模式识别 4 Programming2

自博 16　陈斯杰 2016310721

1. Run the EM algorithm based on data2 provided by hw5em2.mat with m = 2, 3, 4, 5 components. **Select the appropriate model** (number of components) and **give reasons** for your choice.

Note that you may have to rerun the algorithm a few times (and select the model with the highest log-likelihood) for each choice of m as EM can sometimes get stuck in a local minimum.

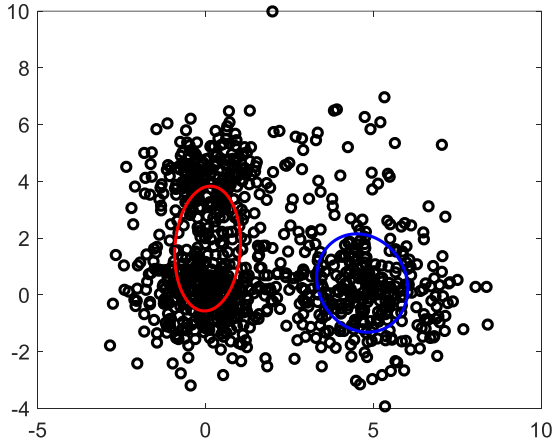
Is the model selection result **sensible** based on what you would expect visually? Why or why not?

选用 1e-8 作为 eps 的收敛阈值

m=2 时：

[param, hist, ll]=em_mix(data2, 2, 1e-8)

对数似然收敛到-4.2309*1e3



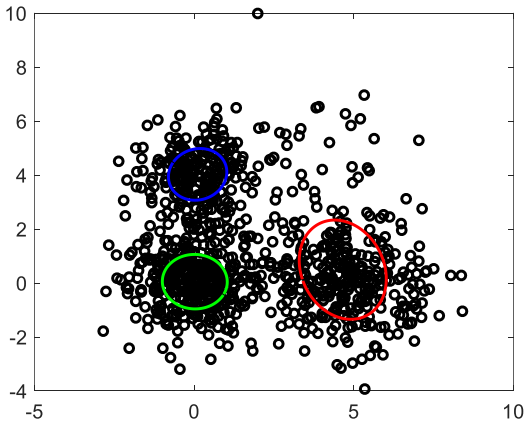
参数估计结果如下：

字段	mean	cov	p	prior_cov	prior_p	prior_n
1	[4.6627,0.4186]	[1.8399,-0.2811;-0.2811,3.0010]	0.3561	[5.3268,0;0,5.3268]	0.5000	1
2	[0.0576,1.6325]	[0.9593,0.1952;0.1952,4.8342]	0.6439	[5.3268,0;0,5.3268]	0.5000	1

m=3 时:

[param, hist, ll]=em_mix(data2, 3, 1e-8)

对数似然收敛到-4.1289*1e3



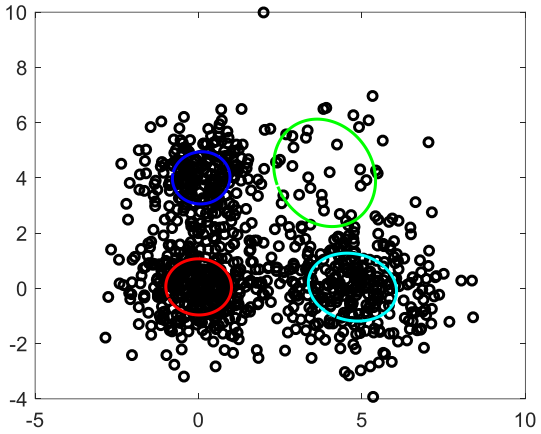
参数估计结果如下：

字段	mean	cov	p	prior_cov	prior_p	prior_n
1	[0.1081,4.0268]	[0.8357,0.0800;0.0800,0.9186]	0.2493	[5.3268,0;0,5.3268]	0.3333	1
2	[4.6538,0.4949]	[1.8588,-0.4654;-0.4654,3.3825]	0.3571	[5.3268,0;0,5.3268]	0.3333	1
3	[0.0194,0.0508]	[1.0338,-0.0098;-0.0098,1.0195]	0.3935	[5.3268,0;0,5.3268]	0.3333	1

m=4 时：

[param, hist, ll]=em_mix(data2, 4, 1e-8)

