

# Analysis of Densities

(AnDe)

By

George Tomlinson and Michael Escobar<sup>1</sup>

---

<sup>1</sup>We greatly acknowledges the support by the National Science and Engineering Research Council of Canada.

## Preamble

- This an example of an active use of the Dirichlet process in an applied setting.
- History
  - Theory: (Begun in 1963). Definition of the Dirichlet process. Many properties have been defined.
  - Applied, passive use (Late 1980's, 1990's). With the development of MCMC methods, Hierarchical models contained Dirichlet process components.
    - \* Usually, this was to control for ill-behaving parts of the model.
    - \* This was similar in spirit to the Cox Proportional Hazard model used in survival analysis.
    - \* The goal was to use the Dirichlet process to adjust for the ill-behaving parts of model and then to make inferences with the “parametric” part of the model.
  - Applied, active use. (Beginning ?) Using the Dirichlet process to make direct inferences.

# Outline

- Motivation
- Some other ideas
- Our Model
- Inference/Example
- Comments and Pictures

## Motivating Data Set

- The size of a person's blood cells.
- Modern diagnostic equipment can quickly and accurately measure blood cell sizes from a vial of blood.
- Clinical interest is in the shape of the distribution of the sizes.
- **Question:** Does the distribution look like a typical distribution for someone from the normal population, or does it look like someone from a disease population?

## Basic Model

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

Let:

- $Y_{ij}$  be the  $j^{\text{th}}$  observation for the  $i^{\text{th}}$  person.
  - $P_i$  be the distribution of the  $i^{\text{th}}$  person.
  - If  $P_i$  can be defined by parameters, let these parameters be  $\xi_i$ .
- 

Let:

- $\mathcal{M}(P)$  be the distribution of the  $P_i$ 's.
  - $m(\xi)$  be the distribution of the  $\xi_i$ 's.
- 

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

## So, now everything is easy

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

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

## Simplest Approach

If  $P_i = N(\mu_i, \sigma_i^2)$

(so,  $\xi_i = \mu_i$ )

Then  $\mu_i \sim N(\mu_0, \sigma_0^2)$

(So,  $m(\cdot) = N(\mu_0, \sigma_0^2)$  )

---

- This is the basic random effect/repeated measures model.
- Need the distributions to be normal. In the example of the blood cell sizes, the distribution looks multi-modal.

# Mixture of Normals

$$P_i = \sum_{k=1}^K w_k N(\mu_k, \sigma_k^2),$$

where  $K$  is small, often 2 or 3.

$$\text{So, } \xi_i = (w_1, \dots, w_{K-1}, \mu_1, \dots, \mu_K, \sigma_1^2, \dots, \sigma_K^2).$$

The distribution of  $\xi_i$  is a large multivariate joint distribution.

Potential problems/difficulties with identifiability:

- Nuisance case: label switching.
- Bad case: Using many normals to fit one component. Example: Age of onset data.
- How many components? It is hard to compare a 2 component normal model with a 3 component normal model.

# Kernel Density Method

If  $P_i$  is unusual enough, then one could fit each  $P_i$  with a kernel density estimator.

$$\text{So, } \hat{P}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} N(Y_{ij}, \sigma^2)$$

Don't really have the parameters  $\xi_i$ 's in this model

What is the distribution,  $\mathcal{M}(P)$ ???? (bootstrapping???)

---

Also, one could consider splines and wavelets, etc., but there will still be a problem with the distribution  $\mathcal{M}(P)$ .

Please note that there seems to be a rift between research on exploratory, nonparametric methods (eg: kernel density estimation, etc.) and formal, detail model building (eg. hierarchical mixed models, etc.)

## Where do we go from here?

- Some might say that a random effect/repeated measurement model is simply a Bayesian Hierarchical model where people sometimes forget the highest level prior.
- So, how about if we use a “Bayesian” kernel density estimator and then our problem is solved.

# “Bayesian” Density Estimation

Ferguson (1983) and Escobar and West (1995) show how to get a “Bayesian density estimate” by modelling the distribution of the data with a mixture of Dirichlet processes. This approach is used in this talk.

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

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

where  $\mathcal{D}$  is a Dirichlet process. The Dirichlet process has parameters  $G_0$ , a distribution, and  $\alpha_0$ , a scalar. A sample from a Dirichlet process is a distribution.

# Bayesian Density Estimation (continued)

The distribution  $G_i$ , which is a sample from the Dirichlet process, has the the following form (see Setharamin, 1994):

$$G_i|G_0, \alpha_0 = \sum_l^{\infty} \delta_{\theta_l^*} p_l$$

where

- $\delta_{\theta_l^*}$  is a the point mass distribution at  $\theta_l^*$ ,
  - the values  $\theta_l^*$  are sampled independently from the distribution  $G_0$ , and
  - the probability weights  $p_l$  are sampled from a distribution which depends on  $\alpha_0$ .
- 

Therefore

$$\begin{aligned} y_{ij}|P_i &\sim P_i \\ dP_i|G_0, \alpha_0 &= \int \phi(y|\theta) dG_i(\theta) \\ &= \sum_{l=1}^{\infty} p_l \phi(y|\theta_l^*) \end{aligned}$$

# Bayesian Density Estimation (continued)

Note the following:

- When all but one  $p_l$  is approximately zero, then  $P \approx N(y|\theta)$ . That is, **one normal component**. This happens when  $\alpha_0$  is very tiny (.1,  $1/n$ , etc).
- When all but one  $p_l$  is approximately zero, then  $P \approx \sum_{l=1}^L p_l N(y|\theta_l^*)$ , with  $L$  small. This is **the finite mixture model**. This happens when  $\alpha_0$  is small (1, 10, or so).
- When there are many values of  $p_l$  which are approximately equal, then  $P \approx \sum_{l=1}^L p_l N(y|\theta_l)$ , with  $L$  very large. After data  $(Y_1, \dots, Y_n)$  are observed, then the random distribution  $P|Y$  will look similar to the **kernel density estimate**. This happens when  $\alpha_0$  is large ( $n$ ,  $n^2$ , etc.).
- When there are many values of  $p_l$  which are approximately equal and since the  $\theta_l^*$  follow the distribution  $G_0$ , then  $G_i$  will be close to  $G_0$ . Therefore,  $G_0$  acts like the location parameter and  $\alpha_0$  acts like a precision parameter.

