Dirichlet Process Mixture Models and their Application in bioinfomatics

Shaun Dowling

Supervisor: Dr. Colin Campbell

Abstract

The genomic era has resulted in an explosion of biological data thanks to huge advances in the fields of molecular biology and genomics. Now, in the post-genomic era, researchers in these areas have been overwhelmed by experimental data. The field of bioinformatics is dedicated to working with this data and can be used to answer some very interesting and important questions.

A key component of all cancer treatment is to correctly diagnose the patient's cancer subtype so that the optimal treatment can be provided.

Since the most effective treatment can vary dramatically between cancer subtypes, giving the correct diagnosis is a very important step in the road to recovery. The aim of this project was to focus on classifying data into breast cancer subtypes.

Bioinformatics

The field of bioinformatics refers to the creation, maintenance, and analysis of biological information such as nucleotide and amino acid sequences. revolves around trying to answer important biological questions by making use of analytical machinery from statistics and computer science, but always being driven by biological understanding.These techniques are used to answer questions such as

- what are the functional roles of different genes and in what cellular processes do they participate;
- how are genes regulated, how do genes and gene products interact, and what are these interaction networks;
- how does gene expression level differ in various cell types and states, how is gene expression changed by various diseases or compound treatments.

As methods for attaining data develop, such as DNA sequencing and measuring gene expression data, there will be an ongoing need for ever more sophisticated methods in order to understand all the data that is produced. genomic era has resulted in an explosion of biological data thanks to huge advances in the fields of molecular biology and genomics.

Nowadays there is a vast amount of bioinformatic data available online from sources such as: Gene Expression Omnibus (GEO), the Cancer Genome Atlas, and Metabric. The amount of readily available data has really accelerated the development of bioinformatics and enabled even individual researches to investigate very rich datasets.

Microarrays

Arrays in which each probe represents a gene are hybridised to fluorescently labelled DNA prepared from different tissue samples. The results from any two samples are compared; if hybridisation is stronger in one sample for a specific gene then that gene is said to be differentially expressed in that sample.

Likewise, by looking at a sample from a person with a disease when compared against a control healthy sample, it is possible to see what genes that disease is causing to be over or under expressed. This second technique becomes particularly useful when looking at cancer patients because different cancer types can often be clearly differentiated by their expression profiles.

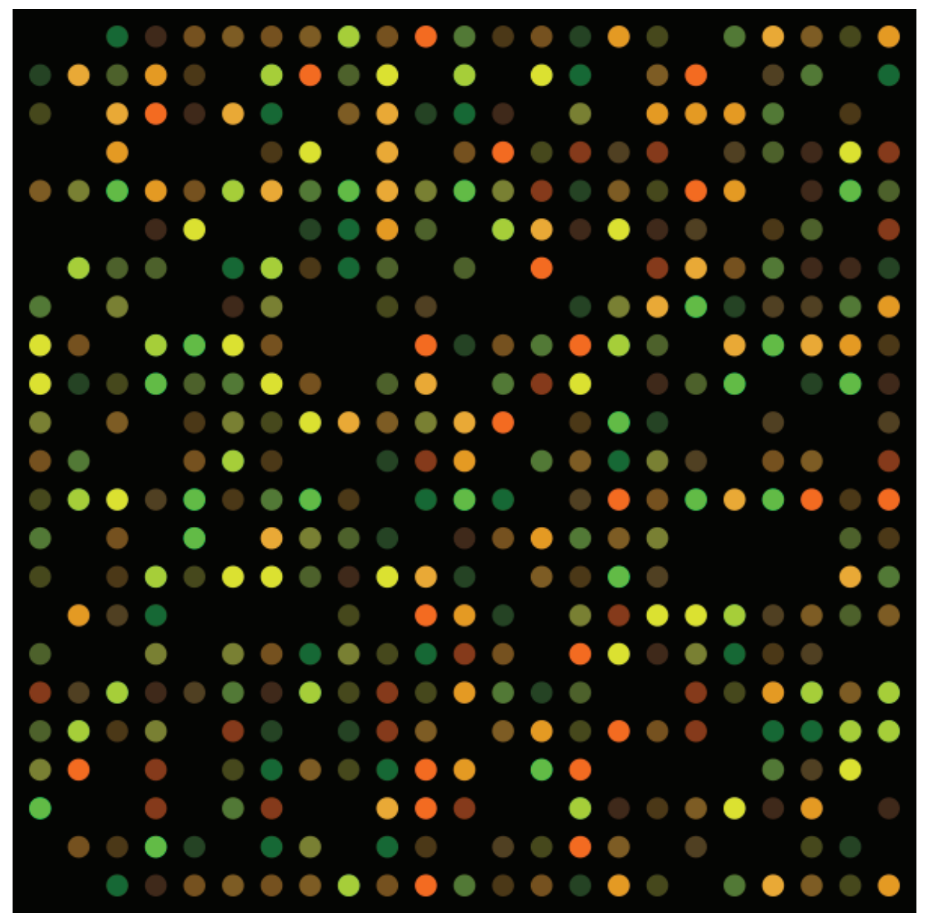

Figure 1: An example of a gene expression array.

