

Hierarchical models with multiple mixtures of Dirichlet processes

George Tomlinson

Michael Escobar†

Christine McLaren‡

† Supported by the National Science and Engineering Research Council of Canada.

‡ Supported by NIH and Burroughs Wellcome Foundation

Motivating Data Set

- Red blood cell volumes.
- Modern diagnostic equipment can quickly and accurately measure blood cell sizes from a vial of blood. From one blood sample, 6 000 cells can be measured. Therefore, one is able to get a very good estimate of the distribution of the cell sizes.
- Clinical interest is in the shape of the distribution of the sizes.
- Different disease states have different distributional shapes from normal states, and from each other.
- **Question:** Does the distribution of blood cell sizes from a particular individual look like a typical distribution for someone from the normal population, or does it look like someone from a diseased population?

The following two figures show the histograms for the cell sizes for 20 normal subjects and for 20 subjects with anemia.

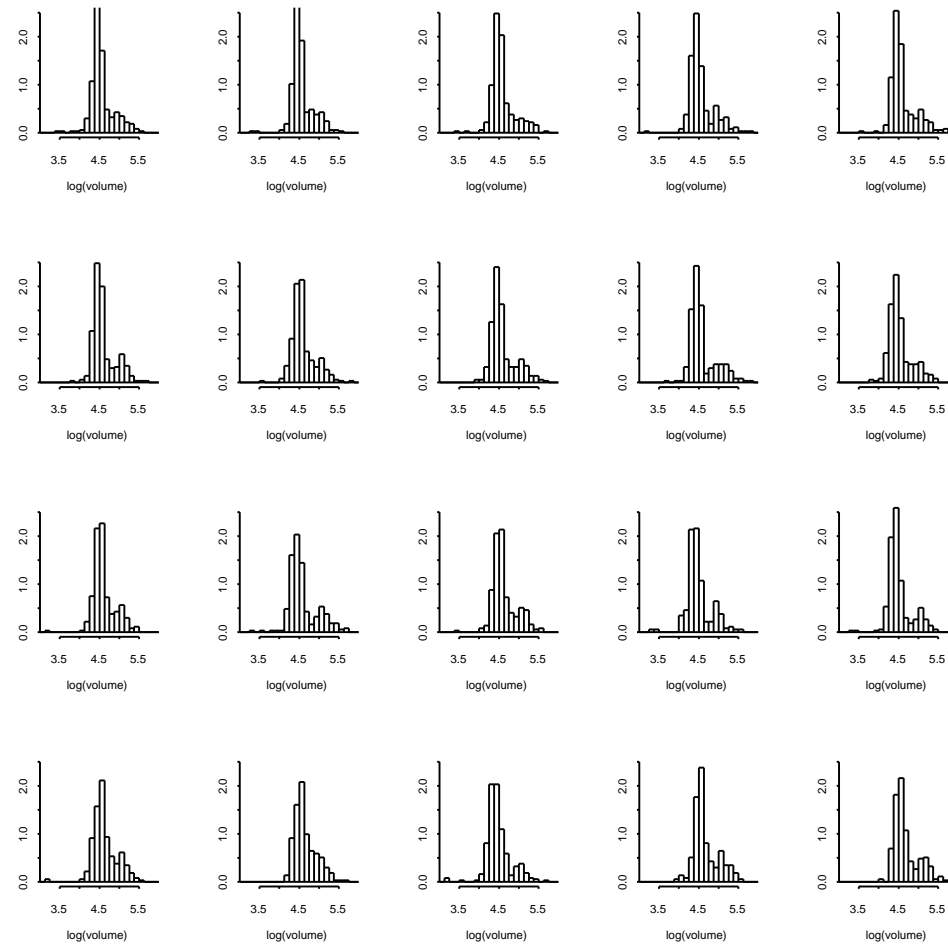


Figure 1: 20 training set normals

