COURSE 02457

Non-Linear Signal Processing: Answers to Exercise 8

Checkpoint 8.1

0.0.1 Regression

In this checkpoint we are going to learn the parameters for the sunspot data using the EM algorithm. This is an example of density estimation based regression.

The EM algorithm is run for $K = 2, 25$, both with the constraint that all clusters should have the same variance and without.

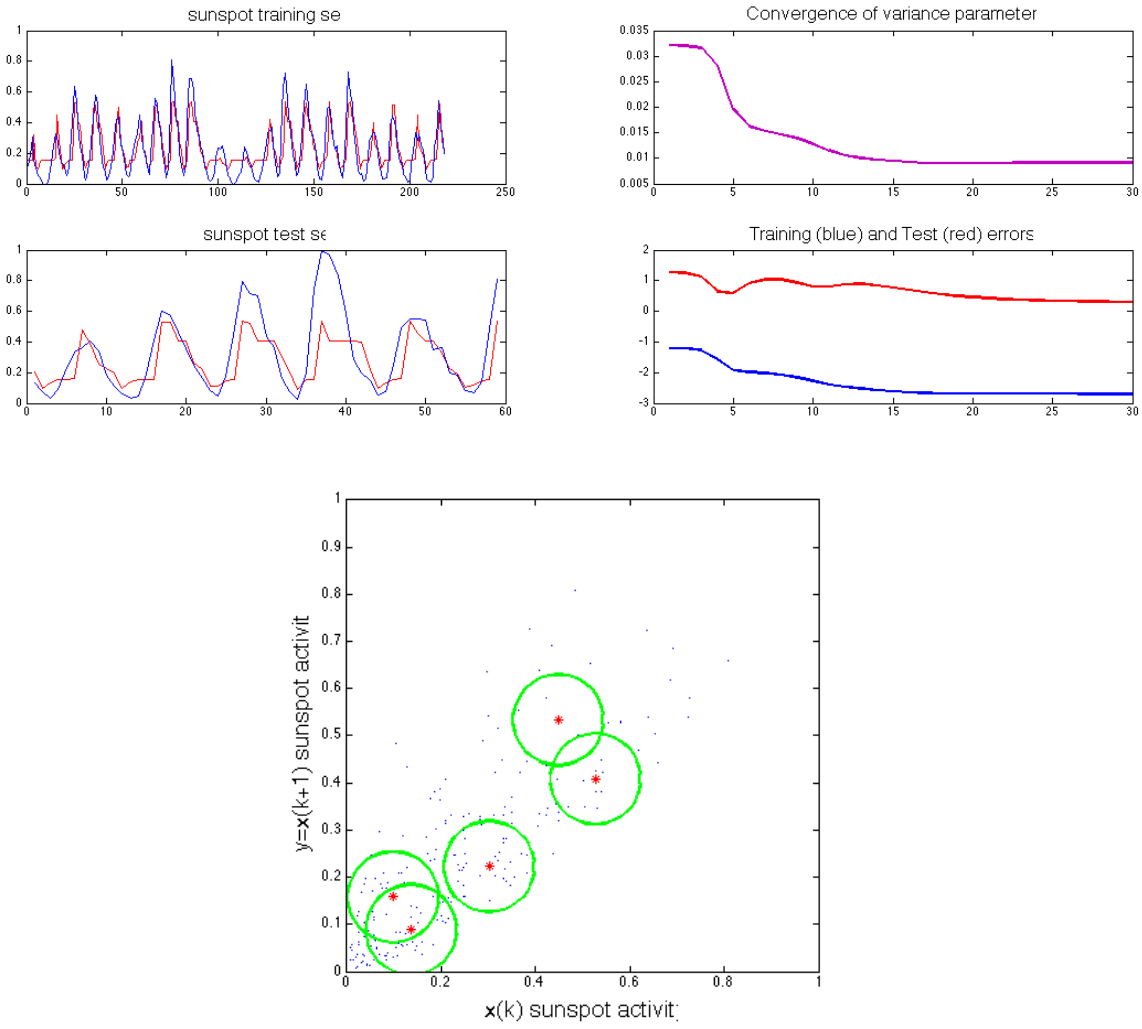

Figure 1: K = 5, Common variance

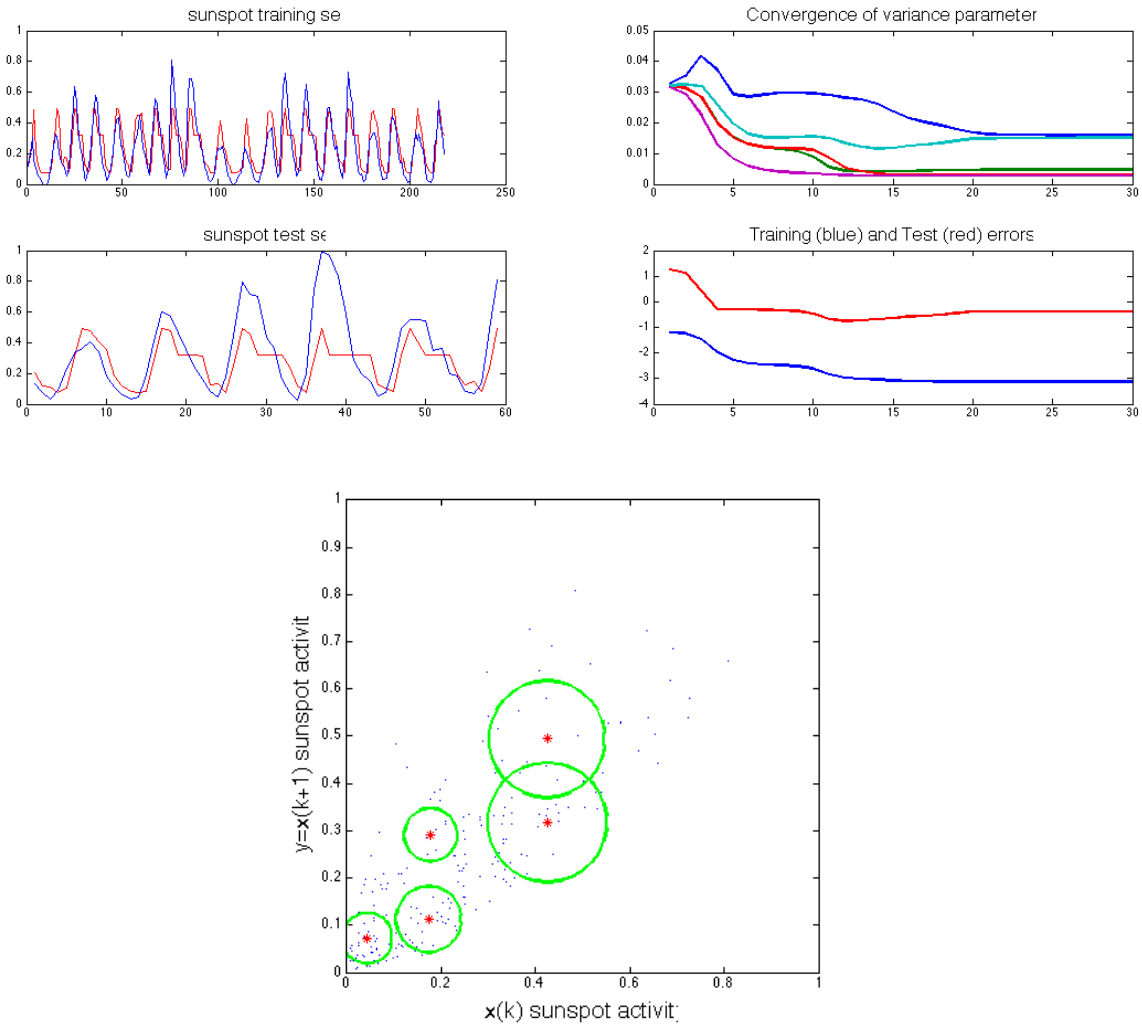