# Examples of Weights for Dirichlet Process

For  $\alpha = .1$ , some sampled weights:

- $p = (.991, .009, .00009, \dots)$
- $p = (.984, .015, .001, .00001, \dots)$
- $p = (.839, .151, .010, .0000005, \dots)$
- $p = (1.000, 10^{-13}, \dots)$
- $p = (.142, .775, .083, .0001, \dots)$

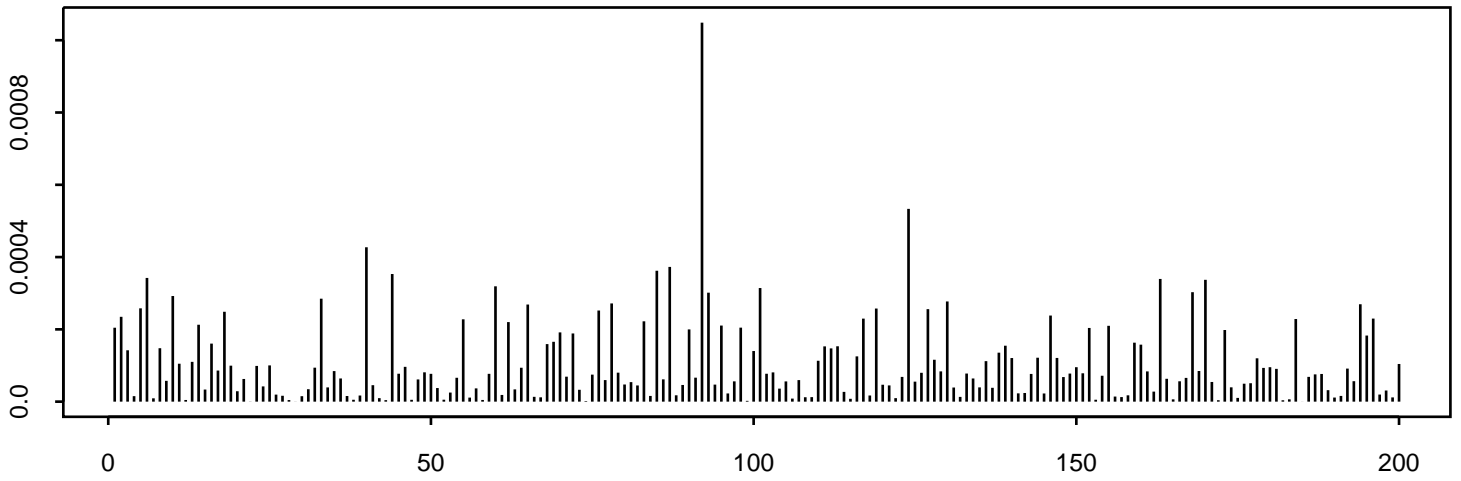
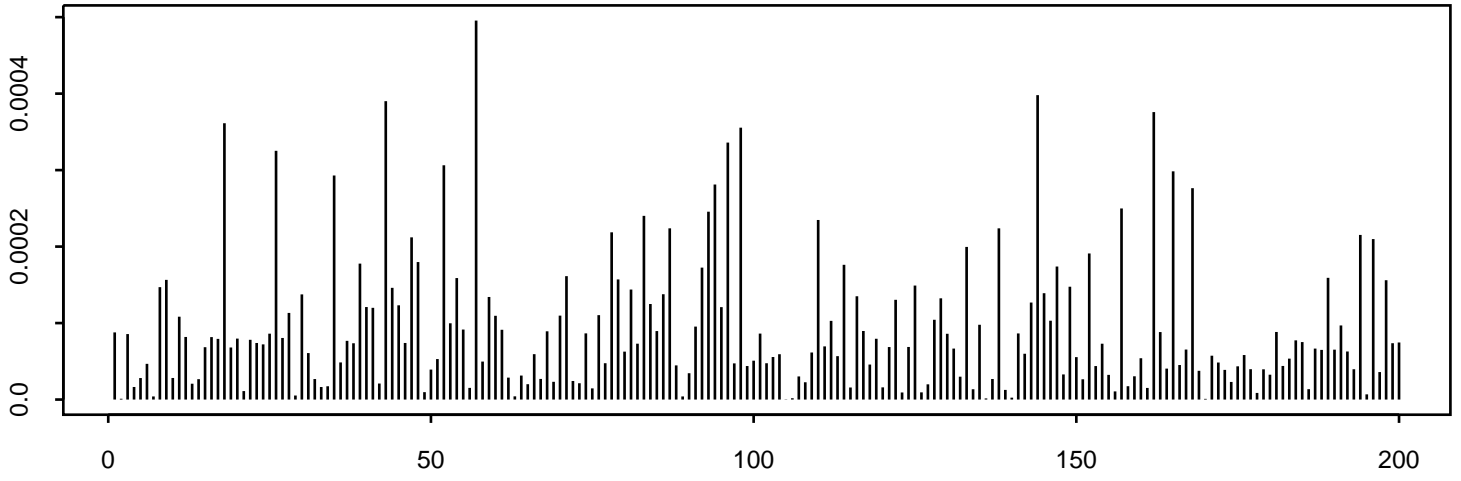
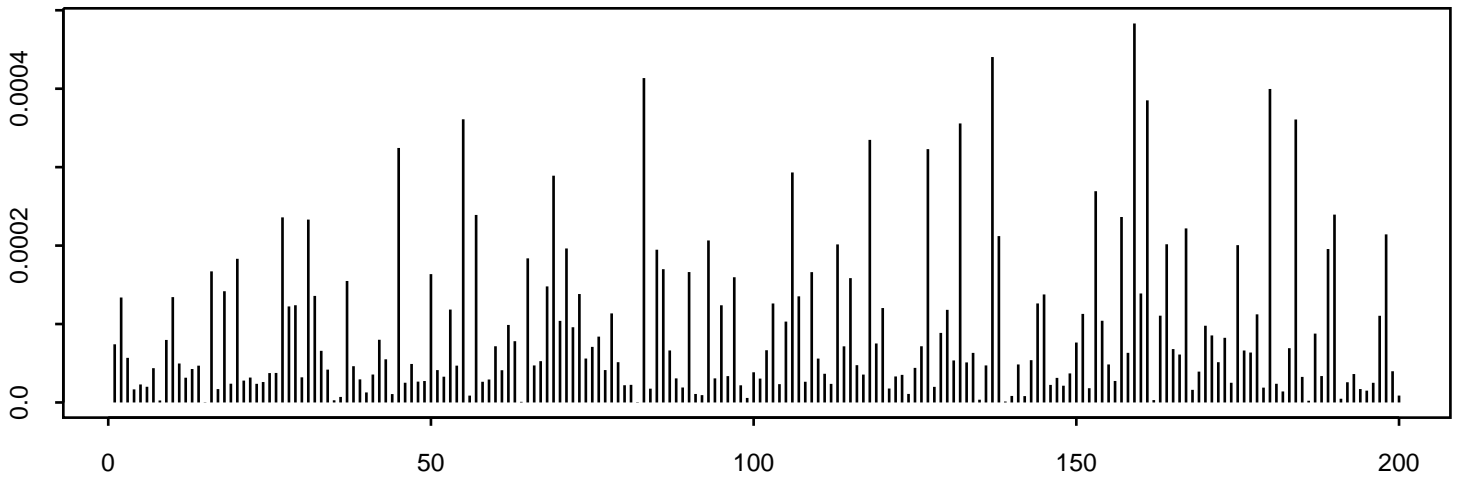
For  $\alpha = 1$ , some sampled weights:

- $p = (.077, .6144, .134, .047, .047, .051, \dots)$
- $p = (.886, .016, .093, .001, .003, .0004, \dots)$
- $p = (.868, .081, .008, .023, .004, .008, \dots)$
- $p = (.497, .429, .027, .023, .013, .001, \dots)$
- $p = (.753, .042, .119, .011, .022, .033, \dots)$

## Examples of Weights for Dirichlet Process (II)

For  $\alpha = 5$ , some sampled weights:

- $p = (.189, .152, .097, .252, .107, .055, \dots)$
- $p = (.213, .087, .034, .030, .122, .133, \dots)$
- $p = (.112, .103, .548, .023, .094, .028, \dots)$
- $p = (.010, .165, .002, .032, .061, .368, \dots)$



# Helpful Definition of Dirichlet Processes

Stick Breaking (Setharamin, 1994)

- This is a constructive definition, which shows how to sample a random distribution  $G$  from  $\mathcal{D}(G_0, \alpha_0)$ .
- “Sample” an infinite number of pairs,  $(p_l, \theta_l^*)$ , where  $\theta_l^*$  is sampled independently from  $G_0$  and  $p_l$  sampled by the following algorithm:

$$\begin{aligned} B_l &\sim \text{Beta}(1, \alpha_0) \\ p_1 &= B_1 \\ p_l &= B_l \prod_{j=1}^{l-1} (1 - B_j), \quad \text{for all } l > 1 \end{aligned}$$

- So, one can now construct the random distribution  $G$  with the following equation:

$$G = \sum_{l=1}^{\infty} p_l \delta_{\theta_l^*}$$

## Priors for the Dirichlet process parameters

- $G_0$ : a distribution and the location parameter. Use another mixture of Dirichlet processes as the prior for this parameter.
- $\alpha_0$ : a scalar and the precision parameter. We will want this to be not too small because we want  $G_i$  to be somewhat near  $G_0$ . Therefore, the prior for  $\alpha_0$  should put most of its mass on large values.

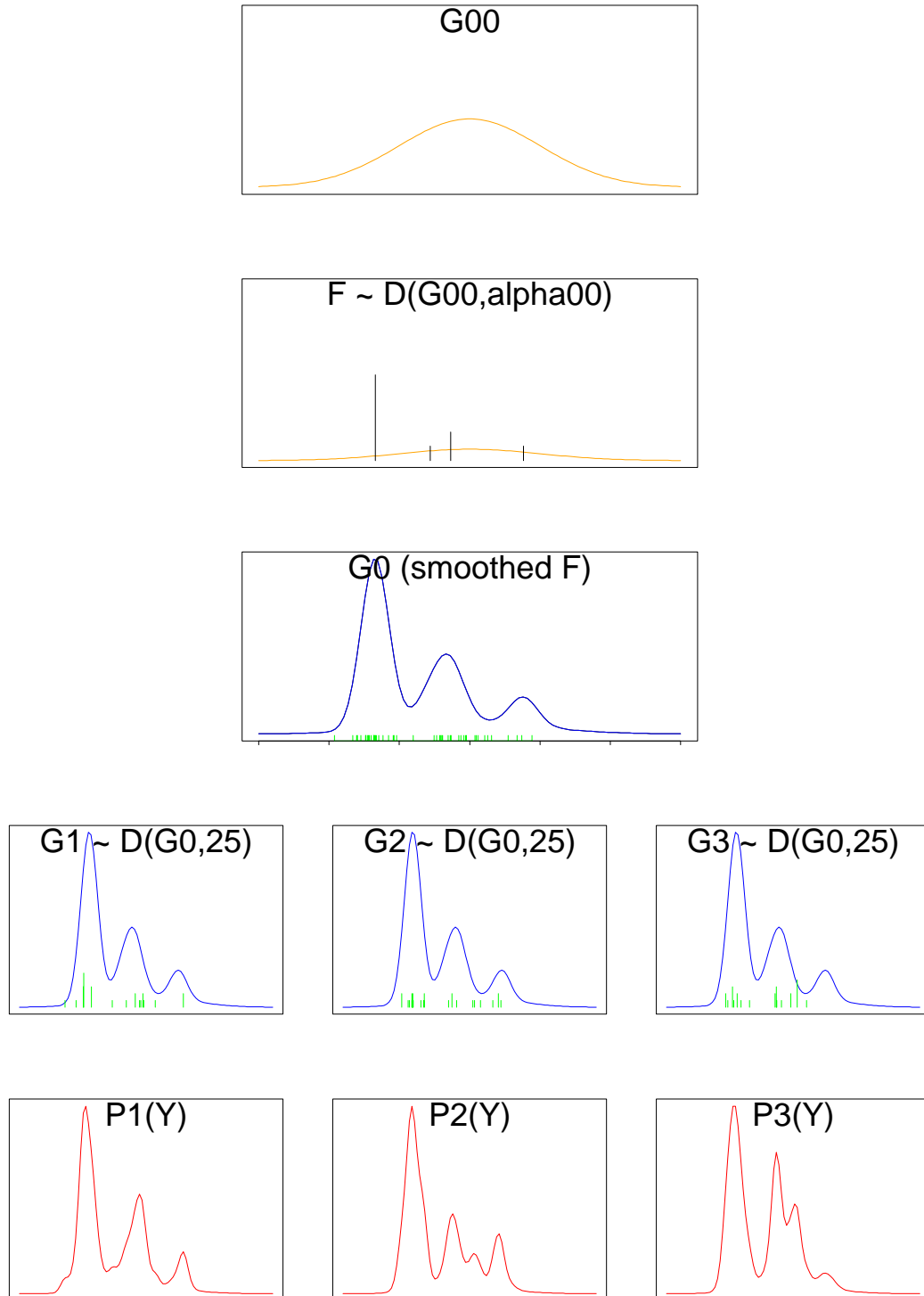
# Full Model

$$\begin{aligned} Y_{ij}|P_i &\sim P_i \\ P_i|G_0, \alpha_0 &\sim \mathcal{MDP}(G_0, \alpha_0) \\ G_0|G_{00}, \alpha_{00} &\sim \mathcal{MDP}(G_{00}, \alpha_{00}) \end{aligned}$$

Expanded model:

$$\begin{aligned} Y_{ij}|\theta_i &\sim N(Y_{ij}|\theta_{ij}) \\ \theta_{ij}|G_i &\sim G_i \\ G_i|G_0, \alpha_0 &\sim \mathcal{D}(G_0, \alpha_0) \\ G_0(\cdot) &\sim \int N(\cdot|\tau)dF(\tau) \\ F|G_{00}, \alpha_{00} &\sim \mathcal{D}(G_{00}, \alpha_{00}) \end{aligned}$$

Figure 1: Estimation of the individual densities and the “parent” distribution  $G_0$  proceeds by iterating the following two steps (1) Using the current estimate of  $G_0$  in  $G_i \sim \mathcal{D}(G_0, \alpha_0)$ , samples the posterior distributions of  $\Pi_i | \mathbf{Y}_i, \mathbf{T}$  for  $i = 1, \dots, n$ . (2) Using the (fixed) distribution  $G_{00}$  in  $F \sim \mathcal{D}(G_{00}, \alpha_{00})$ , sample the posterior distribution of  $\mathbf{T} | \Pi^*$ .



