『学習院大学　経済論集』第46巻　第2号（2009年7月）

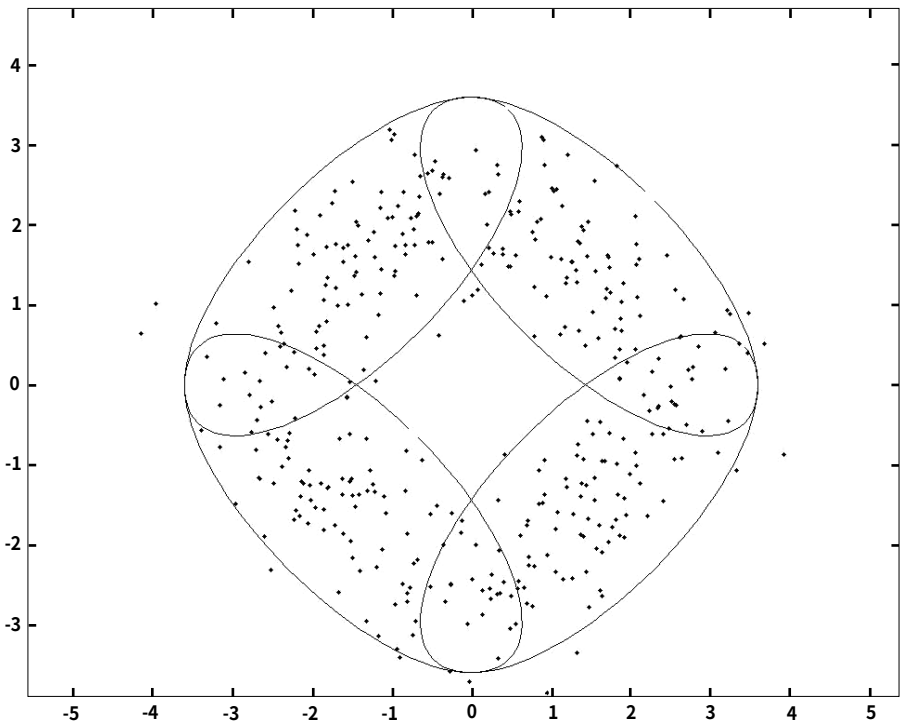
混合分布問題
――その基礎からカーネル降下法まで―― Part 2

金田　尚久、新居　玄武

９．シミュレーション

以下の図の4クラスター・モデルからデータを発生させて，シミュレーションを行った。

図7



各クラスターのパラメターは,

表3

	μ_x	μ_y	σ_x	σ_y	ρ	w
左上	-1.5	1.5	1	1	0.7	0.25
右上	1.5	1.5	1	1	-0.7	0.25
右下	1.5	-1.5	1	1	0.7	0.25
左下	-1.5	-1.5	1	1	-0.7	0.25

それぞれのクラスターに等しいウェイトを置き，400個の観測値を発生させる。これを1データ・セットとして，100データ・セットを用意する。図6は真のモデルの確率90％の等高線と，400個の観測値の実例を示している。

シミュレーションの結果は良好である。実行時間は，procedure一回につき，（データを発生させる時間を除いて）約9分しかかからない。AICとBICの違いに関しては，BICの方が，良好なパフォーマンスであった。

表4

	AIC	BIC
3クラスター		8回
4	45回	85
5	33	7
6	11	
7	8	
8	2	
9	1	

これらの結果は，AICについては図8に，BICについては図9に表わされている。

図８

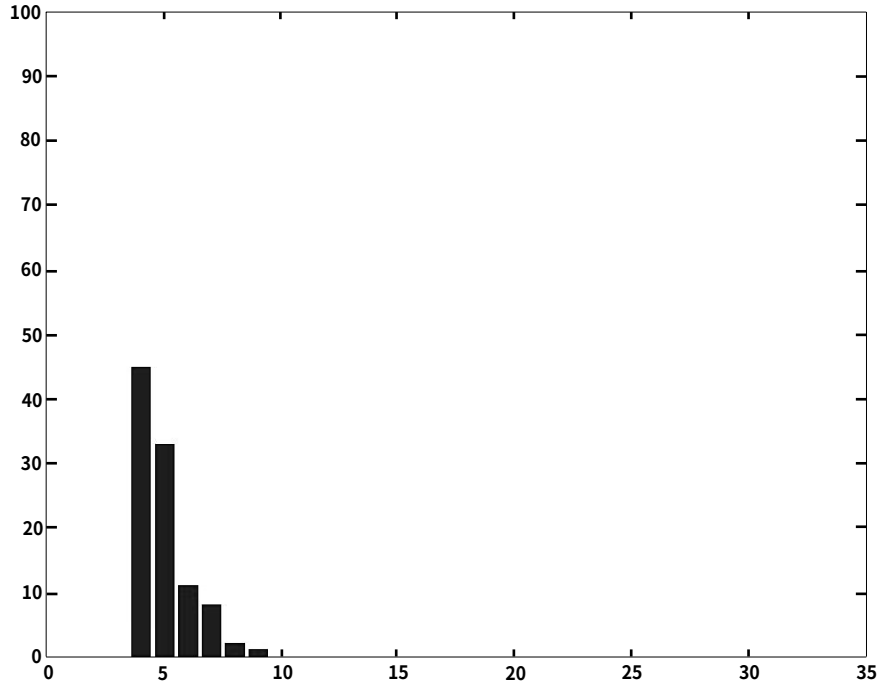
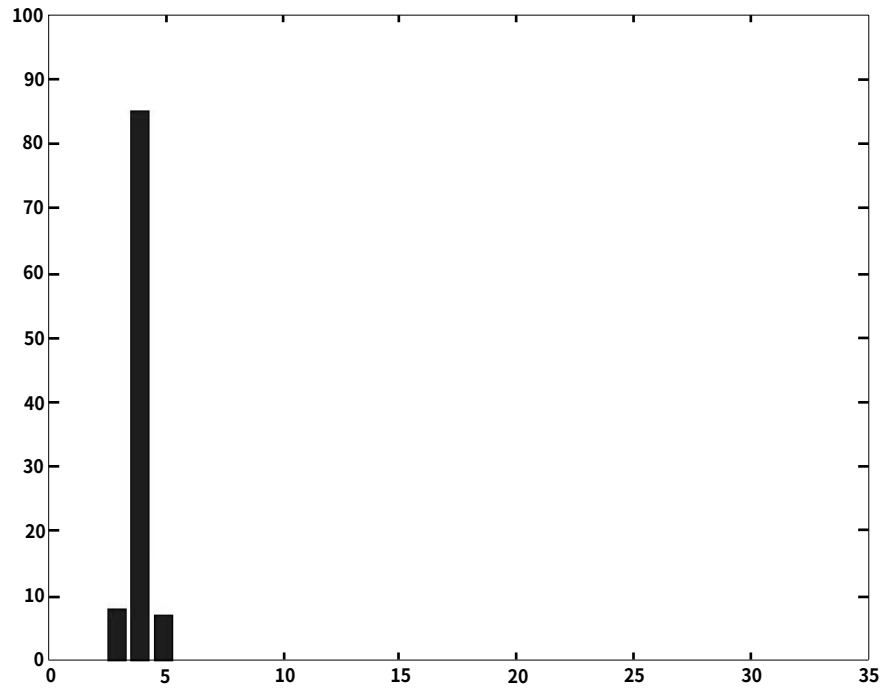

100
90
80
70
60
50
40
30
20
10
0
0
5
10
15
20
25
30
35

図９


100
90
80
70
60
50
40
30
20
10
0
0
5
10
15
20
25
30
35

本論文冒頭

AICはコンポーネント数を多めに推定する傾向がある。SSも，彼らの1次元の実験で，これと同じことを見出した。（彼らは，繰り返し実験はやっていないが，1セットだけデータを発生させる実験をやっている。）400個の観測値は大きなデータ・セットではない。だから，それらはドーナッツ型に散らばっている。4つのクラスターは，必ずしも目によって見分けられない。そのことを考えに入れれば，BICのパフォーマンスは注目に値する。次にMEASURE2を少し変えて，4クラスター vs. 4クラスター相似測度を作る。そして，真の4クラスター・モデルとBICによって正しく選ばれた4クラスター・モデルの間にこの測度を当てはめる。このやり方で，我々は，4クラスター・モデルとして選ばれたものの中の，ベストとワーストを見つけ出すことができる。それらの図とパラメターを示そう。

図10　ワースト・モデル

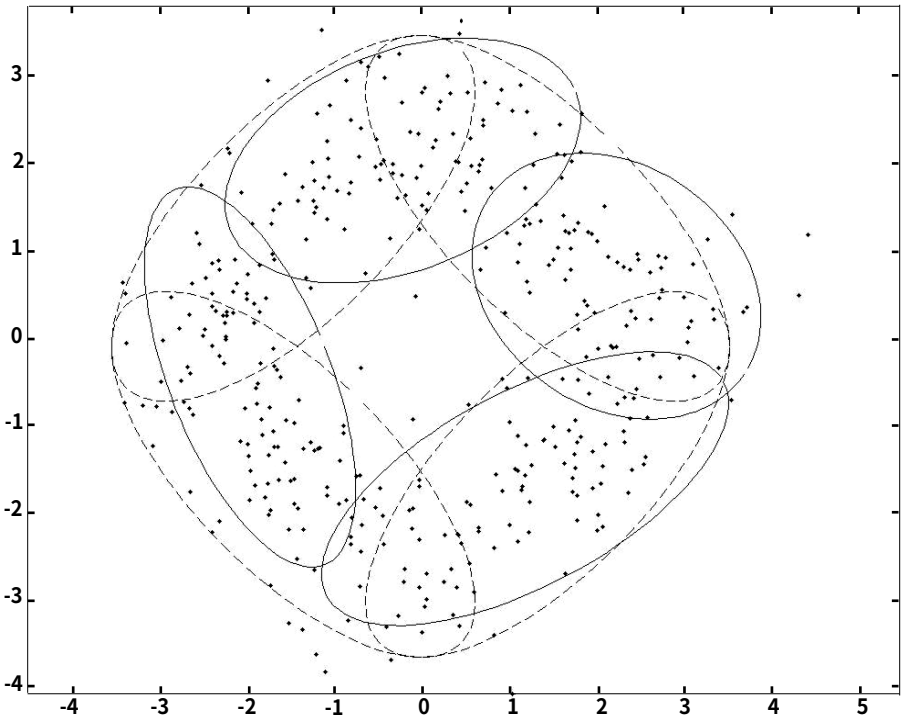


表5　ワースト・モデルのパラメター

	μ_x	μ_y	σ_x	σ_y	ρ	w
左上	-0.21	2.18	0.98	0.67	0.36	0.28
右上	2.31	0.71	0.80	0.73	-0.23	0.21
右下	1.24	-1.67	1.12	0.75	0.62	0.27
左下	-2.01	-0.36	0.58	1.04	-0.58	0.24

図11　ベスト・モデル

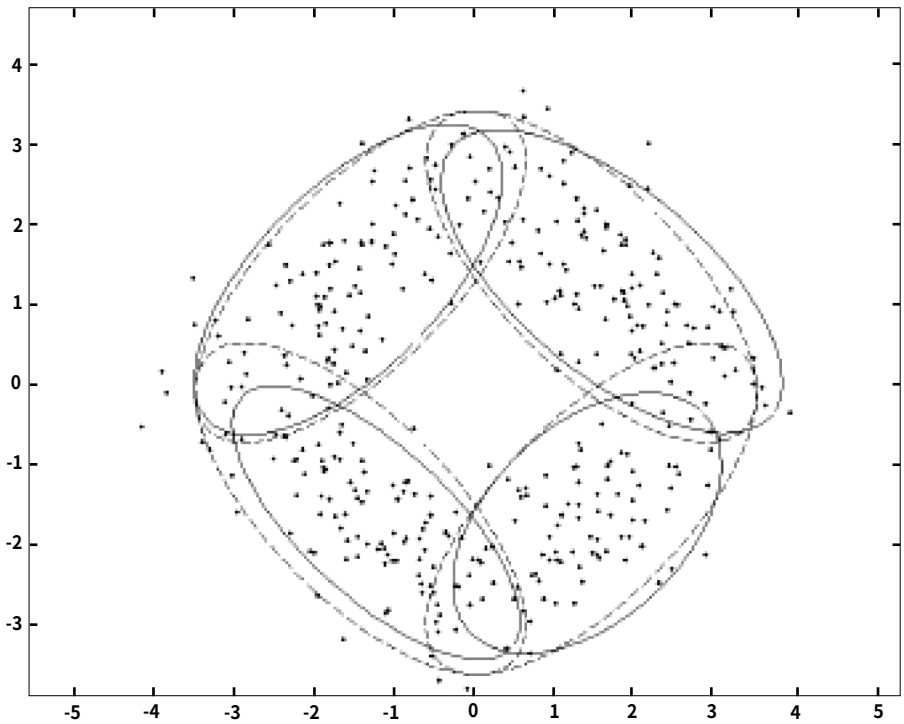


表6　ベスト・モデルのパラメター

	μ_x	μ_y	σ_x	σ_y	ρ	w
左上	-1.66	1.46	0.93	0.94	0.65	0.255
右上	1.77	1.44	1.03	0.91	-0.66	0.2875
右下	1.44	-1.69	0.81	0.78	0.46	0.2375
左下	-1.29	-1.68	0.87	0.82	-0.72	0.22

