Homework 3
M1522.001300 Probabilistic Graphical Models (2016 Fall)

2015-21259 Hyunwoo Lee

Date: October 20 Thursday

1 Parameter Learning

1.1 DGM with hidden variables

1. Assuming all nodes (including H) are binary and all CPDs are tabular, prove that the model on the left has 17 free parameters.

   To decide the parameters for each nodes with k parent nodes, it needs 2^k parameters.

   From the equation

   $$\begin{aligned} p(X_{1:6}) &= \sum_h p(X_1)p(X_2)p(X_3)p(H=h|X_{1:3}) \\ & \quad p(X_4|H=h)p(X_5|H=h)p(X_6|H=h) \end{aligned}$$

   First, We can know X_1, X_2, andX_3 needs only one, summing up 3. Next, the node H gets three parent nodes, it needs $2^3 = 8$ parameters. Finally, the node X_4, X_5, andX_6 has one parent node, i.e. the node H, they need 2 parameters for each, summing up 6. So, the number of overall free parameters is $3 + 8 + 6 = 17$.

2. Assuming all nodes are binary and all CPDs are tabular, prove that the model on the right has 59 free parameters.

   From the equation

   $$p(X_{1:6}) = p(X_1)p(X_2)p(X_3)p(X_4|X_{1:3})p(X_5|X_{1:4})p(X_6|X_{1:5})$$

   First, we can easily know X_1, X_2, X_3 need one parameter for each, summing up 3. Next, X_4 has 3 parent nodes, it has 8 free parameters. And X_5 has 4 parent nodes, it has 16 free parameters. Finally, X_6 has 5 parent nodes, it has 32 free parameters. Thus, the total number of free parameters is $3 + 8 + 16 + 32 = 59$.

3. Suppose we have a data set $\mathcal{D} = X_{1:6}^n$ for n = 1 : N, where we observe the Xs but not H, and we want to estimate the parameters of the CPDs using maximum likelihood. For which model is this easier? Explain your answer.

   Because the first one have smaller number of free variables, it is better. We just need to learn 17 parameters rather than 59 parameters.

1.2 Bayesian Parameter Estimation

We know $N = n_h + n_t$, and

$$
\begin{aligned}
\widehat{\theta}_{Bayes} &= \int_0^1 \theta \cdot P(\theta|\mathcal{D})d\theta \\
&= \int_0^1 \theta^{n_h+\alpha} \cdot (1-\theta)^{n_t+\beta-1} d\theta \\
&= -[\theta^{n_h+\alpha} \cdot \tfrac{(1-\theta)^{n_t+\beta}}{n_t+\beta}]_0^1 + \int_0^1 (n_h+\alpha)\theta^{n_h+\alpha-1} \cdot \tfrac{(1-\theta)^{n_t+\beta}}{n_t+\beta} d\theta \\
&= \tfrac{n_h+\alpha}{n_t+\beta} \int_0^1 (1-\theta) \cdot P(\theta|\mathcal{D})d\theta \\
&= \tfrac{n_h+\alpha}{n_t+\beta} (\int_0^1 P(\theta|\mathcal{D})d\theta - \int_0^1 \theta \cdot P(\theta|\mathcal{D})d\theta) \\
&= \tfrac{n_h+\alpha}{n_t+\beta} - \tfrac{n_h+\alpha}{n_t+\beta} \cdot \widehat{\theta}_{Bayes}
\end{aligned}
$$

From the above expression, finally we get,

$$
\begin{aligned}
(1 + \tfrac{n_h+\alpha}{n_t+\beta})\widehat{\theta}_{Bayes} &= \tfrac{n_h+\alpha}{n_t+\beta} \\
(N+\alpha+\beta)\widehat{\theta}_{Bayes} &= n_h + \alpha \\
\widehat{\theta}_{Bayes} &= \tfrac{n_h+\alpha}{N+\alpha+\beta}
\end{aligned}
$$

which we want.

1.3 Bayesian Approach for Gaussian Distribution

1. We now consider a Bayesian approach for learning the mean of a Gaussian distribution. It turns out that in doing Bayesian inference with Gaussians, it is mathematically easier to use the precision $\tau = \frac{1}{\sigma^2}$ rather than the variance. Note that larger the precision, the narrower the distribution around the mean.

