k(\mathbf{x}_i)$

M step Compute the ML estimates $(\hat{p}_k, \hat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k)$ using the cluster P_k as sub-sample of the kth cluster

It is a *K-means*-like algorithm, converges in a finite number of iterations.

CEM algorithm (Classification EM)

E step Compute $\hat{p}_k(\mathbf{x}_i)$ which are the current conditional probabilities that $\mathbf{x}_i$ from the kth cluster

C step Design a partition $P = (P_1, \dots, P_K)$ of $\mathbf{x}_1, \dots, \mathbf{x}_n$ by assigning $\mathbf{x}_i$ to the cluster maximizing $\hat{p}_k(\mathbf{x}_i)$

M step Compute the ML estimates $(\hat{p}_k, \hat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k)$ using the cluster P_k as sub-sample of the kth cluster

It is a *K-means*-like algorithm, converges in a finite number of iterations.

When z_i is the missing cluster label of $\mathbf{x}_i$, it maximize the complete data log-likelihood

$$Cl(\theta | z_1, \dots, z_n, \mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{k=1}^{K} \sum_{\{i:\, z_i = k\}} \ln[p_k \phi(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]$$

SEM algorithm (Stochastic EM)

E step Compute $\hat{p}_k(\mathbf{x}_i)$ which are the current conditional probabilities that $\mathbf{x}_i$ from the kth cluster

S step Design a partition $P = (P_1, \dots, P_K)$ of $\mathbf{x}_1, \dots, \mathbf{x}_n$ by assigning $\mathbf{x}_i$ at random to one of the clusters according to the Multinomial distribution with parameter $\hat{p}_k(\mathbf{x}_i)$

M step Compute the ML estimates $(\hat{p}_k, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)$ using the cluster P_k as sub-sample of the kth cluster

SEM algorithm (Stochastic EM)

E step Compute $\hat{p}_k(\mathbf{x}_i)$ which are the current conditional probabilities that $\mathbf{x}_i$ from the kth cluster

S step Design a partition $P = (P_1, \dots, P_K)$ of $\mathbf{x}_1, \dots, \mathbf{x}_n$ by assigning $\mathbf{x}_i$ at random to one of the clusters according to the Multinomial distribution with parameter $\hat{p}_k(\mathbf{x}_i)$

M step Compute the ML estimates $(\hat{p}_k, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)$ using the cluster P_k as sub-sample of the kth cluster

SEM algorithm (Stochastic EM)

It is same as the Monte Carlo EM algorithm with a single replication, and generates Markov chain.

Parameter estimate from a SEM sequence $(\theta^r)_{r=1,\dots,R}$:

1. The mean value of the sequence after a burn-in period b

$$\hat{\theta} = \sum_{r=b+1}^{R} \theta^r / (R - b)$$

2. The value leading to the highest likelihood in the sequence

목 차

SECTION 1
도입

SECTION 2
GMM 추정

SECTION 3
실험 계획

SECTION 4
실험 결과

SECTION 5
결론

Methods for Choosing Starting Values

1. Random initialization
2. Using short runs of EM
3. Using the CEM algorithm
4. Using the SEM algorithm

Methods for Choosing Starting Values

1. Random initialization	**EM**
2. Using short runs of EM	**em-EM**
3. Using the CEM algorithm	**CEM-EM**
4. Using the SEM algorithm	**SEM-EM**

Experimental Strategies

1. EM, CEM, SEM 알고리즘을 1번 반복할 때 계산 시간에 큰 차이가 없다.

→ 따라서 총 반복수를 동일하게 설정하여 총 계산 시간을 제한

2. 동일한 총 반복수 하에 각 알고리즘을 10번 반복하는 방법을 포함하여 총 8개의 방법을 비교

(iteration) (repetition)

EX. Total Number of Iterations = 1000

1EM 1000 iterations for EM

10EM 10 repetitions of 100 iterations for each EM run

1em-EM 500 iterations for em and 500 iterations for EM

10em-EM 10 repetitions of 50 iterations for em and 50 iterations for EM

1CEM-EM 500 iterations for CEM and 500 iterations for EM

10CEM-EM 10 repetitions of 50 iterations for CEM and 50 iterations for EM

SEMmean-EM & **SEMmax-EM** 500 iterations for SEM and 500 iterations for EM

목 차

SECTION 1
도입

SECTION 2
GMM 추정

SECTION 3
실험 계획

SECTION 4
실험 결과

SECTION 5
결론

Simulation Data

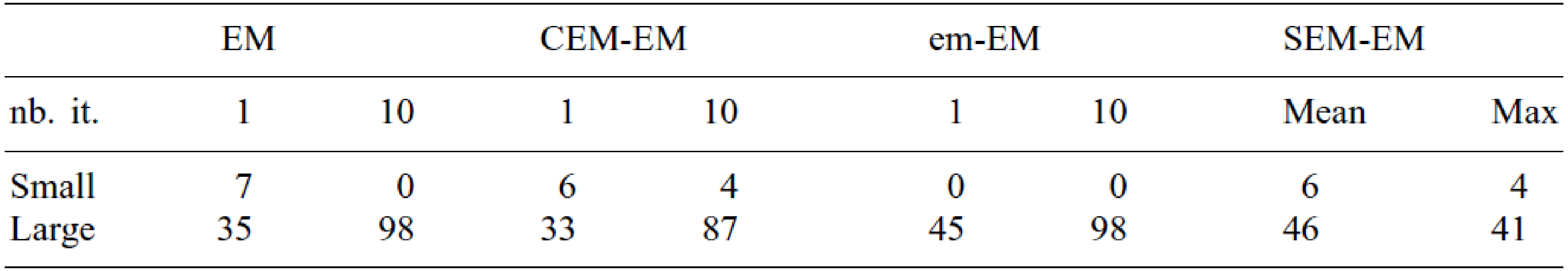
Data set	P1	P2	P3	P1 noise, P2 noise, P3 noise
Dimension	$d = 2$	$d = 2$	$d = 2$	Add noise to P1, P2, P3 from $Uniform$ $[-0.8, 0.8] \times [-0.8, 0.8]$ with proportion 0.2
Number of Clusters	$K = 2$	$K = 2$	$K = 4$	
Mixing Proportion	$p_1 = p_2 = 0.5$	$p_1 = 0.7,\ p_2 = 0.3$	$p_1 = p_2 = p_3 = p_4 = 0.5$	
Parameters	$\mu_1 = (0, 0)'$ $\mu_2 = (2.5, 0)'$ $\mathrm{diag}(\mathbf{\Sigma}_1) = \left(3, \frac{1}{3}\right)$ $\mathrm{diag}(\mathbf{\Sigma}_2) = \left(\frac{1}{3}, 3\right)$	$\mu_1 = \mu_2 = (0, 0)'$ $\mathrm{diag}(\mathbf{\Sigma}_1) = \left(3, \frac{1}{3}\right)$ $\mathrm{diag}(\mathbf{\Sigma}_2) = \left(\frac{1}{3}, 3\right)$	$\mu_1 = (0, -2)',\ \mu_2 = (2.0)'$ $\mu_3 = (0, 2)',\ \mu_4 = (-2, 0)'$ $\mathrm{diag}(\mathbf{\Sigma}_1) = \mathrm{diag}(\mathbf{\Sigma}_2) = \left(3, \frac{1}{3}\right)$ $\mathrm{diag}(\mathbf{\Sigma}_3) = \mathrm{diag}(\mathbf{\Sigma}_4) = \left(\frac{1}{3}, 3\right)$	

Simulation Data

총 반복수 = (60, 120, 240, 480, ···, 15360)