10．MEASURE1と2の比較

MEASURE1と2の違いを明らかにするために，次の二つのprocedureの比較を行う。一つは，前章まで説明し，シミュレーションを行って来たprocedureである。もう一つは，一貫してMEASURE1の用いられるprocedureである。即ち，そのPhase3では，最も良く似たペアがMEASURE1で選ばれ，それを置き換えるコンポーネントが，L2Eで推定される。これまでのprocedureをProcedure A,「MEASURE1のみ」のprocedureをProcedure Bと呼ぼう。図12から17では，Procedure A, Bの降下過程から選ばれた図に，それぞれ添え字a, bを付ける。

図12a. n=14 (PROC.A)

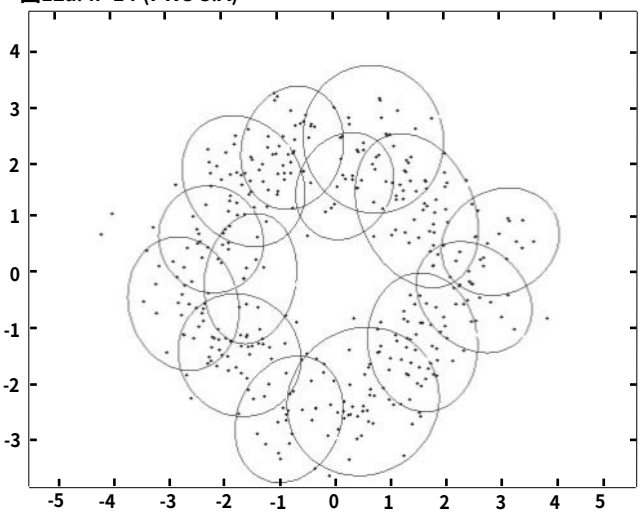
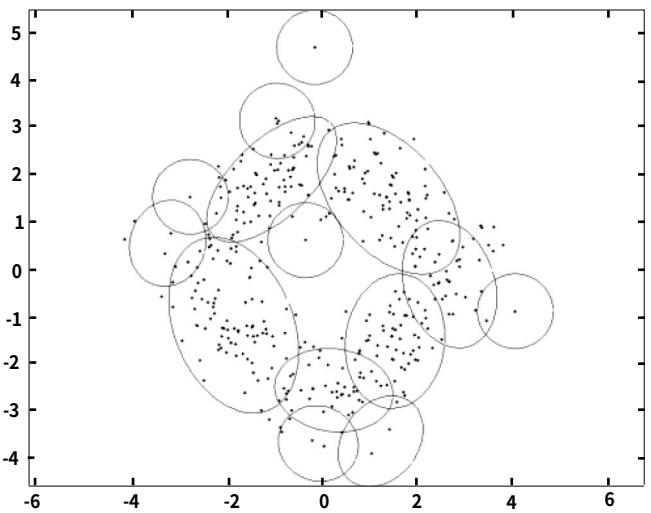


図12b. n=14(PROC.B)



本論文冒頭

図13a. n=12 (PROC.A)

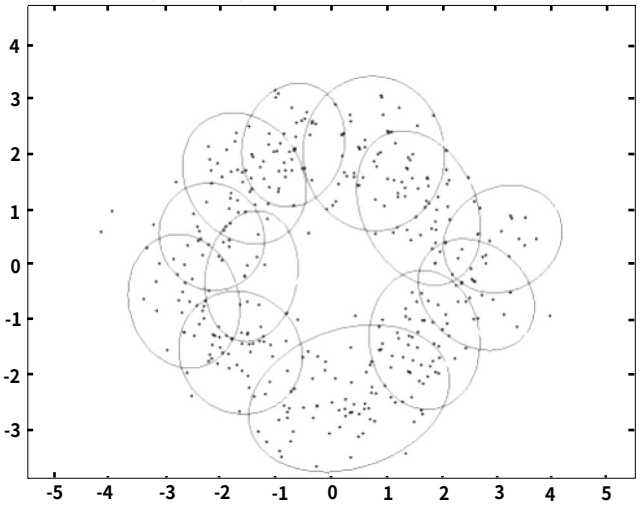
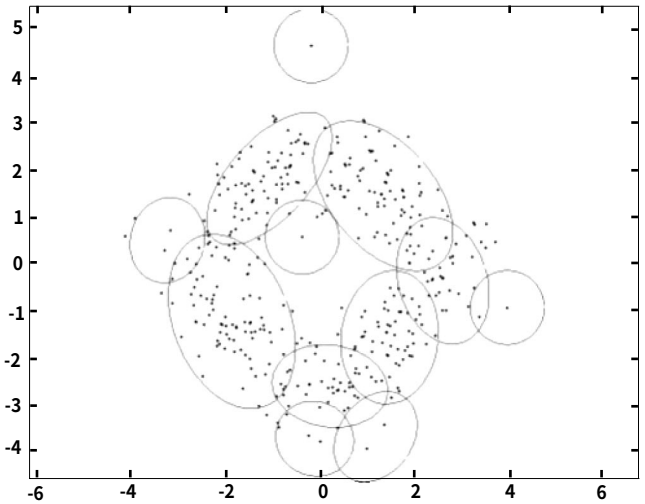


図13b. n=12 (PROC.B)



本論文冒頭

図14a. n=9 (PROC.A)

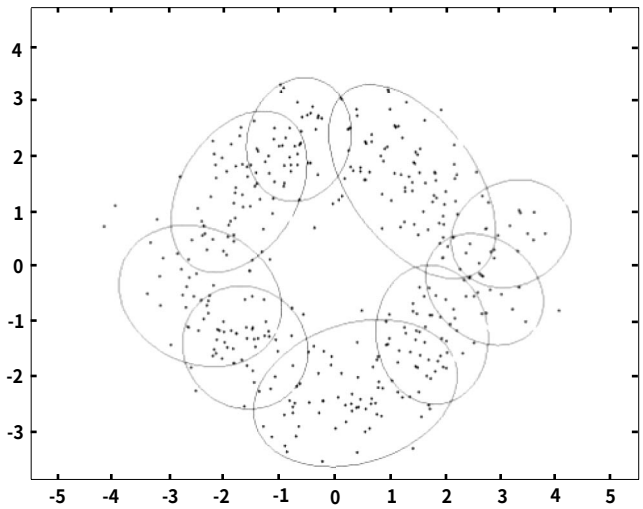
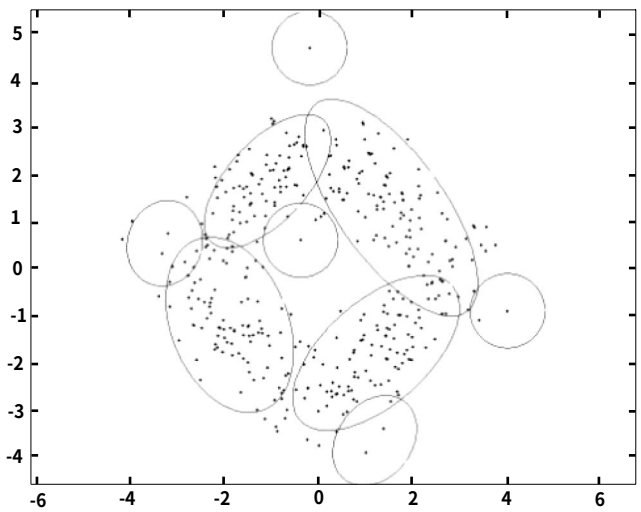


図14b. n=9 (PROC.B)



本論文冒頭

図15a. n=7 (PROC.A)

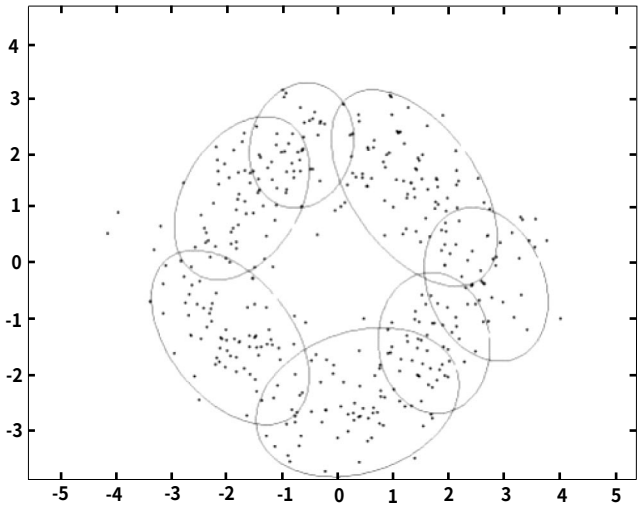


図15b. n=7 (PROC.B)

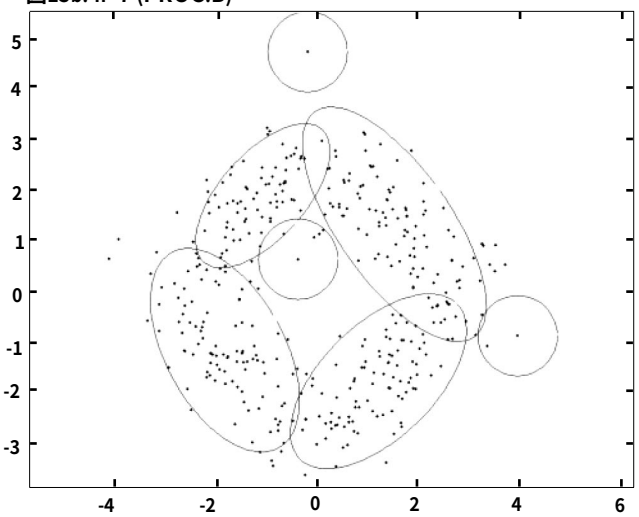


図16a. n=6 (PROC.A)

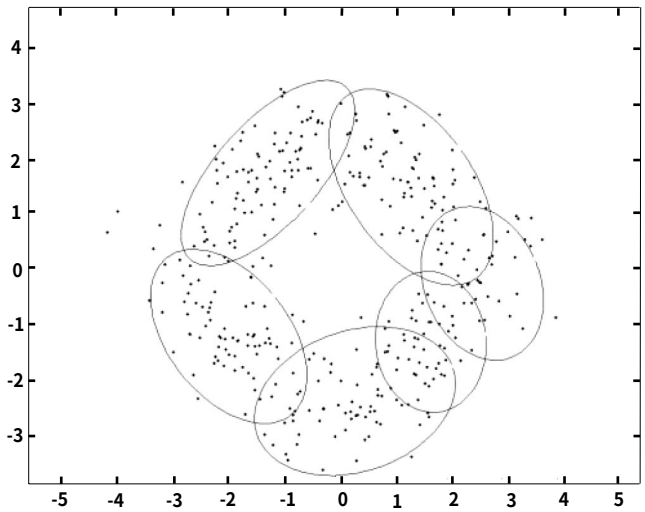
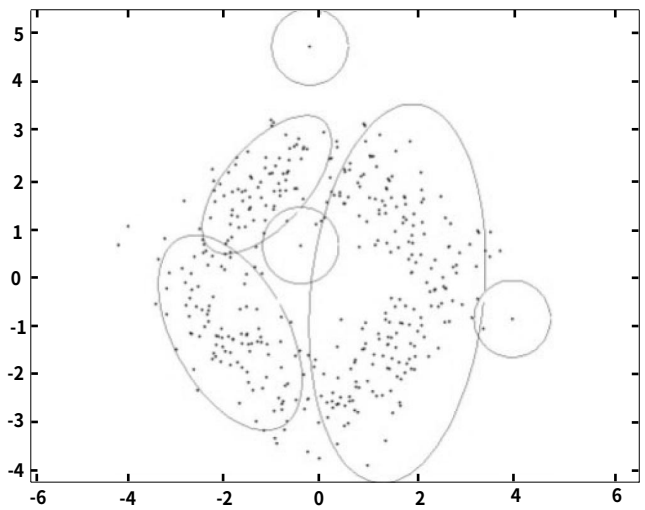


図16b. n=6 (PROC.B)



本論文冒頭

図17a. n=4(PROC.A)

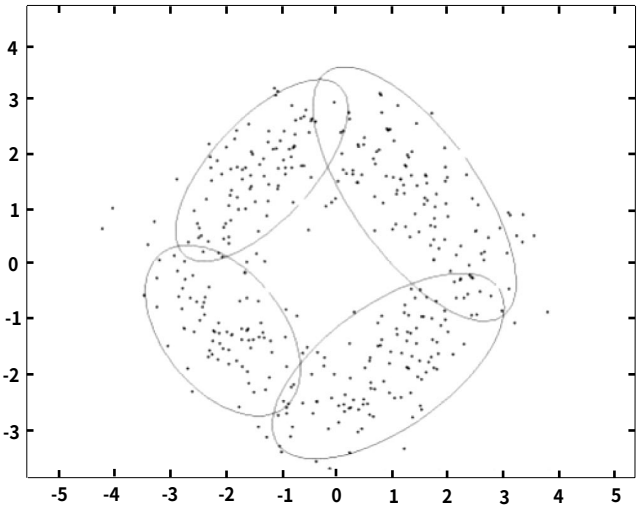
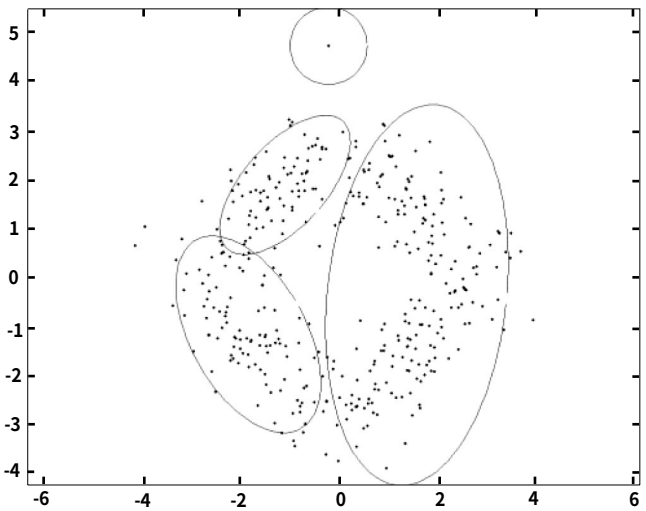