   Suppose that we have M IID samples $x[1], , x[M]$ from $X \sim \mathcal{N}(\theta; \tau_X^{-1})$. Moreover, assume that we know the value of τ_X. Thus, the unknown parameter θ is the mean. Show that if the prior $P(\theta)$ is $\mathcal{N}(\mu; \tau_\theta^{-1})$, then the posterior $P(\theta|\mathcal{D})$ is $\mathcal{N}(\mu'; (\tau'_\theta)^{-1})$ where

$$
\begin{aligned}
\tau_\theta^{`} &= M\tau_X + \tau_\theta \\
\mu' &= \tfrac{M\tau_X}{\tau'_\theta}\mathbb{E}_{\mathcal{D}}[X] + \tfrac{\tau_\theta}{\tau'_\theta}\mu_0
\end{aligned}
$$

   We know that the posterior $P(\theta|D)$ is proportional to the likelihood times the prior probability, that is,

$$
P(\theta|D) \propto P(D|\theta) \cdot P(\theta)
$$

   Because $X \sim \mathcal{N}(\theta; \tau_X^{-1})$ and the samples are i.i.d, the $P(D|\theta) \cdot P(\theta)$ is,

$$
\begin{aligned}
P(D|\theta) \times P(\theta) &= \textstyle\prod_{m=1}^M \mathcal{N}(x[m]|\theta, \tau_X^{-1}) \times \mathcal{N}(\theta|\mu, \tau_\theta^{-1}) \\
&= \textstyle\prod_{m=1}^M (\tfrac{\tau_X}{2\pi})^{\frac{1}{2}} e^{-\frac{\tau_X}{2}(x[m]-\theta)^2} \times (\tfrac{\tau_\theta}{2\pi})^{\frac{1}{2}} e^{-\frac{\tau_\theta}{2}(\theta-\mu)^2} \\
&= (\tfrac{\tau_X}{2\pi})^{\frac{M}{2}} \textstyle\prod_{m=1}^M e^{-\frac{\tau_X}{2}(x[m]-\theta)^2} \times (\tfrac{\tau_\theta}{2\pi})^{\frac{1}{2}} e^{-\frac{\tau_\theta}{2}(\theta-\mu)^2} \\
&= (\tfrac{\tau_X}{2\pi})^{\frac{M}{2}} (\tfrac{\tau_\theta}{2\pi})^{\frac{1}{2}} e^{-\frac{1}{2}(\tau_X \sum_{m=1}^M (x[m]-\theta)^2 + \tau_\theta(\theta-\mu)^2)} \\
&\propto e^{-\frac{1}{2}(\tau_X \sum_{m=1}^M (x[m]-\theta)^2 + \tau_\theta(\theta-\mu)^2)}
\end{aligned}
$$

   From this, we get

$$
P(\theta|D) \propto e^{-\frac{1}{2}(\tau_X \sum_{m=1}^M (x[m]-\theta)^2 + \tau_\theta(\theta-\mu)^2)}
$$

   Because,

$$
\begin{aligned}
(x[m]-\theta)^2 &= ((x[m]-\mathbb{E}_{\mathcal{D}}[X])-(\theta-\mathbb{E}_{\mathcal{D}}[X]))^2 \\
&= (x[m]-\mathbb{E}_{\mathcal{D}}[X])^2 - 2(x[m]-\mathbb{E}_{\mathcal{D}}[X])(\theta-\mathbb{E}_{\mathcal{D}}[X]) + (\theta-\mathbb{E}_{\mathcal{D}}[X])^2
\end{aligned}
$$

By summing up,

$$
\begin{aligned}
&\textstyle\sum_{m=1}^{M}(x[m]-\theta)^2 \\
&\quad = \textstyle\sum_{m=1}^{M}((x[m]-\mathbb{E}_{\mathcal{D}}[X])^2 - 2(x[m]-\mathbb{E}_{\mathcal{D}}[X])(\theta-\mathbb{E}_{\mathcal{D}}[X]) + (\theta-\mathbb{E}_{\mathcal{D}}[X])^2) \\
&\quad = \textstyle\sum_{m=1}^{M}(x[m]-\mathbb{E}_{\mathcal{D}}[X])^2 - 2\sum_{m=1}^{M}(x[m]-\mathbb{E}_{\mathcal{D}}[X])(\theta-\mathbb{E}_{\mathcal{D}}[X]) + \sum_{m=1}^{M}(\theta-\mathbb{E}_{\mathcal{D}}[X])^2 \\
&\quad = \textstyle\sum_{m=1}^{M}(x[m]-\mathbb{E}_{\mathcal{D}}[X])^2 + \sum_{m=1}^{M}(\theta-\mathbb{E}_{\mathcal{D}}[X])^2 \ (\because \sum_{m=1}^{M}(x[m]-\mathbb{E}_{\mathcal{D}}[X]) = 0) \\
&\quad = \textstyle\sum_{m=1}^{M}(x[m]-\mathbb{E}_{\mathcal{D}}[X])^2 + M(\theta-\mathbb{E}_{\mathcal{D}}[X])^2
\end{aligned}
$$