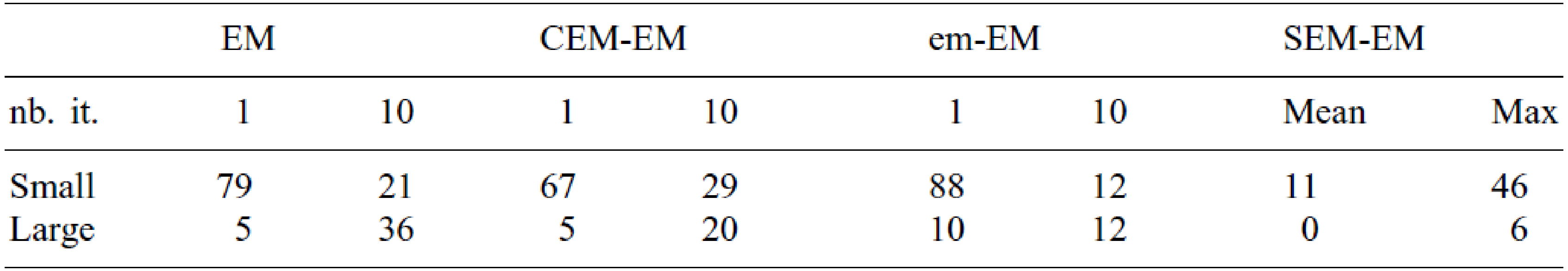
※ 총 반복수 960을 기준으로 Small / Large로 구분

Table1
Small v.s. Large



	EM		CEM-EM		em-EM		SEM-EM	
nb. it.	1	10	1	10	1	10	Mean	Max
Small	7	0	6	4	0	0	6	4
Large	35	98	33	87	45	98	46	41

Table2
Single run
v.s.
Repeated runs



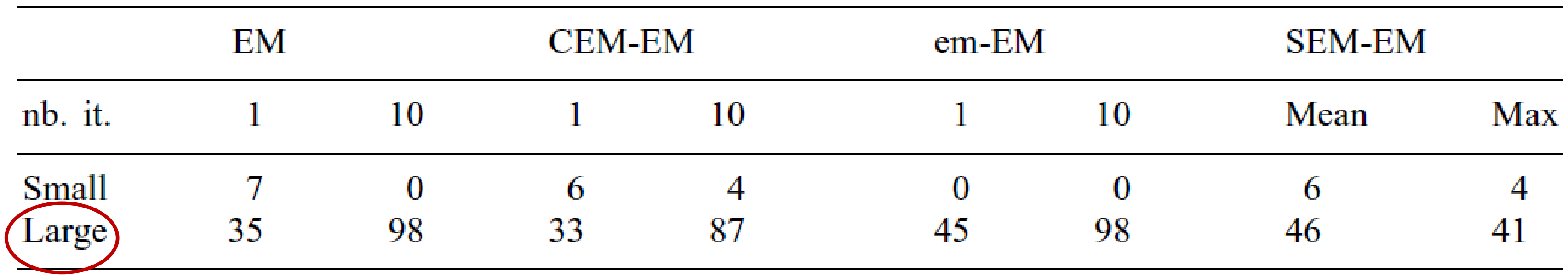
	EM		CEM-EM		em-EM		SEM-EM	
nb. it.	1	10	1	10	1	10	Mean	Max
Small	79	21	67	29	88	12	11	46
Large	5	36	5	20	10	12	0	6

Simulation Data

총 반복수 = (60, 120, 240, 480, ···, 15360)

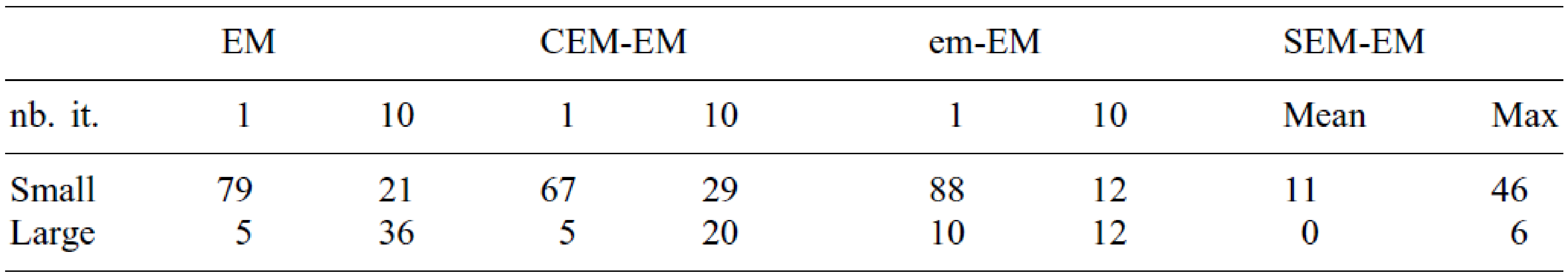
※ 총 반복수 960을 기준으로 Small / Large로 구분

Table1
Small v.s. Large



	EM		CEM-EM		em-EM		SEM-EM	
nb. it.	1	10	1	10	1	10	Mean	Max
Small	7	0	6	4	0	0	6	4
Large	35	98	33	87	45	98	46	41

Table2
Single run
v.s.
Repeated runs



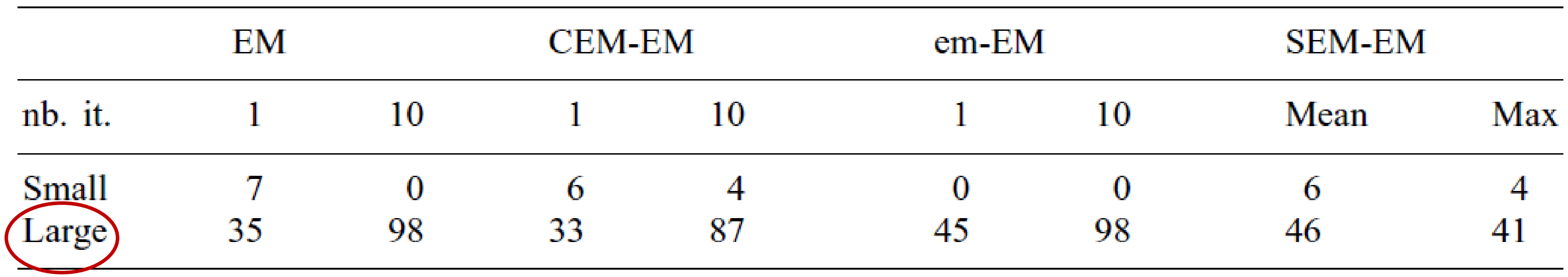
	EM		CEM-EM		em-EM		SEM-EM	
nb. it.	1	10	1	10	1	10	Mean	Max
Small	79	21	67	29	88	12	11	46
Large	5	36	5	20	10	12	0	6

Simulation Data

총 반복수 = (60, 120, 240, 480, ···, 15360)

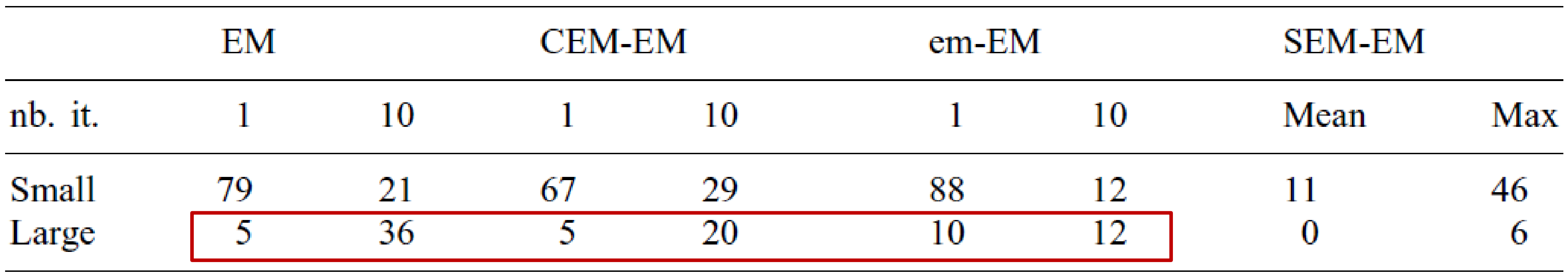
※ 총 반복수 960을 기준으로 Small / Large로 구분

Table1
Small v.s. Large



	EM		CEM-EM		em-EM		SEM-EM	
nb. it.	1	10	1	10	1	10	Mean	Max
Small	7	0	6	4	0	0	6	4
Large	35	98	33	87	45	98	46	41