図17b. n=4 (PROC.B)

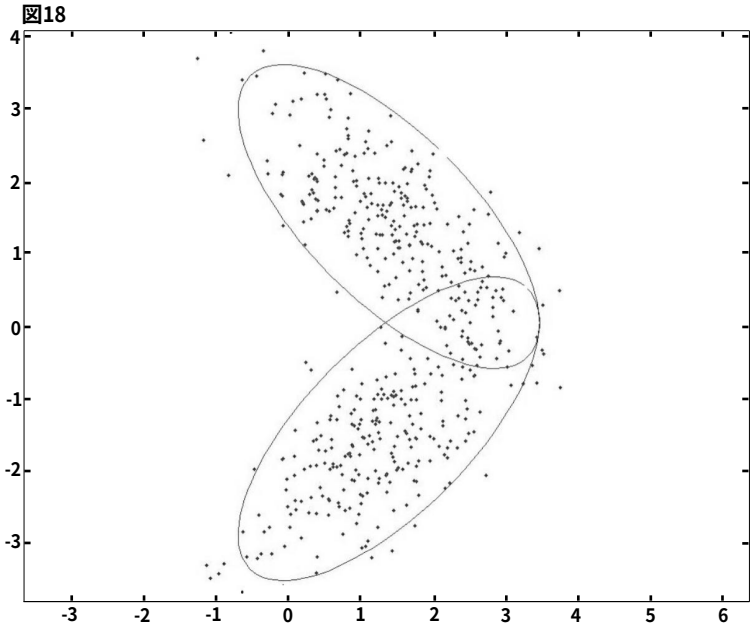


本論文冒頭

二つのprocedureは，Phase2までは完全に同じである。しかし，Phase3をおよそ半分ほど進んだn＝14では，顕著な相違が見られる。Procedure Aでは，どのクラスターでも，確率90％ の等高線が，ほぼ同じ大きさである。しかし，Procedure Bでは，「単カーネル」（カーネル推定のカーネルが，モデルの中に残っている場合，こう呼ぶことにする。英語では”singleton”が適当であろう。）と単カーネル2,3個分のウェイトのクラスターが，まだ残っている。これらは，図上で見分けられる。procedure は，x, y方向に等しい分散のカーネルから出発したのだから，ウェイトが極めて小さいクラスターは，等高線が完全な円か円に近くなっている。この段階以後，どう変わるか，見て行こう。Procedure Aでは，n＝14, 12, 9と降下過程はスムーズに進行する。図は飛び飛びではあるが，どのクラスターが融合したかはわかるし，そこに置き換えられた新しいクラスターが，どれくらい良く当てはまっているかも，確かめられる。一方，同じn＝14, 12, 9で，Procedure Bは，単カーネルを「クラスター」と見なす奇妙な処理法を改められない。このprocedureの7クラスター・モデルは重要である。BICに選ばれた，ベスト・モデルだからである。4つの真のクラスターがかなりうまく捕えられているが，3つの単カーネルがまだ残っている。n＝6で重大な失敗が生ずる。3つの単カーネルがまだあるのに，右側の2つの大きなクラスターが融合してしまうのである。ここに，我々はMEASURE1の主要な欠陥を見る。procedureが進むにつれて，単カーネルは，増々孤立していく。結果として，1 vs. 1相似測度は，大きなクラスター同士のペアを，単カーネルを相手とするペアよりも，相似度が高いと認識してしまうのである。そこで，意味のあるクラスターが，大きくて意味の無いクラスターに融合されてしまう。この欠陥は，この段階以後，修正されることは無い。n＝4でも，まだ単カーネルが残っている。Procedure Bの結果が不振なのとは対称的に，Procedure Aは，真のモデルに向かって，少しずつ近づいていく。そして，最終的に4クラスター・モデルが，ベストとして選ばれる。MEASURE2が，単カーネルを除去する強い力を持っていることが，明らかに，Procedure Aの成功の原因である。実際のところ，全ての単カーネルと，ウェイトが0.01（単カーネル4つ分）以下の小さなクラスターは，Phase3の初めの部分で，除去されている。

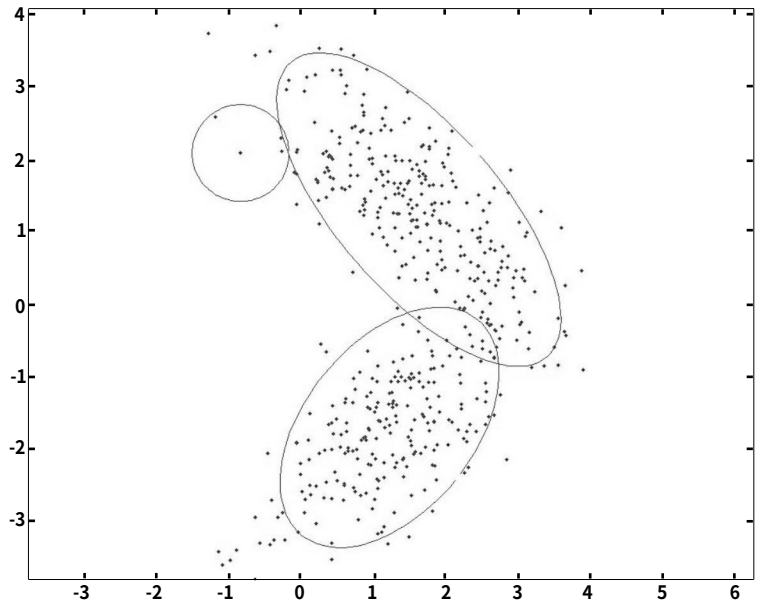
これまでの議論で，MEASURE1が，クラスター数決定の重要な段階で，何故うまくいかないのかが，明らかになった。では，MEASURE2は，同じ単カーネルの問題を，どうやって乗り越えるのだろうか？この問いに答えるために，我々はさらに単純なデータ・セットを用意する。これまで考察して来た4クラスター・モデルの右側の2クラスターからなる，2クラスター・モデルが図18に描かれている。それぞれのクラスターに同じウェイトを置いて，観測値を500個発生させる。

図18



これにProcedure Bを当てはめれば，図19の3クラスター・モデルが最適として選ばれる。

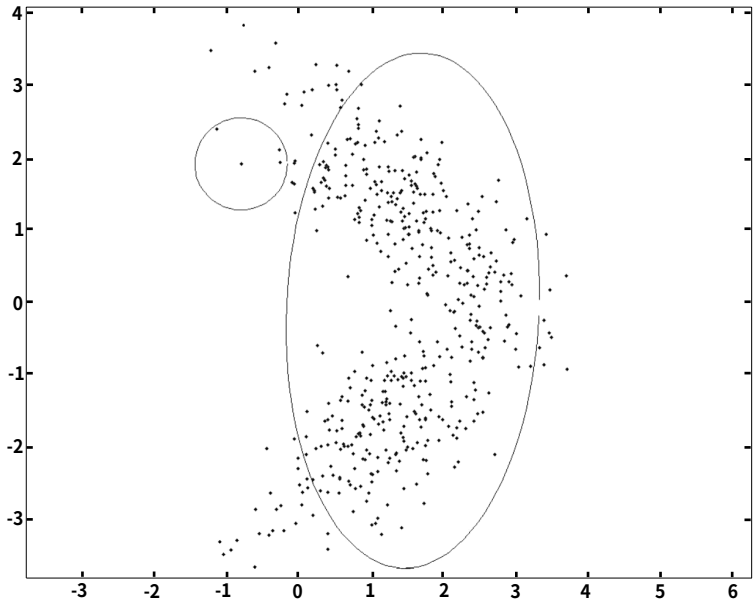
図19



前の例に照らして，この失敗の原因を推測するのは，たやすい。もう1段階降下すると，右側の充分良く当てはまっているクラスターが，一つに融合する（図20）。

本論文冒頭

図20



左上のクラスターは，単カーネルのように見えるが，単カーネル2つ分のウェイトを持った小クラスターである。これを相手とする2つの組み合わせよりも，右側の2つのクラスターの組み合わせの方が相似度が高いと，procedureが認識したのは明らかである。図19の3つのクラスターを

クラスター1 : 右上
クラスター2 : 右下
クラスター3 : 左上

と名付けよう。i番目とj番目のクラスターの間のMEASURE1を $d(i,j)$ と書けば,

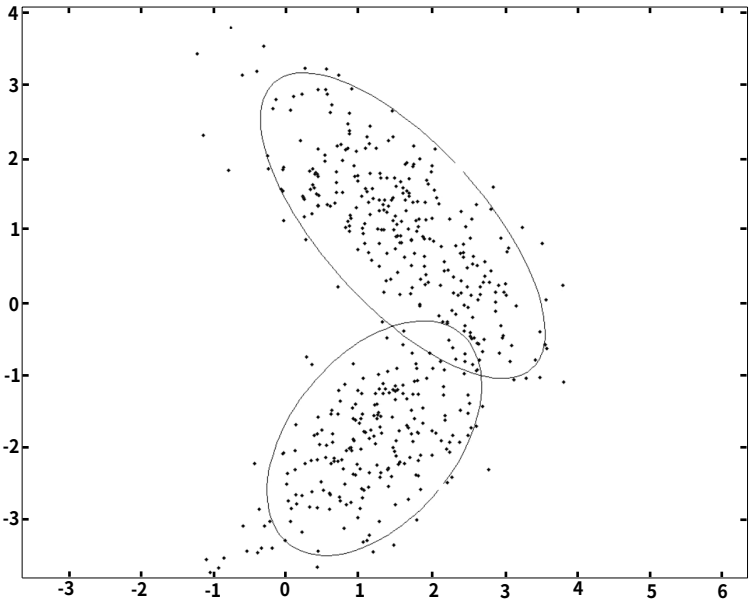
$$d(1,2) = 0.0287 \qquad d(1,3) = 0.0133$$

$$d(2,3) = 2.500 \times 10^{-10}$$

3クラスターから2クラスターに降下するときのみ，Procedure Aを当てはめたら，どうなるだろうか？　結果は図21である。。

本論文冒頭

図21



小クラスターは消え，そのウェイトは新クラスター（クラスター1とほとんど変わらない）に，組みこまれた。MEASURE2の力のもう一つの証明である。この成功をさらによく理解するために，図22-24を準備した。それぞれの図には，3クラスター・モデルと，2クラスター・モデル中の新しいクラスターが描かれている。

本論文冒頭

図22　3クラスター・モデルと *tenta*(1,2)

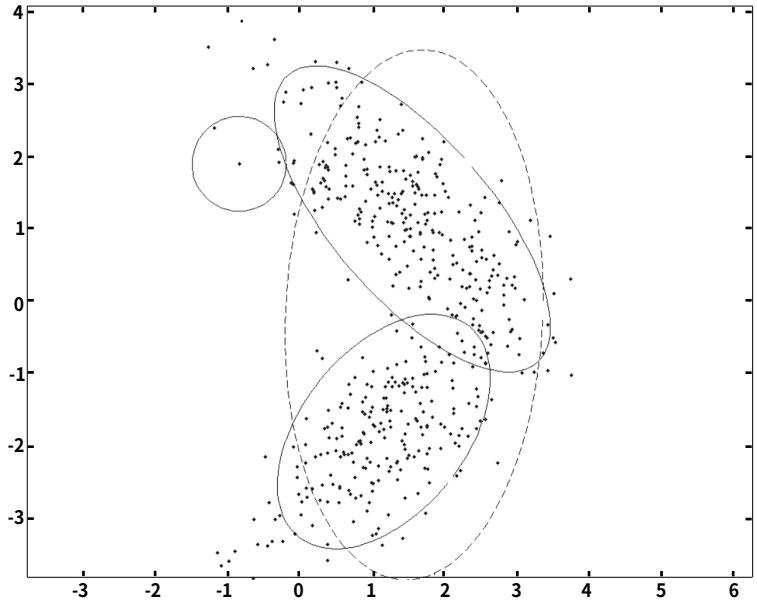
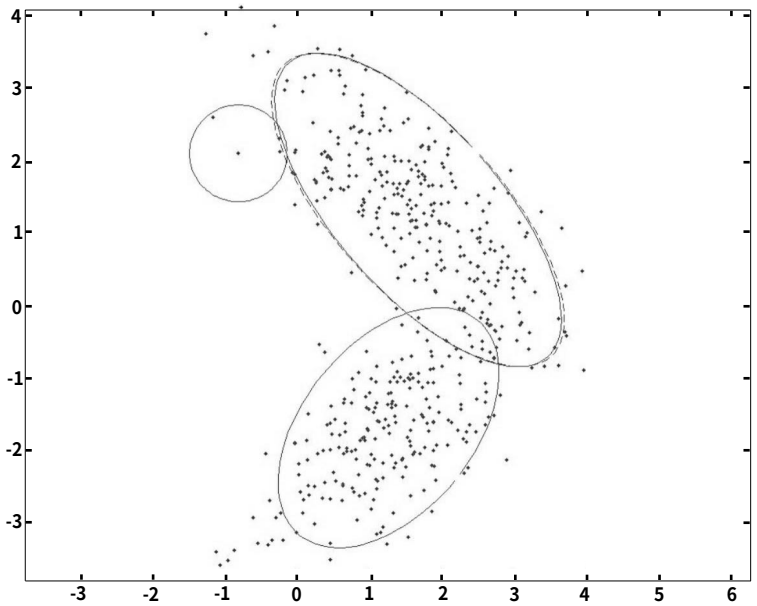
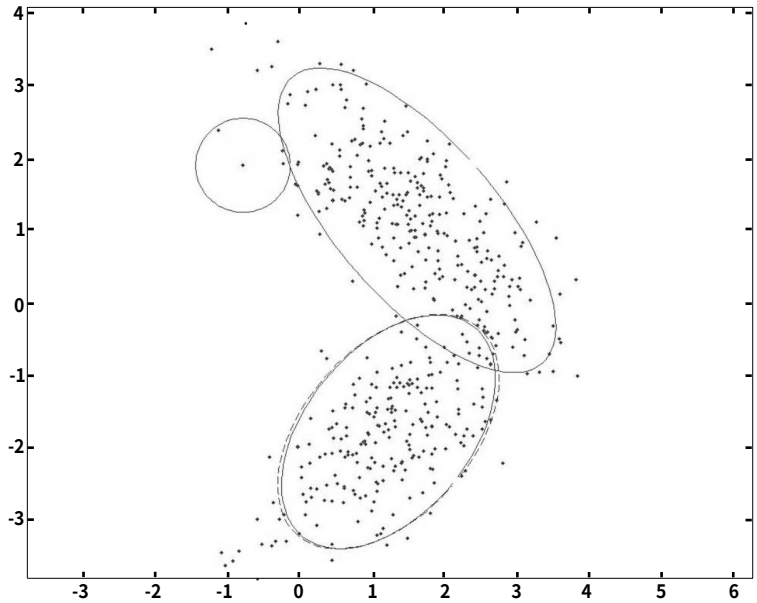