Figure 2: K = 5, Different variance

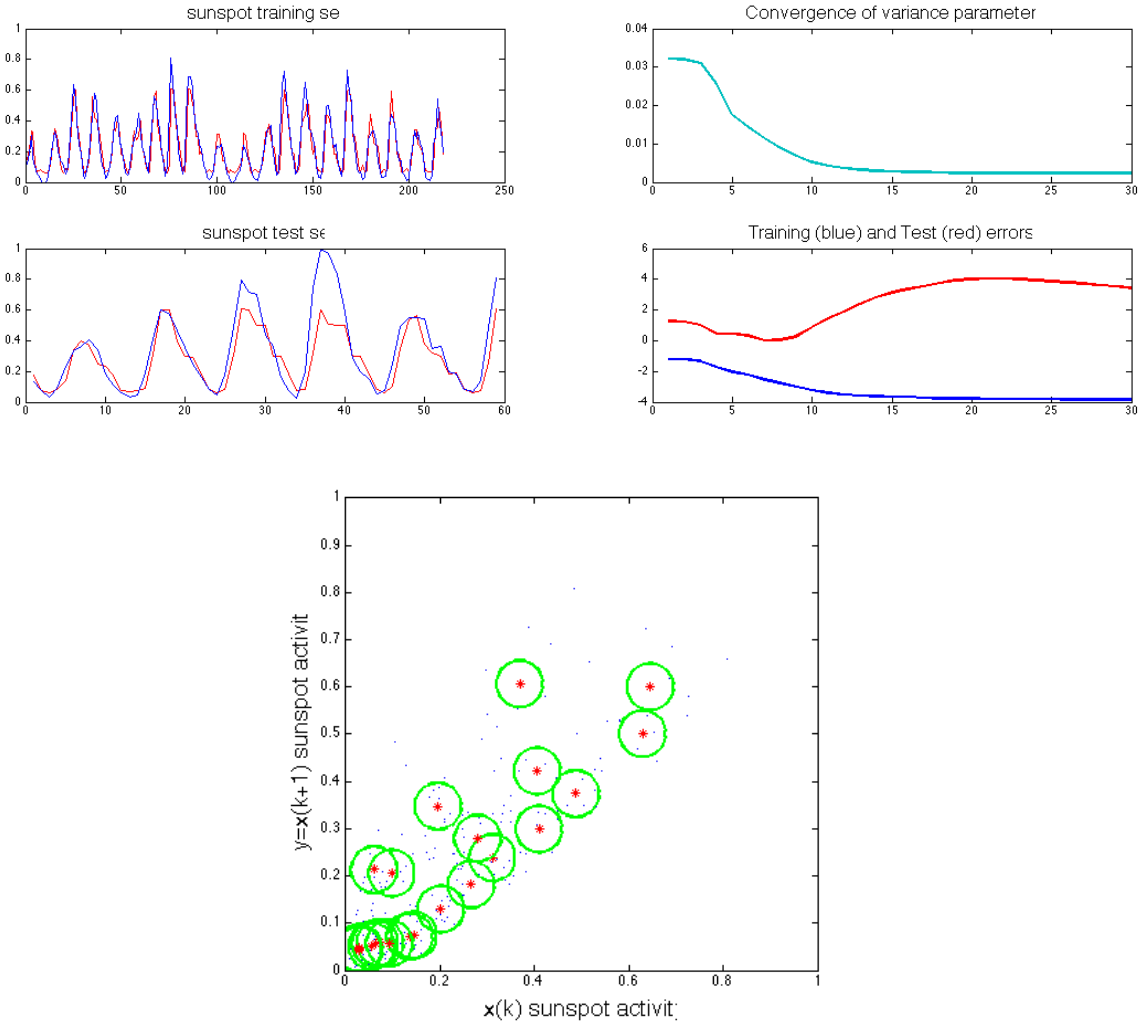

Figure 3: K = 25, Common variance

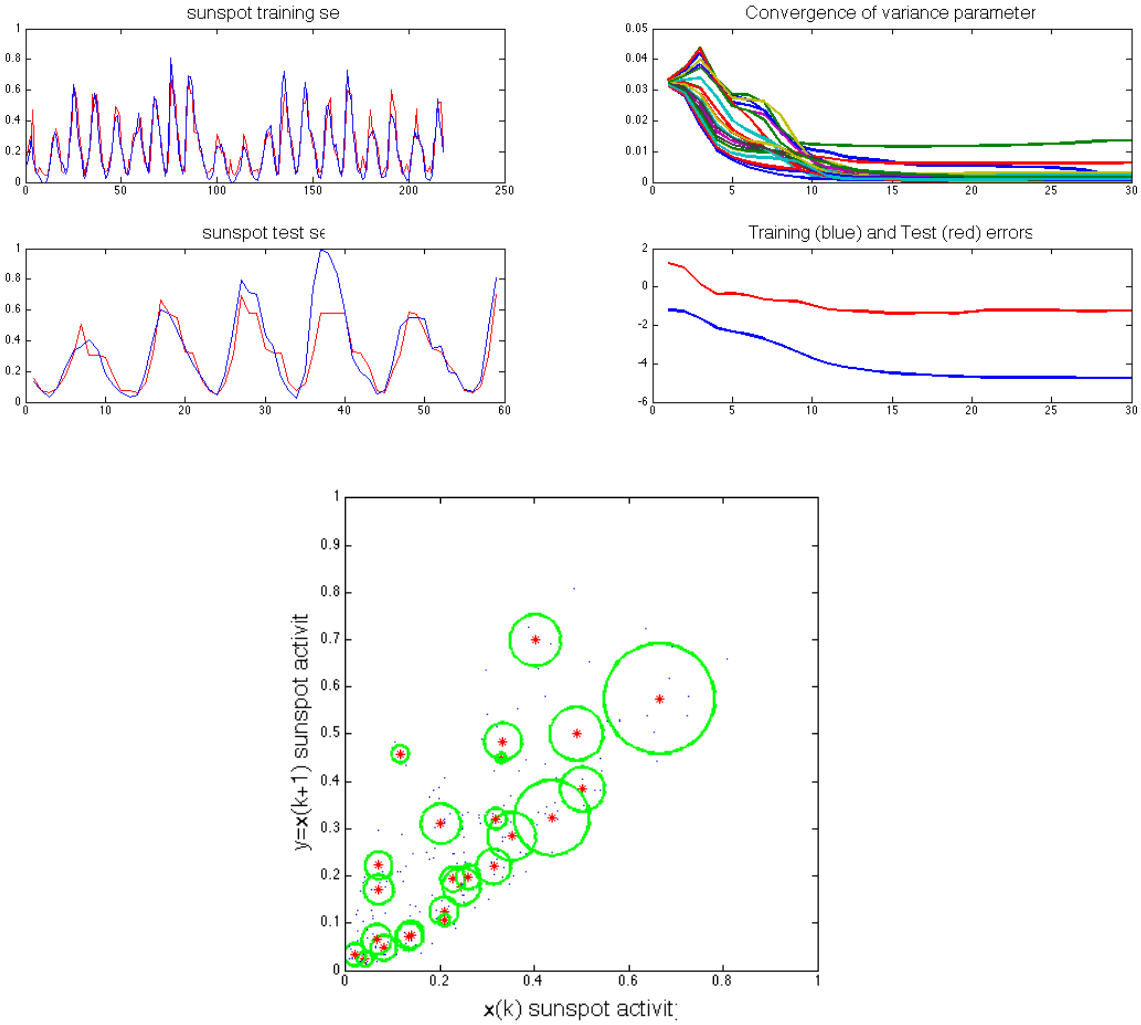

Figure 4: K = 25, Different variance, lower bound 1e-4

The errors in the plots are the negative log likelihood for the test and training set calculated during the EM algorithm,

$$E_l = -\sum_n log\ p(x_n, t_n|w) \tag{1}$$

The predictions errors are calculated as

$$E_p = \sum_n (t_n - y_n)^2 \tag{2}$$

and are given in the table below.

