ECI 2017 Bayesian Models’ Answers

By Julián Bayardo[[1]](#footnote-1)

Table of Contents

[Remark 1](#_Toc491618697)

[Questions 1](#_Toc491618698)

[Answers for the default data 3](#_Toc491618699)

[Question 1 3](#_Toc491618700)

[Question 2 4](#_Toc491618701)

[Question 3 5](#_Toc491618702)

[Question 4 6](#_Toc491618703)

[Answers for the generated data 7](#_Toc491618704)

[References 7](#_Toc491618705)

# Remark

I wrote this document in order to avoid polluting the IPython notebook and making it hard to understand. All of the parameters for plots and experiments can be found on the first cell of the accompanying IPython notebook; plots can be regenerated by changing the parameters and running all cells again.

# Questions

1. What can you say about the obtained posterior distributions? What do they represent? How do these posterior distribution compare to the parameter estimates obtained from the EM algorithm?
2. Sample from the approximate posterior distribution and plot the GMM distributions corresponding to all the samples into a single figure. Comment on this plot. What do the individual GMM distributions represent?
3. Now average all the samples from the previous step. What can you say about the obtained average distribution? What does it represent?
4. How does the posterior predictive distribution compare to:
   1. The true training data distribution
   2. The GMM obtained using ML training (i.e. using EM algorithm)
   3. The average of GMM distributions obtained in the previous step by sampling

Regenerate all the plots with a larger number of training observations and comment on how they change from the previous experiments with a smaller training dataset.

# Answers for the default data

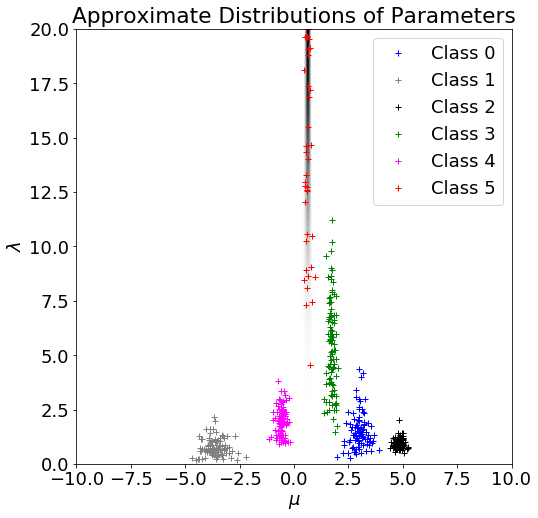
## Question 1

The posterior distributions represent the confidence that we have on a specific parameter (or subset of them) being equal to one value. In the case of the approximate prior distribution of weights q(pi), each of the values represents the probability that one particular gaussian is used to generate the data; while for q(mu, lambda), we have the probability that one particular set of parameters is fed into the normal distribution that is used to generate the data when that specific class is selected.

Since q(pi) is also a proxy to how much data is given for one particular q(mu, lambda) to be approximated, it also means that we will get more confidence (i.e. lower variance) for the q(mu, lambda) distributions that have more data. This can be clearly seen in the plot below: class 2 has the highest probability in the approximate distribution, and accordingly a highly peaked distribution; while other classes have similar lower values, and are less concentrated.

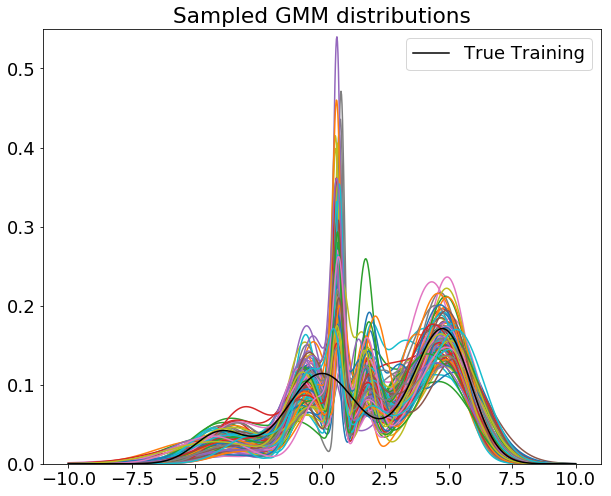
The EM parameters are akin to taking one sample from the approximate posterior distributions: in the EM algorithm, we converge to a single set of parameters for our normal distributions, instead of a distribution over the possible parameters.