Table2
Single run
v.s.
Repeated runs



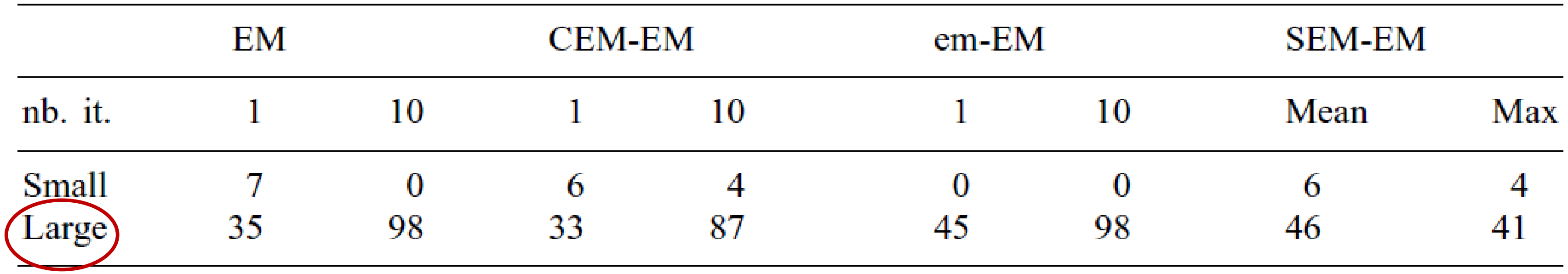
	EM		CEM-EM		em-EM		SEM-EM	
nb. it.	1	10	1	10	1	10	Mean	Max
Small	79	21	67	29	88	12	11	46
Large	5	36	5	20	10	12	0	6

Simulation Data

총 반복수 = (60, 120, 240, 480, ⋯, 15360)

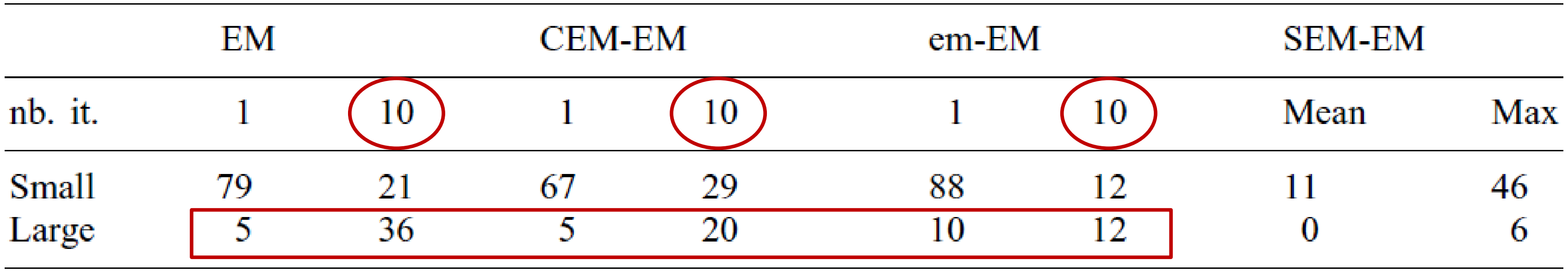
※ 총 반복수 960을 기준으로 Small / Large로 구분

Table1
Small v.s. Large



	EM		CEM-EM		em-EM		SEM-EM	
nb. it.	1	10	1	10	1	10	Mean	Max
Small	7	0	6	4	0	0	6	4
Large	35	98	33	87	45	98	46	41

Table2
Single run
v.s.
Repeated runs



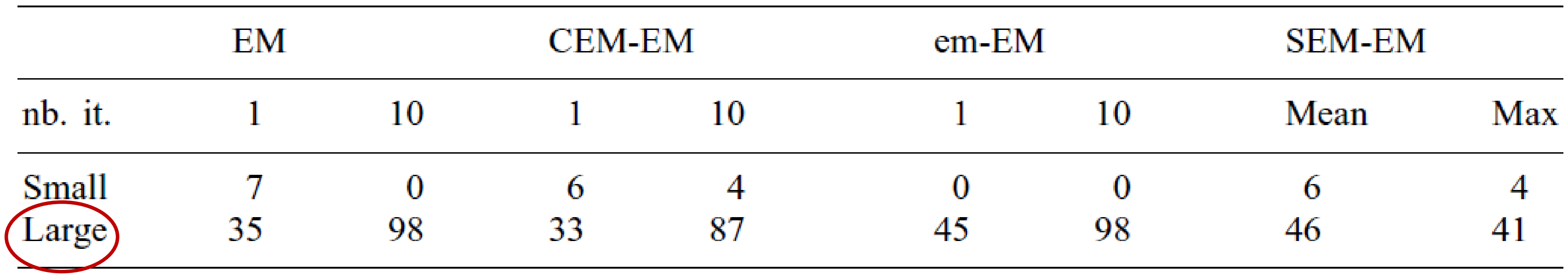
	EM		CEM-EM		em-EM		SEM-EM	
nb. it.	1	10	1	10	1	10	Mean	Max
Small	79	21	67	29	88	12	11	46
Large	5	36	5	20	10	12	0	6

Simulation Data

총 반복수 = (60, 120, 240, 480, ⋯, 15360)

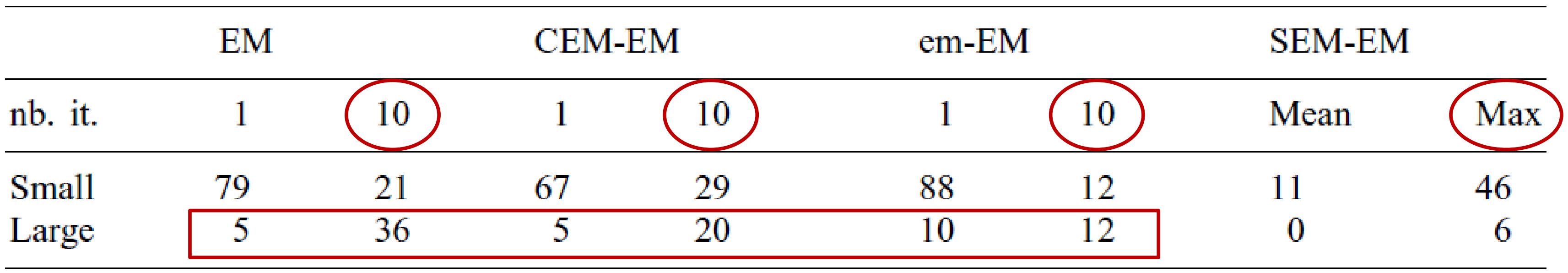
※ 총 반복수 960을 기준으로 Small / Large로 구분

Table1
Small v.s. Large



	EM		CEM-EM		em-EM		SEM-EM	
nb. it.	1	10	1	10	1	10	Mean	Max
Small	7	0	6	4	0	0	6	4
Large	35	98	33	87	45	98	46	41

