

Mixture Models

- Data too complex for single parametric model
 - Multimodality
 - Subgroups
 - Genetic sources of variation
 - Heterogeneous populations zero-inflated Poisson
- Why mixtures?
 - Flexibility To Depict Many Shapes.
 - Example: Mixtures of normal kernels (Density Functions) for continuous distributions

POPULATION MODEL AS HIERARCHICAL MODEL

$$p(y_{ij}|\theta_i), j = 1, \dots, n_i$$

$$p(\theta_i|\phi), i = 1, \dots, I$$

$$p(\phi)$$

- Characterize $p(\theta_i|\phi), i = 1, \dots, I$ as a mixture of simpler distributions
 - Bayesian nonparametrics
 - Finite mixtures or
 - Infinitely many parameters

Finite Mixtures

- Instead of $Y_i \sim f(Y_i|\theta_i)$ consider

$$Y_i \sim \sum_{j=1}^c w_j f_j(Y_i|\theta_{ij}), \sum_{j=1}^c w_j = 1$$

- Here, $f_j(Y_i|\theta_{ij})$ is some parametric pdf
–perhaps normal, with $\theta_{ij} = (\mu_{ij}, V_j)$

Group Membership

- Consider latent indicator $Z_i = j$ if i^{th} unit belongs to group j

- So, one can write out $Y_i \sim \sum_{j=1}^c w_j f_j(Y_i|\theta_{ij})$ as

$$Y_i \sim \begin{cases} f_1(Y_i|\theta_{i1}) & \text{if } Z_i = 1 \\ f_2(Y_i|\theta_{i2}) & \text{if } Z_i = 2 \\ \vdots & \vdots \\ f_c(Y_i|\theta_{ic}) & \text{if } Z_i = C \end{cases} \quad \begin{array}{l} \text{with } \Pr(Z_i = j) = w_j \\ \& \sum_{j=1}^c w_j = 1 \end{array}$$

Full Likelihood

- With latent group indicators $\prod_{j=1}^C \prod_{i=1}^n w_j^{s_{ij}} [f_j(y_i | \alpha_j)]^{s_{ij}}$

$$s_{ij} = \begin{cases} 1 & \text{if } z_i = j \\ 0 & \text{otherwise} \end{cases}$$

- Generalize as hierarchical model
$$Y_i | \underline{\theta}, \underline{s}_i, z_i \sim f_{z_i}(y_i | \theta_{z_i})$$
$$s_{ij} \sim g(s_{ij} | \theta_j)$$
$$\theta_j \sim p(\theta_j)$$

Estimation

- The Z_i are unknown
 - Estimation includes mixing prop'ns
 - Prior for mixing weights

$$(w_1, \dots, w_C) \sim \text{Dir}(\alpha_1, \dots, \alpha_C)$$

$$p_{ij} = \Pr(Z_i = j | \text{data}) \propto w_j f(y_i | \theta_j)$$

- With flat prior ($\alpha_1 = \dots = \alpha_C = 1$), have

$$p_{ij} = \Pr(Z_i = j | \text{data}) \propto f(y_i | \theta_j)$$

Issues

- **Identifiability**
 - Representation in terms of (w, θ_1, θ_2) not unique
 - $p(0) = w(1 - \theta_1)^2 + (1 - w)(1 - \theta_2)^2$
 - $p(1) = 2w\theta_1(1 - \theta_1) + 2(1 - w)\theta_2(1 - \theta_2)$
 - Flat likelihoods
 - Appropriate "objective" priors
- **So-called label switching in MCMC**
 - If unconstrained prior
 - $C!$ ways to permute group labels
 - Possible solutions
 - Order mixing weights, means, vars
- **Empty classes**

Example: Batting Avgs for 18 Baseball Players

- **Predict final batting avg from 1st 45 "at bats" for 18 players in 1970**
 - $Y_i \sim \text{Bin}(\pi_{z_i}, 45)$
 - $Z_i = j$ with prob. = θ_j
 - $R_{ji} = \begin{cases} 1 & \text{if } z_i = j \\ 0 & \text{otherwise} \end{cases}$
 - 2 categories: $\theta_i = \Pr(z_i = j)$
 - Take $(\theta_1, \theta_2) \sim \text{Dir}(\lambda_1, \lambda_2)$
 - $\sum \theta_i = 1$

