LEHIGH UNIVERSITY

CSE 426 PROJECT REPORT

Learning the Number of Components in Mixture Models

Author:
Ziyi GUO

Lecturer:
Prof. Henry BAIRD

November 29, 2014

Learning the Number of Components in Mixture Models

Ziyi Guo
Department of Computer Science and Engineering
Lehigh University
zig312@lehigh.edu

November 29, 2014

Abstract

Mixture models are of great importance in machine learning, but choosing the number of components could be arbitrary and inaccurate, and in this project, we learn to detect it automatically. Taking Gaussian Mixture Model as an example, in the case of maximum likelihood learning, we discuss EM algorithm for parameter estimation and compute Bayesian Information Criteria for model selection; in the case of Bayesian learning, we discuss variational approximation for parameter estimation and calculate the variational lower bound for model selection. Experiments on the 'Old Faithful' dataset show that our methods are both accurate and efficient.

I Introduction

The mixture model is a probabilistic model for representing an overall population via mixture of different subpopulations. It can be viewed as a simple form of latent variable model when $z_i \in \{1, ..., K\}$ represents a discrete latent state since we can assume that observed variables are correlated because they arise from some hidden causes. For the likelihood, we use $p(\mathbf{x}|z_i = k) = p(\mathbf{x}|\theta)$ where $p(\mathbf{x}|\theta)$ is the kth base distribution for the observation variable $\mathbf{x}$. After mixing K base distributions together, we get the mixture model as follows:

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}|\theta) \tag{1}$$

This is a convex combination with constraints $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. Some well-know statistical learning models are examples of mixture models, such as Gaussian Mixture Model(GMM), Probabilistic Principal Component Analysis(PPCA) and Latent Dirichlet Allocation(LDA).

Mixture models play a significant role in the field of pattern recognition and machine learning because it is very difficult to draw and express real-world data via single distribution or model,

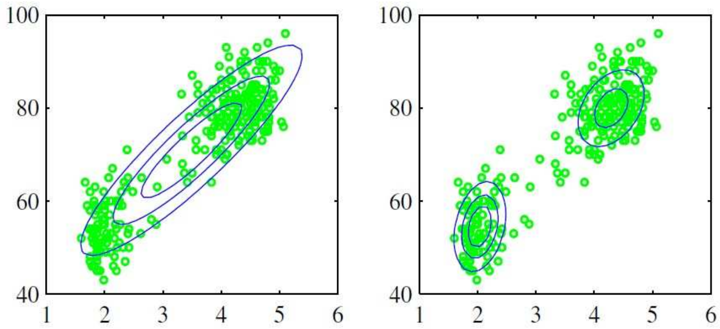

Figure 1: Gaussian Mixture Model Example[1]. Left: a Gaussian distribution failed to capture the data using maximum likelihood since data around the center are relatively sparse. Right: a mixture of two Gaussian distributions fitted the data well using maximum likelihood.

and they are largely used in the application of unsupervised clustering[2], data compression[3] and topic modeling[4]. One simple example of mixture models ia shown in Figure 1.

II Related Works

When using latant variable models(LVM), we need to specify the number of latent states which control the complexity of the model. In the case of mixture models, this means that we need to know K, the number clusters, in advance. Essentially, the choice of K is a problem of model selection since the model becomes more complex as we have more clusters and more parameters in the model should be estimated.

One approach is to use cross validation strategy, which allows a proportion $(S-1)/S$ of all the data to be used for training and the remaining proportion for testing. In particular, the leave-one-out technique is given when S is equal to the number of data points. However, the process of cross validation cound be slow since we have to run a large number of training, and this strategy is especially inpractical for model in which each training run itself is computational expensive.

Alternatively, various information criteria methods have been proposed in which a penalty term is added to the maximum likelihood in order to prevent the over-fitting problem. For example, the Akaike Information Criteria(AIC[5] is defined as:

$$lnp(D|\theta) - M \tag{2}$$

where $lnp(D|\theta)$ indicates the best log likelihood and M indicates the number parameters in the model. Another option called Bayesian Information Criteria(BIC) is a similar technique but gives more weights on the penalty term, and we will discuss it later when attempting to optimize the number of component in GMM via expectation maximization algorithm.