Gaussian Mixture Models

The model used in thie project is an extension of a Gaussian mixture model. A mixture is the term used to describe a linear combination of a number of simple distributions to for a more complex distribution.

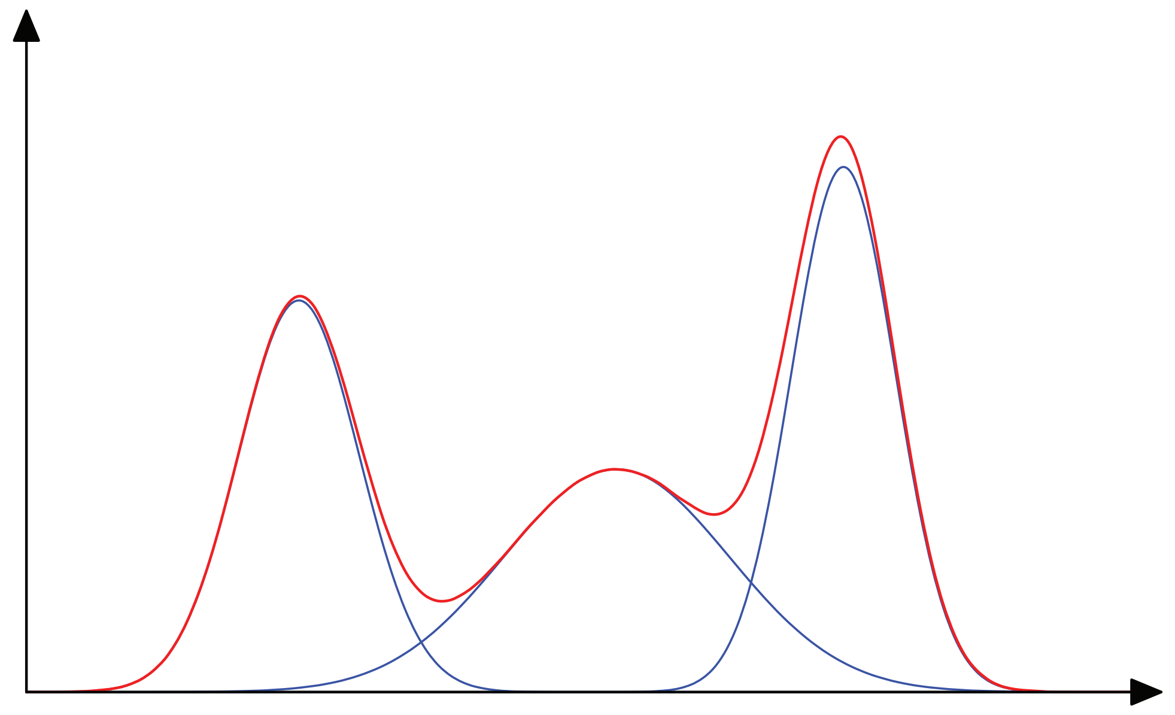

Figure 2: An exmaple of a 1-d mixture model. To cluster gene expression data, mixture models with around 10, 000 dimenions must be used.