## Note on the Precision Parameters

- Use priors for the precision parameters  $\alpha$ , do not make them fixed. One would not use fixed variances in a repeated measures model. (For one approach for priors on  $\alpha$  see Escobar and West, 1995).
- The parameter  $\alpha_0$  is usually large. Typically about  $n$ . The underlying reason to consider the model presented in this talk is because it is thought that the  $P_i$ 's are considered to be “near” some common, underlying distribution. If the prior for  $\alpha_0$  has little weight on large value (such as 10, 100, etc.) then the  $P_i$ 's will not be near each other.
- The parameter  $\alpha_{00}$  is usually small. One wants the  $G_0$  to be flexible enough to capture the shape of the common, underlying density. Therefore,  $G_0$  will be allowed to deviate far from  $G_{00}$ .

# Difference between the Kernel Density and the Bayesian approach

- With kernel density estimation, how does one quantify the variation in the estimate around the “true” value? We know it is consistent and the rate of convergence but what about the variation? The distribution of the variability?
- 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:

$$P_1, \dots, P_n \sim \mathcal{MDP}(G_0, \alpha_0)$$

- 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.

# 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 white blood cell volumes.
- We use the reaction time data for schizophrenics versus non-schizophrenics by Belin and Rubin (1990) to illustrate this application in a small dataset.

## Applications: Belin and Rubin Data

Looking at reaction time differences between schizophrenic versus non-schizophrenic subjects. Use the reaction time data introduced by Belin and Rubin (1990) to illustrate this application in a small dataset.

- There were 6 schizophrenic and 11 non-schizophrenic subjects.
- Belin and Rubin model the data with a 3-component normal model. this model is based on scientific theory.
- Our approach will be more empirical and we will ignore the scientific theory proposing a 3 component normal fit.
- Goal in our analysis: find “outliers” (subjects who have schizophrenia).

# 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 for between “common 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

- What we are interested in is looking for “outliers”. Which subjects are not “usual”.
- Our approach will be to use the Hellinger distance between the common distribution and the subjects fitted distribution. (Note there are many distance measures from which to choose.)
- Two approaches:
  - Fit all the data and look for far points,
  - Fit the normals and then see if the the schizophrenic subjects are very different.

Figure 2: **An example:** We use the data on reaction times for 11 non-schizophrenics and 6 schizophrenics introduced by Belin and Rubin (1990). The model was fit separately for the non-schizophrenic and schizophrenic subjects. Overlaid on the histograms of the log-times are the mean predictive densities for each subject.

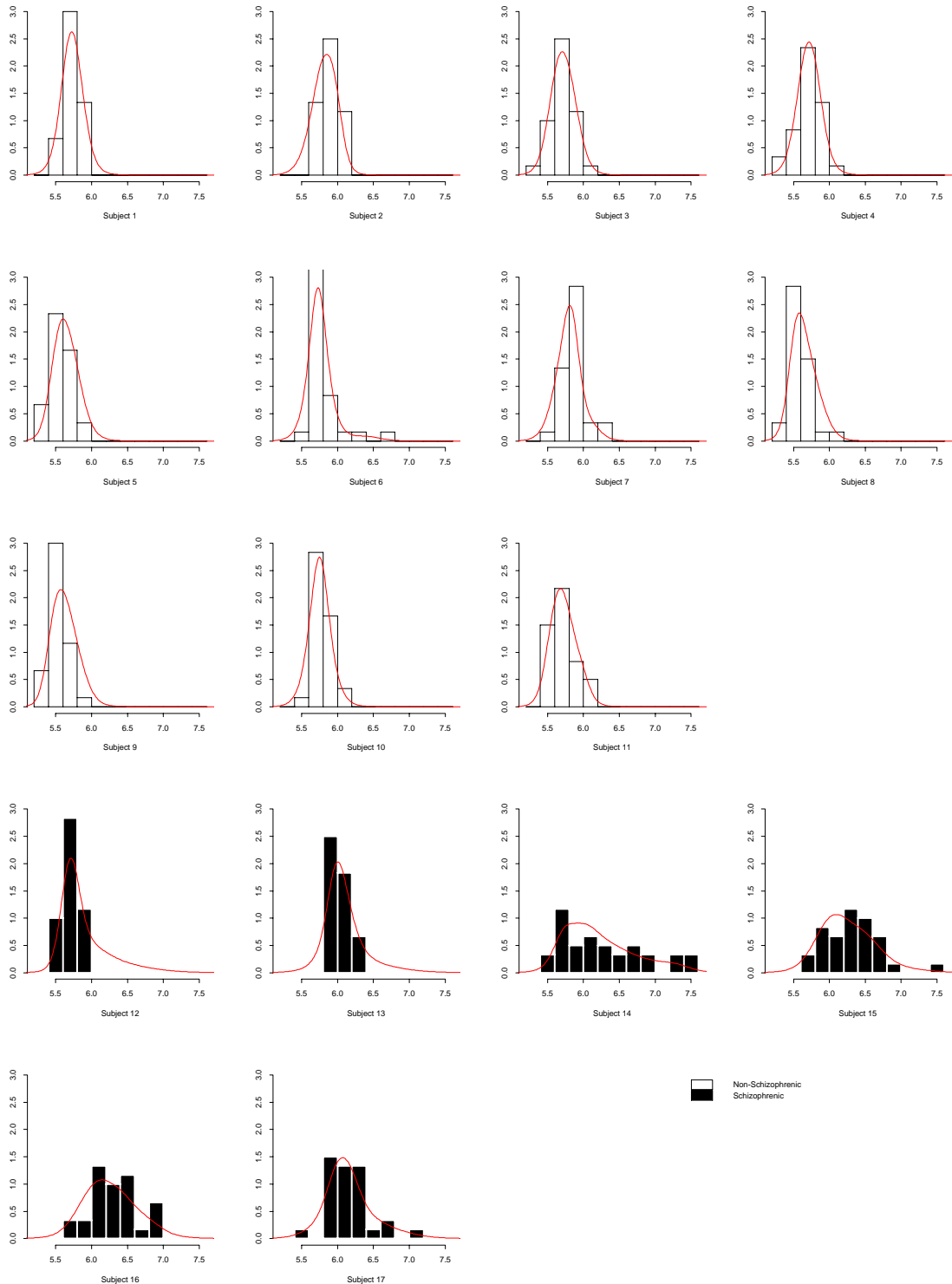
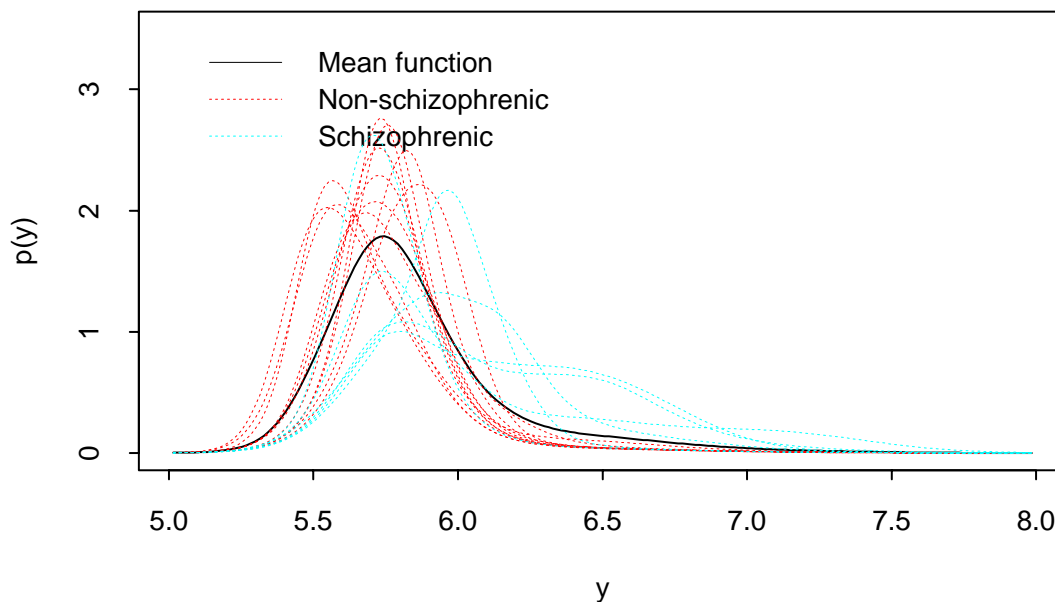