Approximate posterior distribution of weights (q(pi)): [0.09530115, 0.0962085, 0.40421586, 0.10629599, 0.19582768, 0.10215082]



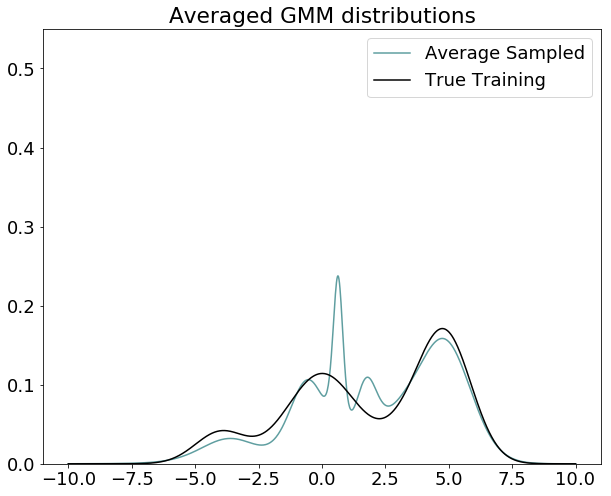
## Question 2

Each sample from q(pi, mu, lambda) is one possible set of parameters for the Bayesian inference network according to the approximate posterior distribution over the parameters. Hence, plotting the distribution given by the sample corresponds to one possible probability distribution over the data generated by the GMM process as learned by the VBGMM algorithm.



## Question 3

If we average the distributions obtained by sampling from q(pi, mu, lambda), we get the following plot:

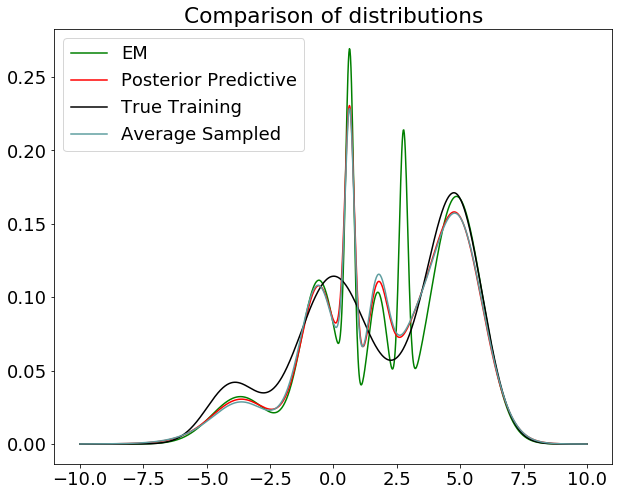


If you compare the previous plot with this one, it is clear that it resembles the distributions from the previous plot, and has the same “mistake” trends (i.e. the humps at ~0.5 and ~2.0).

Intuitively and informally, it makes sense to see this happening: it is what one would expect from the law of large numbers if we applied it to random variables whose range is the space of probability distributions generable by the learning algorithm.

## Question 4

If we plot all of the four distributions together, we get the following plot:



There are a few key things to point out from this plot:

* The average sampled distribution follows the posterior predictive distribution really closely. This makes sense as the posterior predictive distribution is taking into consideration the uncertainty over the parameters that the averaged is not taking by itself.

# Answers for the generated data

# References

**There are no sources in the current document.**

1. [julian@bayardo.info](mailto:julian@bayardo.info). LU: 850/13. DNI: 37 835 049. [↑](#footnote-ref-1)