図23　3クラスター・モデルと *tenta*(1,3)



本論文冒頭

図24　3クラスター・モデルと *tenta*(2,3)



クラスターiとjが融合してできる，新クラスターを $tenta(i,j)$ と書こう。”tenta”とは，”tentative model”（暫定的モデル）の略である。まず，図22とその他の図は，大きく異なっている。図22では，$tenta(1,2)$ が非常に大きく，クラスター1と2をゆるくカバーしている。しかし，その他の図では，新クラスターは旧モデルの大きなクラスターのどちらかとほとんど一致している。これは自然な結果と考えられる。3クラスター・モデルでは，全確率質量の57％がクラスター1に，42.6％がクラスター2に，0.4％がクラスター3に置かれている。クラスター3が，大きなクラスターと組み合わされたとき，クラスター3は新クラスターの位置と形に大きな影響力を持ち得ない。しかし，二つの大きなクラスターが融合するときは，融合は「タイ」である。新クラスターは，両方をおおいこまなければならない。$tenta(i,j)$ のpdfも $tenta(i,j)$ と書くことにしよう。これまで通りの記号法で，図22および23に描かれた2クラスター・モデルと3クラスター・モデルの組み合わせは，以下のように書ける。

図22

$$w_1 f_1 + w_2 f_2 + w_3 f_3$$
$$(w_1 + w_2)\,tenta(1,2) + w_3 f_3$$

図23

$$w_1 f_1 + w_2 f_2 + w_3 f_3$$
$$(w_1 + w_3)\,tenta(1,3) + w_2 f_2$$

本論文冒頭

w_3を無視し，$tenta(1,3)$がf_1と極めて相似していることを考えに入れるなら，図23の二つのモデルはほとんど同一であり，したがって，相似測度はほとんど1となる。一方，w_3を無視し，$tenta(1,2)$がf_1ともf_2とも相似していないことを考えに入れるなら，図22の二つのモデルの相似度は図23よりも，ずっと低くなる。この推論を確かめるために，MEASURE2の値を示そう。第iクラスターと第jクラスターを融合したときの，3クラスター・モデルと2クラスター・モデルの相似測度を$S(i, j)$と書けば，

$$S(1,2) = 0.82516 \qquad S(1,3) = 0.99983$$

$$S(2,3) = 0.99976$$

それでは，何故ペア$(2,3)$ではなく，ペア$(1,3)$が選ばれたのだろうか？　目には分からないが，実際のところ，クラスター2から$tenta(2,3)$への動きは，クラスター1から$tenta(1,3)$への動きよりも大きいのである。クラスター1から$tenta(1,3)$とクラスター2から$tenta(2,3)$へのパラメターの変化は以下のようである。

	μ_x	μ_y	σ_x	σ_y	ρ
$cluster1 \to tenta(1,3)$	-0.0008	-0.0002	0.0199	-0.0011	-0.0002
$cluster2 \to tenta(2,3)$	-0.0016	0.0034	0.0246	0.0024	0.0007

全てのパラメターにおいて，後者の変化は前者よりも大きい。さらに，主な変化は平均ではなく，σ_xに起こっていることがわかる。何故こうなったのかは，充分に明らかでないが，大よそ次のように考えられる。旧クラスターは，観測値が集中している領域に当てはまっているから，その平均を動かすことは，そのような領域におけるフィットの減少を招く。クラスター3が無くなったことに対する，可能な調整は，分散を張り出すことだけである。仮に，クラスター3が，大きな分布に近ければ，その分布の高い所に位置しているのだから，クラスター3が失われたことによる尤度の減少は，大きな分布がわずかに分散を伸ばすことによって（全部ではなくとも）補われる。ところが，クラスター3が大きな分布から遠ければ，その分布の低い所に位置しているのだから，同じ尤度の減少を補うのに，大きな分布はより遠くまで分散を伸ばさなければならない。このような理由から，MEASURE2は，単カーネルまたはそれに近い小クラスターを，最も近くの大きなクラスターに吸収するものと思われる。

11．EMアルゴリズムとの比較

KRのパフォーマンスを評価するために，標準的なEMアルゴリズムを，9章で用いたデータ・セットに当てはめてみよう。100個のデータ・セットに，クラスター数を固定しながら，EMアルゴリズムを当てはめ，AICとBICを計算する。各データ・セットにおいてAICまたはBICが最小となるモデルが，それぞれの基準の下での最適なモデルである。我々の用いたアルゴリズムは，Dr. Patrick Tsuiによって製作され，MATLAB Exchange [13]を通して入手できる。

しかし，残念ながら，このアルゴリズムは，固有値の計算が取り扱えない場合には，停止してしまう。この種の失敗は，クラスターの数が多くなるにつれて，より頻繁となる。そこで，試みるのは10クラスターまでとした。（KRの場合，BICは29クラスター・モデルまで計算してある。）さらに，このアルゴリズムでは，k-meansによって初期値が与えられることになっている。k-meansは1クラスター・モデルでは定義不可能であるから，この場合は省略する。

結果は驚くべきであった。100回のシミュレーション中，4クラスター・モデルが最適となったことは1回も無い。この結果は，一見したところ，非常に極端である。しかし，慎重に検討してみると，Dr. Tsuiのアルゴリズムにも，我々の利用法にも，問題はないことがわかった。100回のシミュレーションを通して，EMの限界は一様に現れている。何よりもまず，全てのシミュレーションで選択されたのは，2クラスター・モデルであった。何故このような，画一的とも言える失敗が起こったのだろうか？　まず，一つのデータの推定結果を見ることから，この問題を考えてみよう。我々が取り上げるのは，図7に描かれ，前章まで考察を続けてきたデータである。この場合に，EMは14クラスターまで停止しなかったので，14-2クラスター・モデルを図25a-37aに挙げる。EMは降下過程ではないが，KRとの比較を容易にするために，クラスター数の下がる順に図を並べてある。KRによる14-2クラスター・モデルは図25b-37bである。EMとKRそれぞれによって推定されたモデルのICとlog likelihoodを表にまとめた（表7, 8, 9）。図38と39は，これらの表をグラフにしたものである。

図25a n=14 (EM)

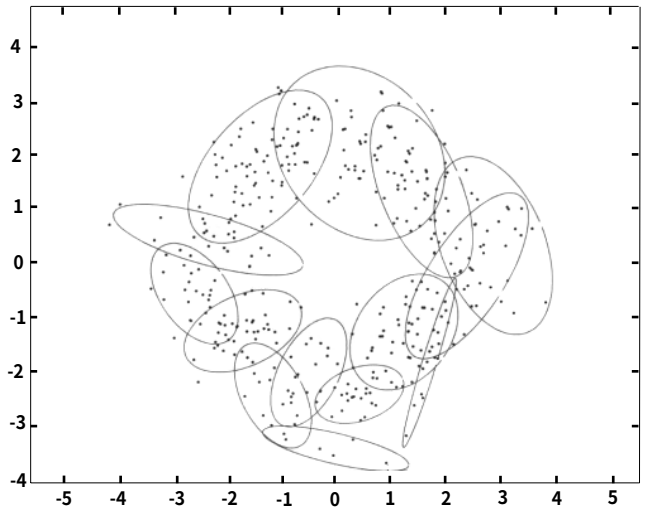
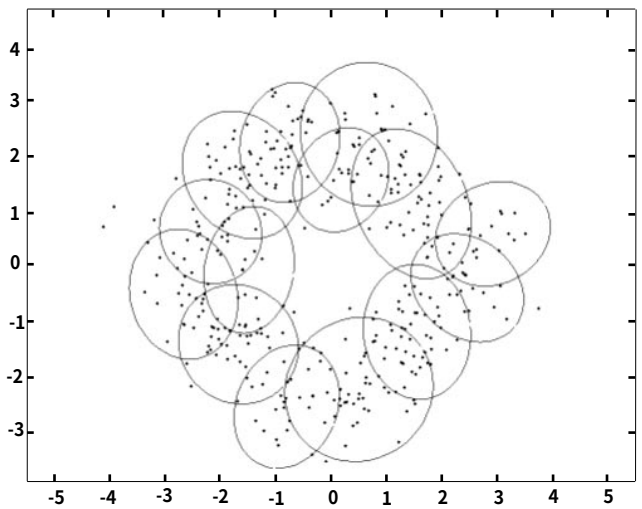


図25b n=14 (KR)



本論文冒頭

図26a n=13（EM）

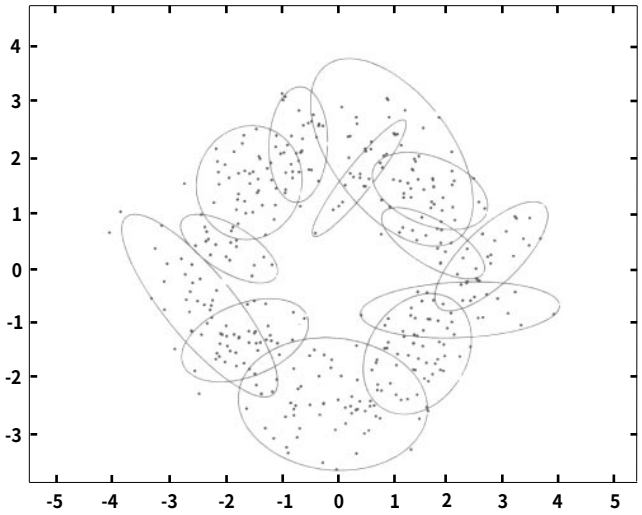
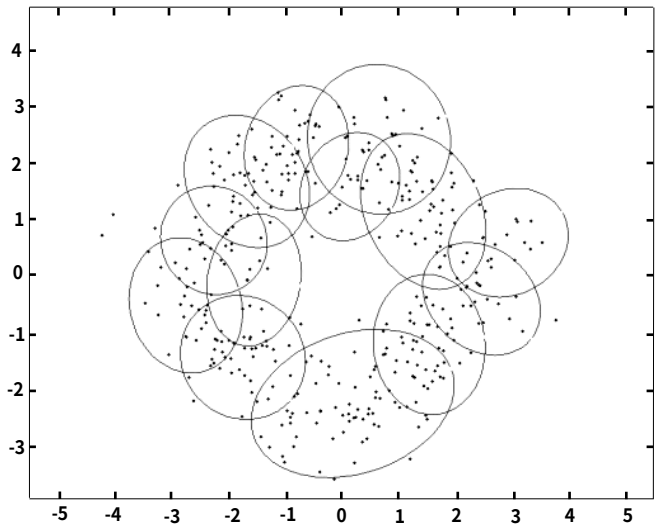


図26b n=13（KR）



本論文冒頭

図27a n=12 (EM)

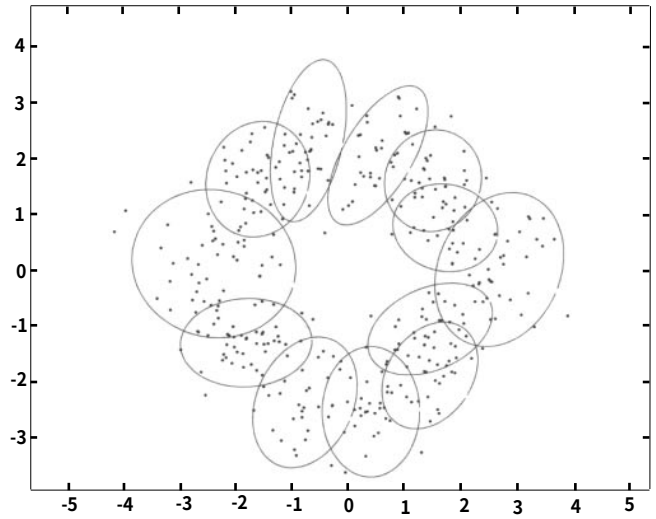


図27b n=12 (KR)