According to the E_l in the plots, it looks like the models with 25 clusters are over fitting, because the test error increases. This error does not tell us anything about the predictions however, and the prediction errors does not indicate over fitting. In fact, according to the prediction error, the models with 25 clusters seem to be doing the best.

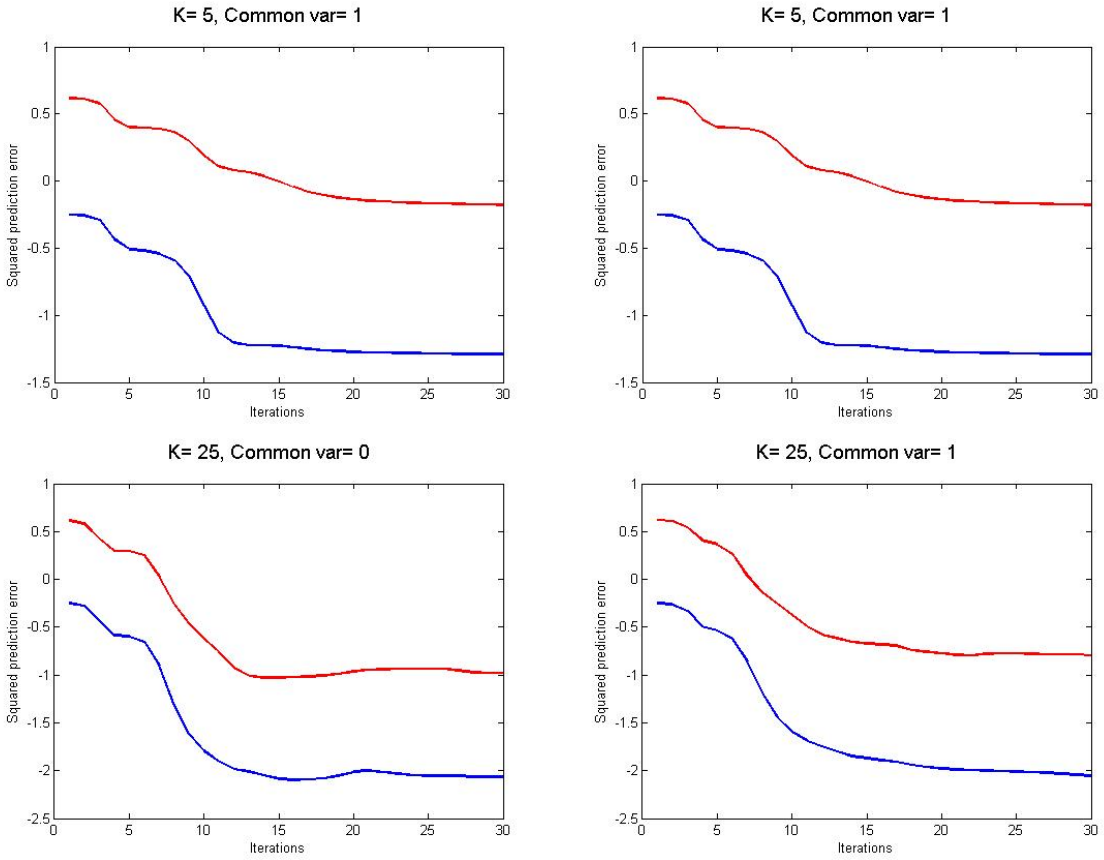

Figure 5: The prediction error during EM training. **Red:** Test. **Blue:** Training.

The prediction errors after EM is shown in the table below.

	K=5	K=25
Common variance = 1	**Etr**= 0.27496 **Ete**=0.83755	**Etr**=0.12851 **Ete**=0.45184
Different variance = 0	**Etr**=0.27666 **Ete**=0.92182	**Etr**=0.12697 **Ete**=0.37511

In comparison, the neural network from Exercise 5 was able to produce a much better prediction with $d = 9$ and $\alpha^{(1)} = \alpha^{(2)} = 0.001$ that gave a prediction error normalized in the same way as in this Exercise of $Epred_{NN} = 0.1961$.