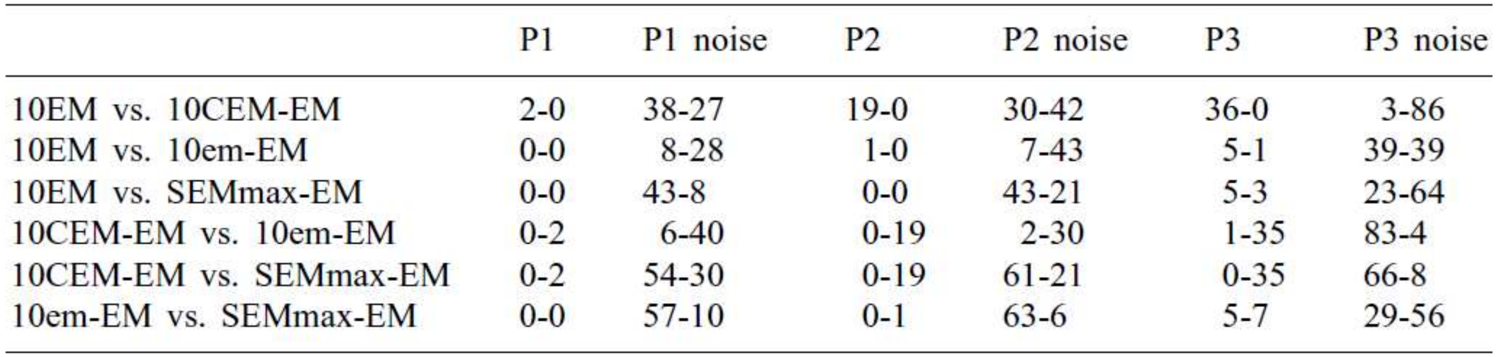
Table2
Single run
v.s.
Repeated runs



	EM		CEM-EM		em-EM		SEM-EM	
nb. it.	1	10	1	10	1	10	Mean	Max
Small	79	21	67	29	88	12	11	46
Large	5	36	5	20	10	12	0	6

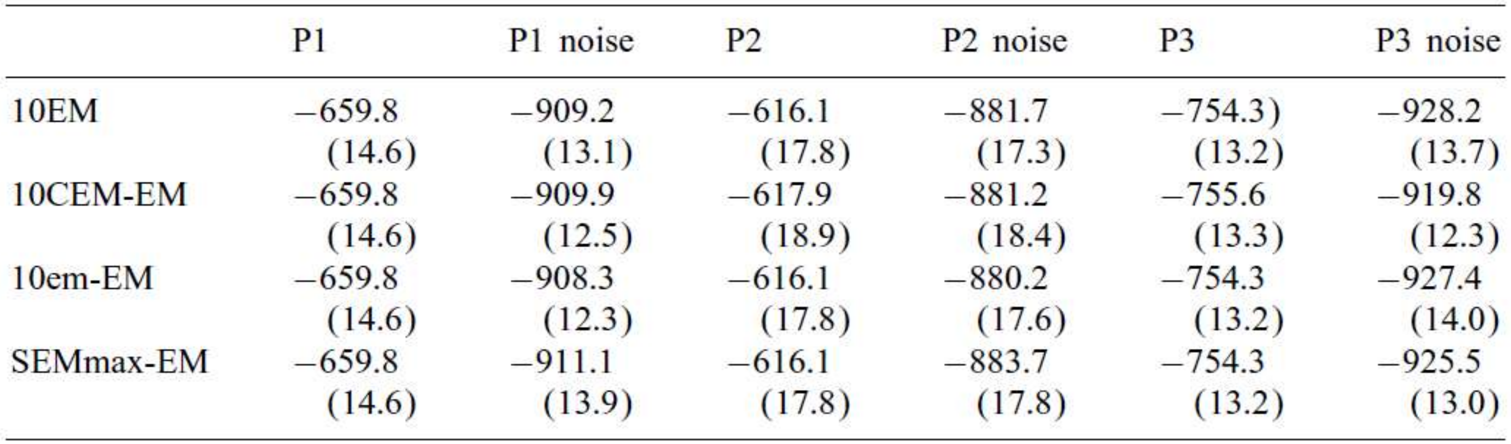
Simulation Data

Table3
Comparison by pairs with a large number of iterations



	P1	P1 noise	P2	P2 noise	P3	P3 noise
10EM vs. 10CEM-EM	2-0	38-27	19-0	30-42	36-0	3-86
10EM vs. 10em-EM	0-0	8-28	1-0	7-43	5-1	39-39
10EM vs. SEMmax-EM	0-0	43-8	0-0	43-21	5-3	23-64
10CEM-EM vs. 10em-EM	0-2	6-40	0-19	2-30	1-35	83-4
10CEM-EM vs. SEMmax-EM	0-2	54-30	0-19	61-21	0-35	66-8
10em-EM vs. SEMmax-EM	0-0	57-10	0-1	63-6	5-7	29-56

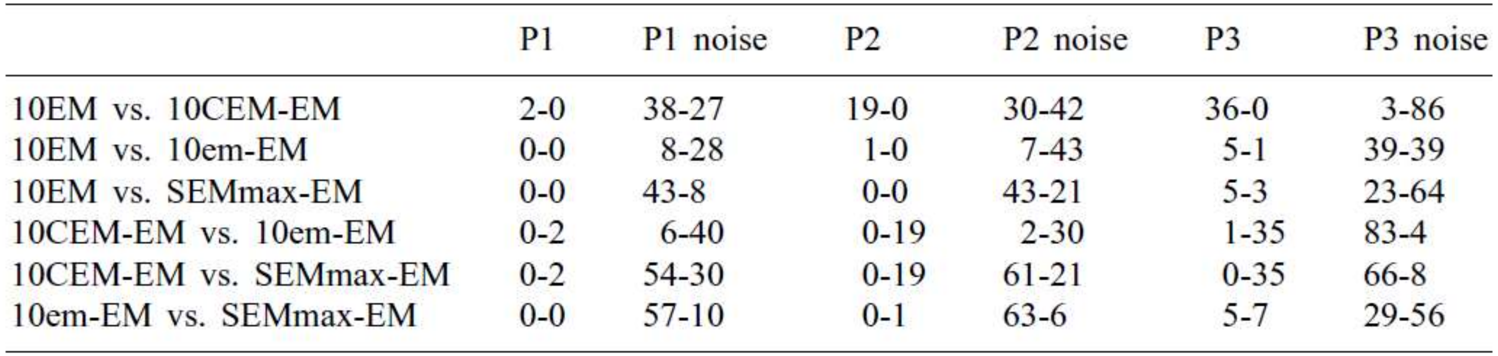
Table4
Means & standard deviations of maximum log-likelihood



	P1	P1 noise	P2	P2 noise	P3	P3 noise
10EM	−659.8 (14.6)	−909.2 (13.1)	−616.1 (17.8)	−881.7 (17.3)	−754.3) (13.2)	−928.2 (13.7)
10CEM-EM	−659.8 (14.6)	−909.9 (12.5)	−617.9 (18.9)	−881.2 (18.4)	−755.6 (13.3)	−919.8 (12.3)
10em-EM	−659.8 (14.6)	−908.3 (12.3)	−616.1 (17.8)	−880.2 (17.6)	−754.3 (13.2)	−927.4 (14.0)
SEMmax-EM	−659.8 (14.6)	−911.1 (13.9)	−616.1 (17.8)	−883.7 (17.8)	−754.3 (13.2)	−925.5 (13.0)