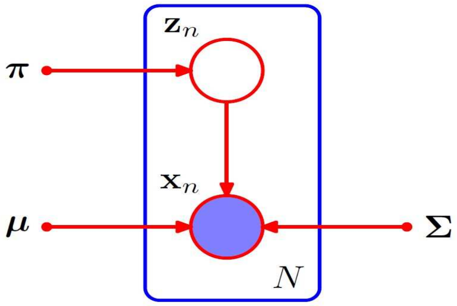

Figure 2: Graphical Representation of the Gaussian Mixture Model[1].

III Gaussian Mixture Model

In this section, we would like to use Gaussian Mixture Model(GMM) as an example and discuss how to detect the optimal number of components automatically.

III.I Model Definition

First, we give a 1-of-K representation of latent variable $\mathbf{z}$ which satisfies $z_k \in \{0,1\}$ and $\sum_k z_k = 1$. The marginal distribution over $\mathbf{z}$ is given in terms of mixing coeffiients π_k with constraints $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$:

$$p\left(z_k = 1\right) = \pi_k \tag{3}$$

vectorizing this equation, we get:

$$p\left(\mathbf{z}\right) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{4}$$

Similarly, in the form of Gaussian, the conditional distribution can be written as,

$$p\left(\mathbf{x}|\mathbf{z}\right) = \prod_{k=1}^{K} N\left(\mathbf{x}|\mu_{\mathbf{k}}, \Sigma_k\right)^{z_k} \tag{5}$$

Then by summing the joint distribution over all latent variables, the marginal distribution with respect to $\mathbf{x}$ can be written as:

$$p\left(\mathbf{x}\right) = \sum_{\mathbf{z}} p\left(\mathbf{z}\right) p\left(\mathbf{x}|\mathbf{z}\right) = \sum_{k=1}^{K} \pi_k N\left(\mathbf{x}|\mu_{\mathbf{k}}, \Sigma_k\right) \tag{6}$$

The graphical representation of GMM is given in Figure 2.

Also, suppose we have a set of observations $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ along with i.i.d assumption, the log of the likelihood function, which is the objective in maximum likelihood estimation, is given by:

$$lnp\left(\mathbf{X}|\pi,\mu,\Sigma\right)=\sum_{n=1}^{N}ln\left\{\sum_{k=1}^{K}\pi_k N\left(\mathbf{x}_n|\mu_k,\Sigma_k\right)\right\} \tag{7}$$

Usually, it is very difficult to derive close-form solutions by maximizing this objective because summation over k appears inside of log function. Therefore, in the following section, we alternatively consider an approach called expectation maximization(EM) algorithm for the solution.

III.II Expectation Maximization

III.II.1 Auxiliary Function

Now, we can alternatively consider the likelihood for the complete data set $\{\mathbf{X},\mathbf{Z}\}$:

$$p\left(\mathbf{X},\mathbf{Z}|\mu,\Sigma,\pi\right)=\prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}N\left(\mathbf{x}_n|\mu_k,\Sigma_k\right)^{z_{nk}} \tag{8}$$

we obtain the log of likelihood by taking the logarithm:

$$lnp\left(\mathbf{X},\mathbf{Z}|\mu,\Sigma,\pi\right)=\sum_{n=1}^{N}\sum_{k=1}^{K}z_{nk}\left\{ln\pi_k+lnN\left(\mathbf{x}_n|\mu_k,\Sigma_k\right)\right\} \tag{9}$$

However, z_{nk} is unknown and we cannot maximize the likelihood directly. Instead, we define an auxiliary function $Q\left(\theta,\theta^{old}\right)$ which represents the expected complete data log likelihood:

$$Q\left(\theta,\theta^{old}\right)=\sum_{\mathbf{Z}}p\left(\mathbf{Z}|\mathbf{X},\theta^{old}\right)lnp\left(\mathbf{X},\mathbf{Z}|\theta\right) \tag{10}$$

where $\theta=\{\pi,\mu,\Sigma\}$. In the context of GMM, the auxiliary function is given by:

$$\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\left\{ln\pi_k+lnN\left(\mathbf{x}_n|\mu_k,\Sigma_k\right)\right\} \tag{11}$$

where $r_{nk}=\mathbb{E}\left[z_{nk}\right]$ which can be explained as the responsibility of component k for data $\mathbf{x}_n$.