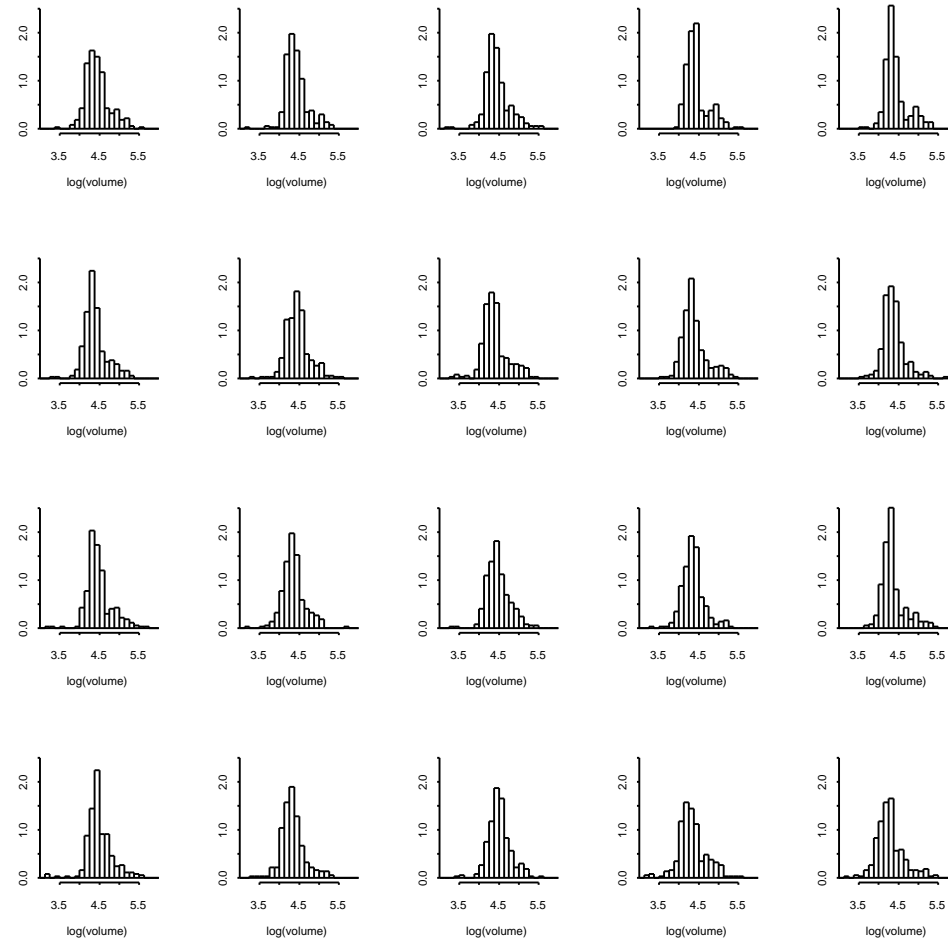


Figure 2: 20 training set anemics

Basic Model

Repeated measures model set up. (or, random effects, hierarchical model, etc.)

Let:

- Y_{ij} be the j^{th} observation for the i^{th} person.
 - f_i be the distribution of the i^{th} person.
 - If f_i can be defined by parameters, let these parameters be θ_i .
-

Let:

- $\mathcal{M}(f)$ be the distribution of the f_i 's.
 - $m(\theta)$ be the distribution of the θ_i 's.
-

Assume that the $m(\cdot)$ will define $\mathcal{M}(\cdot)$ when θ_i defines f_i .

So, now everything is “easy”.

- Use Y_{ij} to learn about f_i (or θ_i).
- Use f_i 's (or θ_i 's) to learn about $\mathcal{M}(\cdot)$ (or $m(\cdot)$).
- Then, use $\mathcal{M}(\cdot)$ to see which subjects have f_i far from the non-diseased population and label these subjects as diseased.

Aside: often study f_i 's by directly studying the parameters θ_i , and, therefore, study $\mathcal{M}(\cdot)$ by directly studying the distribution $m(\cdot)$.

However, in our study, it is not easy to fit the f_i 's with a parametric model, so we will model with nonparametric methods.