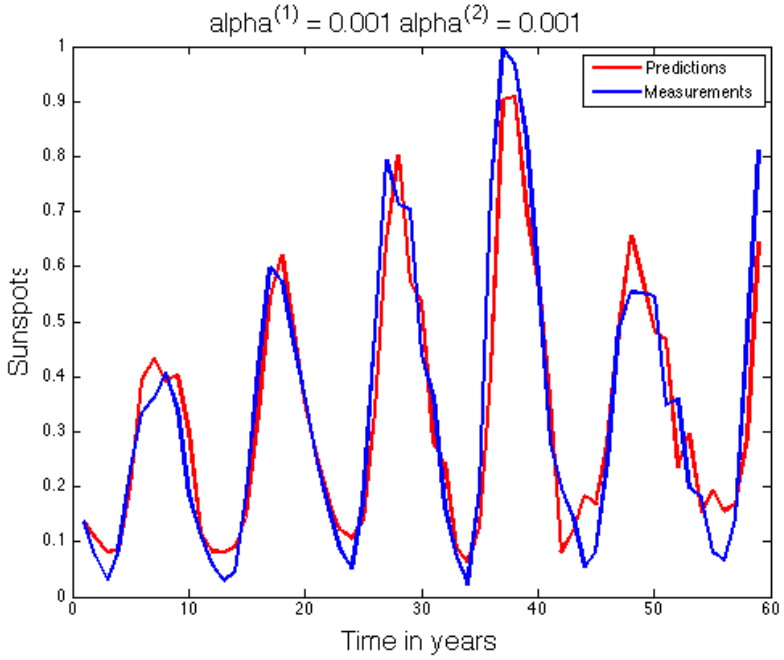

Figure 6: Prediction from the neural network in Ex. 5.

Checkpoint 8.2

The prediction errors for the test and training sets are plotted as a function of the number of clusters both with and without common variance. The result can be seen in Fig. 7. For common variance, the optimal number of clusters is 74, but when the clusters are allowed to have different variances, the optimal number is only 23.

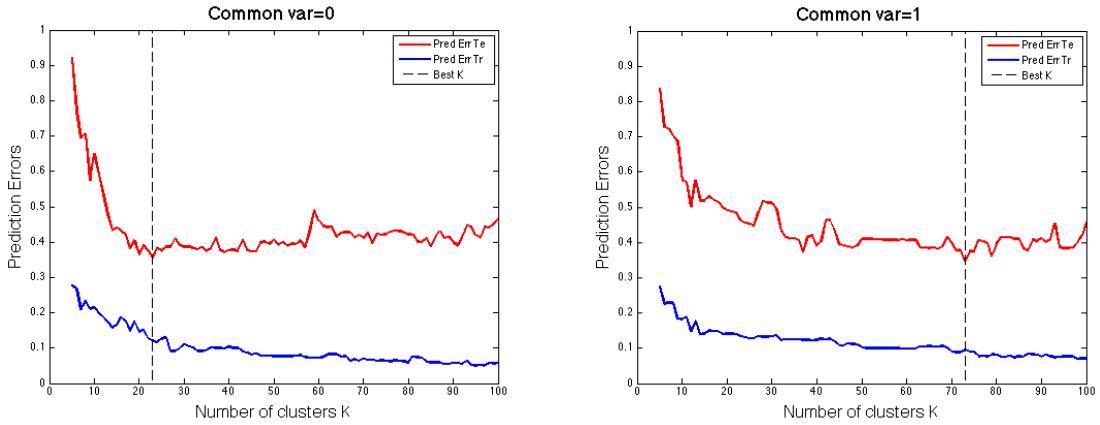

Figure 7: The test and training prediction errors as a function of the number of clusters.

Checkpoint 8.3

Classification

We now turn to a classification problem in which we try classify a pima indian woman as having diabetes or not based on her age and plasma glucose concentration. This is done in a bayesian setting where we wish to model the conditional density $p(C_k|x)$:

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} \tag{3}$$

where the parameters of $p(x|C_k)$ which are μ_k, Σ_k, π_k are estimated using the EM algorithm with K clusters. We want to investigate the optimal number of clusters in the EM algorithm when estimating the likelihood so as to get the lowest misclassification error from the posterior density $p(C_k|x)$.