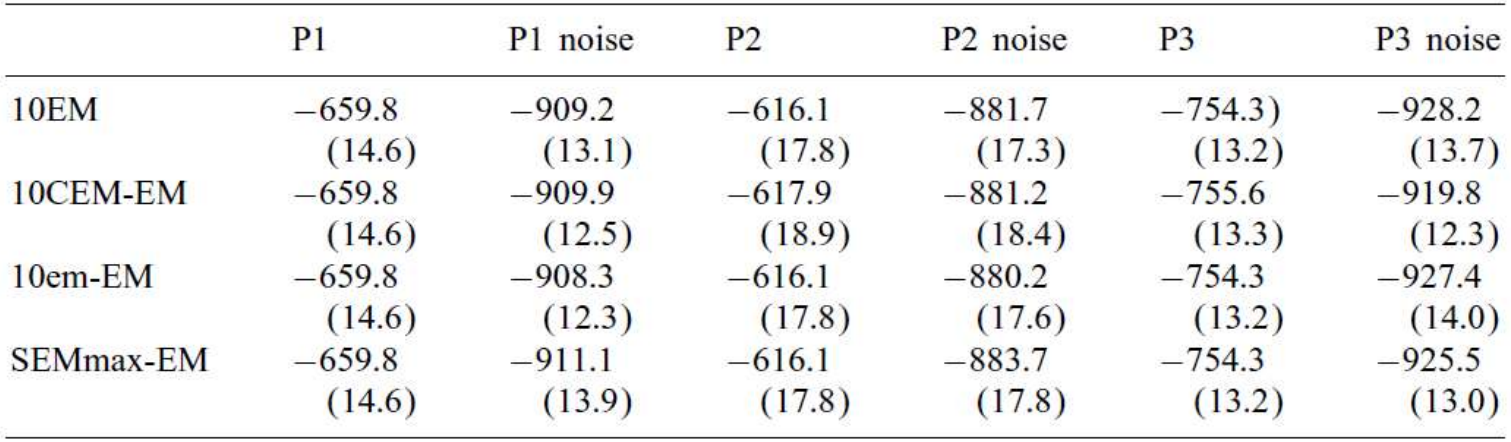
Simulation Data

Table3
Comparison by pairs with a large number of iterations



	P1	P1 noise	P2	P2 noise	P3	P3 noise
10EM vs. 10CEM-EM	2-0	38-27	19-0	30-42	36-0	3-86
10EM vs. 10em-EM	0-0	8-28	1-0	7-43	5-1	39-39
10EM vs. SEMmax-EM	0-0	43-8	0-0	43-21	5-3	23-64
10CEM-EM vs. 10em-EM	0-2	6-40	0-19	2-30	1-35	83-4
10CEM-EM vs. SEMmax-EM	0-2	54-30	0-19	61-21	0-35	66-8
10em-EM vs. SEMmax-EM	0-0	57-10	0-1	63-6	5-7	29-56

Table4
Means & standard deviations of maximum log-likelihood



	P1	P1 noise	P2	P2 noise	P3	P3 noise
10EM	−659.8 (14.6)	−909.2 (13.1)	−616.1 (17.8)	−881.7 (17.3)	−754.3) (13.2)	−928.2 (13.7)
10CEM-EM	−659.8 (14.6)	−909.9 (12.5)	−617.9 (18.9)	−881.2 (18.4)	−755.6 (13.3)	−919.8 (12.3)
10em-EM	−659.8 (14.6)	−908.3 (12.3)	−616.1 (17.8)	−880.2 (17.6)	−754.3 (13.2)	−927.4 (14.0)
SEMmax-EM	−659.8 (14.6)	−911.1 (13.9)	−616.1 (17.8)	−883.7 (17.8)	−754.3 (13.2)	−925.5 (13.0)

➡ Noise가 없는 data에서는 10CEM-EM이 가장 안 좋지만 cluster의 개수가 적을 때는 차이가 크지 않다.
Noisy data에서는 결론이 명확하지 않다.

Real Data Sets

Data set	Stars	Geyser	Biological Profiles
Number of Observations	$n = 2370$	$n = 272$	$n = 3641$
Dimension	$d = 2$	$d = 2$	$d = 5$
Number of Clusters	$K = 3$	$K = 3$	$K = 10$
Explanation	Stars described by their velocity U toward the galactic center and velocity V toward the galactic rotation (Soubiran, 1993)	Eruptions of the Old Faithful Geyser in Yellowstone National Park measured by the *duration* of the eruption, and the *waiting time* before the next eruption (Venables and Ripley, 1994)	Biological profiles of patients (Sandor, 1976)

Real Data Sets – Stars & Geyser

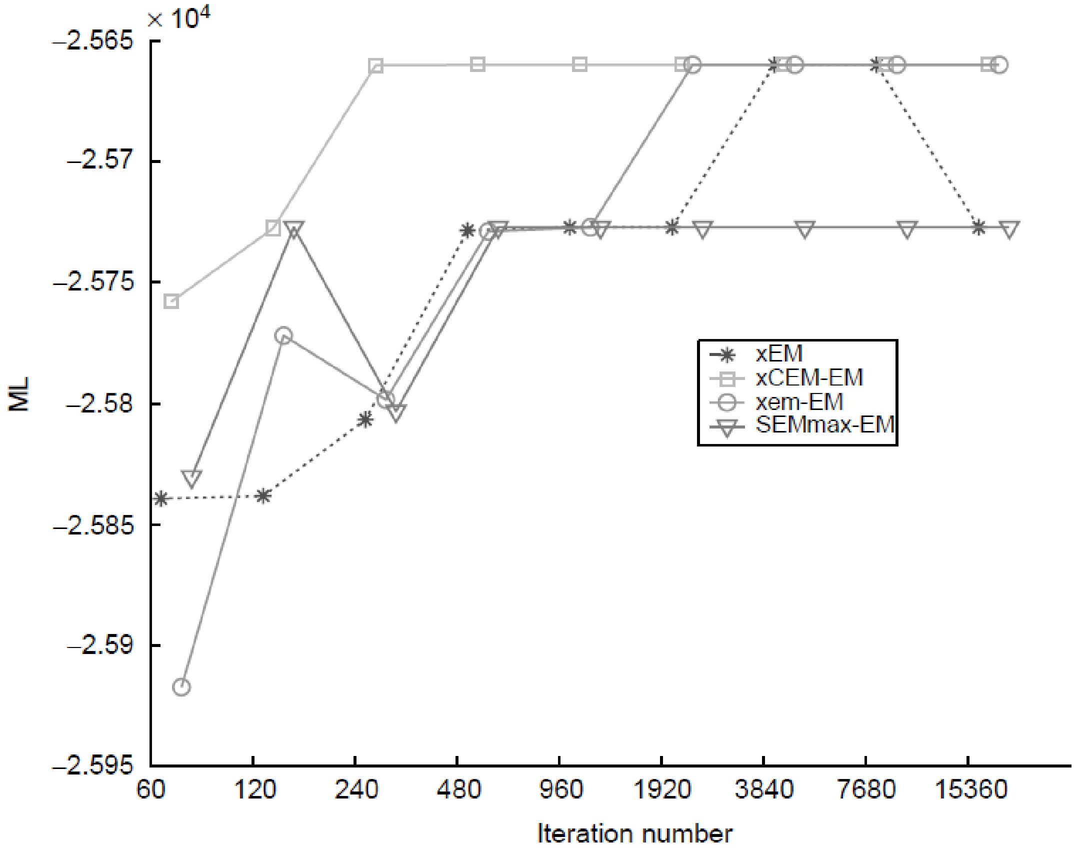


Figure1. Stars (x=10)

