**Bayesian Statistics and Marketing**

Notes taken from the book found at: https://www.wiley.com/en-us/Bayesian+Statistics+and+Marketing-p-9780470863671

**Chapter 2 – Bayesian Essentials**

**2.0 Essential Concepts from Distribution Theory**

Bayesian inference relies heavily on probability theory, specifically on the different distributions that inform what probability to give to an outcome that you are attempting to predict.

Some of the more common distributions within Bayesian applications include:

* *Discrete random variable*

In this case, the variable X can take on a countable number of values, each with some probability associated with it. An example of this is a *Bernoulli* random variable that has only two values – a 1 with a probability p and a value of 0 with a probability 1-p. Examples of Bernoulli variables include whether a company will sell a product tomorrow, with p representing the probability of a sale.

Another examples of a discrete random variable is the number of units sold tomorrow. Here, the value of X can be 0, 1, 2, etc. Each value of X has an associated probability, with the sum of probability of all possible values adding up to 1.

If we know that the distribution of our variable is discrete, we can answer questions such as the probability that there will be at least one sale tomorrow, probability there will be between 1 and 10 units sold, etc.

Another question we can answer is the *expectation* on the number of units sold which is the sum of all the values of X multiplied by each value’s probability. So, let’s say X has 4 distinct values – 0, 1, 2, and 3. If there is a 20% chance that 1 unit is sold, 30% chance 2 units are sold, 10% chance 3 units are sold (and 40% chance 0 units are sold), the expectation of the number of units sold tomorrow is 1.1.

* *Continuous random variable*

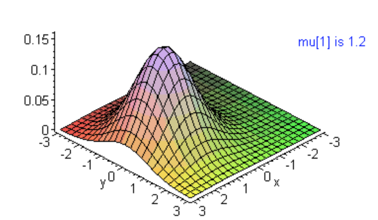
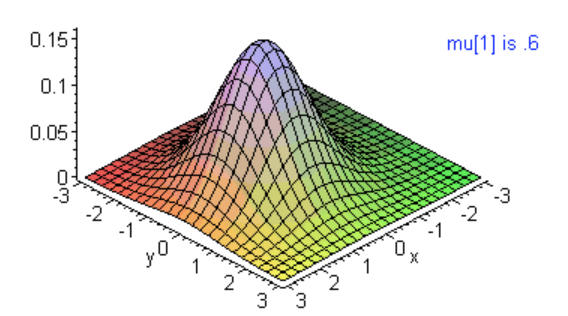
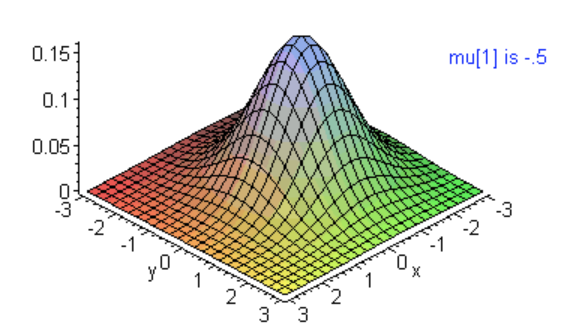
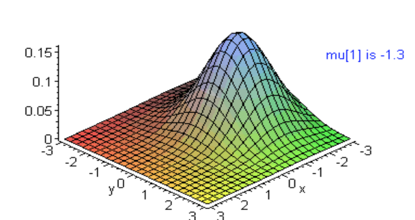
These are variables that can take any nonnegative number. We can summarize the probability distribution of X by the probability density function (PDF) – an example of which is the bell curve. This PDF is a *rate* *function* – meaning it is a function used to quantify the probability of rare events – which *integrate* to 1 (integrating to 1 means that the sum of probabilities come very close but never reach 1). For continuous variables, the probability that X is any specific value is usually very low, but similar values have similar probabilities, which is why we use a rate function to describe the distribution of such variables.

Discrete random variables have a similar concept known as the ­*probability mass function*, which simply shows the probability of all the unique values of that variable.

* *Joint distribution*

In many cases, we want to consider the joint distribution of two or more random variables, X and Y. As an example – we could consider the number of sales tomorrow in two different markets. In these cases, there is a bivariate density function which gives the probability rate per unit of area in the two-dimensional plane. We then use this bivariate density function to compute the probability of any set of (X,Y) values. For example, we can compute the probability that both X and Y are positive by calculating the area under the bivariate density curve in the positive quadrant of the two-dimensional plane.