Apply it to the posterior probability,

$$
\begin{aligned}
P(\theta|D) &\propto e^{-\frac{1}{2}(\tau_X \sum_{m=1}^{M}(x[m]-\theta)^2+\tau_\theta(\theta-\mu)^2)} \\
&= e^{-\frac{1}{2}(\tau_X(\sum_{m=1}^{M}(x[m]-\mathbb{E}_{\mathcal{D}}[X])^2+M(\theta-\mathbb{E}_{\mathcal{D}}[X])^2)+\tau_\theta(\theta-\mu)^2)} \\
&\propto e^{-\frac{1}{2}(M\tau_X(\theta-\mathbb{E}_{\mathcal{D}}[X])^2+\tau_\theta(\theta-\mu)^2)} \\
&= e^{-\frac{1}{2}(M\tau_X\theta^2-2\tau_X M\mathbb{E}_{\mathcal{D}}[X]\theta+M\tau_X\mathbb{E}_{\mathcal{D}}[X]^2+\tau_\theta\theta^2-2\tau_\theta\mu\theta+\tau_\theta\mu^2)} \\
&= e^{-\frac{1}{2}((M\tau_X+\tau_\theta)\theta^2-2(M\tau_X\mathbb{E}_{\mathcal{D}}[X]+\tau_\theta\mu)\theta+C_1)} \\
&= e^{-\frac{1}{2}(M\tau_X+\tau_\theta)(\theta-\frac{M\tau_X\mathbb{E}_{\mathcal{D}}[X]+\tau_\theta\mu}{M\tau_X+\tau_\theta})^2+C_2} \\
&\propto e^{-\frac{1}{2}(M\tau_X+\tau_\theta)(\theta-\frac{M\tau_X\mathbb{E}_{\mathcal{D}}[X]+\tau_\theta\mu}{M\tau_X+\tau_\theta})^2}
\end{aligned}
$$

Finally, from the equation, now we get

$$
\begin{aligned}
\tau'_\theta &= M\tau_X + \tau_\theta \\
\mu' &= \tfrac{M\tau_X}{\tau'_\theta}\mathbb{E}_{\mathcal{D}}[X] + \tfrac{\tau_\theta}{\tau'_\theta}\mu
\end{aligned}
$$

2. We now consider making predictions with the posterior of previous problem. Suppose we now want to compute the probability

$$
P(x[M+1]|\mathcal{D}) = \int P(x[M+1]|\theta)P(\theta|\mathcal{D})d\theta
$$

Show that this distribution is Gaussian. What is the mean and precision of this distribution?

Because $X \sim \mathcal{N}(\theta; \tau_X^{-1})$,

$$
\begin{aligned}
P(x[M+1]|\theta) &= \mathcal{N}(x[M+1]|\theta, \tau_X^{-1}) \\
P(\theta|\mathcal{D}) &= \mathcal{N}(\theta|\mu', (\tau'_\theta)^{-1}) \\
P(x[M+1]|\theta)P(\theta|\mathcal{D}) &= \mathcal{N}(x[M+1]|\theta, \tau_X^{-1}) \times \mathcal{N}(\theta|\mu', (\tau'_\theta)^{-1})
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&P(x[M+1]|\theta)P(\theta|\mathcal{D}) \\
&\quad = \mathcal{N}(x[M+1]|\theta, \tau_X^{-1}) \times \mathcal{N}(\theta|\mu', (\tau'_\theta)^{-1}) \\
&\quad = (\tfrac{\tau_X}{2\pi})^{\frac{1}{2}} e^{-\frac{\tau_X}{2}(x[M+1]-\theta)^2} \times (\tfrac{\tau'_\theta}{2\pi})^{\frac{1}{2}} e^{-\frac{\tau'_\theta}{2}(\theta-\mu')^2} \\
&\quad \propto e^{-\frac{\tau_X}{2}(x[M+1]-\theta)^2-\frac{\tau'_\theta}{2}(\theta-\mu')^2}
\end{aligned}
$$

2 Iterative Proprotional Fitting

(a) The partition function $Z^{(t)}$ is constant over all IPF iteration.