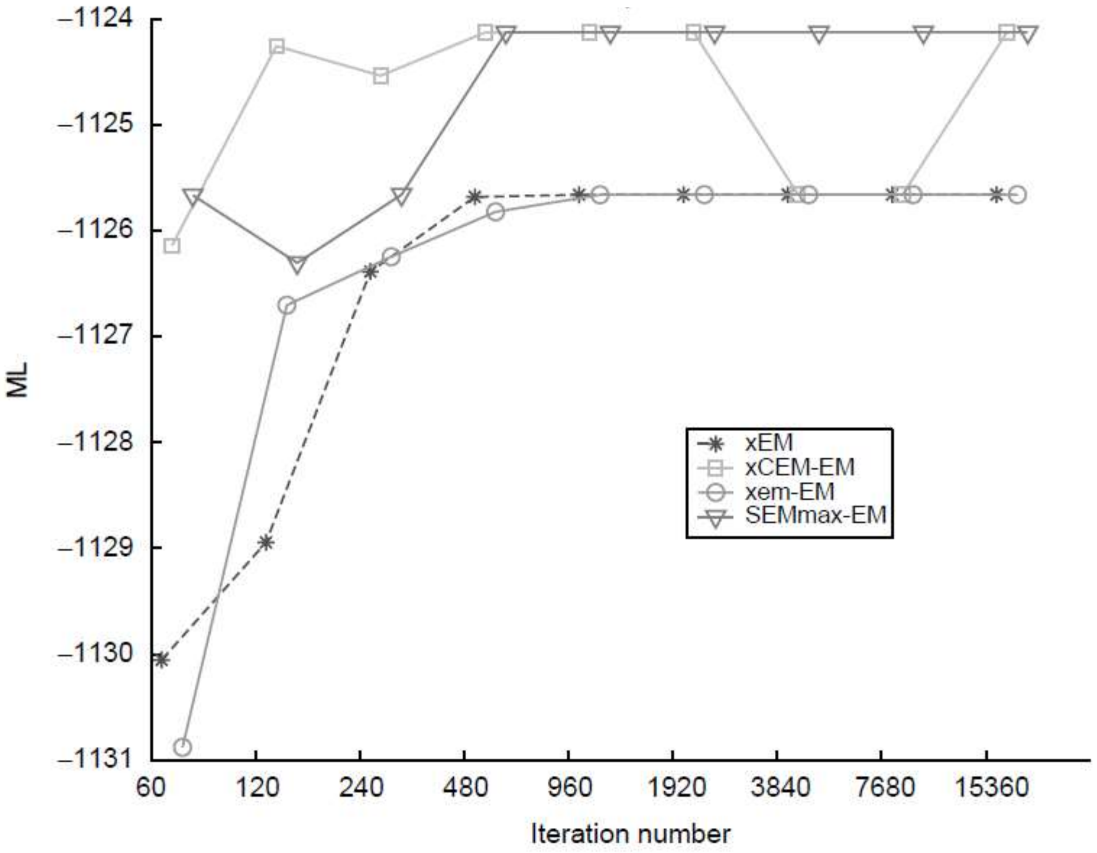


Figure2. Geyser (x=10)

Real Data Sets – Stars & Geyser

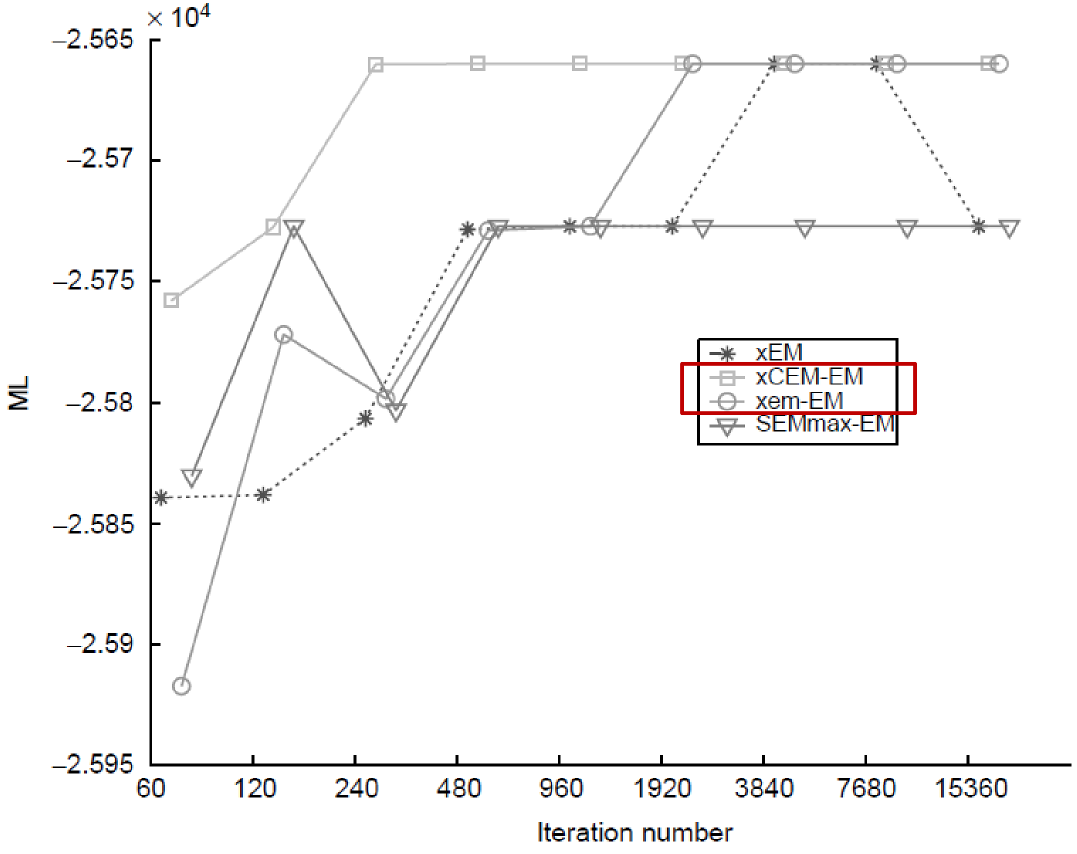


Figure1. Stars (x=10)

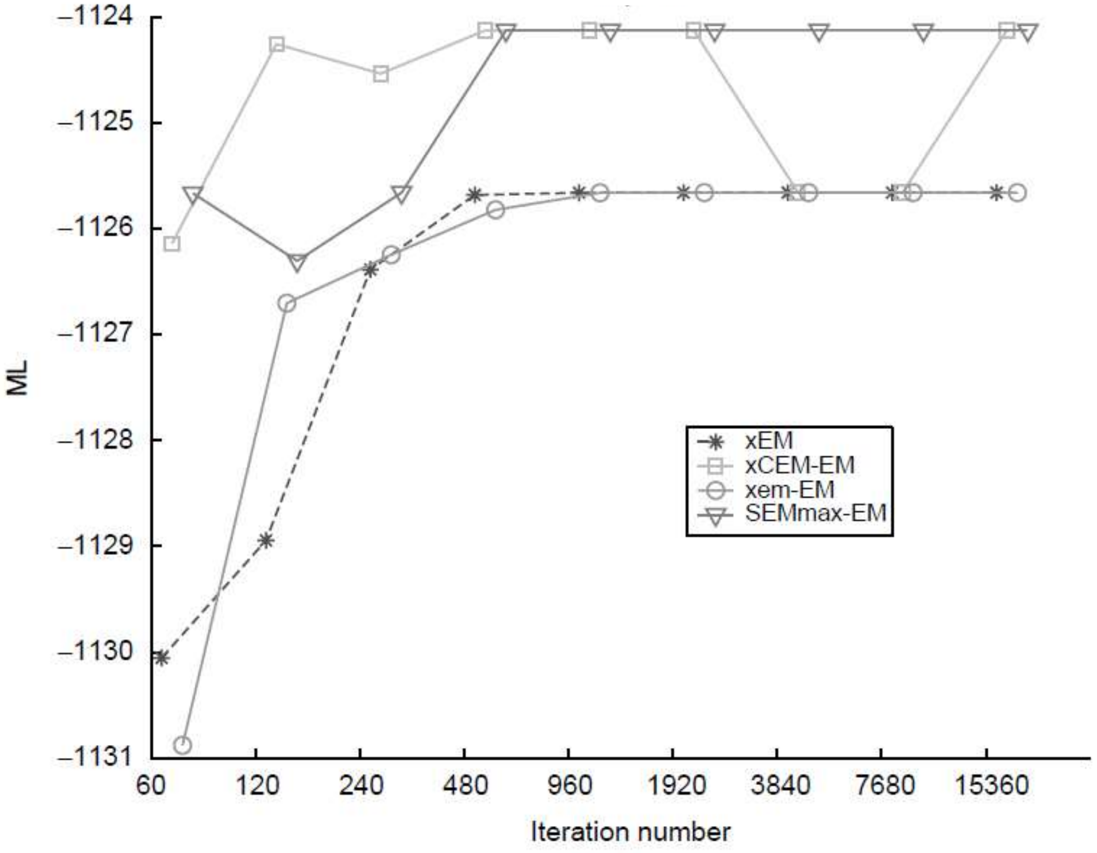


Figure2. Geyser (x=10)