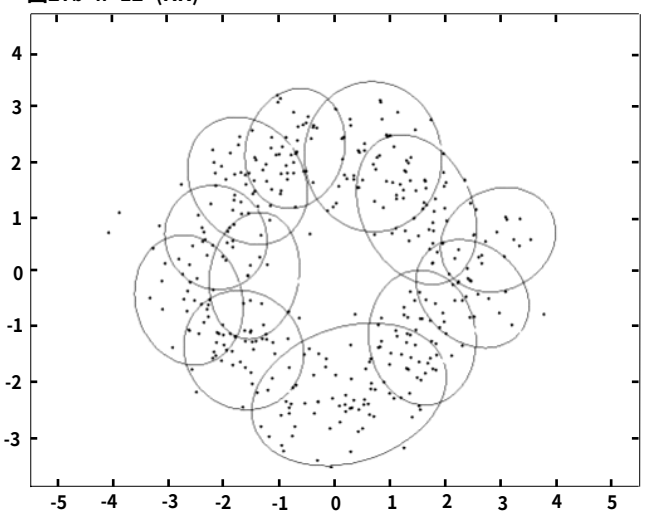


図28a n=11 (EM)

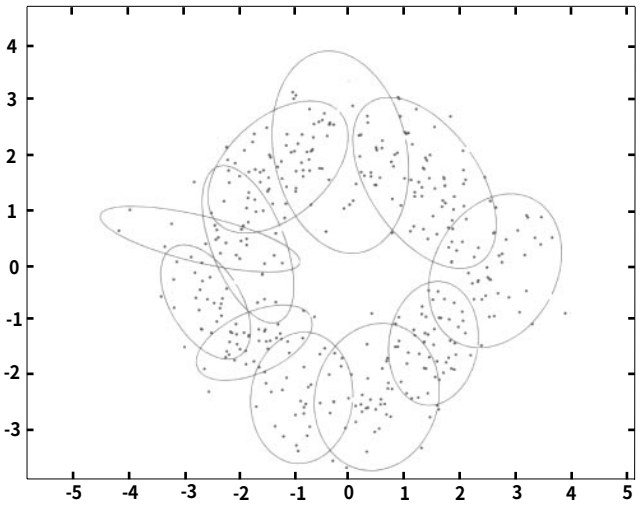
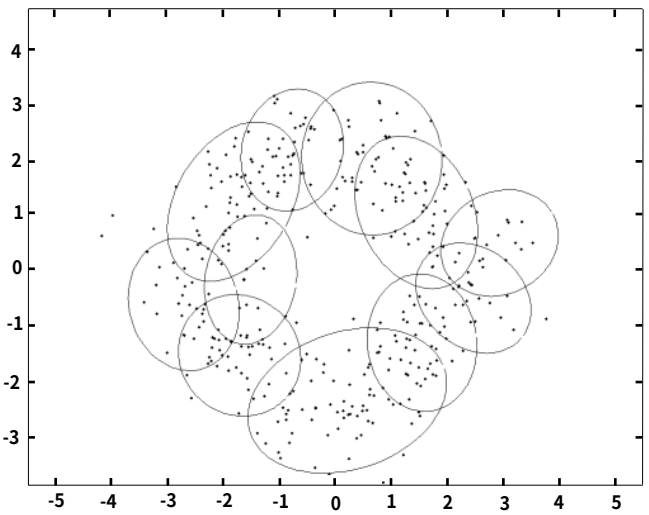


図28b n=11 (KR)



本論文冒頭

図29a n=10（EM)

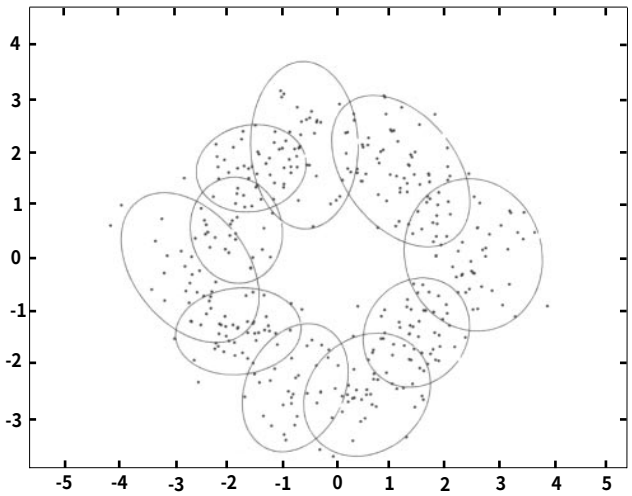
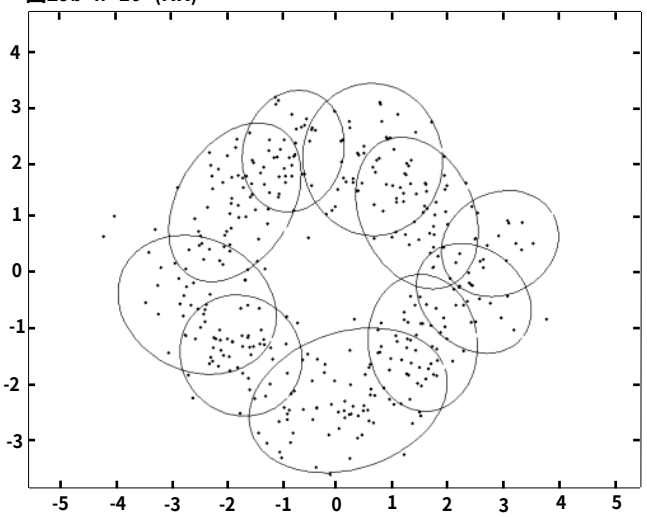


図29b n=10（KR)



本論文冒頭

図30a n=9 (EM)

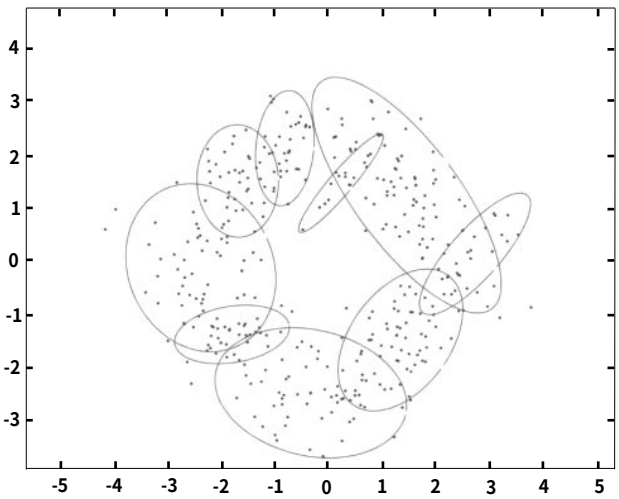


図30b n=9 (KR)

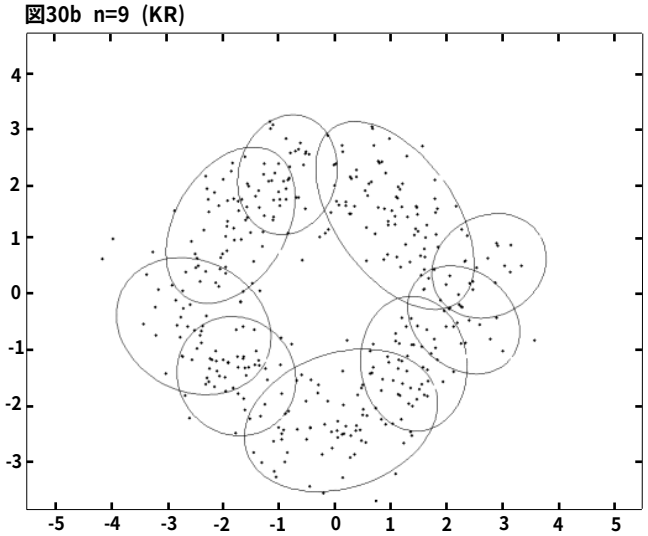


図31a n=8 (EM)

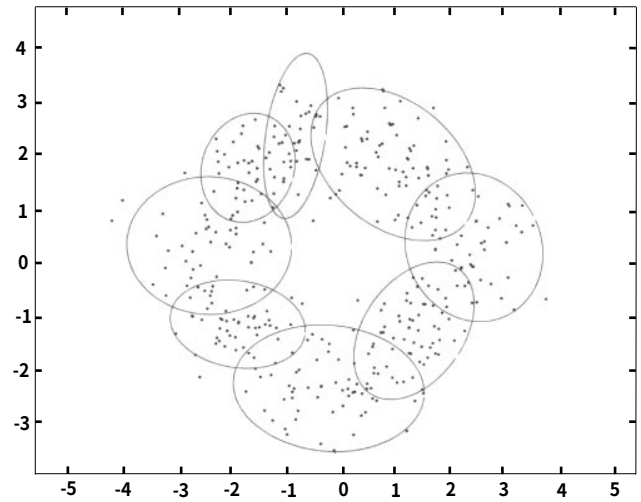
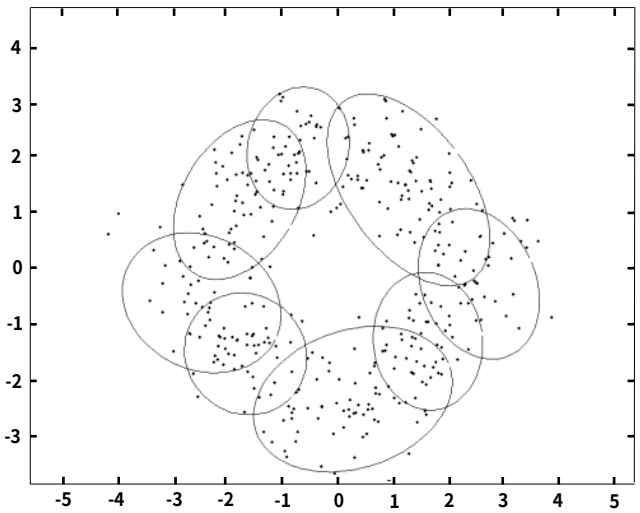


図31b n=8 (KR)



本論文冒頭

図32a n=7（EM)

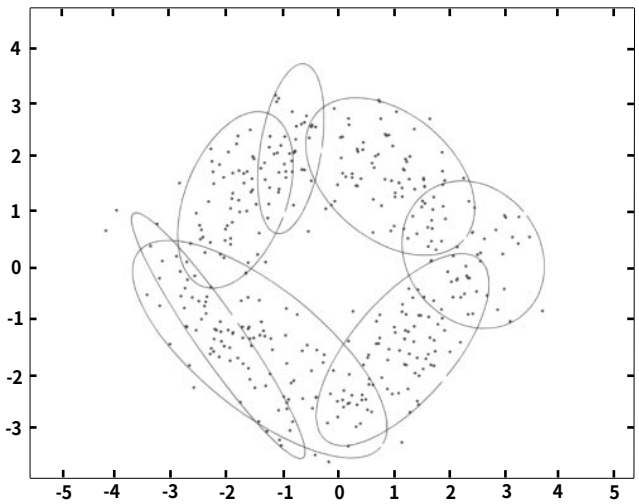
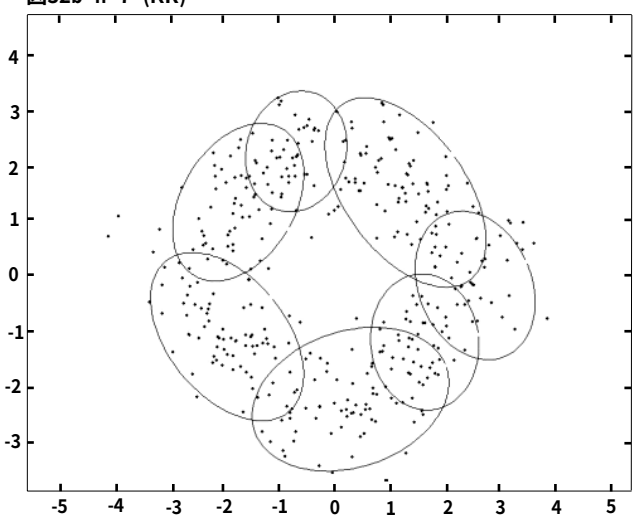


図32b n=7（KR)



本論文冒頭

図33a n=6 (EM)