Example of joint density with changing values of the mean of X



*Marginal Distribution*

If we have the joint density of X and Y, we can also compute the *marginal* densities of each of the variables X and Y. The marginal density of any variable within a joint distribution is simply the distribution of that variable without reference to the values of the other variables within that joint distribution.

In other words, when we know the joint distribution of X and Y, we can find the distributions of X and Y independent of one another. We can derive the joint distribution based upon observed values of X and Y through simulation; through this derived joint distribution we can look at only the values of X or Y to get the marginal distribution of those two variables, respectively.

*Conditional Distribution*

Once we have the joint distribution, we can also define the concept of *conditional* distribution. Whereas marginal distribution refers to the density of any variable included in the joint distribution without regard to the values of the other variables, *conditional* density refers to the density of a variable when the other variable(s) in the joint distribution are fixed.

If two continuous variables, X and Y, are at least somewhat correlated to one another (i.e., are not completely independent), then both X and Y have standard normal *marginal* distributions, along with standard normal *conditional* distributions. Because they are correlated, however, the mean of tge conditional distributions change along with the value of the other variable included in the joint distribution.

**2.1 The Goal of Inference and Bayes’ Theorem**

The goal of statistical inference is to use information to make inferences about unknown quantities. One important source of information is data, but there is a role for information that is not necessarily based on data. These sources of information include theories of behavior (e.g., demand slopes downward), subjective views that there is a structure underlying the unknown quantities, and expectations that the unknown quantity to be finite or within some range.

The term *unknown quantity* is nebulous for a reason. It can refer to both the quantity being predicted as well as the parameters to the model that generate those predictions.

The goal, then, is to make inferences about unknown quantities *given* the information we have available, whether that information is obtained through data or independently *prior* to the data. Bayesian inference uses probability statements as the basis of inferring values of the unknown quantities *conditional* on the data and prior information.

In order to take advantage of this concept of conditional probability, we need to encode the prior information as a probability distribution, which therefore requires the view that probability can represent subjective beliefs. This is where the controversy of Bayesian theory arises; some disagree this can/should be done.

*Bayes’ Theorem*

Some explanation of various notation in regards to Bayes Theorem:

* = Prior beliefs expressed as a probability distribution
* = Generic notation for the appropriate density
* – Information provided by the data, or the probability distribution of the data
* – *Posterior* distribution reflecting the data and prior information
* – Likelihood function, which is any function that is proportional to the probability distribution of the data,

To infer the values of the unknown quantities, we need to combine the prior and the likelihood to produce the distribution of the observable quantities given what was *actually* observed in the data. The term likelihood is differentiated from probability in this context in that the likelihood describes the plausibility of a *parameter* value of the model (e.g., the true ) which described the observed data, given the data we actually observed.

If the likelihood function equals the probability distribution of the data, , then the constant of proportionality is equal to the marginal distribution of the observed data. This assumes there is some constant that we can multiply by in order to produce . If the prior distribution represents a proper distribution, meaning that the area underneath its curve integrates to 1, then the constant of proportionality does exist.

How can a constant equal the marginal distribution? In other words, how do you boil down a distribution into a constant?

**2.2 Conditioning and the Likelihood Principle**

The likelihood principle states that the likelihood function, , contains all relevant information from the data. As mentioned above (repeated here because this is confusing), two likelihood functions are equivalent if one is a scalar equivalent to the other (i.e., they are *proportional* to one another).

This discussion illustrates the differences in the schools of thought of Bayesian inference and so-called frequentist inference which is based on experimental design, statistical hypothesis testing and confidence intervals.

In frequentist thought, parameters of a model are often treated as fixed but unknown – there is no concept of the probability of different values of these unknown parameters, because there is only one true value (e.g., The likelihood of the flip of a fair coin landing heads is 50% - there is no probability that the likelihood of that parameter is anything other than 50%).

In Bayesian inference, a probability distribution is given to the unknown parameters based on the data that has been observed.

Adhering to the likelihood principle means that any inferences are conditional on the observed data, since the likelihood function is defined based on the data.

In some frequentist approaches, the observed data is irrelevant in estimating the unknown parameters of the model since the design of the sampling experiment results in the distribution of the data; in other words, the data summarizes the properties of the estimator that were determined *before* the data was observed through the design of the experiment in which the data was collected.