WinBUGS

$Y_i | Z_i \sim \text{Bin}(p_i, 45)$

$Z_i \sim \text{Cat}(\theta_1, \theta_2)$

$(\theta_1, \theta_2) \sim \text{Dir}(\lambda_1, \lambda_2)$

$p_i = \pi(Z_i)$

```
for (i in 1:18) {  
  Y[i] ~ dbin(p[i],45)  
  Z[i] ~ dcat(theta[1:2])  
  R[1,i] <- equals(Z[i],1)  
  R[2,i] <- equals(Z[i],2)  
  p[i] <- pi[Z[i]]  
}
```

- For Dirichlet dist'n, use gammas

```
for (j in 1:2) {  
  Rs[j] <- sum(R[j,])+1  
  Ts[j] ~ dgamma(Rs[j],1)  
  theta[j] <- Ts[j]/sum(Ts[])  
}
```

$\lambda_j = 1$
or
theta[] ~ ddirch(lambda[])

WinBUGS

$Y_i | Z_i \sim \text{Bin}(p_i, 45)$

$Z_i \sim \text{Cat}(\theta_1, \theta_2)$

$(\theta_1, \theta_2) \sim \text{Dir}(\lambda_1, \lambda_2)$

$p_i = \pi(Z_i)$

```
for (i in 1:18) {  
  Y[i] ~ dbin(p[i],45)  
  Z[i] ~ dcat(theta[1:2])  
  R[1,i] <- equals(Z[i],1)  
  R[2,i] <- equals(Z[i],2)  
  prob[i] <- pi[Z[i]]  
}
```

- Need to enforce identifiability

$\text{logit}(\pi_1) = \beta$ and

$\text{logit}(\pi_2) = \beta + \delta, \delta > 0$

```
for (j in 1:2) {  
  logit(pi[j]) <- beta[j]  
}
```

beta[1] ~ dnorm(-1.33,0.001)

beta[2] <- beta[1]+delta

delta ~ dnorm(0,1) I(0,)

Results

- Burn-in = 5000; Every 10th to 15,000

node	mean	sd	2.5%	median
pi[1]	0.231	0.044	0.098	0.242
pi[2]	0.315	0.082	0.245	0.290
delta	0.441	0.413	0.015	0.345
beta[1]	-1.230	0.314	-2.217	-1.141
beta[2]	-0.789	0.372	-1.125	-0.896
MSE	0.038	0.018	0.022	0.032

$\text{logit}(\pi_1) = \beta$ and

$\text{logit}(\pi_2) = \beta + \delta, \delta > 0$

$$\text{MSE} = \sum_i (P_i - \phi_i)^2 = \sum_i \left(\frac{Y_i}{45} - \phi_i \right)^2$$

Dirichlet Process Priors

- Does not prespecify no. of groups
- What is a Dirichlet Process?
 - "Distribution on space of distributions"
 - Related to Dirichlet distribution
 - Divide up measure on real line according to some c.d.f.
 - Dirichlet dist'n according to intervals

Dirichlet Distribution

- Suppose $(n_1, \dots, n_k) \sim \text{Mult}(\theta_1, \dots, \theta_k)$
- Likelihood: $\binom{n}{n_1, \dots, n_k} \sim \prod_{j=1}^k \theta_j^{n_j}, n = \sum_{j=1}^k n_j$
- Conjugate prior is Dirichlet dist'n

$$(\theta_1, \dots, \theta_k) \sim \text{Dir}(\alpha_1, \dots, \alpha_k)$$

- Generalization of Beta dist'n

$$p(\theta) = \frac{\prod_{j=1}^k \theta_j^{\alpha_j - 1}}{B(\alpha)}, B(\alpha) = \frac{\prod_{j=1}^k \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^k \alpha_j\right)}$$

Combine or split categories,
still Dirichlet

Dist'n of Dist'ns

- Let $y_1, \dots, y_n \sim F(y)$, i.i.d.
- Group into classes
 - Corresponding discrete rand var
- Let $x_i = j$ if $b_{j-1} < x_i \leq b_j, -\infty < b_1 < \dots < b_k = \infty$
 - Then $(x_1, \dots, x_k) \sim \text{Mult}(\theta_1, \dots, \theta_k)$ with

$$\theta_j = F(b_j) - F(b_{j-1})$$

- Now, consider Dirichlet prior

$$p(\theta) = \frac{\prod_{j=1}^k \theta_j^{\alpha_j - 1}}{B(\alpha)}$$

Dirichlet Process Definition

- Split categories further $b_{j-1} < b_j \leq b_j$
- When split to effectively infinite-dimensional distribution, have Dirichlet process.
- Definition:
 - The unknown dist'n function $F(y)$ has Dirichlet process distribution with parameter $G(y)$, written DP(G), if the distribution of $\underline{\theta}$ is always Dirichlet, no matter how we define class boundaries.