Figure 11.3: PDFs for Reaction Time Data. These are the estimates based on the priors in section 11.2. Figure (a) shows the mean predictive density for each subject and the overall mean across subjects. Figure (b) shows the four predictive densities corresponding to the largest Hellinger distances.

(a) Mean Predictive Distribution and Samples



(b) Selected Sample PDFs

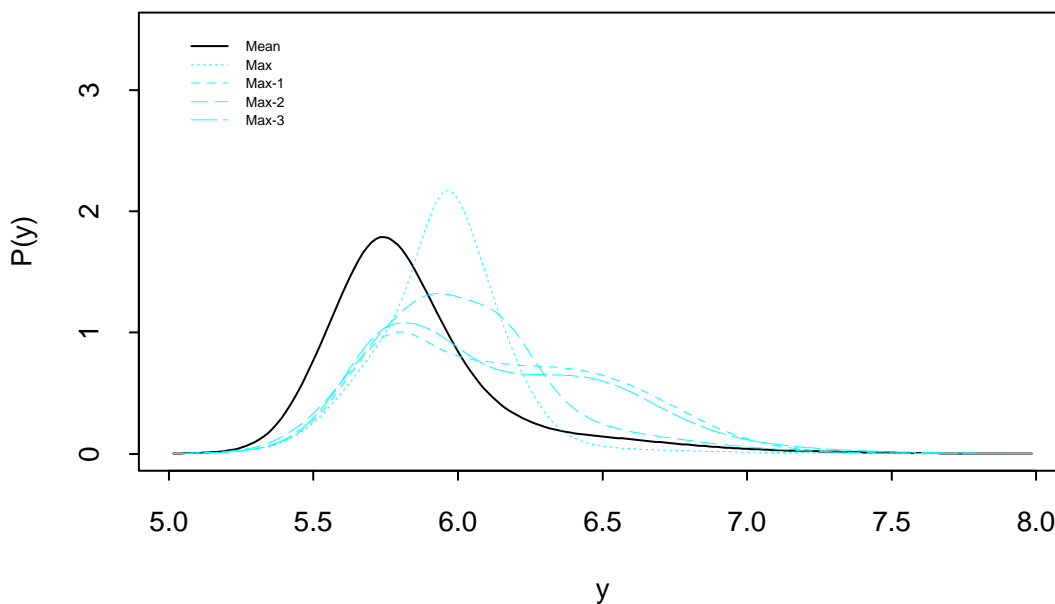
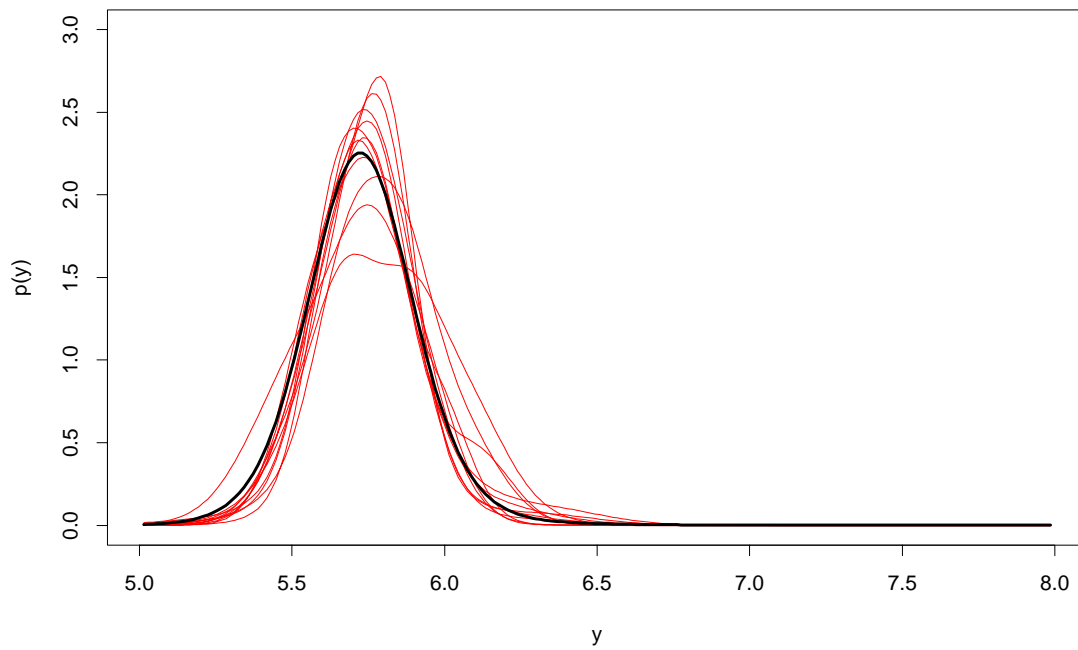


Figure 3: Mean PDFs for Reaction Time Data: these figures show the mean predictive densities for each subject (red) and the overall mean across subjects (black). Figure (a) shows the non-schizophrenic subjects. Figure (b) shows the schizophrenic subjects.