Example

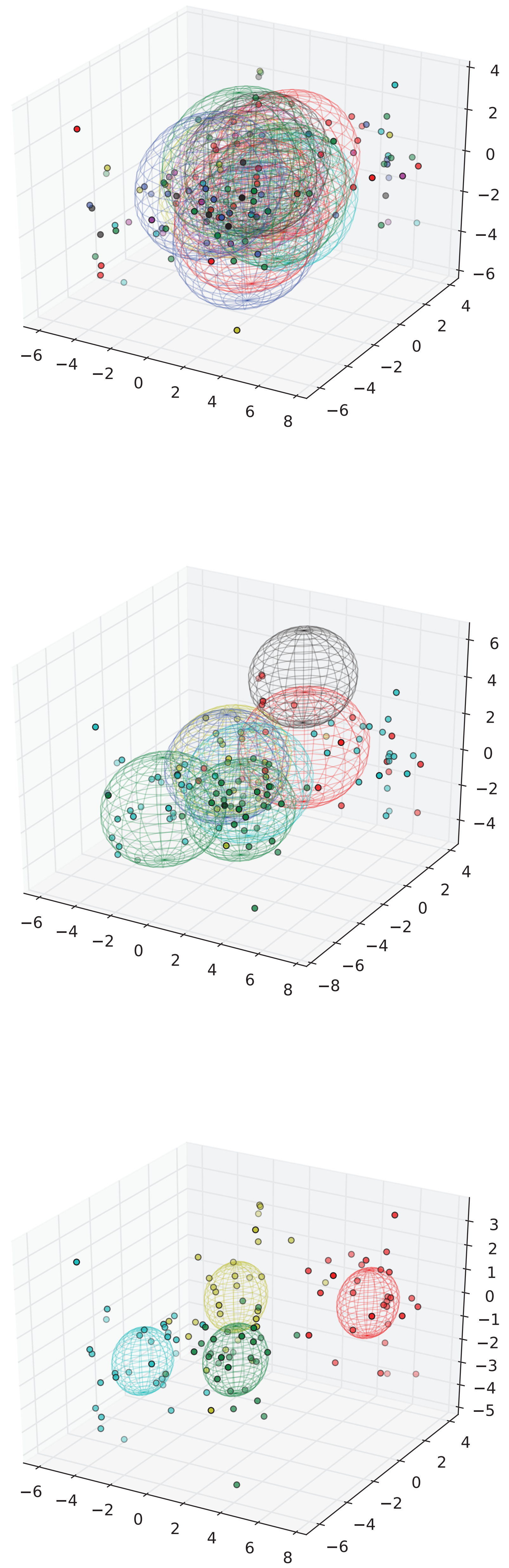

Figure 3: An example of the clustering algoritm working with three dimenions. It is easy to see how the number of clusters varies as the model runs

Dirichlet process

Making use of a Dirichlet process is what differentiates this model from a standard mixture model. By combining the two to create a Dirichlet process Gaussian mixture model, it is possible to infer the clustering in a set of data without setting the number of clusters a priori. The Dirichlet process is used as a prior over the cluster components, including how each component is weighted in the model and as a result of being used, allows for easy inference of the number of clusters.

Results

Figure 4 shows the results of running the model on some breast cancer gene expression data. There is a clearly visible peak at 5 clusters indicating that there are most likely 5 cancer subtypes in the data.

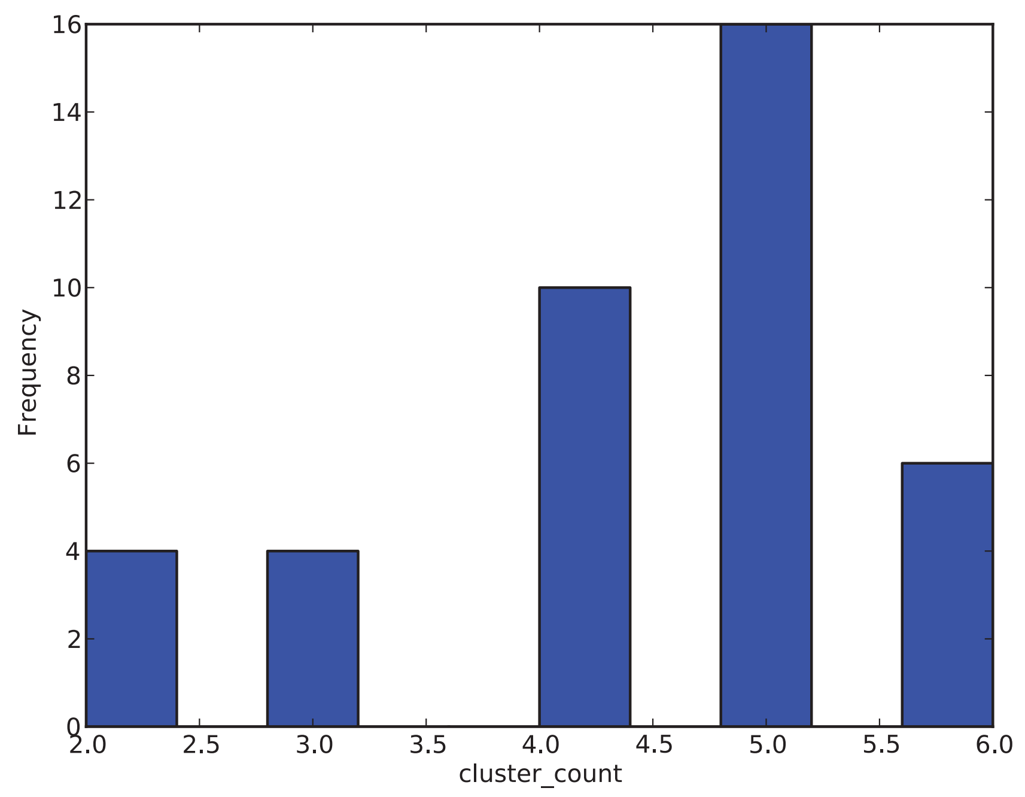

Figure 2: The results of running the model on real breast cancer data

Conclusion

Data from microarrays inherently provide a number of challenges when performing clustering of any sort. Dirichlet process mixture models undoubtedly provide a very elegant and flexible method to potentially sidestep the need for model selection and averaging techniques. Although performing well with synthesised data, all the model sometimes had issues when classifying real data.

In particular, one shortcoming of the Dirchlet process as it was applied in these models is that is appears to lose the dynamic addition and removal of cluster components that make it so effective at lower dimension.