Bayesian Density Estimation

- The distribution $\mathcal{M}(f)$ will be a mixture of Dirichlet processes.
- Let f_i be a sample of a normal mixture of Dirichlet processes. So, f_i can be defined by:

$$\begin{aligned} Y_{ij}|f_i &\sim f_i \\ f_i &= \text{Normal} \otimes G_i \\ dF_i &= \int \phi(y|\theta) dG_i(\theta) \\ G_i|G_0, \alpha_0 &\sim \mathcal{D}(G_0, \alpha_0) \end{aligned}$$

where \mathcal{D} is a Dirichlet process.

- The Dirichlet process has parameters G_0 , a distribution, and α_0 , a scalar. A sample from a Dirichlet process is a distribution.

Extending to Multiple Samples

- So far, we have described how to compute a set of Bayesian density estimates f_i with a fixed G_0 .
- We want the f_i 's to be sampled “near” the same template distribution. Therefore, one needs to be “serious” about modelling G_0 .
- Therefore, we put a prior distribution on G_0 , and obtain the distribution of $G_0|\{G_1, G_2, \dots\}$
- The prior for G_0 will be a sample from a mixture of a Dirichlet processes. That is, the prior is a Dirichlet processes $\mathcal{D}(G_{00}, \alpha_{00})$ which is convolved with a normal-gamma.

Full Model

$$Y_{ij}|f_i \sim f_i$$

$$F_i|G_0, \alpha_0 \sim \mathcal{MDP}(G_0, \alpha_0)$$

$$G_0|G_{00}, \alpha_{00} \sim \mathcal{MDP}(G_{00}, \alpha_{00})$$

Expanded model:

$$Y_{ij}|\theta_i \sim N(Y_{ij}|\theta_{ij})$$

$$\theta_{ij}|G_i \sim G_i$$

$$G_i|G_0, \alpha_0 \sim \mathcal{D}(G_0, \alpha_0)$$

$$G_0(\cdot) \sim \int N(\cdot|\tau)dF(\tau)$$

$$F|G_{00}, \alpha_{00} \sim \mathcal{D}(G_{00}, \alpha_{00})$$

Features of the Bayesian approach

- Using a mixture of Dirichlet processes, the model is a full, proper Bayesian model.
- It is conceptually easy to consider several samples from the distribution: $f_1, \dots, f_n \sim \mathcal{MDP}(G_0, \alpha_0)$
- The \mathcal{MDP} defines the likelihood of this model.
- Inference: use the standard Bayesian methods. State the questions of interest and calculate the posterior distributions. Calculations via MCMC methods.
- Conceptual Simplicity. Our model is similar to the simple random effects model. We have the model in terms of a “location” parameter and a precision parameter.

The following figures show an example of how this model “generates” the data.

- The top plot shows the distribution G_{00} which is the prior distribution for the shape of the “template” distribution.
- The second plot show the distribution F which is sampled from a Dirichlet process with parameters G_{00} and α_{00} .
- The next plot is the distribution G_0 . This is the template distribution. Individual subjects which is sampled from a mixture of a Dirichlet process which is centered at this template distribution. Therefore, individual subjects are similar to the template distribution.
- The template distribution G_0 is formed by mixing the distribution F with a normal distribution.

- The next two row of plots show how an individual distribution is generated. First, for the i th individual, a distribution G_i is drawn from a Dirichlet distribution with parameters G_0 and α_0 . Then, in the last row, we see P_i which is G_i mixed with a normal distribution.