**2.3 Prediction and Bayes**

One of the appeals of the Bayesian approach is that all unknown are treated the same. Prediction is defined as making probability statements about the distribution of as yet unobserved data, denoted by . The only distinction between parameters and unobserved data is that the unobserved data is potentially *observable*. In many cases, we assume that the observed data is independent of the unobserved data, , conditional on the parameters .

To generate predictions for unobserved data, we average the likelihood for the unobserved data over the posterior distribution of . This averaging accounts for the uncertainty in when forming predictive statements about .

Note that in the above equation, means the “delta of ”. In turn, this means a rectangle with an infinitely small width underneath the probability curve of . For more info see this [link](http://mathforum.org/library/drmath/view/60949.html).

**2.4 Summarizing the Posterior**

For most problems of practical interest, the posterior distribution has many different variables included. Summaries of the posterior play an important role in Bayesian statistics in that they allow us to generate interpretable predictions for unobserved data and evaluate those predictions.

In frequentist approaches, parameter estimates and standard errors serve this purpose. In Bayesian methods, moments of the marginal distribution of the parameters (i.e., quantitative measures like the mean and standard deviation that describe the shape of the probability distribution of the parameters) serve this purpose. In turn, simulation methods are especially well suited to produce the marginal distributions of the posterior. However, if the marginal distributions are non-normal, the mean and standard deviation are not particularly useful.

Before simulation methods, the *expectation* of a function of (e.g., its mean, standard deviation, etc.) was limited to marginal posteriors that have a known distributional form and have integrals that can be evaluated analytically (by, for example, taking various asymptotic approximations to these integrals).

I spent about a half hour trying to figure out what “taking various asymptotic approximations to these integrals” means in English, but unfortunately couldn’t find anything

Because of modern simulation methods, we do not have to rely on asymptotic approximations in modern Bayesian inference.

**2.5 Decision Theory, Risk, and the Sampling Properties of Bayes Estimators**

The general problem to solve within decision theory and prediction is to search among possible actions for the action which minimizes the expected loss. Here, we use the term *action* but in the context of parameter estimation, we need to search for the parameter value that minimizes the expected loss across all possible parameter values. In order to find this *Bayes action* or *Bayes estimator*, we can look at the posterior distribution of the action/parameter in question to find the value that minimizes the loss function.

There are many different loss functions for different types of problems. Likewise, different parts of the posterior distribution will be emphasized in calculating the Bayes estimator depending on which loss function is chosen for a particular problem. Mean squared error (MSE) is a common loss function; in problems that use MSE as the loss function, the mean of the posterior distribution is used as the Bayes estimator. For nonlinear loss functions, however, the spread or uncertainty is just as important as location. In either case, particular aspects/moments of the posterior distribution is used to find the Bayes estimator that minimizes the loss function.

We used the term *loss function* multiple times here, but we could have just used the term *risk function* as well; they are synonymous. The *risk* is determined as a function of a given parameter value, , and so there is a different risk at every point in the parameter space. Bayes estimators minimize the loss function across *the entire parameter space*, which is a crucial distinction. There can be estimators that outperform the Bayes estimator within certain regions (but not all of) the parameter space. Bayes estimators perform very well if you are in the region of the parameter space that you expect to be in based on your prior.

**2.6 Identification and Bayesian Inference**

Within statistical modeling, the problem of *identification* arises when there is a set of different parameter values that give rise to the same distribution of data. In such cases, the set of parameter values are said to be *observationally equivalent* since the distribution of the data is the same for any member of this set. Lack of identification implies that there will be regions within the likelihood function that is constant for any given data set.

The reason why this is less of a problem in Bayesian analysis is because the prior can modify the shape of the likelihood in the posterior. Other methods of inference such as maximum likelihood will encounter problems with unidentified models; the maximizing function may be unstable in regions of constancy in the likelihood. With an informative prior, Bayesian analysis can overcome the issues associated with lack of identification since the posterior may not have any region of constancy (since the prior “overtakes” the posterior in that section of the distribution. However, this means that the lack of identification implies that there are certain functions of the parameters for which the posterior is entirely driven by the prior.

**2.7 Conjugacy, Sufficiency and Exponential Families**

Prior to modern simulation methods, it was only practical to use Bayesian methods for models that would allow you to summarize the posterior in an analytical expression that is relatively straightforward to solve. One approach to ensuring this occurs is to require that the prior distribution be *conjugate* to the likelihood; thus means that the posterior distribution is in the same family of distributions as the prior.