Let see the relation between $p^{(t+1)}(x_C)$ and $p^{(t)}(x_C)$, when we update in terms of x_C.

$$\begin{aligned}
p^{(t+1)}(x_C) &= \textstyle\sum_{x_{\mathcal{V}\setminus C}} p^{(t+1)}(x) \\
&= \textstyle\sum_{x_{\mathcal{V}\setminus C}} \frac{1}{Z^{(t+1)}} \prod_D \psi_D^{(t+1)}(x_D) \\
&= \textstyle\sum_{x_{\mathcal{V}\setminus C}} \frac{1}{Z^{(t+1)}} \psi_C^{(t+1)}(x_C) \prod_{D\neq C} \psi_D^{(t)}(x_D) \\
&\quad (\because \psi_D^{(t+1)}(x_D) = \psi_D^{(t)}(x_D), \text{ only } C \text{ is updated by assumption}) \\
&= \textstyle\frac{1}{Z^{(t+1)}} \sum_{x_{\mathcal{V}\setminus C}} \psi_C^{(t)}(x_C) \frac{\widetilde{p}(x_C)}{p^{(t)}(x_C)} \prod_{D\neq C} \psi_D^{(t)}(x_D) \\
&\quad (\because \psi_D^{(t+1)}(x_D) = \psi_C^{(t)}(x_C) \tfrac{\widetilde{p}(x_C)}{p^{(t)}(x_C)}) \\
&= \textstyle\frac{1}{Z^{(t+1)}} \frac{\widetilde{p}(x_C)}{p^{(t)}(x_C)} \sum_{x_{\mathcal{V}\setminus C}} \prod_D \psi_D^{(t)}(x_D) \\
&= \textstyle\frac{Z^{(t)}}{Z^{(t+1)}} \frac{\widetilde{p}(x_C)}{p^{(t)}(x_C)} \sum_{x_{\mathcal{V}\setminus C}} \frac{1}{Z^{(t)}} \prod_D \psi_D^{(t)}(x_D) \\
&\quad (\because multiply \tfrac{Z^{(t)}}{Z^{(t)}}) \\
&= \textstyle\frac{Z^{(t)}}{Z^{(t+1)}} \frac{\widetilde{p}(x_C)}{p^{(t)}(x_C)} \sum_{x_{\mathcal{V}\setminus C}} p^{(t)}(x) \\
&= \textstyle\frac{Z^{(t)}}{Z^{(t+1)}} \frac{\widetilde{p}(x_C)}{p^{(t)}(x_C)} p^{(t)}(x_C) \\
&= \textstyle\frac{Z^{(t)}}{Z^{(t+1)}} \widetilde{p}(x_C)
\end{aligned}$$

Now, we get

$$p^{(t+1)}(x_C) = \tfrac{Z^{(t)}}{Z^{(t+1)}} \widetilde{p}(x_C) \tag{1}$$

By summing up in terms of C,

$$\textstyle\sum_C p^{(t+1)}(x_C) = \sum_C \frac{Z^{(t)}}{Z^{(t+1)}} \widetilde{p}(x_C)$$

Because $\sum_C p^{(t+1)}(x_C) = 1$ and $\sum_C \widetilde{p}(x_C) = 1$, we get

$$Z^{(t+1)} = Z^{(t)} \tag{2}$$

From this equation, we know the partition function $Z^{(t)}$ is constant over IPF iteration.

(b) At each iteration the model marginal $p^{(t+1)}(x_C)$ is equal to the observed marginal $\widetilde{p}(x_C)$.