(a) PDFs for non-schizophrenics



(b) PDFs for schizophrenics

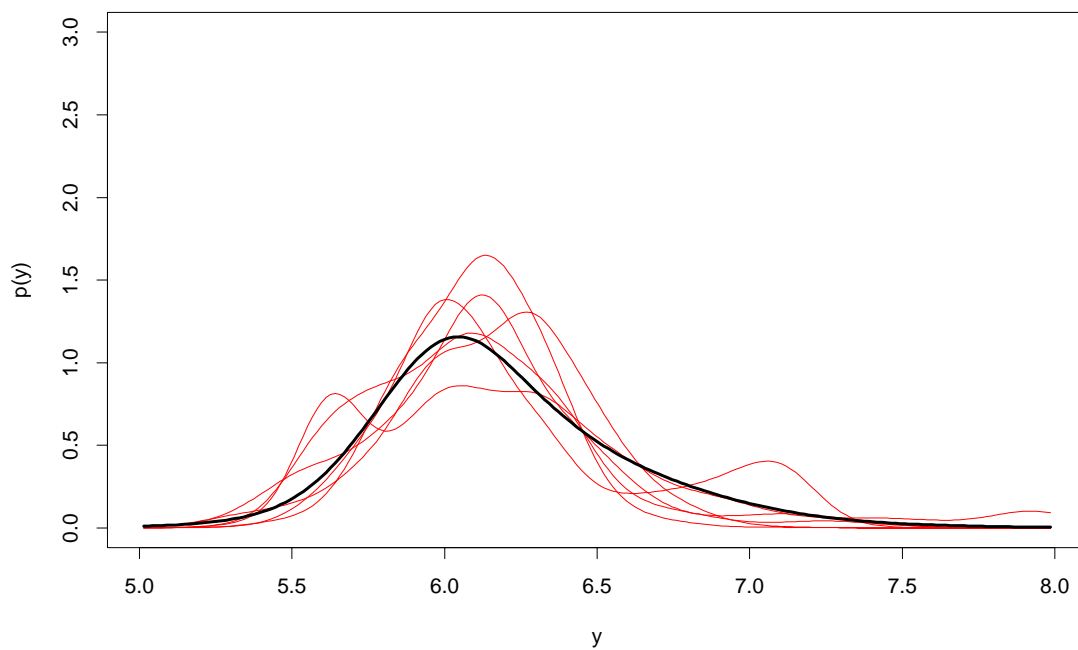


Figure 11.4: Empirical Cumulative Distribution Function for the Hellinger Distances  $D(f_i^B)$