DP

- For any boundaries $-\infty < b_1 < \dots < b_k = \infty$, the random variables $\underline{\theta}$ have a Dirichlet distribution $Dir(\underline{\alpha})$, where

$$\theta_1 = F(b_1), \theta_2 = F(b_2) - F(b_1), \dots, \theta_k = 1 - F(b_{k-1})$$

$$\alpha_1 = G(b_1), \alpha_2 = G(b_2) - G(b_1), \dots, \alpha_k = g - G(b_{k-1})$$

where $G(y)$ nondecreasing with

$$\lim_{y \rightarrow -\infty} G(y) = 0 \text{ and } \lim_{y \rightarrow \infty} G(y) = g$$

Distribution on Distributions

- Consider 1 boundary $-\infty < b_1 < \infty$ and

$$\theta_1 = F(b_1) \sim \text{Beta}[G(b_1), g - G(b_1)]$$

$$E[F(b_1)] = \frac{G(b_1)}{g}.$$

Some Properties

- In general,

$$E[F(y)] = \frac{G(y)}{g}$$

$$\text{var}[F(y), F(y')] = \frac{G(y)[g - G(y)]}{g^2(g+1)}$$

and

$$\text{cov}[F(y), F(y')] = \frac{G(y)[g - G(y')]}{g^2(g+1)}, y < y'$$

Effect of data? Posterior of DP

- Prior $F(y) \sim DP(G)$
- Data $y_1, \dots, y_n | F \sim F(\bullet), i.i.d.$
- Consider k class boundaries and Dirichlet dist'n (as before)
 $(\theta_1, \dots, \theta_k) \sim Dir(\alpha_1, \dots, \alpha_k) \quad -\infty < b_1 < \dots < b_k = \infty$
- If $n_i = \#\{y's \in (b_{i-1}, b_i]\}$, then have Dirichlet posterior
 $(\theta_1, \dots, \theta_k) | y_1, \dots, y_n \sim Dir(\alpha_1 + n_1, \dots, \alpha_k + n_k)$

In Limit Have Dirichlet Process

- As number of intervals grows, posterior is DP

$$F | y_1, \dots, y_n \sim DP(G_n),$$

$$G_n(x) = G(x) + \frac{\#\{y \leq x\}}{n} = G(x) + F_n(x)$$

In WinBUGS, choose large enough upper limit

Refs: Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *Annals of Statistics* 1: 209-230.

Escobar, M. D. (1994). "Estimating Normal Means With A Dirichlet Process Prior." *JASA* 89(425): 268-277.

Escobar, M. D. and M. West (1995). "Bayesian Density-Estimation And Inference Using Mixtures." *JASA* 90(430): 577-588.

Dirichlet Process Mixtures

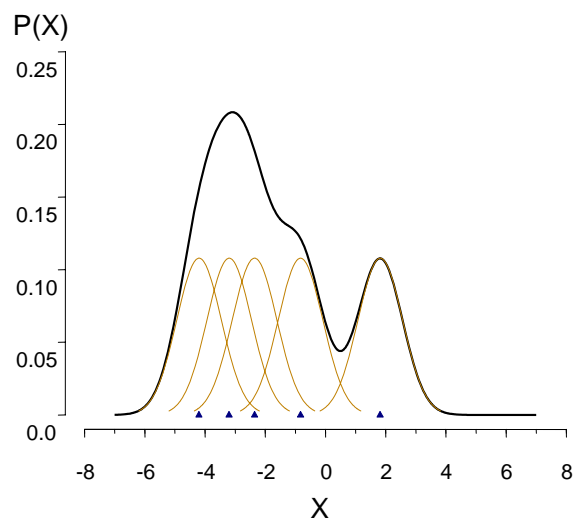
- Posterior mean $E[F|y_1, \dots, y_n]$ has jumps and positive probability

$$y_{n+1} = y_i \quad (i = 1, \dots, n)$$

$$F \sim DP(G) \Rightarrow F \text{ a.s. discrete}$$

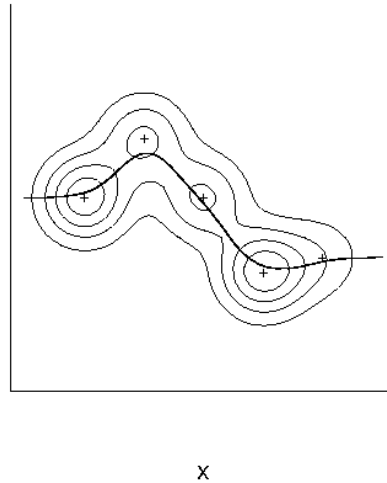
- Suppose believe Y continuous
- Model with DP mixture

Like Kernel Smoothing



Semiparametric Regression

- Locally-weighted, piecewise linear regression



DP Mixture Distribution

- DP prior on dist'n of parameter in dist'n function, such as location

$$F(y) = \int F_1(y - \mu) dF_2(\mu),$$

$F_2 \sim DP(G)$ and F_1 is known continuous kernel.

$$F(y) = \sum_j f_j F_1(y - \mu_j)$$

F_2 assigns point mass f_j to point μ_j .