对数似然收敛到-4.0885*1e3



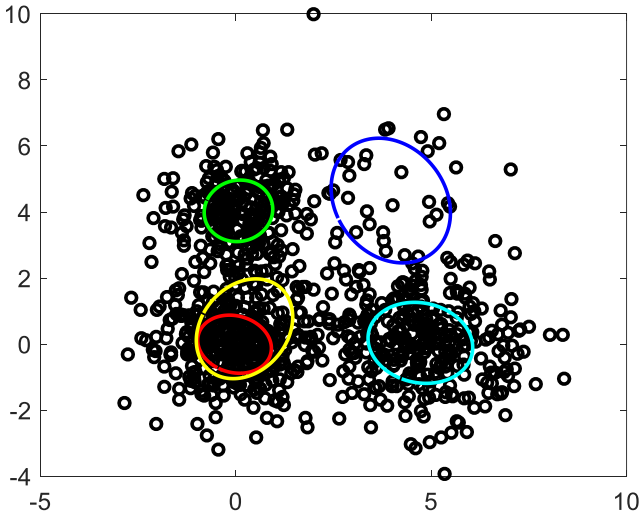
参数估计结果如下：

字段	mean	cov	p	prior_cov	prior_p	prior_n
1	[0.0818,4.0011]	[0.7767,0.0231;0.0231,0.8928]	0.2447	[5.3268,0;0,5.3268]	0.2500	1
2	[0.0037,0.0531]	[1.0092,-0.0134;-0.0134,1.0329]	0.3914	[5.3268,0;0,5.3268]	0.2500	1
3	[3.8584,4.1824]	[2.4285,-0.4470;-0.4470,3.7825]	0.0447	[5.3268,0;0,5.3268]	0.2500	1
4	[4.7102,0.0448]	[1.8198,-0.2019;-0.2019,1.5116]	0.3191	[5.3268,0;0,5.3268]	0.2500	1

m=5 时：

[param, hist, ll]=em_mix(data2, 4, 1e-8)

对数似然收敛到-4.0912*1e3



参数估计结果如下：

字段	mean	cov	p	prior_cov	prior_p	prior_n
1	[0.0818,4.0011]	[0.7767,0.0231;0.0231,0.8928]	0.2447	[5.3268,0;0,5.3268]	0.2500	1
2	[0.0037,0.0531]	[1.0092,-0.0134;-0.0134,1.0329]	0.3914	[5.3268,0;0,5.3268]	0.2500	1
3	[3.8584,4.1824]	[2.4285,-0.4470;-0.4470,3.7825]	0.0447	[5.3268,0;0,5.3268]	0.2500	1
4	[4.7102,0.0448]	[1.8198,-0.2019;-0.2019,1.5116]	0.3191	[5.3268,0;0,5.3268]	0.2500	1

模型的选择

计算这四个模型的 BIC

BIC=ln(PointsCount)*ParameterCount-2*ln(Likelihood)

m=2, BIC= ln(1000)*(2*3) -2*(1e3*-4.2309)= 8503.246531673894

m=3, BIC= ln(1000)*(3*3) -2*(1e3*-4.1289)= 8319.96979751084

m=4, BIC= ln(1000)*(4*3) -2*(1e3*-4.0885)= 8259.893063347785

m=5, BIC= ln(1000)*(5*3) -2*(1e3*-4.0912)= 8286.016329184731

当 m=4 时， BIC 最小，所以我们有理由认为四个高斯分量是最好的选择。

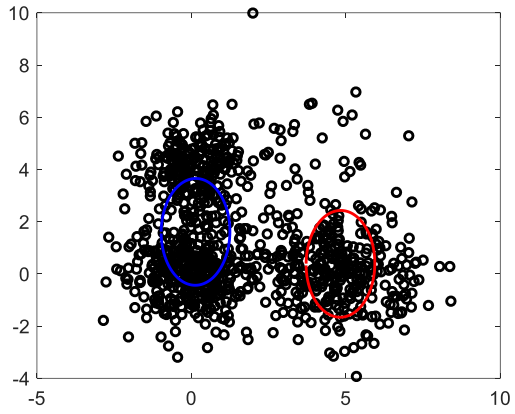
这个结果与视觉上的结果是一致的，四个高斯分量时，右上角的较为稀疏的一坨点能够被较好地覆盖到。

2. Modify the M-step of the EM code so that the covariance matrices of the Gaussian components are constrained to be equal. Give detailed derivation. Rerun the code and then select a appropriate model. **Would we select a different number of components** in this case?

等方差情形推导见纸质文件。

m=2 时,

对数似然为-4.2583*1e3

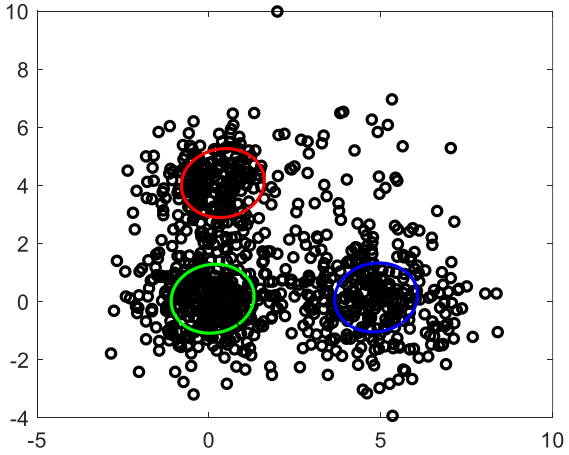


参数估计结果如下：

字段	mean	cov	p	prior_cov	prior_p	prior_n
1	[0.1347,1.6082]	[1.2378,0.0203;0.0203,4.1836]	0.6669	[5.3268,0;0,5.3268]	0.5000	1
2	[4.8263,0.3834]	[1.2378,0.0203;0.0203,4.1836]	0.3331	[5.3268,0;0,5.3268]	0.5000	1