III.II.2 E Step

In E step, we use current values for model parameters to compute responsibilities by:

$$r_{nk}=\frac{\pi_k N\left(\mathbf{x}_n|\mu_k,\Sigma_k\right)}{\sum_{j=1}^{K}\pi_j N\left(\mathbf{x}_n|\mu_j,\Sigma_j\right)} \tag{12}$$

III.II.3 M Step

In M step, we fix values for responsibilities and optimize $Q\left(\theta,\theta^{old}\right)$ in Equ(10) with respect to model parameters and get:

$$\begin{aligned} \pi_k^{new} &= \frac{N_k}{N} && (13) \\ \mu_k^{new} &= \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n && (14) \\ \Sigma_k^{new} &= \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\left(\mathbf{x}_n-\mu_k^{new}\right)\left(\mathbf{x}_n-\mu_k^{new}\right)^T && (15) \\ N_k &= \sum_{n=1}^{N} r_{nk} && (16) \end{aligned}$$

III.II.4 EM Procedure

In summary, EM algorithm on GMM is composed of the following steps:
1. Initialize model parameters $\{\pi_k,\mu_k,\Sigma_k\}$ and evaluate the initial value of log likelihood
2. **E Step.** Evaluate the responsibilities using Equ(12).
3. **M Step.** Re-Compute model parameters using Equ(13-16).
4. Evalute the log likelihood using Equ(7) and check for convergence. If the algorithm has not converged yet, goes to step 2.

III.III Model Selection

Following the steps of EM, we are capable of efficiently estimating parameters in GMM. However, it is still required to specify the number of K manually. To choose the optimal value of K, Bayesian Information Criterion(BIC) is given by $p\left(\mathbf{X},\mathbf{Z}|\mu,\Sigma,\pi\right)$ and

$$lnp\left(\mathbf{X}|\theta\right)-\frac{1}{2}MlnN \qquad (17)$$

where $lnp\left(\mathbf{X}|\theta\right)$ is the maximum log likelihood obtained from EM, M is the number of total model parameters and N is number of data points. Compared with AIC in Equ(2), we find that BIC penalizes model complexity more heavily.

Then, given a reasonable range of K, we prefer the model with the highest BIC value.

IV Variational Bayesian Mixture Model

IV.I Introduction

Although EM algorithm opens the possibility to learn GMM with great ease and efficiency, it has some fatal weaknesses.

First, using EM to learn GMM is a maximum likelihood approach because model parameters are viewed as fixed constant rather than stochastic variables. In this way, the joint distribution in EM is given by $p(\mathbf{X},\mathbf{Z}|\mu,\Sigma,\pi)$ rathe than $p(\mathbf{X},\mathbf{Z},\mu,\Sigma,\pi)$. Therefore, the results of EM could be misleading because distributions over model parameters are not necessarily peak.

Second, in the process of EM, it is possible that one of the Gaussian component collapses onto one data point which contribute an ever increasing additive value to the log likelihood, causing singularity problem.

Given these two issues, we instead seek Bayesian version of EM, and in this context, how to automatically find the the number of components in mixture models could be an interesting topic.

IV.II Model Definition

In this section, we also use GMM as an example but we define the model in the Bayesian sense.

Similarly, starting from a set of observations $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ along with i.i.d assumption, the joint distribution of mixture model is given by:

$$p(\mathbf{X},\mathbf{Z},\mu,\Lambda,\pi) = p(\mathbf{X}|\mathbf{Z},\mu,\Lambda)\,p(\mathbf{Z}|\pi)\,p(\pi)\,p(\mu|\Lambda)\,p(\Lambda) \tag{18}$$

where

$$\begin{aligned}
p(\mathbf{X}|\mathbf{Z},\mu,\Lambda) &= \prod_{n=1}^{N}\prod_{k=1}^{K} N\left(\mathbf{x}_n|\mu_k,\Lambda_k^{-1}\right)^{z_{nk}} && (19)\\
p(\mathbf{Z}|\pi) &= \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} && (20)\\
p(\pi) &= Dir(\pi|\alpha_{\mathbf{0}}) = C(\alpha_{\mathbf{0}})\prod_{k=1}^{K}\pi_k^{\alpha_0-1} && (21)\\
p(\mu|\Lambda)\,p(\Lambda) &= \prod_{k=1}^{K} N\left(\mu_k|\mathbf{m}_0,(\beta_0\Lambda_k)^{-1}\right)\mathrm{Wis}\left(\Lambda_k|W_0,\nu_0\right) && (22)
\end{aligned}$$