As has been discussed in the notes prior to this, Bayes theorem is about making inferences about the posterior distribution of a parameter which is the product of the likelihood function and prior , normalized (divided) by the probability of some data :

The denominator is the area underneath the curve (symbolized by ) of the derivative of the probability curve indicated by the data (symbolized by - the represents the change in slope of the probability density of the information provided by the data itself).

For certain choices of the prior distribution – symbolized by – the posterior has the same distribution as the prior, albeit with different parameter values (in most cases). These choices are the *conjugate priors*. Conjugate priors are an algebraic convenience. While it is possible to calculate the posterior from the prior that is of a different distributional family as one another, complicated math involving numerical integration and calculus is required. With conjugate priors, only algebra is used. There is the practical consideration that with conjugate priors, the likelihood function is more intuitive in how it updates the prior to produce the posterior. All members of the [exponential family of distributions](https://en.wikipedia.org/wiki/Exponential_family) have conjugate priors.

Regarding the term *sufficiency*, a statistic that describes a family of probability distributions is sufficient if the sample from which it is calculated gives no additional information than does the statistic.

**Elements of Statistical Learning**

Notes taken from the book found at: <https://web.stanford.edu/~hastie/ElemStatLearn/>. There is some overlap in the concepts covered by the *Bayesian Statistics and Marketing* book above; in these cases I don’t necessarily repeat myself in the notes.

**Chapter 8 – Model Inference and Averaging**

* 1. **Bayesian Methods**

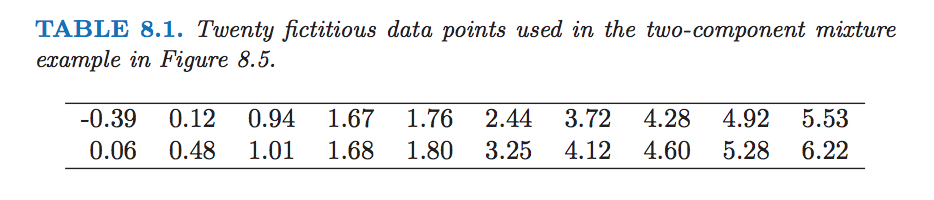
To understand the posterior distribution, we can draw samples from it or summarize by computing its mean or mode. The Bayesian approach differs from the standard (aka *frequentist*) method for inference because of its emphasis on uncertainty of point estimates for parameters – both in its use of the prior distribution expresses uncertainty present before seeing the data, and its use of the posterior distribution to express uncertainty after seeing the data.

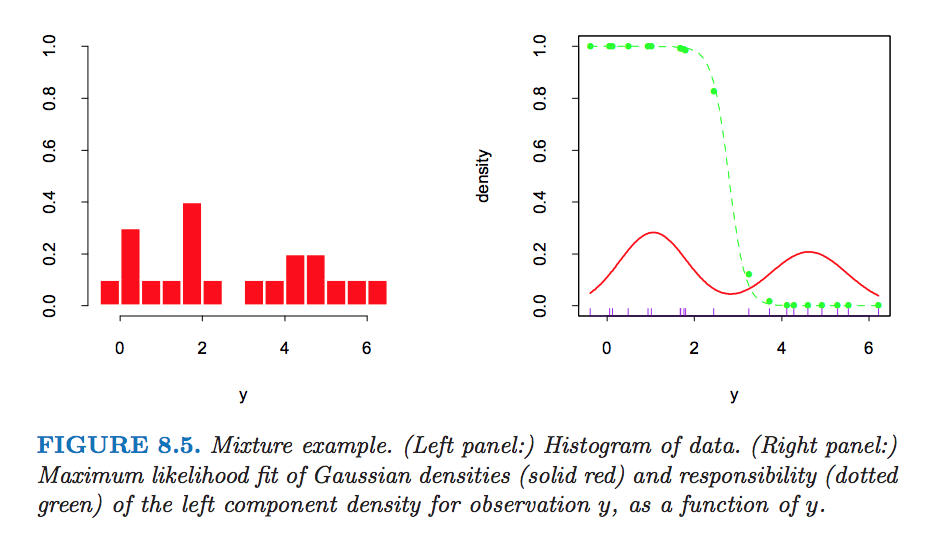
**8.5 The EM Algorithm**

In real life, posterior distributions do not fall neatly into a specific type of distribution (whether that distribution be uniform, Gaussian, etc.). There are cases where there are local maximum and minimums and the density “meanders” away from a traditional distribution.