Real Data Sets – Stars & Geyser

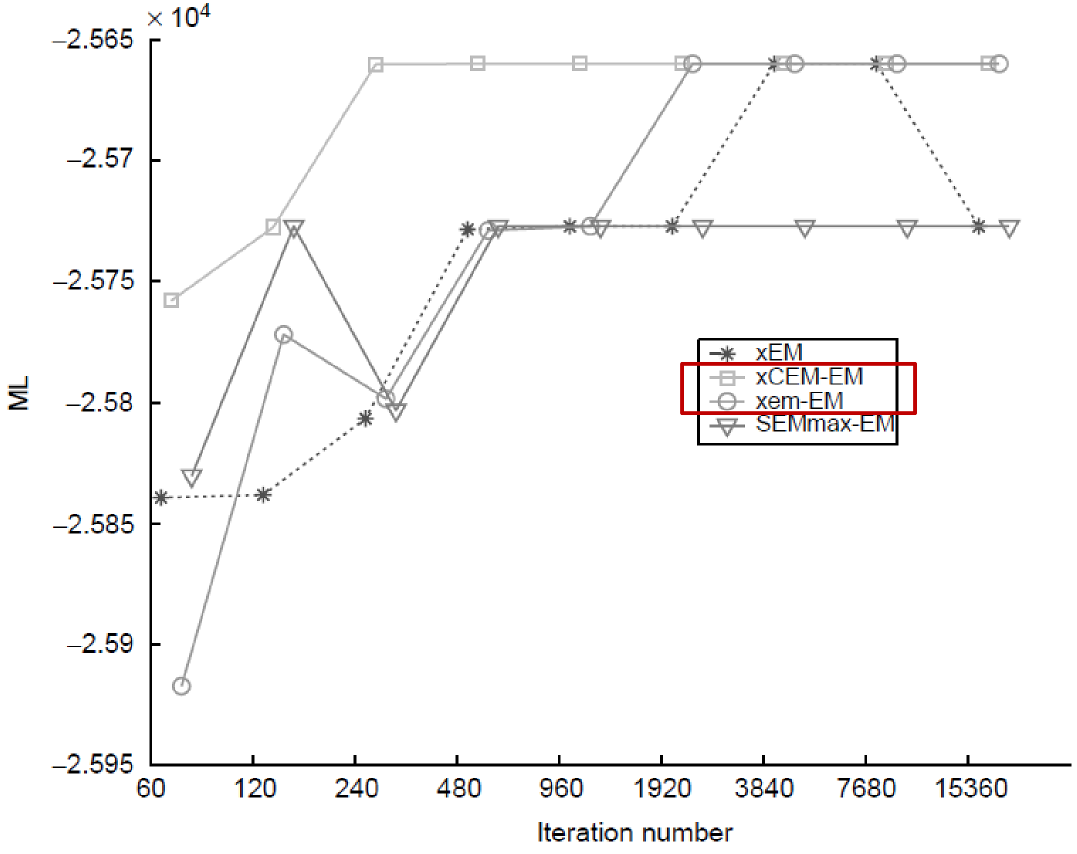


Figure1. Stars (x=10)

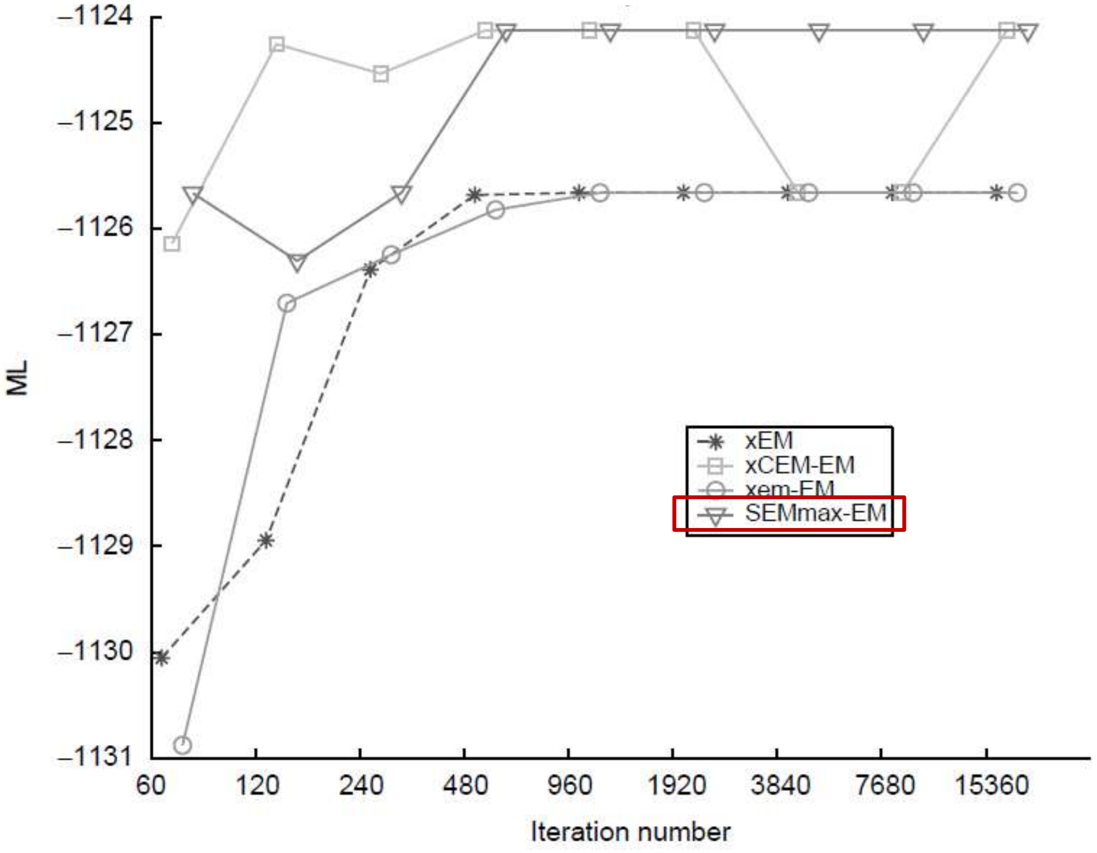
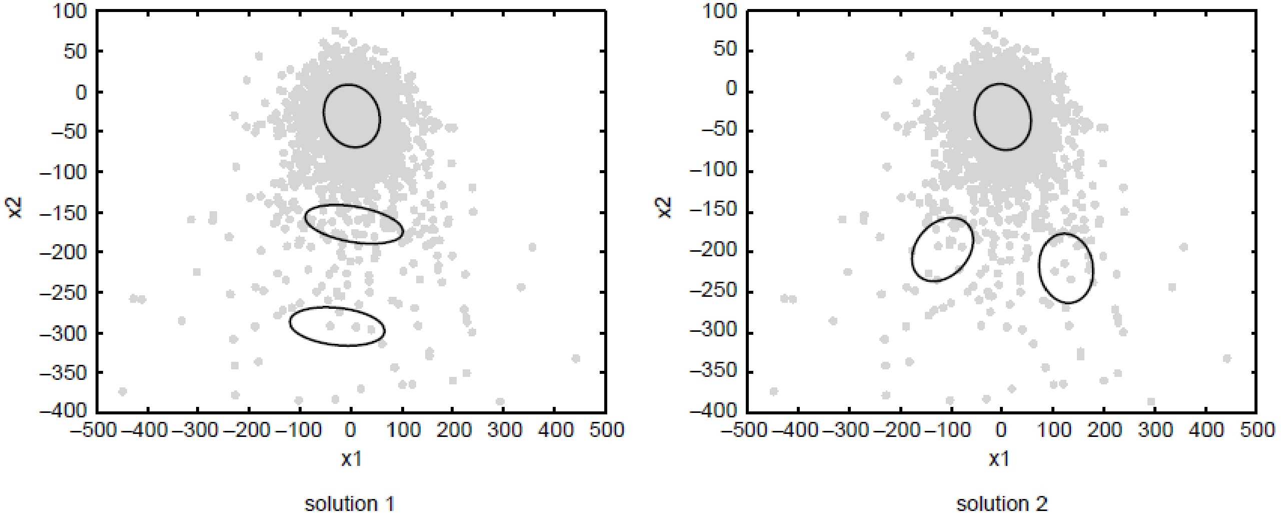