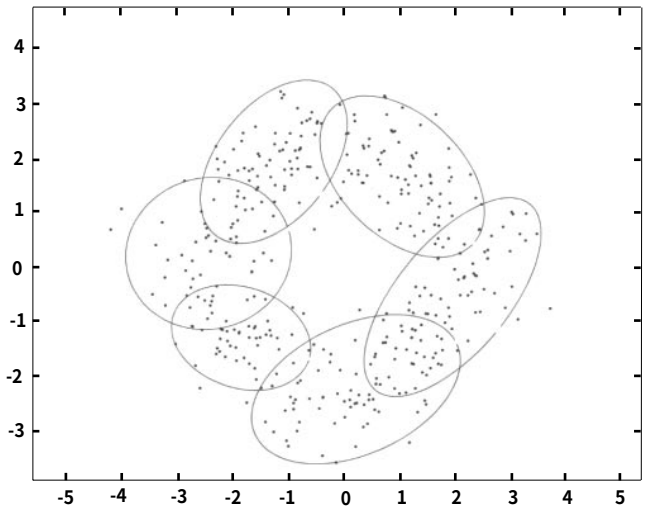
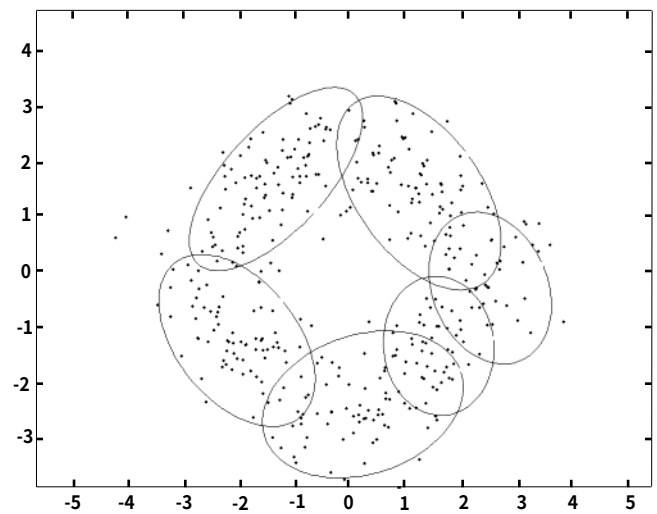


図33b n=6 (KR)



本論文冒頭

図34a n=5（EM）

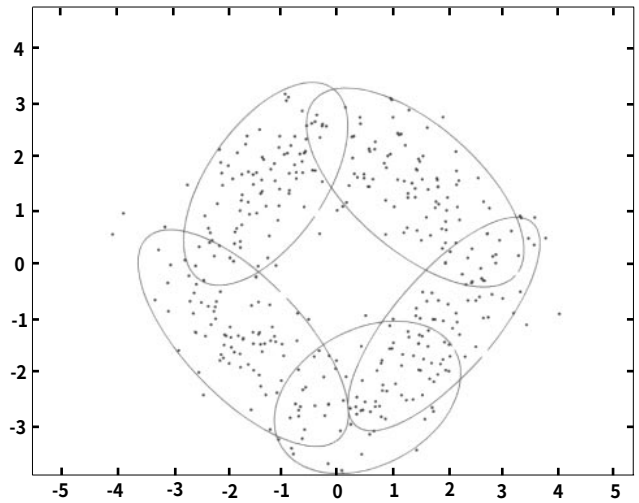
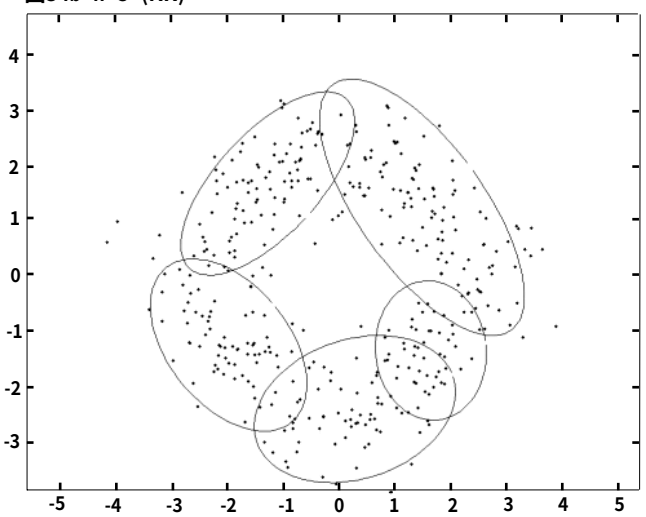


図34b n=5（KR）



本論文冒頭

図35a n=4 (EM)

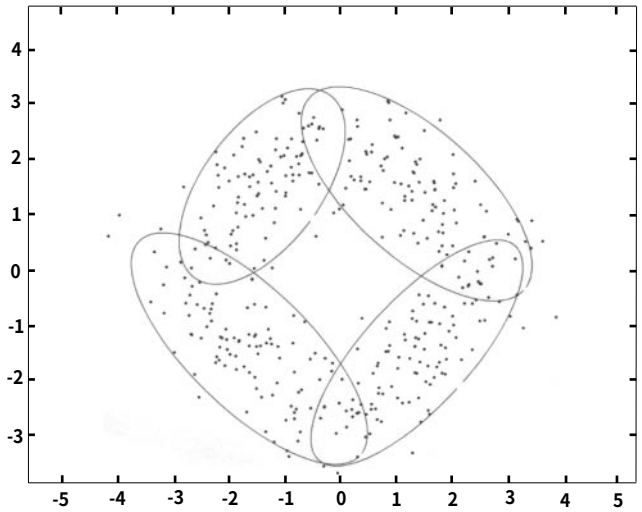
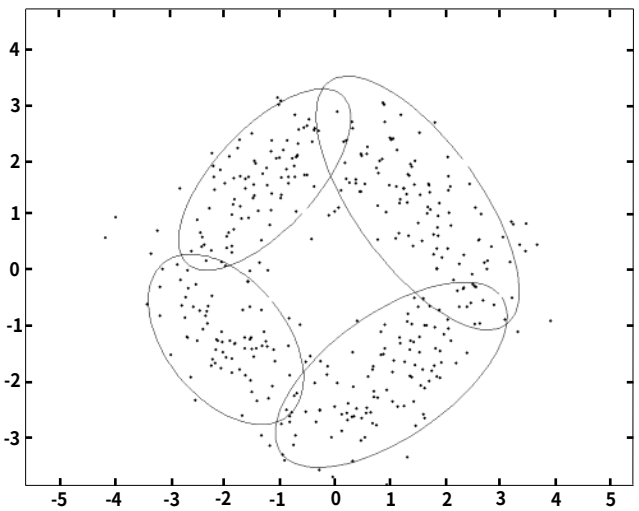


図35b n=4 (KR)



本論文冒頭

図36a n=3 (EM)

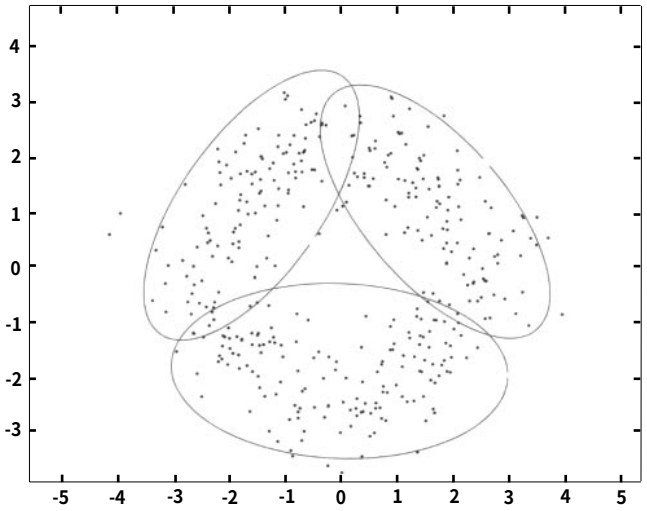
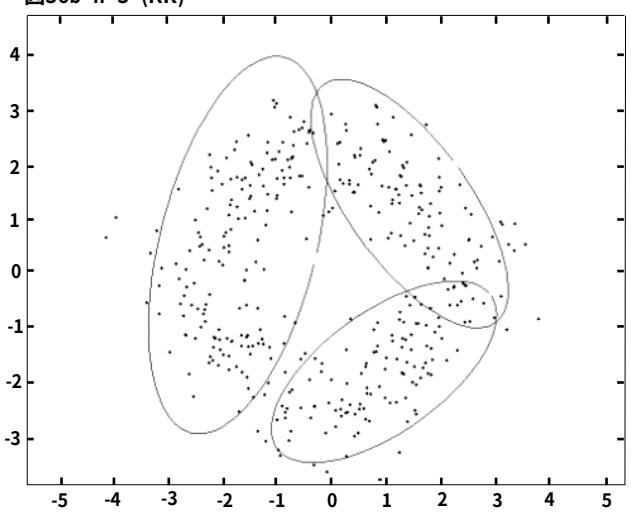


図36b n=3 (KR)



本論文冒頭

図37a n=2 (EM)

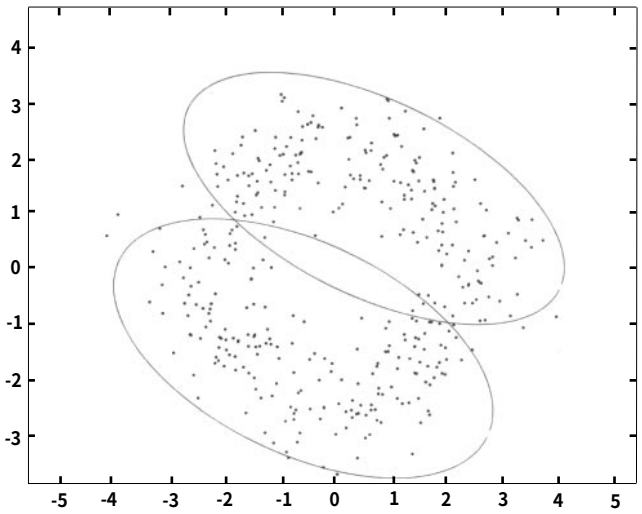
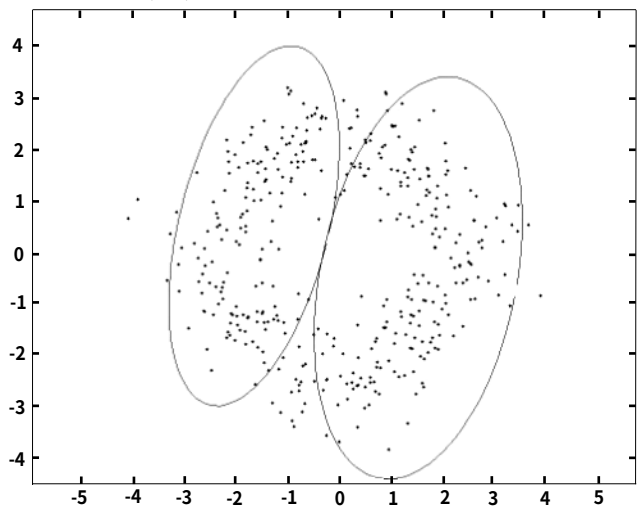


図37b n=2 (KR)



本論文冒頭

表7

Y1 =1.0e+004×	
0	0
0.5090	0.5134
0.8719	0.8787
1.3566	1.3658
1.4027	1.4143
1.4122	1.4262
2.0990	2.1153
1.7318	1.7506
2.5069	2.5280
1.8118	1.8354
2.1416	2.1676
2.2006	2.2289
3.1901	3.2209
5.5161	5.5492
AIC	BIC

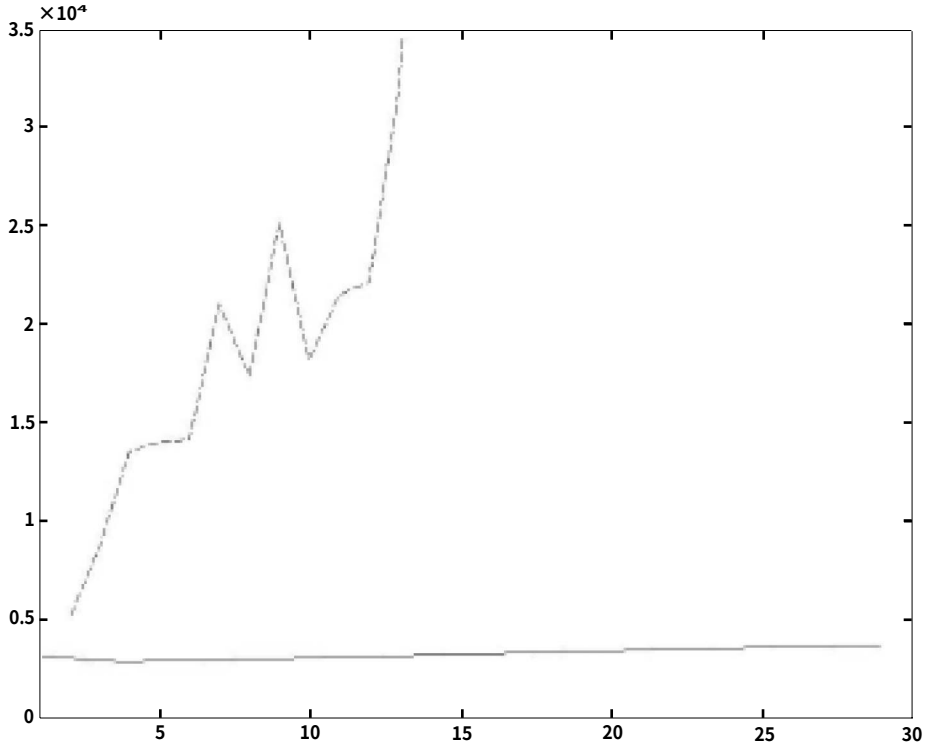
表8

Y2 = 1.0e+003×	
3.1812	3.2012
3.1446	3.1885
2.9667	3.0345
2.8969	2.9887
2.9030	3.0188
2.9082	3.0479
2.9117	3.0753
2.9208	3.1084
2.9168	3.1284
2.9290	3.1645
2.9456	3.2051
2.9567	3.2401
2.9674	3.2748
2.9760	3.3073
2.9913	3.3466
3.0045	3.3837
3.0137	3.4168
3.0250	3.4521
3.0379	3.4889
3.0515	3.5265
3.0484	3.5474
3.0601	3.5830
3.0715	3.6184
3.0821	3.6529
3.0941	3.6888
3.1063	3.7250
3.1169	3.7595
3.1151	3.7816
3.1277	3.8182
AIC	BIC