From the equation (1) and (2), we get,

$$p^{(t+1)}(x_C) = \widetilde{p}(x_C)$$

3 EM for Robot Mapping

1. Assume we perform the E-step for each step x_m by defining

$$\widetilde{P}(\mathrm{x}_m|C_m = k : \theta_{\mathbf{k}}) = \mathcal{N}(d(\mathrm{x}_m, p_k)|0; \sigma^2)$$

and $\widetilde{P}(\mathrm{x}_m|C_m = 0 : \theta_{\mathbf{k}}) = C$ for some constant C. Why is this formula not a correct application of EM?

Because $\sum_m \sum_k \widetilde{P}(\mathrm{x}_m|C_m = k : \theta_{\mathbf{k}}) \neq 1$, so this formula is not a correct.
It needs to be normalized as $\widetilde{P}(\mathrm{x}_m|C_m = k : \theta_{\mathbf{k}}) = \frac{\mathcal{N}(d(\mathrm{x}_m, p_k)|0;\sigma^2)}{\sum_m \sum_k \mathcal{N}(d(\mathrm{x}_m, p_k)|0;\sigma^2)}$

2. Given a solution to the E-step, show how to perform maximum likelihood estimation of the model parameters α_k, β_k, subject to the constraint that α_k be a unit-vector, that is, that $\alpha_k \cdot \alpha_k = 1$.

Let e_{ik} be $E[C_i = k|\theta^{(n)}, x_i] = p(C_i = k|\theta^{(n)}, x_i)$, where $\theta^{(n)}$ is the θ in the step n.

Given e_{ik}, $\theta^{(n+1)} = \arg\min_{\theta} \sum_i \sum_j e_{ij}^{(n)} (\alpha_j \cdot x_i - \beta_j)^2$

This means, $(\alpha^{(n+1)}, \beta^{(n+1)}) = \arg\min_{\alpha,\beta} \sum_i \sum_j e_{ij}^{(n)} (\alpha_j \cdot x_i - \beta_j)^2$

This has a constraint, $\alpha_j \cdot \alpha_j = 1$

This can be solved by introducing the Lagrange multipliers λ_k for $k = 1, \cdots, K$

Let $L := \sum_i \sum_k e_{ik}^{(n)} (\alpha_k \cdot x_i - \beta_k)^2 + \sum_k \lambda_k \alpha_k \cdot \alpha_k$ This is the cost function with the Lagrange multipliers added. By minimizing it, we can solve the problem subject to the constraint, $\alpha_k \cdot \alpha_k = 1$

4 EM-GMM and K-means Clustering

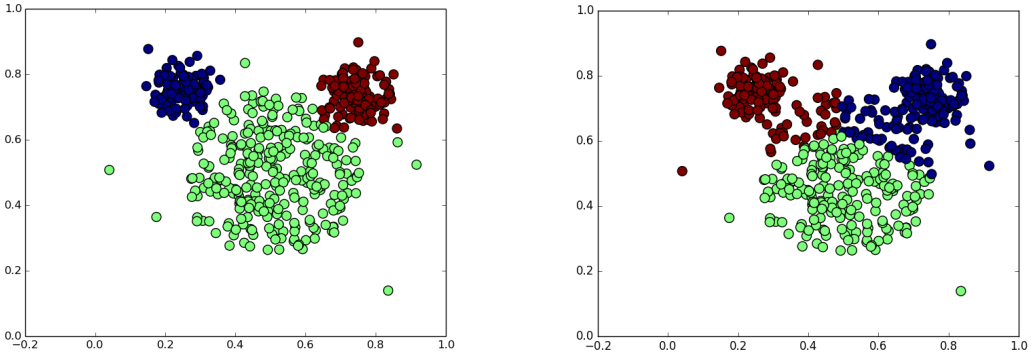
1. Implement theses algorithms

(The solution is separated)

2. Run your EM-GMM algorithm and K-means algorithm on the mouse dataset for K = 3 and compare the result. If two algorithms' results dier, What do you think the reason is?

As from the Figure 1, they are somewhat different. EM-GMM shows more accurate in dividing the clusters. Because the covariance in the K-means algorithm is fixed, the coverage of each Gaussian distribution cannot be adapted unlink that of EM-GMM.