$p(\pi)$ is defined to be a Dirichlet distribution, which is a conjugate prior of $p(\mathbf{Z}|\pi)$, in which normalization term $C(\alpha) = \frac{\Gamma(\hat{\alpha})}{\Gamma(\alpha_1)...\Gamma(\alpha_K)}$ and $\hat{\alpha} = \sum_{k=1}^{K}\alpha_k$

$p(\mu_k|\Lambda_k)\,p(\Lambda_k)$ is defined to be a Gaussian-Wishart distribution, which is a conjugate prior of $N\left(\mathbf{x}_n|\mu_k,\Lambda_k^{-1}\right)$ if both of mean and precision matrix are unknown. Each Wishart distribution is defined by

$$\text{Wis}(\Lambda|W,\nu) = B(W,\nu)|\Lambda|^{(\nu-D-1)/2}exp(-\frac{1}{2}Tr(W^{-1}\Lambda))$$
$$B(W,\nu) = |W|^{-\nu/2}(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma(\frac{\nu+1-i}{2}))^{-1} \tag{23}$$

The graphical representation of Bayesian GMM is given in Figure 3. Compared with Figure 2., we find that model parameters are represented in a circular node, indicating that they are stochastic variables.

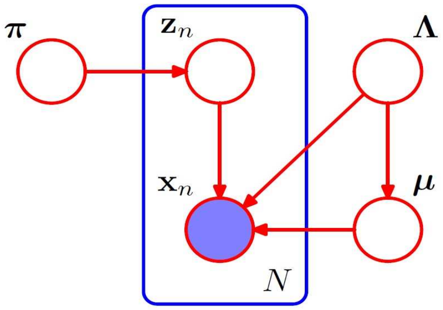

Figure 3: Graphical Representation of the Bayesian Gaussian Mixture Model[1].

IV.III Variational Inference

IV.III.1 Variational Distribution

Following the strategy in EM, we would like to derive posterior distribution $p(\mathbf{Z}|\mathbf{X})$, but it is usually intractable due to high-dimensional parametric space and model complexity. Therefore, we consider a variational distribution $q(\mathbf{Z},\pi,\mu,\Lambda)$ to approximate the true posterior distribution, and it is assumed to factorize between latent variables and model parameters as follows:

$$q(\mathbf{Z},\pi,\mu,\Lambda) = q(\mathbf{Z})q(\pi,\mu,\Lambda) \tag{24}$$

IV.III.2 Variational E Step

Given Equ(24), using mean field recipe on variational Bayesian analysis[6], the log of optimal solution of $q(\mathbf{Z})$ is given by:

$$lnq^*(\mathbf{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}ln\rho_{nk} + const \tag{25}$$

where:

$$ln\rho_{nk} = \mathbb{E}[ln\pi_k] + \frac{1}{2}\mathbb{E}[ln|\Lambda_k|] - \frac{D}{2}ln(2\pi) - \frac{1}{2}\mathbb{E}_{\mu_{\mathbf{k}},\Lambda_k}[(\mathbf{x}_n - \mu_k)^T\Lambda_k(\mathbf{x}_n - \mu_k)] \tag{26}$$

$$\mathbb{E}_{\mu_{\mathbf{k}},\Lambda_k}[(\mathbf{x}_n - \mu_k)^T\Lambda_k(\mathbf{x}_n - \mu_k) = D\beta_k^{-1} + \nu_k(\mathbf{x}_n - \mathbf{m}_k)^T W_k(\mathbf{x}_n - \mathbf{m}_k) \tag{27}$$

$$\mathbb{E}[ln|\Lambda_k|] = \sum_{i=1}^{D}\psi(\frac{\nu_k + 1 - i}{2}) + Dln2 + ln|W_k| \tag{28}$$

$$\mathbb{E}[ln\pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha}) \tag{29}$$

and $\psi(\cdot)$ is digamma function.