表９

LL =1.0e+004×	
0	-0.1586
-0.2534	-0.1561
-0.4343	-0.1466
-0.6760	-0.1425
-0.6985	-0.1423
-0.7026	-0.1419
-1.0454	-0.1415
-0.8612	-0.1413
-1.2482	-0.1405
-0.9000	-0.1405
-1.0643	-0.1408
-1.0932	-0.1407
-1.5874	-0.1407
-2.7498	-0.1405
EM	KR

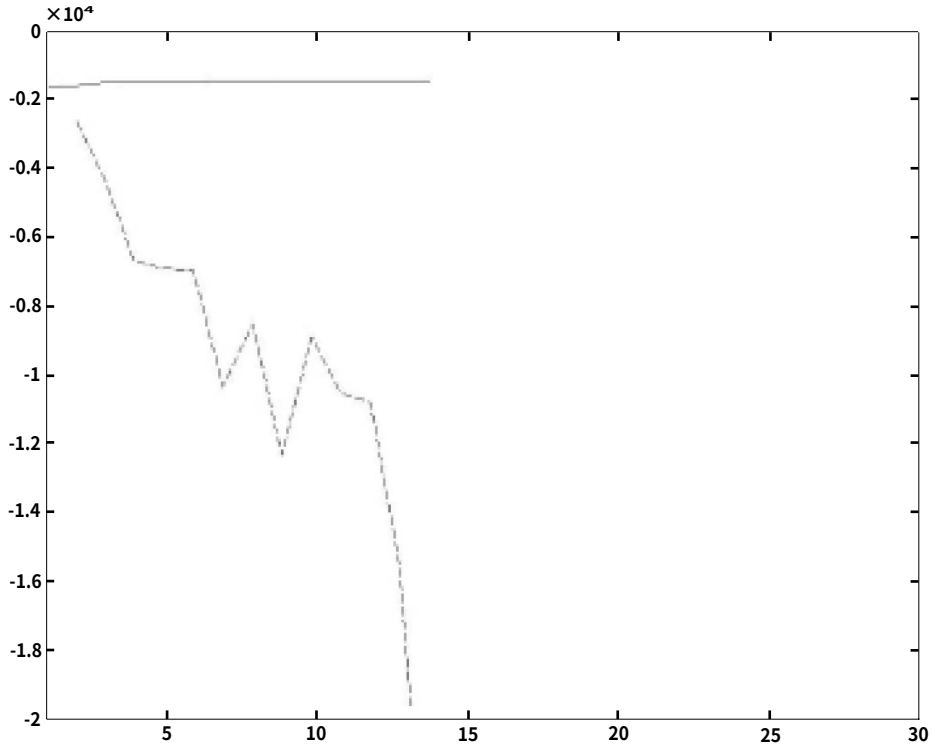
図38



EMによるBIC、点線（表7の第2列）　　KRによるBIC、実線（表8の第2列）

本論文冒頭

図39



EMによるlog likelihood、点線（表9の第1列）　　KRによるlog likelihood、実線（表9の第2列）
後続の図との比較のために、横軸は1から30まで取ってある。

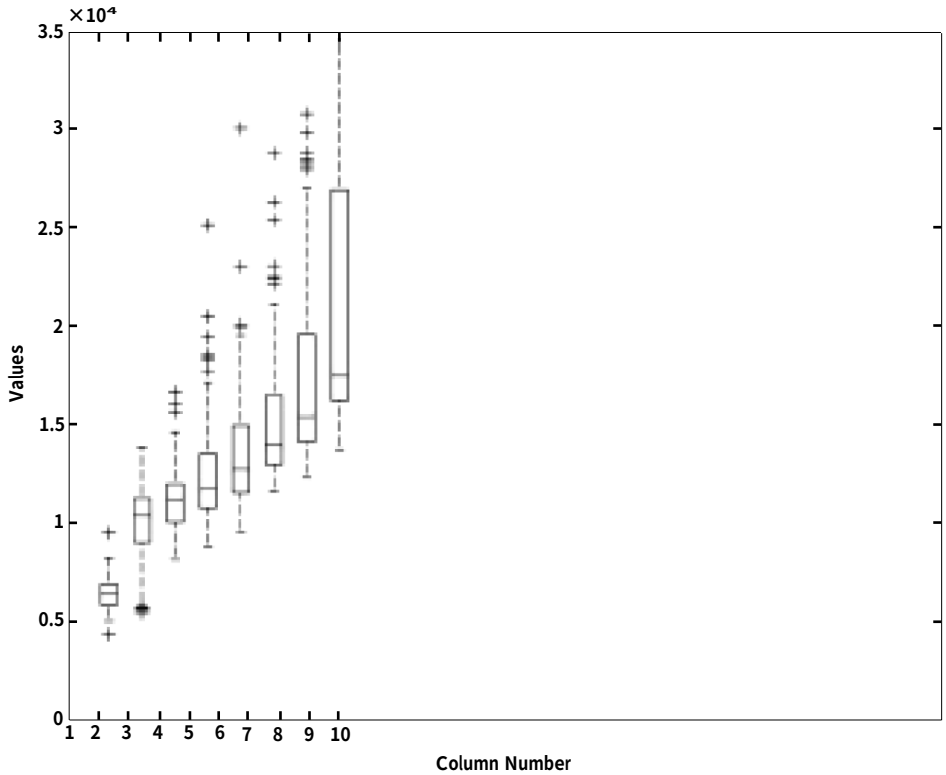
本論文冒頭

図25a-37aを通して見ると，異なったクラスター数のモデルの間には，ほとんど関連が無いことがわかる。クラスター数が変わるごとに，新しい初期値がk-meansによって与えられる。それから，EMは期待値関数の最大化に向かって動き出すが，大域的な最大値は保証されない。一方，KRでは，継続するクラスター数において，モデルは，あまり大きな違いの無いクラスターの集合体である。この違いは，二つの方法によるICの系列に正直に反映されている。Y1と名付けた表6の行列は，EMのICが激しく上下に動くことを示している。例えば，BICはn＝14から10まで減少し，突然2.5280に飛び上り，また減少する。次に飛び上がるときには2.1153であり，その後は最後まで単調減少する。これに対して，KRのBICは真のクラスター数の前後で単調であって，このパターンに逆転は無い。n＝29から真のクラスター数n＝4までは単調減少，n＝4からn＝1までは単調増大である。全てのnで，EMのBICはKRのBICよりも高いのだから（Y1, Y2における小数点は，それぞれ1.0e+004, 1.0e+003であることに注意），EMの尤度はKRの尤度よりもずっと小さいことが予想される。実際にその通りであって，LL行列（表9）の同じ行の二つの値を比べてみれば，KRはEMよりも，常に良いlog likelihoodを出していることがわかる。（LLの（1, 1）要素は0としておいた。EMでは1クラスター・モデルの推定は行なわれないからである。）さらに，nが大きくなるにつれて，KRのlog likelihoodは徐々に改善するが，EMは悪くなる！（ただし，KRには小さな逆転が，EMには大きな逆転が伴う。）EMはlog likelihoodを直接に最大化するものではないが，大きなnに対してlog likelihoodが悪くなるということは，大きなnにおいて，局所的な最大値に捕われる問題が，深刻になることを示している。BICの激しい上下動は，EMが，当てはめの際に，他のnにとらわれない柔軟性を持っていることを意味する。それでは，小さなnについては，どうなっているのだろう？　何故このprocedureは，いつも同じ2クラスター・モデルを選んだのか？　その答えは，小さなnにおける最適化のやさしさと，各nにおけるフィットの柔軟性であろう。これについて考えるために，4クラスターから3クラスターへの降下を，EMとKRとで比較してみよう。EMでは，BICは1.3658×10^4から0.8787×10^4に変わり，0.4871×10^4の改善である（値においては減少）。一方，KRでは，BICは0.29887×10^4から0.30345×10^4に変わり，0.00458×10^4の悪化である（値においては増加）。図を比べれば，この違いの理由は明らかである。EMにおいては，3クラスター・モデルへの適応は柔軟である。4クラスター・モデルの形には全く構わない。さらに，小さなクラスター数では，最適化はやさしくなる。そこで，log likelihoodとBICの大きな改善が可能になった。しかし，これはクラスター数決定にとっては，良くないことである。なぜならば，正しくないクラスター数での改善だからである。これに比べて，KRによる3クラスター・モデルへの適応は柔軟性に欠けている。このやり方では，2つの旧クラスターを残さなければならない。我々の例では，3クラスター・モデルの新クラスターは，観測値の配置に，きつくフィットしているとは見えない。そこで，log likelihoodとBICは悪化する。しかし，これは良いことである。なぜならば，正しくないクラスター数での悪化だからである。3クラスター・モデルから2クラスター・モデルへの降下にも，同じ説明が当てはまる。ここで，我々は，KR固有の硬直性が，欠点としてではなく，長所として働いているのがわかる。

100回のシミュレーションを1つの図に描いたら，どうなるか，予想を立ててみよう。つまり，n＝1,……,29を横軸に取り，各nに100個のBICまたはlog likelihoodの値をプロットするのである。（KRでは，Phase3はn＝30から始まる。即ち，Phase3の最初の結果であり，最初に

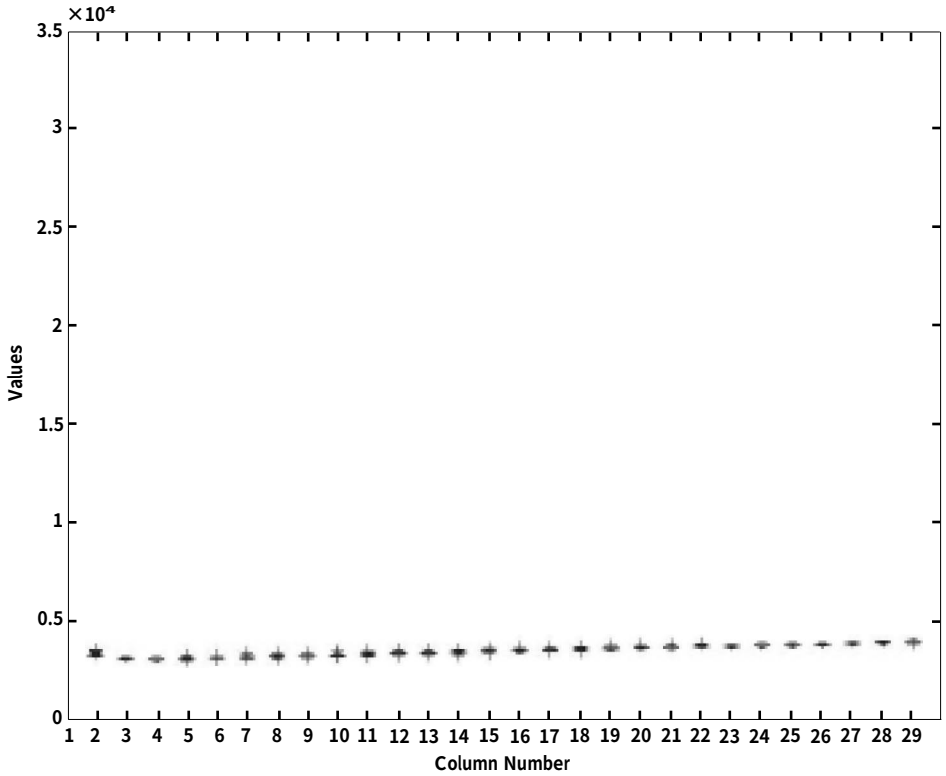
BICを計算されるのは，29クラスター・モデルである。）EMのBICの上昇トレンドは，最適化の困難が増していくことによるのだから，100回のシミュレーションをまとめた図でも，同じトレンドを見出せるだろう。しかし，激しい上下動の起こるnは決まっていない。そのようなnはデータセット中の観測値の散らばり方によるのである。そこで，新しい図では，激しい上下動は，nの増加に伴うBICの変動性の増加として現れるだろう。トレンドと変動性を考えに入れると，新しい図は，若干右に傾いた扇形になるはずである。一方，KRのBICは各nにおいて，ずっと小さな変動性しか示さないだろう。log likelihoodの状況も同様に推論される。下降トレンドと増加する変動性から，EMのlog likelihoodの図は，水平線より下に扇を傾けた形になるだろう。KRの場合，log likelihoodもまた，各nでの変動性は小さいはずである。29本の縦線のそれぞれに100個の値が張り付いていては，見やすい図とは言えない。box-plotによって，視覚的にわかりやすくしよう。我々の予想が正しかったことは，以下の4枚の図で確かめられる。