*8.5.1 Two-Component Mixture Model*

One such instance is a so-called mixture model, which probabilistically represents sub-populations within an overall population density. Take the below table/graphs as an example. With the histogram of data on the left side showing actual observations, we can create a mixture model of two Gaussian distributions to represent the variable . Each of the two components of the mixture model are Gaussian, each with a different mean and standard deviation.





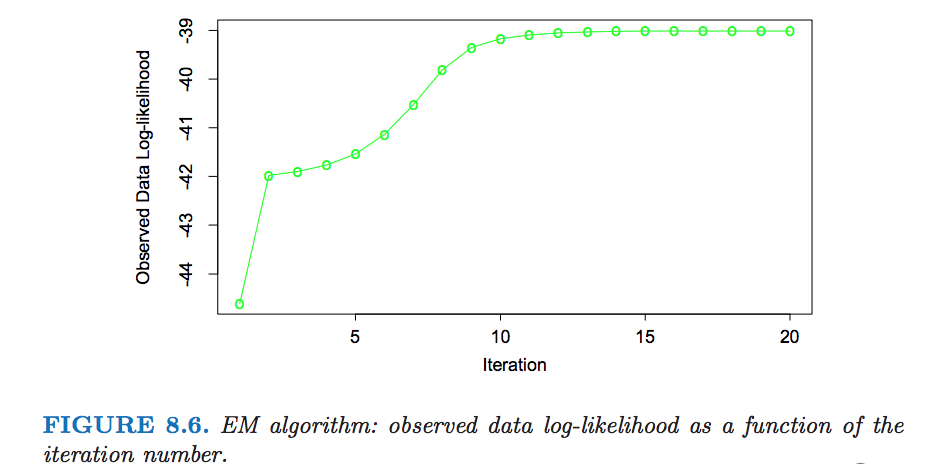
Creating a density distribution can be thought of as finding the maximum likelihood at every value of the variable of interest. The *Expectation-Maximization (EM) Algorithm* simplifies difficult maximum likelihood problems to create complex density distributions given the observed data at hand.

In the case of the mixture model above, direct maximization is difficult. However, if we consider unobserved *latent* variables to determine which of the two components of the mixture model an observation is a part of, solving this maximum likelihood problem becomes easier. In the example, the latent variable is one of two values (say, 0 and 1) , since the density mixture model contains two components. If , the observation belongs to the Gaussian component on the left-hand side of the graph above. If , the observation belongs to the Gaussian component on the right-hand side of the graph above.

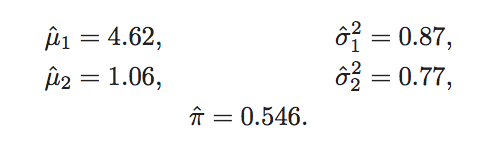
However, is actually unobservable and unknown, so we can proceed in an iterative fashion to find the *expected* value of for each observation. The expected value of the latent variable is called the *responsibility* for observation . The green curve in the plot above represents the responsibility of the left Gaussian component. This is the purpose of the *expectation* step in the EM algorithm, to (probabilistically) assign the observation to one of the models included in the distribution.

The *maximization* step then uses the responsibilities output from the expectation step to determine the estimates of the parameters that have the most likely of values given the mixed model’s component membership of the observation.

As mentioned, the EM algorithm is iterative and so initial values are needed for each component’s mean and variance included in the mixture model. As there are two components in the mixture model in our example, initial values of and could just be values if selected randomly. Values for and can be both be set to the sample variance as a whole. The mixing proportion (percentage of observations that belong to each component) could be uniform; in this case that would be .

As we run through iterations of the component, the actual maximizer of the likelihood happens when a spike of infinite height is found at any given data point; in other words and for a particular observation. This is not useful; we actually are looking for a good local maximum of the likelihood, one for which and . If we run the EM algorithm 20 times with a number of different initial guesses for the parameters, each with , we can choose the run that gave us the maximum likelihood.

In frequentist statistics, there isn’t any probability associated with this individual run which produced the maximum likelihood; this run is used to generate new predictions. In this particular case, the maximum likelihood estimates are:

**8.6 MCMC for Sampling from the Posterior**

Having defined a Bayesian model, one would like to draw samples from the resulting posterior distribution in order to make inferences about the parameters. This is often a difficult computational problem that is helped by the *Markov chain Monte Carlo* (MCMC) approach to posterior sampling. One method to carry out MCMC is called Gibbs sampling and is closely related to the EM algorithm – the main difference is that it samples from the conditional distributions rather than maximizing over them.

