**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?

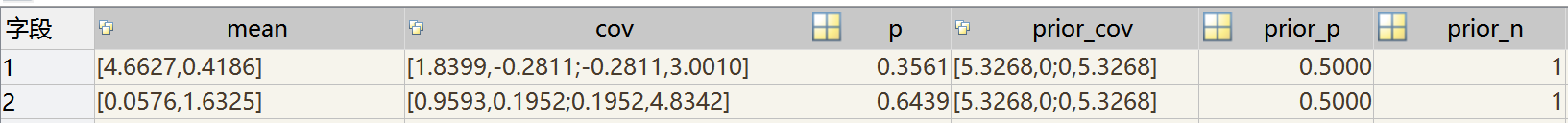
选用1e-8作为eps的收敛阈值

**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： 

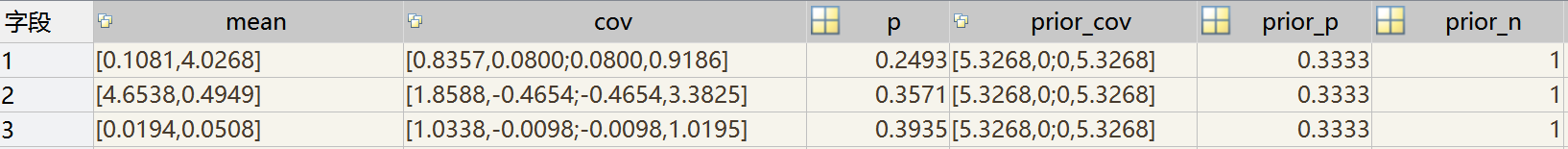
**m=3时:**

[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：



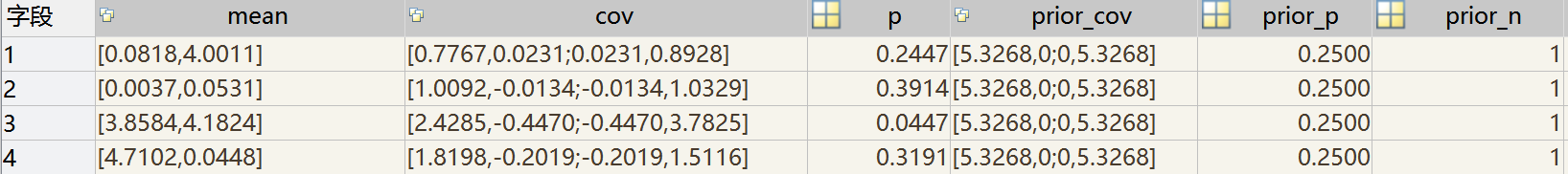
**m=4时：**

[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：



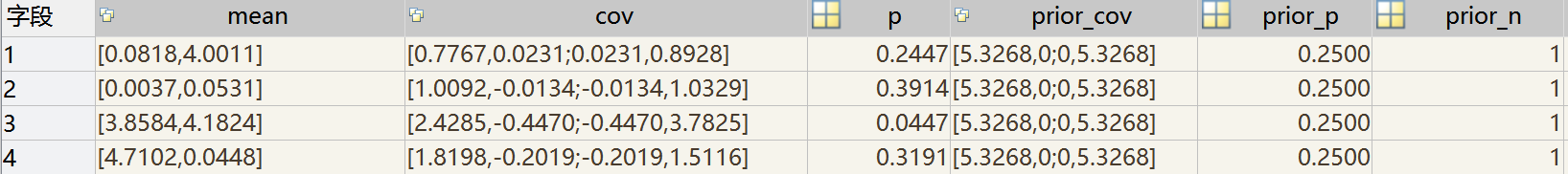
**m=5时：**

[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：



**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。

1. 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?

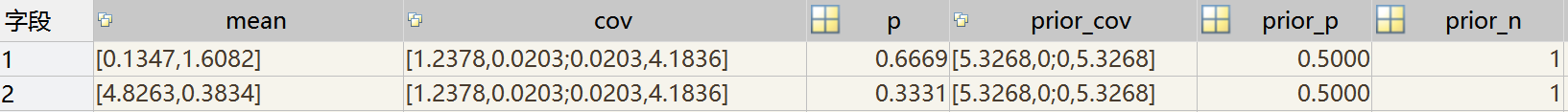
等方差情形推导见纸质文件。

**m=2时,**

对数似然为-4.2583\*1e3



参数估计结果如下：

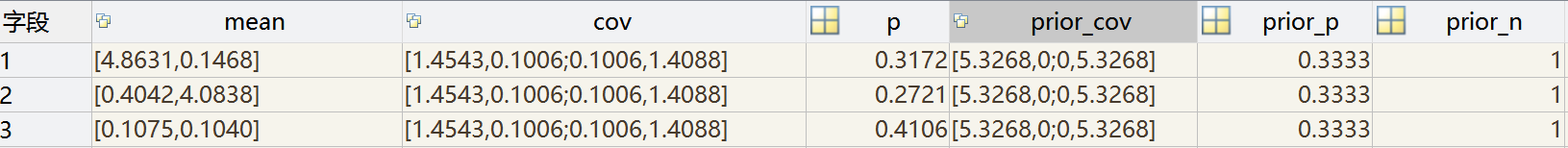


**m=3时,**

对数似然为-4.1725\*1e3



参数估计结果如下：

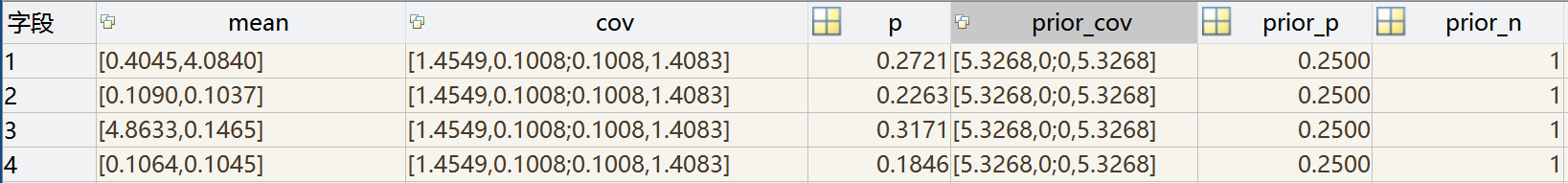


**m=4时,**

对数似然为-4.1768\*1e3



参数估计结果如下：

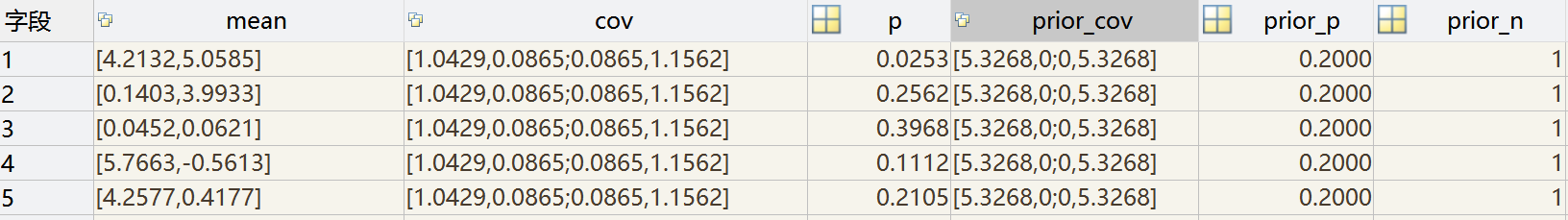


**m=5时,**

对数似然为-4.1139\*1e3



参数估计结果如下：



**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。