Similarities with kernel density estimation

Mixture of Normals Population Model

$p(\theta_i, x_i | \phi) = \text{Mixture of MV Normals}$

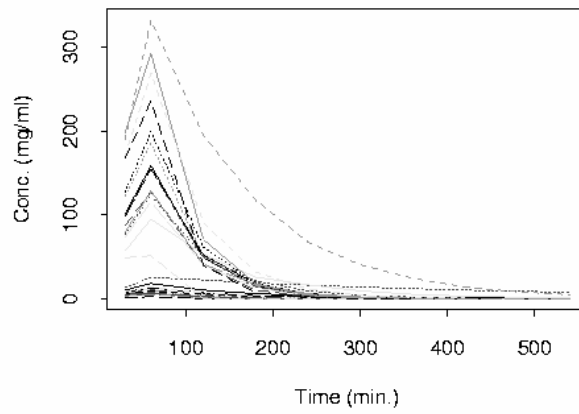
$$(\theta_i, x_i) \sim M_G(\theta, x) = \sum_k w_k N(\mu_k, V)$$

- Multivariate normal kernel in mixture centered at discrete locations
- Mixing measure $G = \sum_k w_k \delta_k \sim DP(\alpha G_0)$
 - Dirichlet process

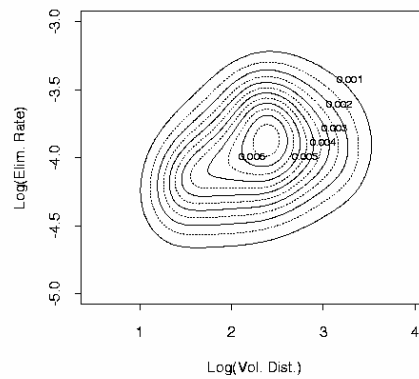
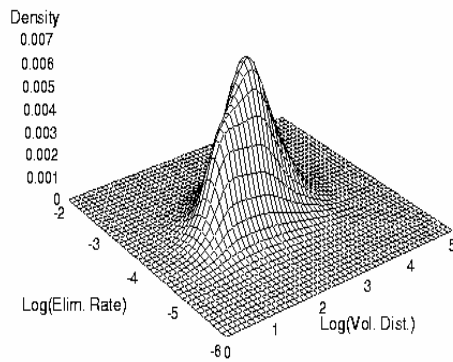
Example (synthetic)

- One-Compartment Model
 - 1-hour i.v. infusion
- Two populations (15 each)
 - Same mean volume of dist'n;
different mean elimination rates
- Population: $\log(V_d)$ & $\log(k_{10})$
multivariate normal dist'n

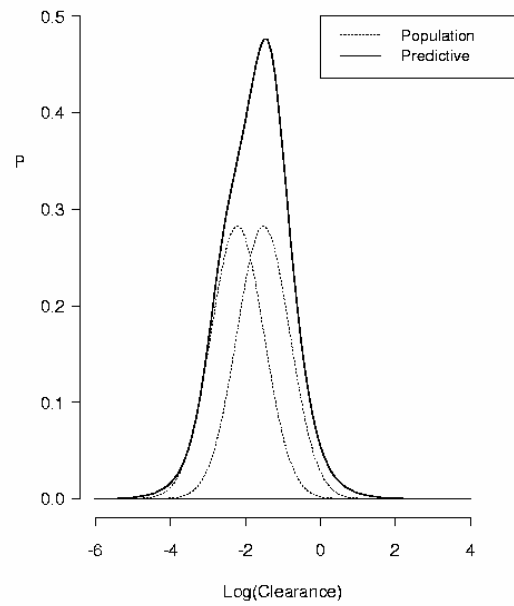
Simulated Data: (Extreme!) Sampling: 30, 60, 90, 120, 180 min.



Posterior Density Estimate



Predicted
Ln(Cl_{tb})
with
2 Pop'ns



Include Covariates with
Parameters (θ_i) to Get
Semiparam. Regression