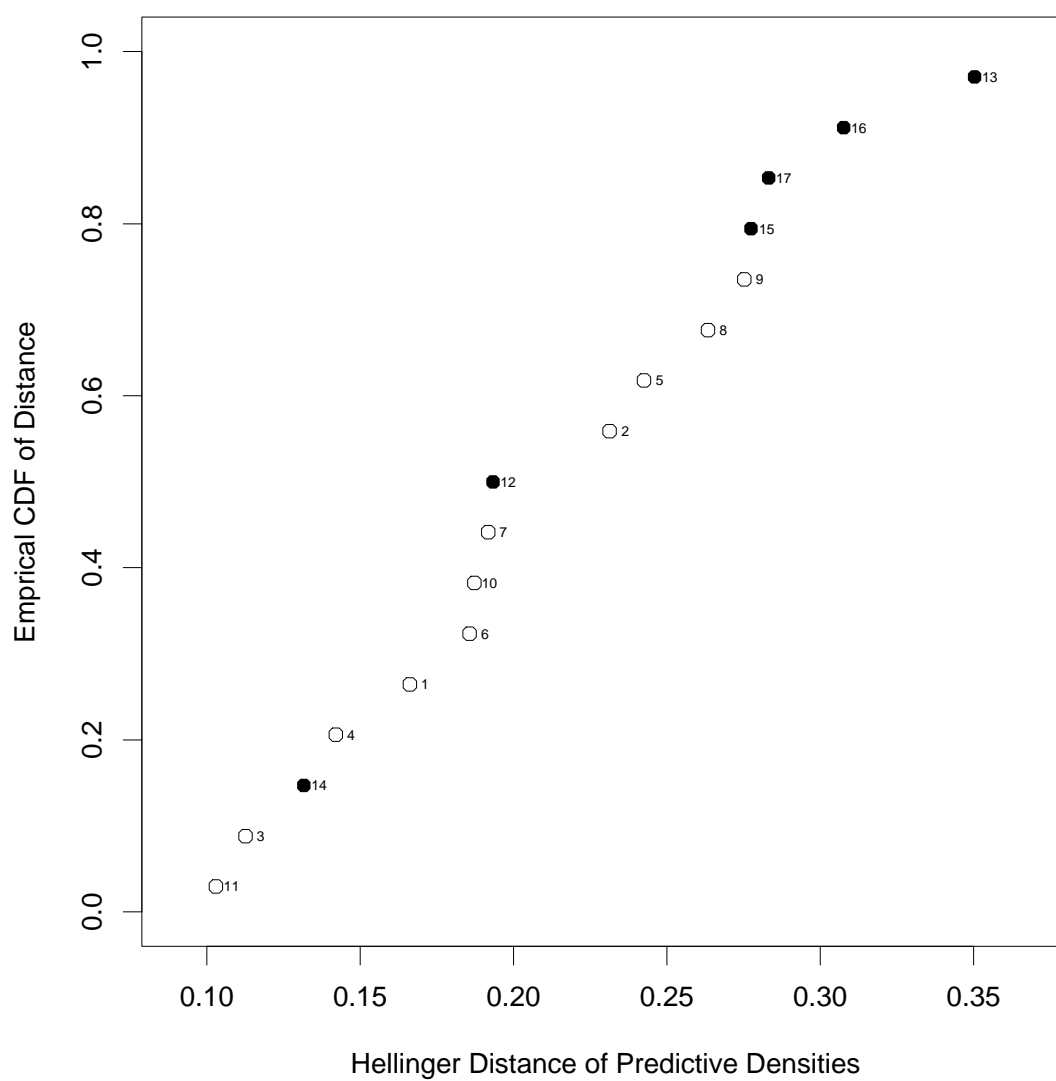


Figure 11.5: Boxplots of Hellinger Distances by Group for Reaction Time Data

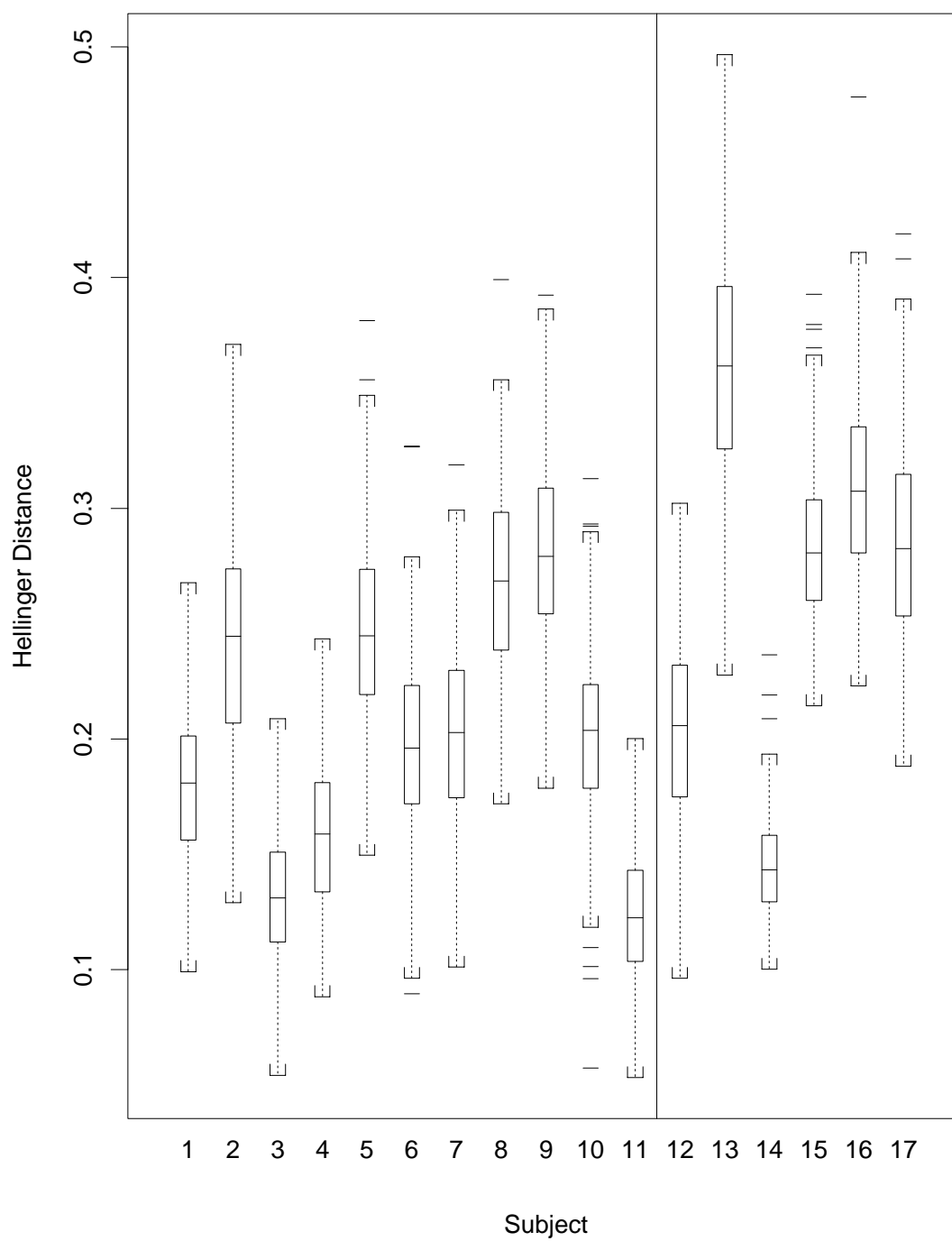
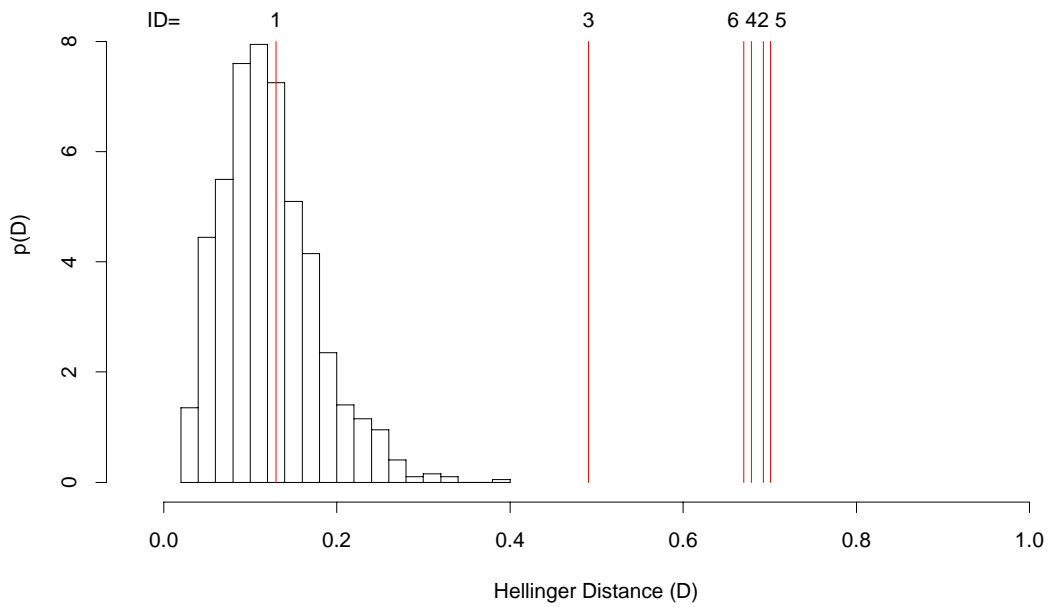


Figure 4: Histograms of Hellinger distances computed on densities simulated from the posterior. Figure (a) shows distances between 1000 simulated non-schizophrenic subjects and the mean predictive density for the non-schizophrenics. The vertical segments are the mean Hellinger distances for the schizophrenics. Figure (b) shows distances between 1000 simulated schizophrenic subjects and the mean predictive density for the non-schizophrenics.