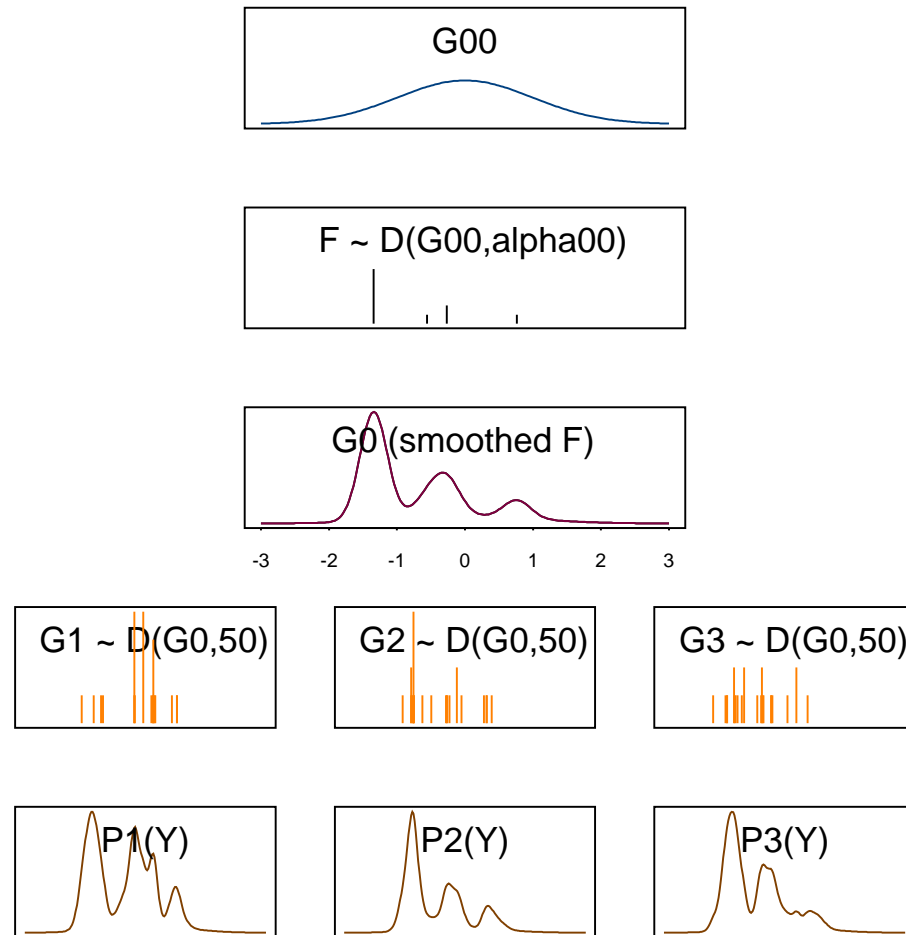


Figure 3: Simulating from model with $\alpha_0 = 50 = n_i$

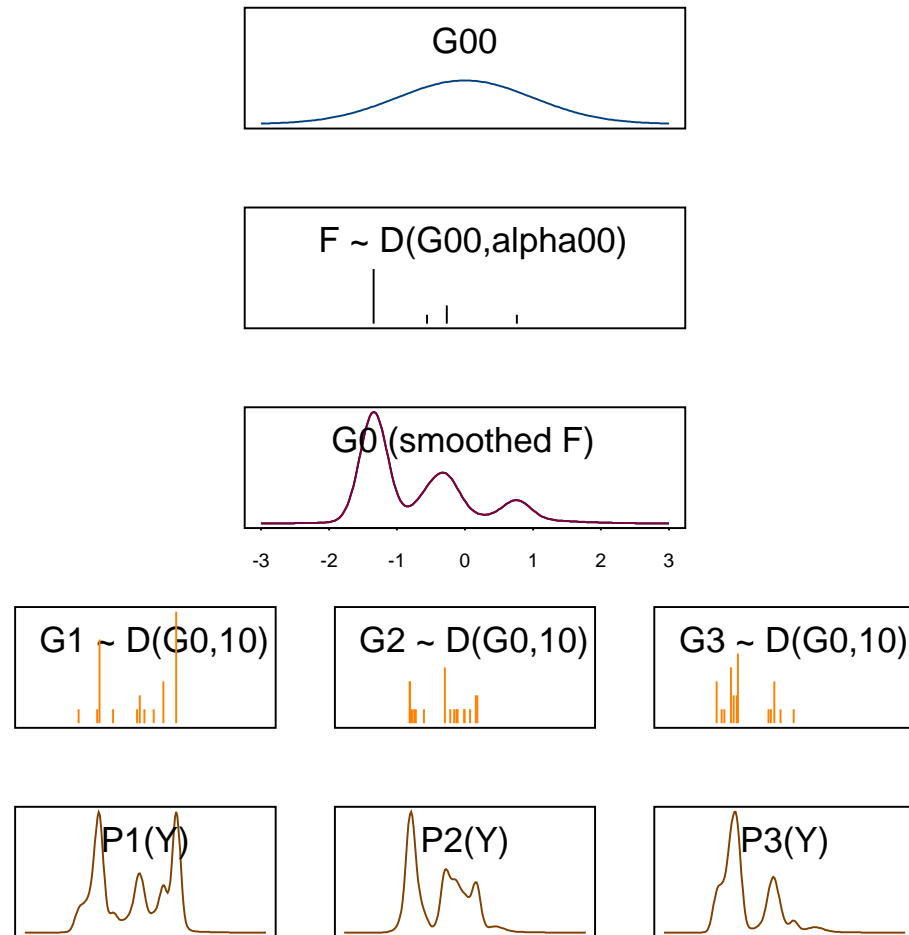


Figure 4: Simulating from model with $\alpha_0 = 10$

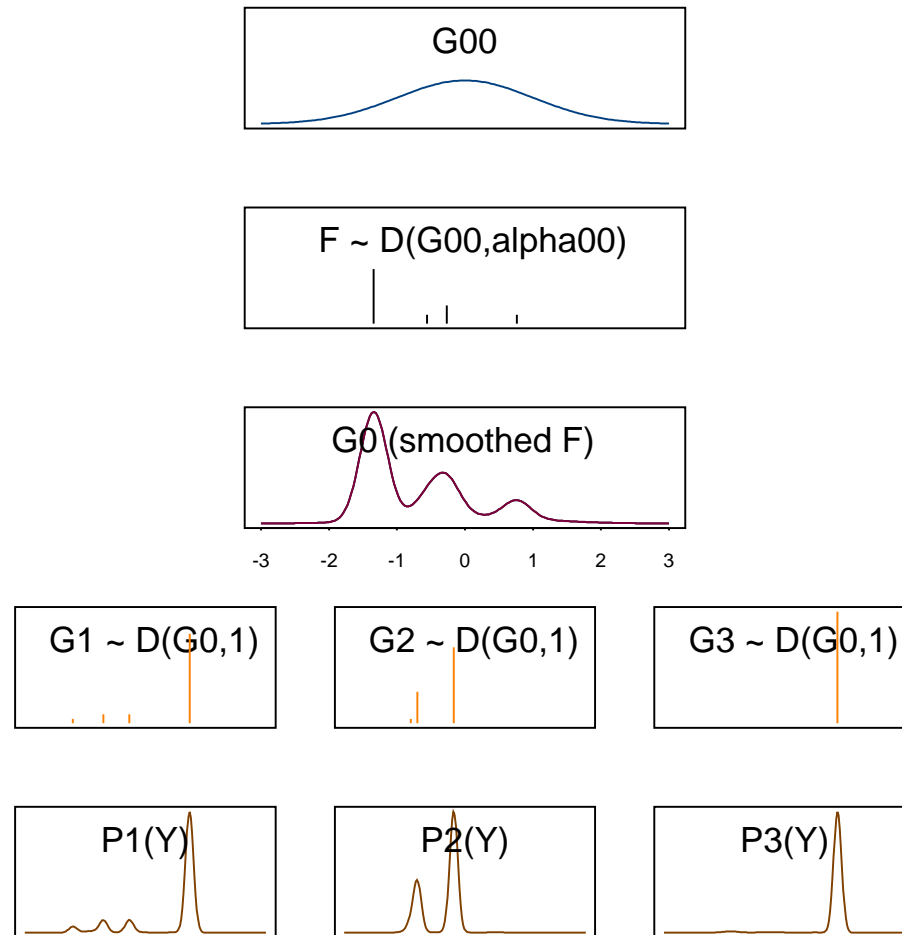


Figure 5: Simulating from model with $\alpha_0 = 1$

Applications: Determining Outlying Subjects

- In some situations, the shape of the density function may be the best tool for determining subjects who are unusual.
- For example, certain haematological problems are diagnosed by examination of a histogram of a subject's red blood cell volumes.