After taking exponential of Equ(25) and normalization, we obtain $q^*(\mathbf{Z})$:

$$q^*(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}} \tag{30}$$

and r_{nk}, the responsibility, is given by:

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K}\rho_{nj}} \tag{31}$$

IV.III.3 Variational M Step

Also, using mean field approach[6], the log of optimal solution of $q(\pi, \mu, \Lambda)$ is given by:

$$\begin{aligned} lnq^*(\pi,\mu,\Lambda) &= lnp(\pi) + \sum_{k=1}^{K} lnp(\mu_{\mathbf{k}},\Lambda_k) + \mathbb{E}_{\mathbf{Z}}[lnp(\mathbf{Z}|\pi)] \\ &\quad + \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]lnN(\mathbf{x}_n|\mu_k,\Lambda_k^{-1}) + const \end{aligned} \tag{32}$$

And this equation implies further factorization:

$$q^*(\pi,\mu,\Lambda) = q^*(\pi)\prod_k q^*(\mu_{\mathbf{k}},\Lambda_k) \tag{33}$$

Exponentiating terms in Equ(32) only dependent on π, we recognize $q^*(\pi)$ as a Dirichlet distribution:

$$q^*(\pi) = Dir(\pi|\alpha) \tag{34}$$

$$\alpha_k = \alpha_0 + N_k \tag{35}$$

$$N_k = \sum_{n=1}^{N} r_{nk} \tag{36}$$

Focusing on terms in Equ(32) dependent on $\mu_{\mathbf{k}}, \Lambda_k$, we get a Gaussian-Wishart distribution:

$$\begin{aligned}
q^*(\mu_{\mathbf{k}}, \Lambda_k) &= N(\mu_{\mathbf{k}}|\mathbf{m_k}, (\beta_k\Lambda_k)^{-1})\text{Wis}(\Lambda_k|W_k, v_k) && (37)\\
\beta_k &= \beta_0 + N_k && (38)\\
\mathbf{m_k} &= \frac{1}{\beta_k}(\beta_0\mathbf{m_0} + \mathbf{N_k}\bar{\mathbf{x}}_k) && (39)\\
W_k^{-1} &= W_0^{-1} + N_k\mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{\mathbf{x}}_k - \mathbf{m}_0)(\bar{\mathbf{x}}_k - \mathbf{m}_0)^T && (40)\\
v_k &= v_0 + N_k && (41)\\
\bar{\mathbf{x}}_k &= \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n && (42)\\
\mathbf{S}_k &= \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^T && (43)
\end{aligned}$$

IV.III.4 Lower Bound

In variational Bayesian framework, the lower bound of the log likelihodd is defined by:

$$L(q) = \int q(\mathbf{Z}) ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z} \leq lnp(\mathbf{X}) \qquad (44)$$

In the case of Bayesian GMM, this lower bound is given:

$$\begin{aligned}
L &= \sum_{\mathbf{Z}} \int\int\int q(\mathbf{Z}, \pi, \mu, \Lambda) ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Lambda)}{q(\mathbf{Z}, \pi, \mu, \Lambda)} \right\} d\pi d\mu d\Lambda \\
&= \mathbb{E}[lnp(\mathbf{X}|\mathbf{Z}, \mu, \Lambda)] + \mathbb{E}[lnp(\mathbf{Z}|\pi)] + \mathbb{E}[lnp(\pi)] + \mathbb{E}[lnp(\mu, \Lambda)] \\
&\quad -\mathbb{E}[lnq(\mathbf{Z})] - \mathbb{E}[lnq(\pi)] - \mathbb{E}[lnq(\mu, \Lambda)] && (45)
\end{aligned}$$