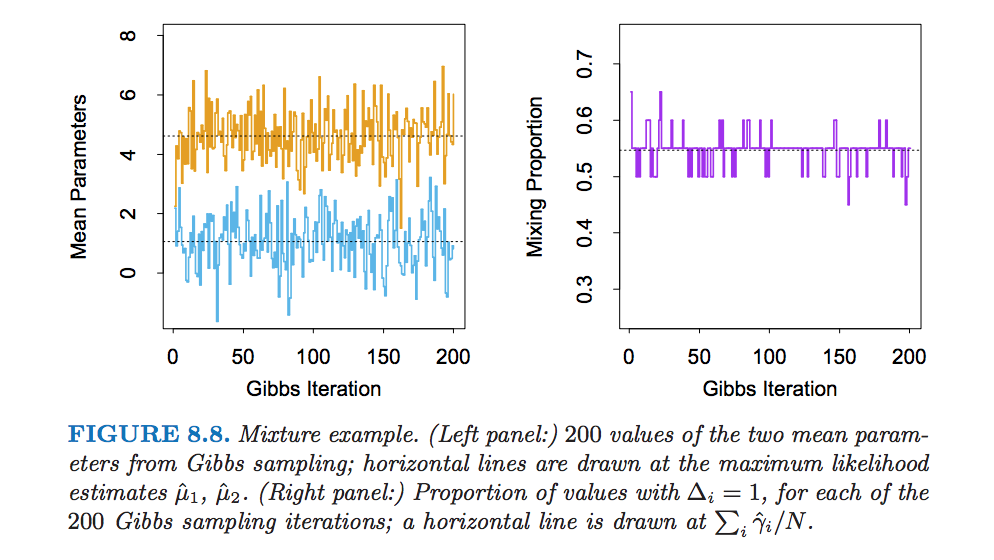
*Gibbs Sampler*

For the Gibbs sampling technique, say we have three random variables, , and . We wish to draw from the joint distribution of these three variables – this can be difficult to do, so we divide the problem into manageable chunks to simulate from the conditional distributions of each variable with the values of the other two variables fixed. In other words, we can successfully figure out the values of the follow probabilities:

The first step is to simulate from the conditional distribution of to find the likeliest value of given the initial values of and . This results in the value . Given this new value and the initial value of , we can find the value of , etc. Here the superscript 1 represents step 1 in the iterative process of the Gibbs sampler. Once we complete this step for all three variables, we move onto step 2 with values of , and fixed at the values output from step 1, and find conditional distributions of each individual variable. This process is repeated over and over again until the joint distribution of , and are *stationary*, meaning they don’t change from one iteration to the next. As the initial steps of this process depend heavily on the initial values (subjectively) set for , and respectively, there is a so-called *burn-in* number of iterations before stationarity is reached in the joint distributions (and therefore the conditional and marginal distributions of each variable included). This example only included three variables to make things somewhat easier to explain, but the same process could be extended to many, many more variables.

Getting back to Bayesian inference, the goal is to draw a sample from the joint posterior of the parameters given the data . Gibbs sampling is helpful to sample from the conditional distribution given the values of the other parameters and the data .

Often times, the parameter values are given through , but not always. To go back to the mixture model example from the **8.5 The EM Algorithm** section, one parameter not included in would be the latent variable which assigns the expected value of which component of the mixture model an observation belongs to. This can be thought of as another parameter for the Gibbs sampler. Rather than compute the responsibilities of the latent variable as in the EM algorithm, the Gibbs sampler simulates data from the distribution of given values of the other parameters and . In this way, the Gibbs sampler and EM algorithm are similar to one another.

Again using the mixture model example from the section above, we can generate plots which summarize the simulations from a Gibbs sampler to the same problem.

The above mixture model was simplified to illustrate the connection between the Gibbs sampler and the EM algorithm, but more realistically, you would put prior distributions on the variances of each component of the mixture model, and along with the *mixing proportion* which specifies the proportion of observations belonging to each component. Then you could include separate Gibbs sampling steps to get to their respective posterior distributions. Additionally, you can also incorporate proper informative priors for the mean parameters of each component.

Gibbs sampling is just one of many procedures to sample from a posterior distribution. Metropolis-Hastins is another ways, as is the No-U-Turn Sampler (NUTS).