(a) Distances for simulated non-schizophrenic subjects



(b) Distances for simulated schizophrenic subjects

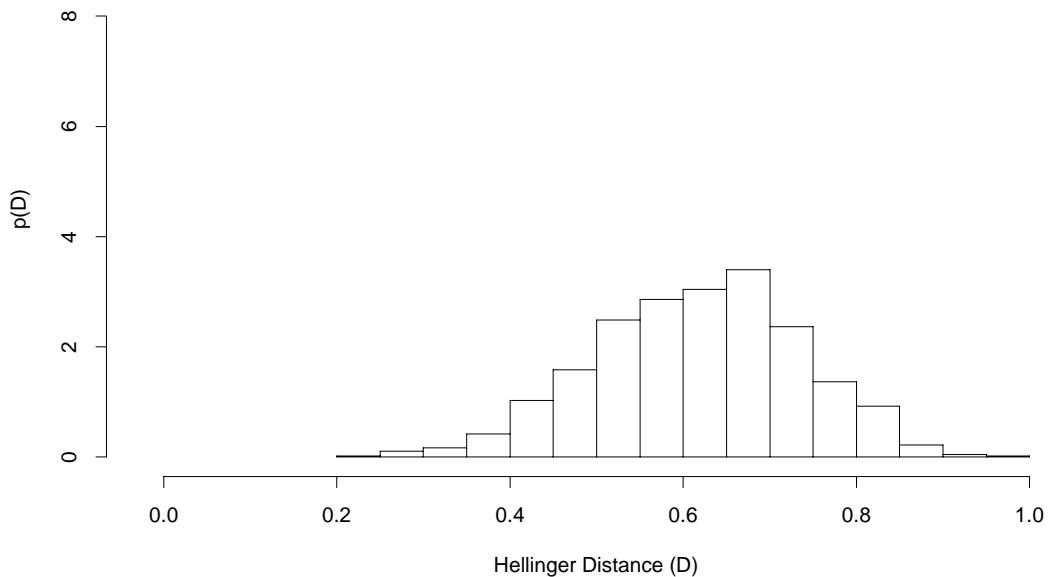
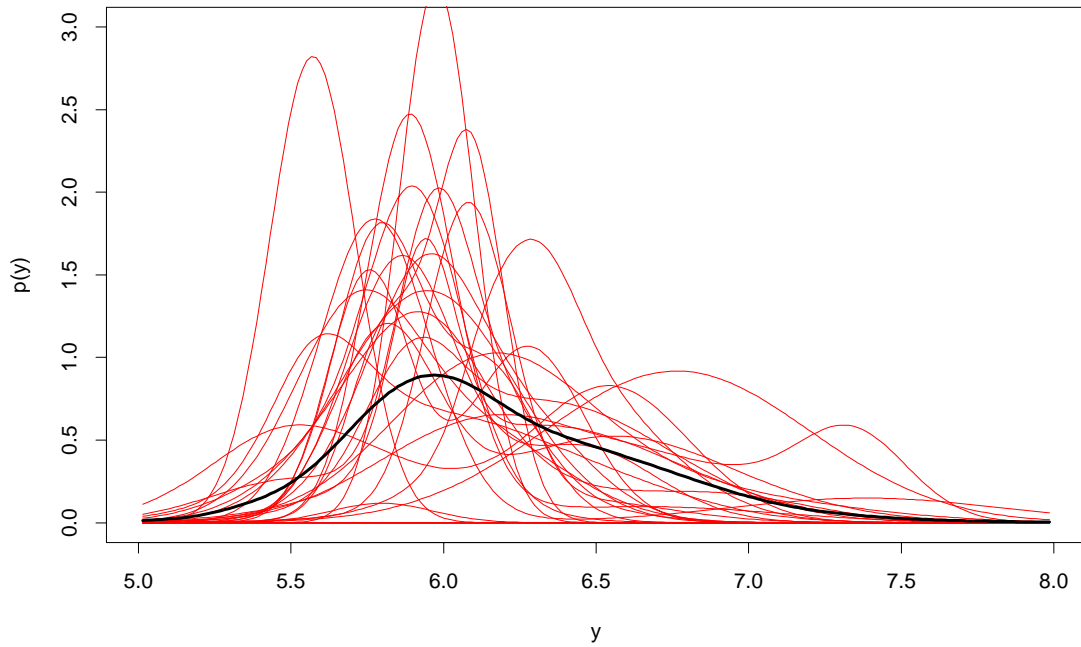
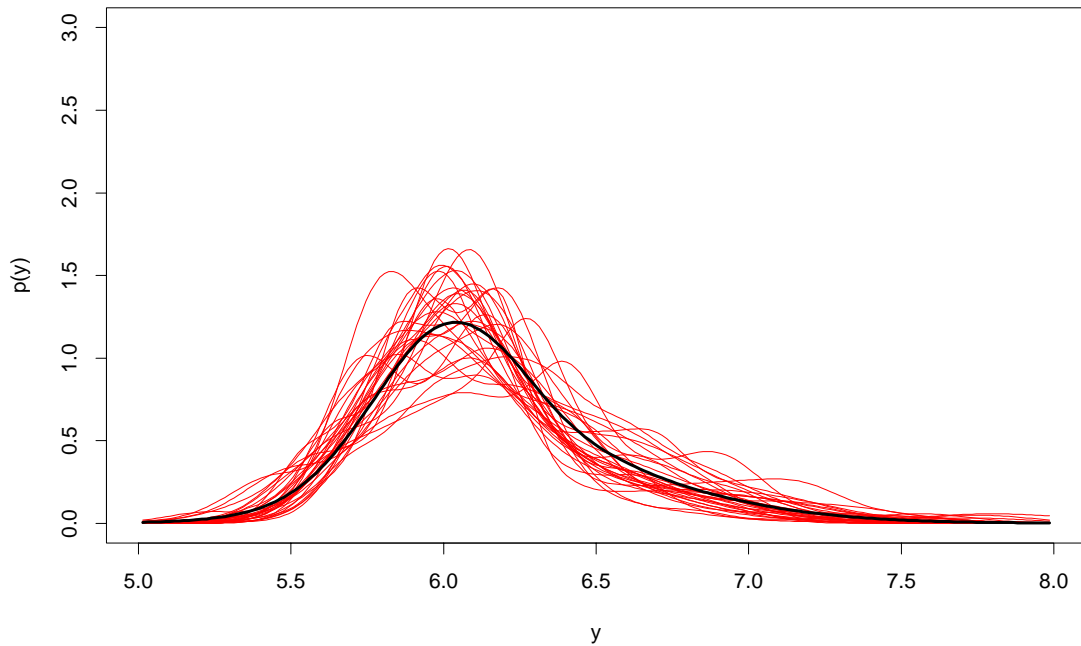


Figure 5: Effect of size of  $\alpha_0$  in  $G_i \sim \mathcal{D}(G_0, \alpha_0)$  on the posterior distributions of  $p(Y)$ . These figures show PDFs generated from the posterior distribution for the schizophrenics. Figure (a) uses  $\alpha_0 = 3$  and Figure (b) uses  $\alpha_0 = 90$ .

(a) PDFs sampled from posterior with small alpha0



(b) PDFs sampled from posterior with large alpha0



## Comments about Model

- This is a fully Bayesian model. For inference, one needs simply to “turn the Bayesian crank”.
- The distribution  $\mathcal{MDP}(G_0, \alpha_0)$  gives the “likelihood” for this model.
- For frequentist who hate the idea of Bayesian models, note that respectable “frequentist repeated measurement” model can be obtained by simple removing the hyperprior,  $\mathcal{MDP}(G_{00}, \alpha_{00})$  and using a frequentist method of estimating  $G_0$ .
- Typically, we will want  $\alpha_0$  to be big and  $\alpha_{00}$  to be small.
- Calculating the MCMC steps for the first MDP (the distribution on  $P_i$ ) is fairly easy, but the calculation for the second MDP (the distribution on  $G_0$ ) is fairly hard.

# Closing Comments

- This type of data is very plausible. We see more of it with the popularity of estimation such as kernel density estimates and we will need methods of doing inference on these problems.
- Mixture of Dirichlet processes is a distribution for structures which are very KDE-like. Therefore, 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.
- The mixture of Dirichlet processes provides the likelihood function for this model.