m=3 时,

对数似然为-4.1725*1e3

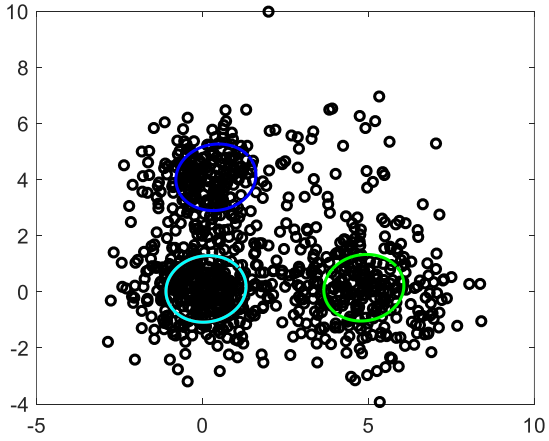


参数估计结果如下：

字段	mean	cov	p	prior_cov	prior_p	prior_n
1	[4.8631,0.1468]	[1.4543,0.1006;0.1006,1.4088]	0.3172	[5.3268,0;0,5.3268]	0.3333	1
2	[0.4042,4.0838]	[1.4543,0.1006;0.1006,1.4088]	0.2721	[5.3268,0;0,5.3268]	0.3333	1
3	[0.1075,0.1040]	[1.4543,0.1006;0.1006,1.4088]	0.4106	[5.3268,0;0,5.3268]	0.3333	1

m=4 时,

对数似然为-4.1768*1e3

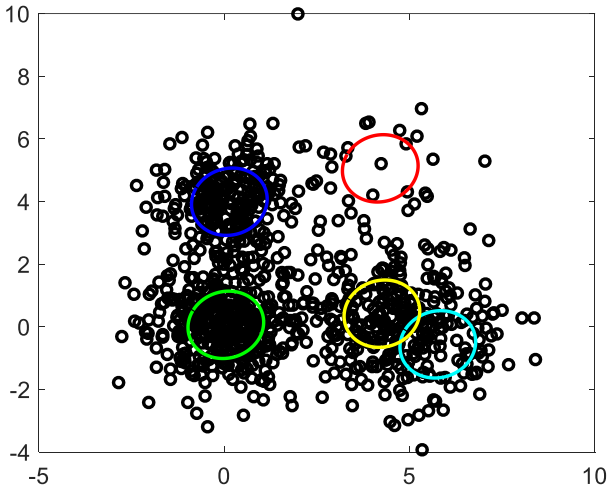


参数估计结果如下：

字段	mean	cov	p	prior_cov	prior_p	prior_n
1	[0.4045,4.0840]	[1.4549,0.1008;0.1008,1.4083]	0.2721	[5.3268,0;0,5.3268]	0.2500	1
2	[0.1090,0.1037]	[1.4549,0.1008;0.1008,1.4083]	0.2263	[5.3268,0;0,5.3268]	0.2500	1
3	[4.8633,0.1465]	[1.4549,0.1008;0.1008,1.4083]	0.3171	[5.3268,0;0,5.3268]	0.2500	1
4	[0.1064,0.1045]	[1.4549,0.1008;0.1008,1.4083]	0.1846	[5.3268,0;0,5.3268]	0.2500	1

m=5 时,

对数似然为-4.1139*1e3



参数估计结果如下：

字段	mean	cov	p	prior_cov	prior_p	prior_n
1	[4.2132,5.0585]	[1.0429,0.0865;0.0865,1.1562]	0.0253	[5.3268,0;0,5.3268]	0.2000	1
2	[0.1403,3.9933]	[1.0429,0.0865;0.0865,1.1562]	0.2562	[5.3268,0;0,5.3268]	0.2000	1
3	[0.0452,0.0621]	[1.0429,0.0865;0.0865,1.1562]	0.3968	[5.3268,0;0,5.3268]	0.2000	1
4	[5.7663,-0.5613]	[1.0429,0.0865;0.0865,1.1562]	0.1112	[5.3268,0;0,5.3268]	0.2000	1
5	[4.2577,0.4177]	[1.0429,0.0865;0.0865,1.1562]	0.2105	[5.3268,0;0,5.3268]	0.2000	1

模型的选择

计算这四个模型的 BIC

BIC=ln(PointsCount)*ParameterCount-2*ln(Likelihood)

m=2, BIC= ln(1000)*(2*3) -2*(-4.2583*1e3)= 8558.046531673894

m=3, BIC= ln(1000)*(3*3) -2*(-4.1725*1e3)= 8407.16979751084

m=4, BIC= ln(1000)*(4*3) -2*(-4.1768*1e3)= 8436.493063347785

m=5, BIC= ln(1000)*(5*3) -2*(-4.1139*1e3)= 8331.416329184733

在等方差假设下，当 m=5 时， BIC 最小，根据 BIC 准则应该选取五个高斯分量。