General Strategy

- Fit the model with the data.
- Choose a measure: a) distance between two densities or b) goodness of fit (which is the distance between a density and a sample of data).
- Compute the distance between “mean density” and either a) the individual fitted density or b) sampled data points.
- For inference, generate a) new densities from the model or b) new data points. This will give the posterior distribution of the distance measure.

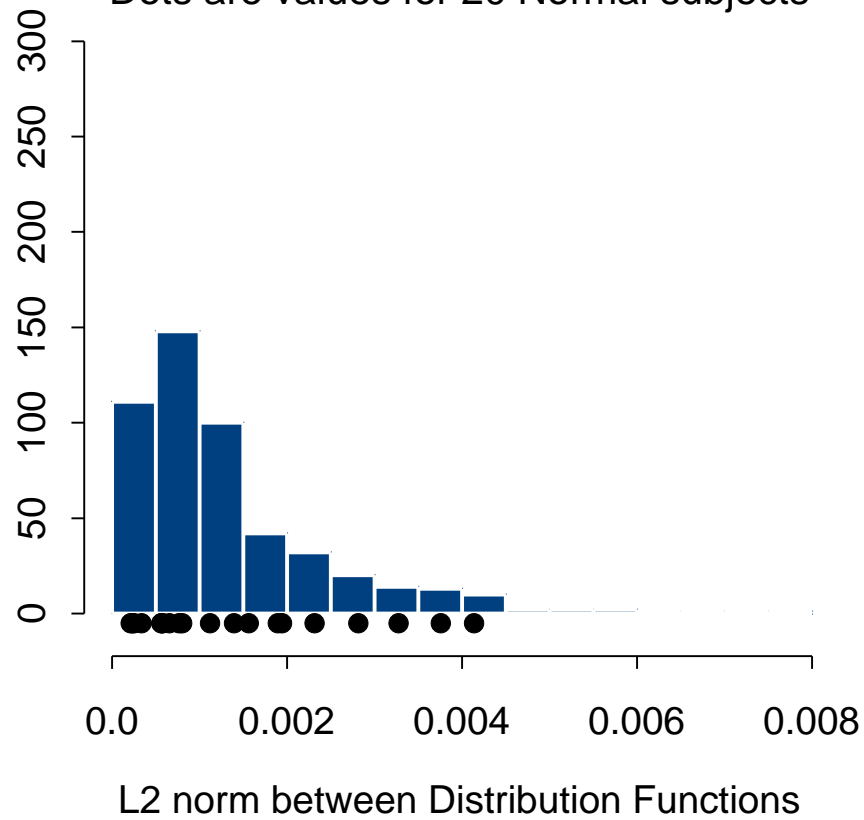
Our Strategy

- For classification, we will calculate the probability of being in one group versus the other group.
- To see how far away the data is from the group center, $f_0(y)$
 - We will use distance measures such as the Kolmogorov-Smirnov statistic for new data or the L_2 norm between CDF's.
 - Calibration is obtained by sampling data from new predicted densities from the model and seeing how far away the empirical distribution of this sample is from f_0 .
- The present analysis is still preliminary.