Here we define the test misclassification error to be

$$E_{misCl} = \sum_n \left(p(C_k = 1|t(x_n) = 1) < 0.5\right) p(C_1) + \sum_n \left(p(C_k = 2|t(x_n) = 2) < 0.5\right) p(C_2) \tag{4}$$

As the clusters are initialized randomly, we get slightly different results every time, but when plotting the misclassification error as a function of the number of clusters, we usually get a local minima around 2-4 clusters and a minima around 10-12 clusters. An example is shown in the figures below.

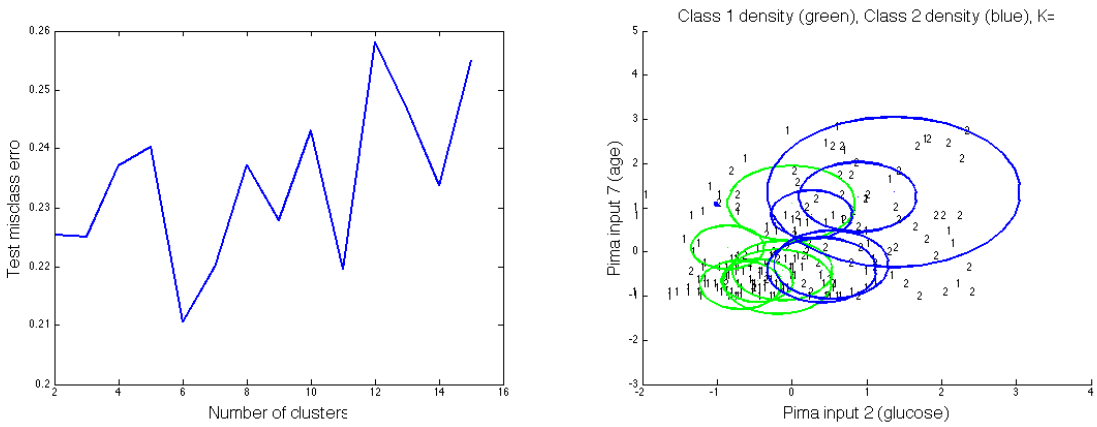

Figure 8: The misclassification error for the pima indian set. The location of the clusters are shown for the optimal number of clusters.

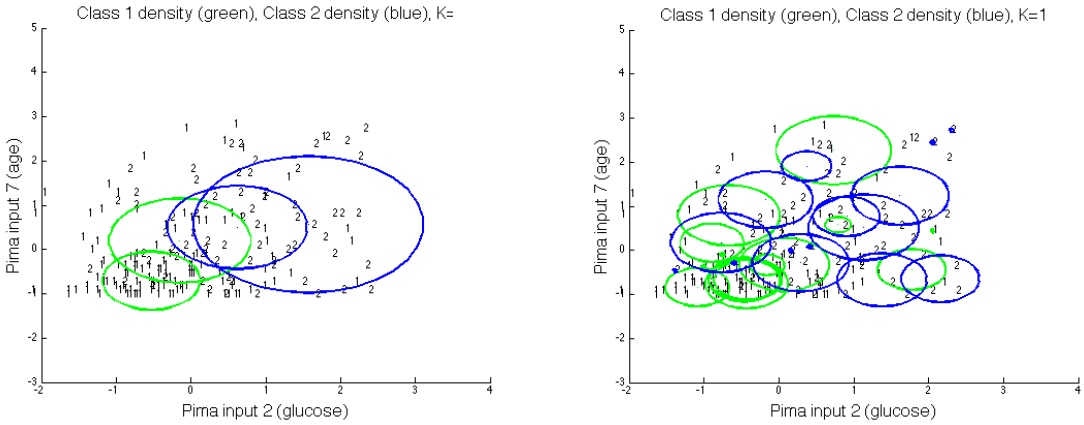

Figure 9: Examples of bad choices for the number of clusters.