The main factor of the difference is whether the "covariance" is fixed or not. That's why EM-GMM can cluster the data finely. It appears significantly in this densely distributed dataset



(a) EM-GMM (b) K-means

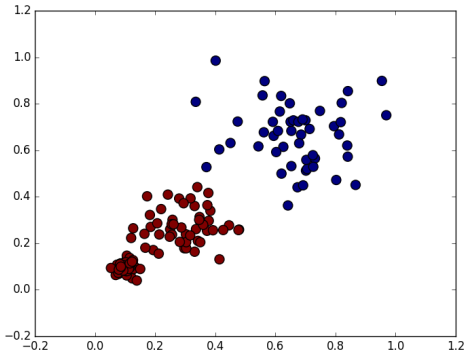
Figure 1: EM-GMM and K-means applying to the mouse point data with K=3

3. Run your EM-GMM algorithm and K-means algorithm on the vary density dataset for some different values of K and give some explanation on the results.

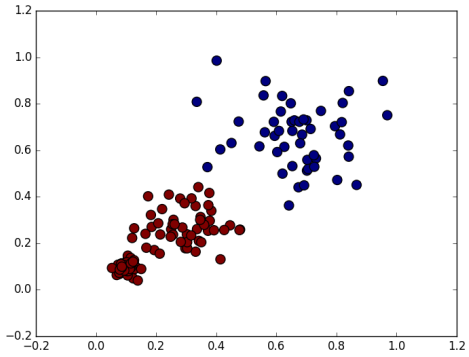
   The result with the vary density dataset applying EM-GMM and K-means in the Figure 2. Varying the number of clusters (K), two algorithms show different results.

   In case of K=2, two algorithms shows exactly same result. But the difference is caused after K=2. Similar to the prior problem, EM-GMM shows the finer clustering, due to the responsibilities and the covariance. Unlike only considering the Euclidean distance, EM-GMM considers the responsibility of each Gaussian distributions and the covariance among the variables.

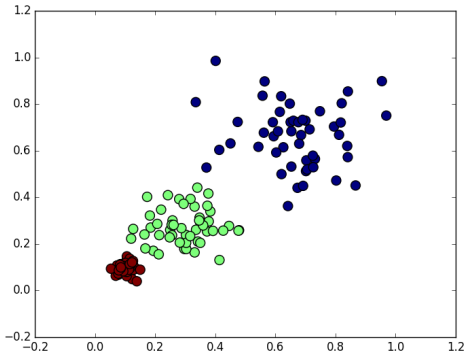
   It looks that the means of the Gaussian distribution in the EM-GMM places on the centroid of each clusters, but K-means doesn't.



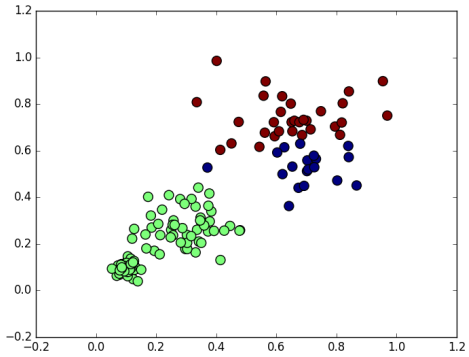
(a) EM-GMM (K=2)



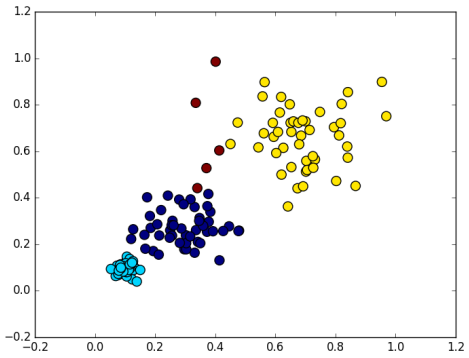
(b) K-means (K=2)



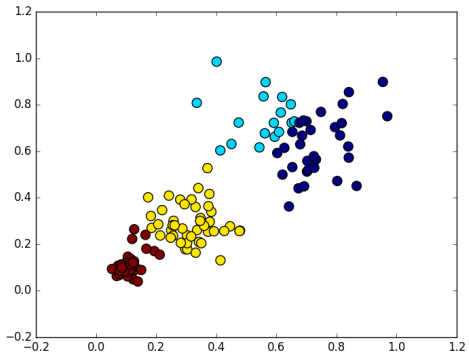
(c) EM-GMM (K=3)



(d) K-means (K=3)



(e) EM-GMM (K=4)



(f) K-means (K=4)

Figure 2: EM-GMM and K-means applying to the vary-density data