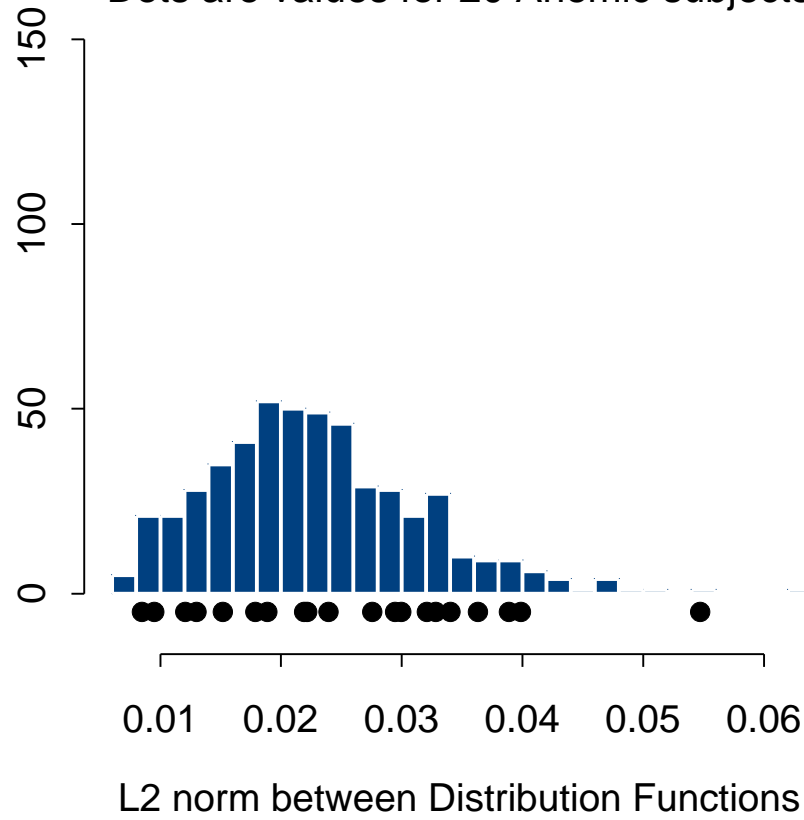
Using L_2 Distance

- Obtain CDF on grid of y-values for (a) mean anemic and (b) mean normal
- Obtain CDF for (c) B simulated new normals and (d) B simulated new anemic densities
- Obtain B posterior samples of L_2 norm between $[a,c]$, $[a,d]$, $[b,c]$, $[b,d]$

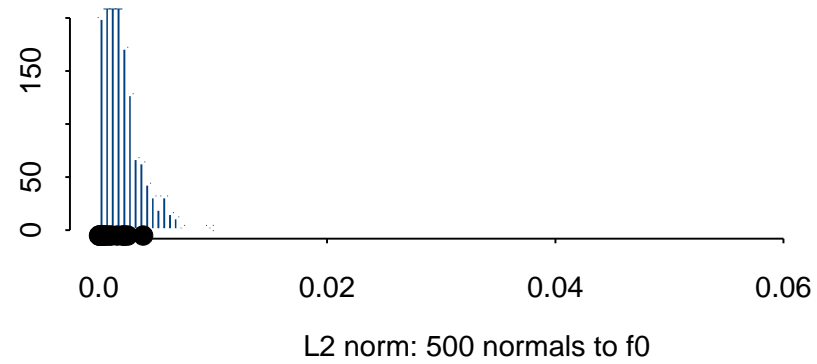
L2 norm: Average Normal to 500 Simulated Normals
Dots are values for 20 Normal subjects



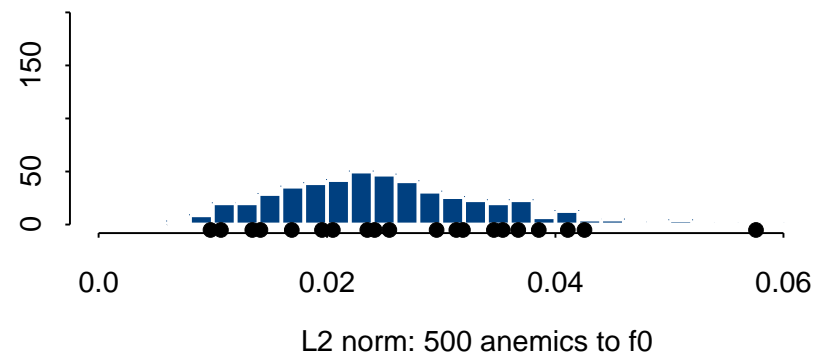
L2 norm: Average Normal to 500 Simulated Anemics
 Dots are values for 20 Anemic subjects



L2 norm: Average Normal to 500 Simulated Normals



L2 norm: Average Normal to 500 Simulated Anemics



Closing Comments

- This type of data is becoming more common. We will need methods of doing inference on these problems.
- The mixture of Dirichlet processes provides a very natural way to do inference on these types of problems.
- Inference via simulation. Since we are using MCMC, we can easily sample the posteriors of interest.