And each expectation term is given by:

$$\begin{aligned}
\mathbb{E}[lnp(\mathbf{X}|\mathbf{Z},\mu,\Lambda)] &= \frac{1}{2}\sum_{k=1}^{K} N_k \{\ \mathbb{E}[ln|\Lambda_k|] - D\beta_k^{-1} - \nu_k Tr(\mathbf{S_k}\mathbf{W_k}) \\
&\quad -\nu_k(\overline{\mathbf{x}}_k - \mathbf{m}_k)^T \mathbf{W_k}(\overline{\mathbf{x}}_k - \mathbf{m}_k) - Dln(2\pi)\ \} \qquad (46) \\
\mathbb{E}[lnp(\mathbf{Z}|\pi)] &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\mathbb{E}[ln\pi_k] \qquad (47) \\
\mathbb{E}[lnp(\pi)] &= lnC(\alpha_0) + (\alpha_0)\sum_{k=1}^{K}\mathbb{E}[ln\pi_k] \qquad (48) \\
\mathbb{E}[lnp(\mu,\Lambda)] &= \frac{1}{2}\sum_{k=1}^{K}\{\ Dln(\beta_0/2\pi) + \mathbb{E}[ln|\Lambda_k|] - \frac{D\beta_0}{\beta_k} \\
&\quad -\beta_0\nu_k(\mathbf{m}_k - \mathbf{m}_0)^T\mathbf{W_k}(\mathbf{m}_k - \mathbf{m}_0)\ \} + KlnB(\mathbf{W_0},\nu_0) \\
&\quad + \frac{(\nu_0 - D - 1)}{2}\sum_{k=1}^{K}\mathbb{E}[ln|\Lambda_k|] - \frac{1}{2}\sum_{k=1}^{K}\nu_k Tr(\mathbf{W}_0^{-1}\mathbf{W}_k) \qquad (49) \\
\mathbb{E}[lnq(\mathbf{Z})] &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}lnr_{nk} \qquad (50) \\
\mathbb{E}[lnq(\pi)] &= \sum_{k=1}^{K}(\alpha_k - 1)\mathbb{E}[ln\pi_k] + lnC(\alpha) \qquad (51) \\
\mathbb{E}[lnq(\mu,\Lambda)] &= \sum_{k=1}^{K}\left\{\frac{1}{2}\mathbb{E}[ln|\Lambda_k|] + \frac{D}{2}ln\left(\frac{\beta_k}{2\pi}\right) - \frac{D}{2} - H[q(\Lambda_k)]\right\} \qquad (52)
\end{aligned}$$

where $H[q(\Lambda_k)]$ is the entropy of the Wishart distribution, defined as:

$$H[q(\Lambda_k)] = -lnB(\mathbf{W},\nu) - \frac{(\nu - D - 1)}{2}\mathbb{E}[ln|\Lambda|] + \frac{\nu D}{2} \qquad (53)$$

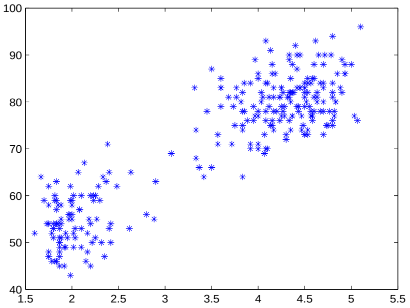

Figure 4: Old Faithful Data Plot.

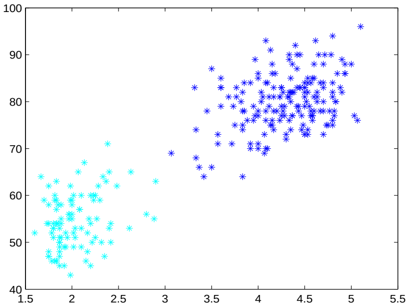

Figure 5: EM Clustering Result.

IV.III.5 Variational Bayesian GMM Procedure

Variational Bayesian approximation to learn GMM is composed of the following steps:
1. Initialize model prior hyper-parameters $\{\alpha_0, \mathbf{m}_0, \beta_0, \mathbf{W}_0, \nu_0\}$ and responsibilities $\mathbf{Z}$.
2. **Variational M Step.** Compute model parameters using Equ(34-43).
3. **Variational E Step.** Evaluate the responsibilities using Equ(31).
4. Evalute the lower bound given in Equ(45) and check for convergence. If the algorithm has not converged, goes to step 2

IV.IV Model Selection

In this project, we choose the optimal number of components by plotting variational lower bound versus the number of K. In this way, using Bayesian approach, the number of K can be determinated directly from data points without resorting to techniques such as cross validation or information criteria.