Figure2. Geyser (x=10)

Real Data Sets – Stars & Geyser



solution 1

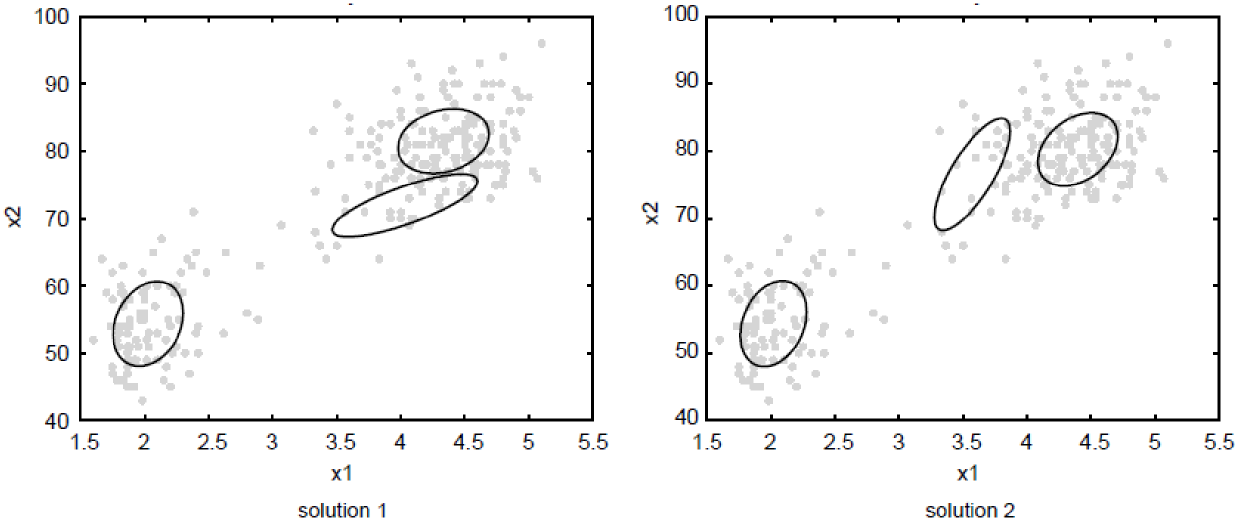
solution 2

< Two solutions in competition >

They provide isodensity ellipses for each cluster.

Solution 1 provides the highest likelihood.

Figure3. Solutions of data set "Stars"



solution 1

solution 2

Figure4. Solutions of data set "Geyser"

Real Data Sets – Biological Profiles

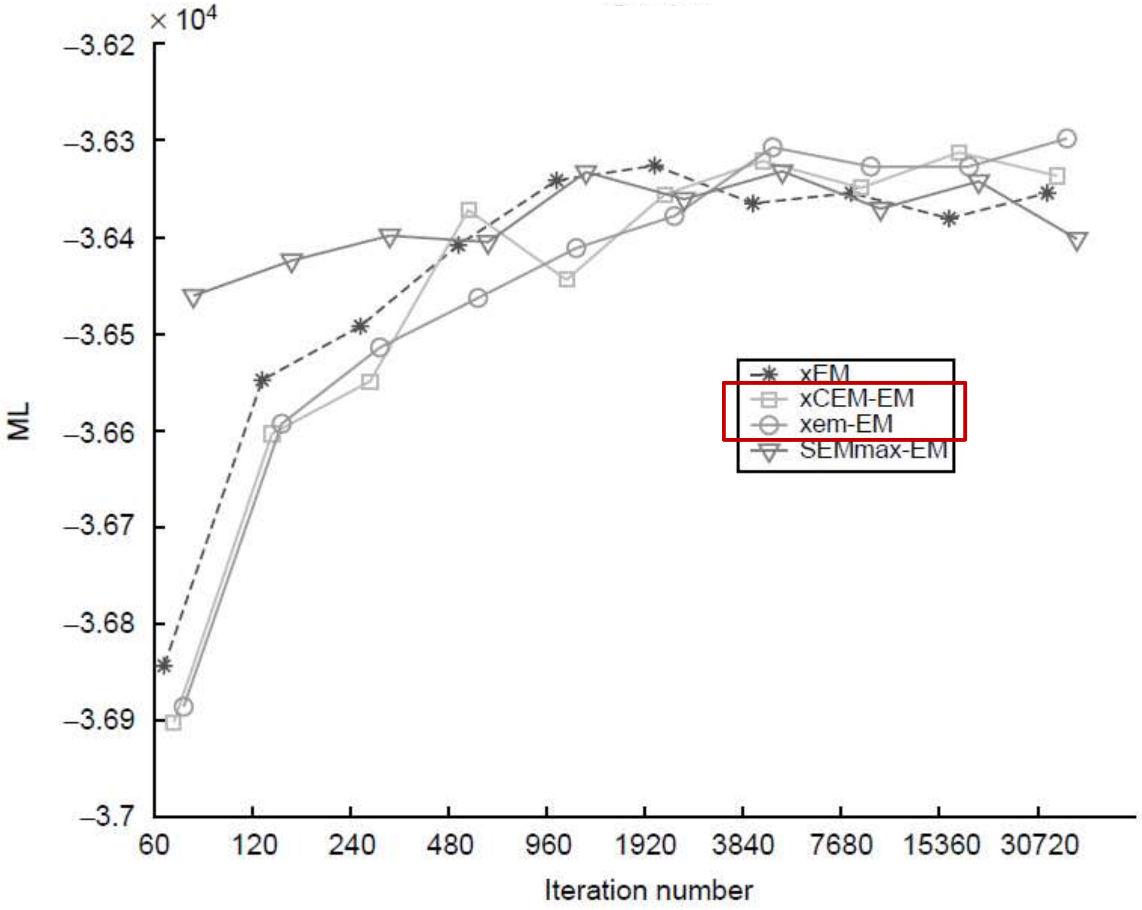


Figure5. Biological Profiles (x=10)

목 차

SECTION 1
도입

SECTION 2
GMM 추정

SECTION 3
실험 계획

SECTION 4
실험 결과

SECTION 5
결론

Discussion

1. 총 반복수가 충분해야 한다.
2. 알고리즘을 여러 번 반복 (repetition) 하는 방법이 좋다. (ex. 10em-EM)
3. short runs of EM 방법의 성능이 다른 방법들에 비해 약간 더 높으며, 기본적인 EM 방법이 가장 나쁘다.
4. short runs of EM 방법은 단순하고 특별한 가정이 필요하지 않으며, noisy data에 덜 민감하기 때문에 CEM 또는 SEM 알고리즘보다 더 적절하다.
5. 결론적으로 총 반복수가 충분히 크다면, short runs of EM의 반복을 통해 초기값을 선택하는 것을 추천한다.

감사합니다