図40



EMの100回のシミュレーションによって算出されたBICのbox-plot
boxの中の横棒はメディアンである。boxの上端は Q_3（第3四分位）、下端は Q_1（第1四分位）である。 $Q_3 - Q_1 = IQR$ （四分位間距離）であり、区間 $\left(Q_1 - 1.5\times IQR,\quad Q_3 + 1.5\times IQR\right)$ の外側の観測値は全てoutlierとみなす。上方または下方のどちらかで、outlierが無いならば、その方向のヒゲの端は最大値または最小値に置かれる。どちらかの方向にoutlierがあるならば、その方向のヒゲの端は、上で定義した区間内の最大値または最小値に置かれる。全てのoutlierは＋で表わす。n=9と10では、図の枠の上にさらにBICの値がある。

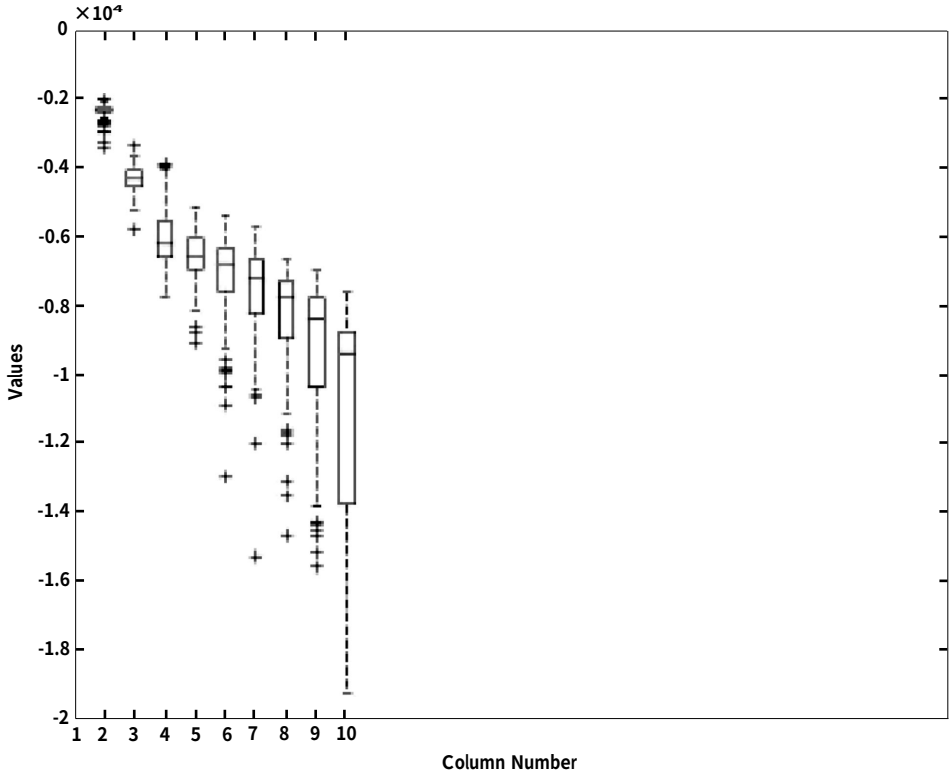
図41



KRの100回のシミュレーションによって算出されたBICのbox-plot

本論文冒頭

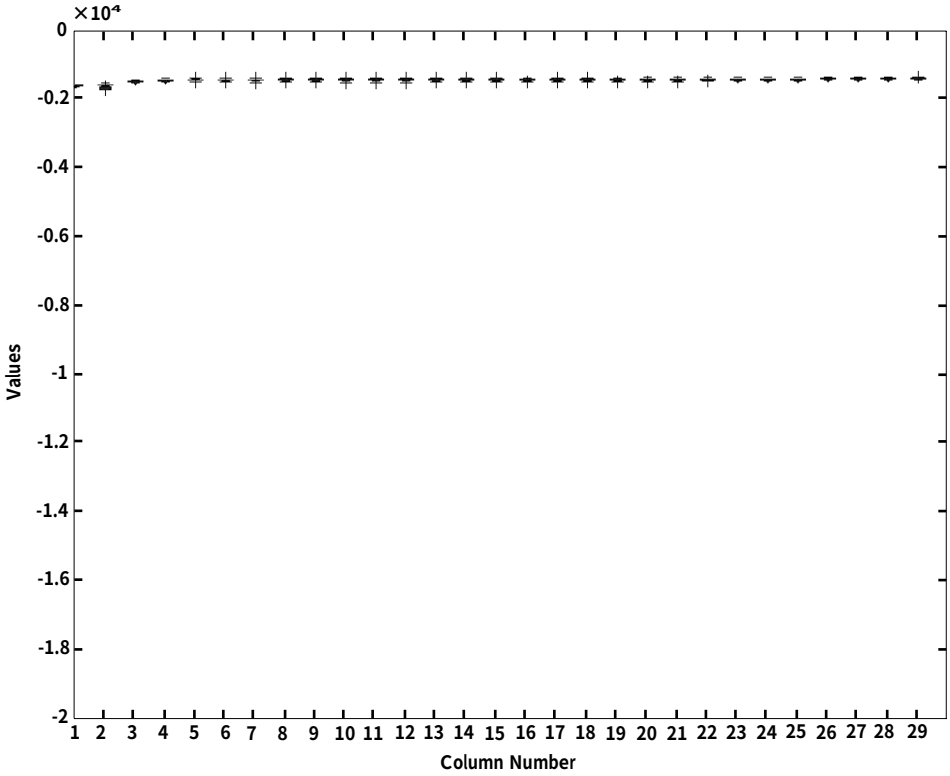
図42



EMの100回のシミュレーションによって算出されたlog likelihoodのbox-plot
n=10では図の枠の下にさらにlog likelihoodの値がある。

本論文冒頭

図43



KRの100回のシミュレーションによって算出されたlog likelihoodのbox-plot

本論文冒頭

KRのBICは，各nにおいて変動性が小さいだけでなく，全てのnを通じて，ほとんど一定の変動性を示している。この特徴によって，我々は，n＝29からn＝4までは下降し，n＝4からn＝1までは上昇するトレンドを読むことができる。KRのlog likelihoodも小さく一定の変動性を示している。1データの場合，表9から読みとれた特徴，即ち，ほとんど一定ではあるが，nが増加するとともに，若干改善するということが，ここでも読み取れる。4つのbox-plotのうち，最も興味深いのは，図42であろう。EMのlog likelihoodのメディアンは，ゆっくりと下降している。しかし，ヒゲの下端は，これよりも速く下降する。nごとの垂直線に沿って，紙面から立ち上がるように棒グラフを描けば，nの増加に伴って，棒グラフの歪度は増していく。即ち，下方のtailが長くなる。この現象は，procedureに統計理論を当てはめることによって，説明すべきではあるまい。大半のシミュレーションでは，最適化の困難はゆっくりと増して行くが，例外的に悪い場合の数と程度は，より急速に悪化する。そう説明するのが，経験分布の変化にふさわしい。

図41は，表7で見出した完全に単調なパターンが，多くのシミュレーションで出ていることを期待させる。即ち，n＝29からn＝4までは単調減少し，n＝4からn＝1までは単調増加するパターンである。「逆転数」を

（逆転数）＝（n＝29からn＝4まで降下する間に，BICが増えているステップの数）
＋（n＝4からn＝1まで降下する間に，BICが減っているステップの数）

と定義すると，数え上げの結果は次のようになる。

表10

	全100回のうち	成功85回のうち
逆転0	39	39
1	48	44
2	10	2
3	3	0

逆転数1-3は28のステップの中のわずかな部分に過ぎないから，逆転数0のパターンを基本パターンと見なして良いだろう。この事実が，KRの性質として，どこまで普遍性を持つかは，もちろん，一例だけのシミュレーションからはわからない。しかし，このような完全なパターンが数多く出て来るということは，クラスター数決定にとっては有利である。そこで，これを例えば，「両側単調性」と名付けて，理論的に研究するのは，意義あることであろう。

12．要約および将来の研究の方向

この章では，本論文の内容を要約し，将来の研究の方向について考察する。1-5章では，混合分布の意味，既存の方法の概要，ScottとSzewczykの研究のアウトラインについて述べた。このうち，混合分布の意味については，マーケティングと心理学から例を取り，人文社会科学系の研究者に親しみ安い解説を心がけた。しかし，3-5章の内容については，それ自体簡潔な

要約なので，ここではくり返さない。6-11章に含まれる，新しい発見・考察・提案について要約する。

まず，SSが発表したのは4-phase procedureであった。しかし，本論文では3 phaseに縮めた形で考察している。SSは彼らのprocedureに含まれる，新しいアイディアを充分に説明しなかったので，その意味を明らかにすることが，第一の仕事である。3-phase procedureは，この目的にふさわしいように作られた。さらに，2次元正規混合分布の図は，SSの新概念を理解するのに，1次元正規混合分布の図よりも，わかりやすい。そこで，我々の第二の仕事である，2次元正規混合分布への拡張は，新概念の説明と同時に行われる。融合公式の導出とMEASURE1（1 vs. 1クラスター相似測度）とMEASURE2（n vs. n-1 クラスター相似測度）の相違点は，最もていねいに考察されている。後者はprocedureの成功にとって，決定的に重要なので，独立の章を当てている。本論文では，SSのprocedureと，我々による，その拡張は，カーネル降下法（KR）と呼ばれる。KRを2次元に拡張するには，様々な変更と追加が必要である。6-8章で述べられた，2次元procedureでは，以下の点が新しい。Phase1のカーネル推定は$\sigma_x = \sigma_y$, $\rho = 0$を仮定して行ったこと。2次元用の融合公式・相似測度・L2Eを開発したこと。我々は，SSのアプローチから大きく離れないように努めたが，6-8章で展開した方法は，以下の点で，彼らの方法と異なっている。MEASURE1は彼らの論文で提案されているが，実際には，procedureで使われていない。彼らは，それを省略計算で置き換えてしまった。しかし，我々の2次元procedureでは，Phase2で用いている。この変更によって，彼らの本来のねらいとMEASURE1と2の違いが明らかになる。

新しいprocedureは，シミュレーションによってテストし，非常に良い結果が得られた。100回のシミュレーション中，正しいクラスター数は，85回であった。既に述べたとおり，10章はMEASURE1と2の比較に当てられている。11章は重要である。9章のシミュレーションで用いたデータに，EMアルゴリズムを当てはめた。この比較の結果は，KRがEMに優っていた。BICおよびlog likelihoodを検討してみると，EMでは，最適化アルゴリズムが失敗する確率が，クラスター数nに強く相関していることがわかる。この確率はクラスター数が小さければ低く，大きければ高い。これによって，クラスター数の決定は深刻な影響を受ける。これに対して，KRは，そのような問題から免れている。

本論文で，KRのパフォーマンスが良好であることがわかったので，将来は，このprocedureに関して様々な研究テーマが考えられる。1. phase間の境界の自動設定 本論文で構築したprocedureでは，Phase 2と3の間の境界は，天下り的に設定した。SSが，彼らの試行錯誤から提案した境界を，そのまま用いたのである。しかし，データが大きかったり，複雑な混合分布が予想される場合に，境界は柔軟に動かせる方がよい。その他のテーマとしては，2. KRの理論的背景を明らかにする。3. さらに高次元に拡張する。4. 正規分布以外の分布形の混合分布に拡張する。などが考えられる。尚，本論文では紙幅の関係で含められなかったが，Kaneda [6]では，4. の最初の試みとして，ガンマ混合分布への拡張を行なっている。

参考文献

[1] Dempster, A.P., Laird, N.M, and Rubin, D.B. “Maximum likelihood from incomplete data via the EM algorithm” *J. Roy. Statist. Soc. Ser. B.* v39. 1-22 (1977)

[2] Dillon, W. and Kumar, A. “Latent structure and other mixture models in marketting : an integrative survey and overview” in *Advanced Methods in Marketing Research,* Richard P. Bagozzi (ed.), pp.295-351,Blackwell(1994)

[3] Everitt, B. S. “An introduction to finite mixture distributions”, *Statistical Methods in Medical Research,* vol.5, pp.107-127

[4] Gilks, W.R., Richardson, S. and Spiegelhalter. D.J. (ed.), *Markov Chain Monte Carlo in Practice,* Chapman and Hall (1996)

[5] P. G. ホーエル『入門数理統計学』培風館（1978）

[6] Kaneda, N. “Fitting mixture models from kernel estimators”, Ph.D.dissertation, U. C. Santa Barbara (2007) (available through ProQuest)

[7] Lewine, R. R. J. “Sex differences in schizophrenia : timing or subtypes?”, Psychological Bulletin, vol. 90, pp. 432-444 (1981)

[8] McLachlan, G. and Peel, D. *Finite Mixture Models,* Wiley(2000)

[9] Pearson, K. (1894) “Contributions to the Mathematical Theory of Evolution,” *Philosophical Transactions of the Royal Society of London, Ser. A*, 185, pp.71-78

[10] Scott, D.W. “Parametric Statistical Modeling by Minimum Integrated Square Error”, *Technometrics,* vol.43, pp. 274-285 (2001)

[11] Scott, D.W. and Szewczyk, W.F. “From Kernels to Mixtures”, *Technometrics*, vol.43, pp. 323-335 (2001)

[12] Titterington, D.M., Smith, A.F.M., and Makov, U.E. *Statistical Analysis of Finite Mixture Distributions,* Wiley(1985)

[13] Tsui, Patrick “EM_GM algorithm” available at http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objecteId=8636&objectType=file

[14] Wand, M.P. and Jones, M.C. *Kernel Smoothing,* Chapman and Hall (1995)

[15] Wedel, M. and Kamakura, W. A. *Market Segmentation : Conceptual and Methodological Foundations,* Kluwer (1998)

本論文冒頭