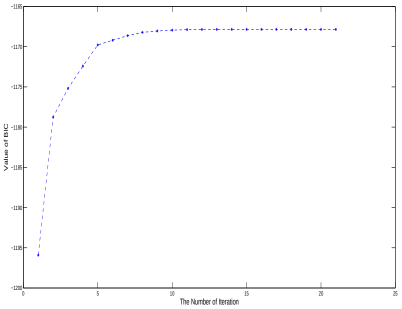

Figure 6: EM Convergence Plot.

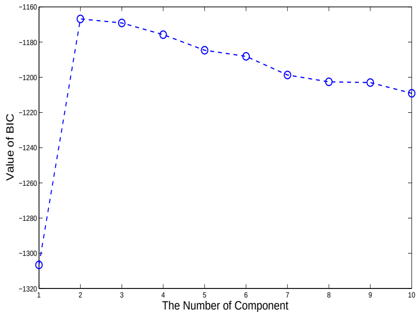

Figure 7: Plot of Average BIC vs. the Number of K in the Range of (1,10).

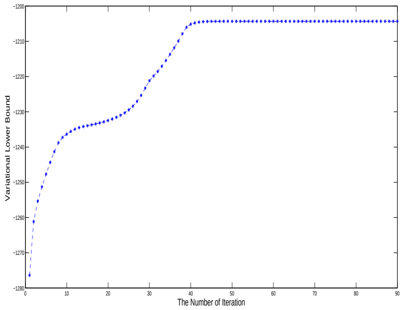

Figure 8: Variational Bayesian Convergence Plot.

V Experimental Evaluation

V.I Data Set

In this project, we use the simple ’Old Faithful’ data set. It compreises 272 measurements of the eruption of the Old Faithful geyser at Yellowstone National Park. Each meansurement comprises the duration of the eruption in minutes and the time in minute to the next eruption. So, the set is actually composed of 272 2D points with two main clusters and is used as demo dataset in several popular textbooks[1, 6]. The data plot is shown in Figure 4.

V.II EM Result

In this section, we firstly show the EM clustering results in Figure 5., and find that the algorithm is able to capture two clusters. Actually, GMM with EM can be viewed as a generalization of Kmeans[1], and can be used as an effective clustering techinique. Then, we plot the convergence of EM on Old Faithful data in Figure 6., noting that the algorithm will converge very fast with little computational cost.

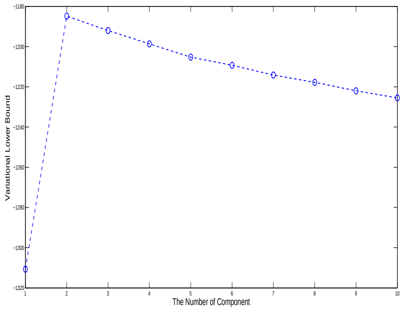

Figure 9: Plot of Average Lower Bound vs. the Number of K in the Range of (1,10)..

Next, we try to detect the optimal number of component in GMM using BIC discussed in the section of EM model selection by plotting BIC versus a range of K in Figure 7. To eliminate the influence of random initialization, each BIC value is averaged over 100 runs for each model.

V.III Variational Bayesian Result

Similarly, we plot the convergence of variational Bayesian on Old Faithful data in Figure 8., indicating that although the Bayesian model is more complicated there is little computational overhead. Then, we try to detect the optimal number of component in Bayesian GMM by plotting variational lower bound versus a range of K in Figure 9.

VI Discussion

In this project, we have briefly discussed how to find the best number of components in mixture models and we find that some criteria, such as BIC and lower bound of Bayesian log likelihood , could be useful. In the future, we would like to explore other alternative solutions. For example, a simple approach is to use Dirichlet process mixture model[7], which allows for an unbounded number of mixture components, and can be fit using Gibbs sampling

References

[1] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. springer New York, 2006, vol. 1.

[2] A. K. Jain, “Data clustering: 50 years beyond k-means,” *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[3] N. M. Nasrabadi and R. A. King, “Image coding using vector quantization: A review,” *Communications, IEEE Transactions on*, vol. 36, no. 8, pp. 957–971, 1988.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[5] H. Akaike, “A new look at the statistical model identification,” *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, 1974.

[6] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT Press, 2012.

[7] S. N. MacEachern and P. Müller, “Estimating mixture of dirichlet process models,” *Journal of Computational and Graphical Statistics*, vol. 7, no. 2, pp. 223–238, 1998.