The Dirichlet Process

The *Dirichlet Process* is a probability distribution on the space of discrete distributions on $\mathbb{R}$ characterized by

- a *base distribution*, G_X, which is some probability measure on $\mathbb{R}$, from which are drawn random variables $X_1, X_2, \ldots$.
- a countable collection of *random probabilities* $\pi_1, \pi_2, \ldots$ with $\pi_i > 0$ such that

$$\sum_{i=1}^{\infty} \pi_i = 1$$

with $\pi_1 = V_1 \sim Beta(1, \alpha)$ and for $i = 2, 3, \ldots,$

$$\pi_i = V_i \prod_{j=1}^{i-1} (1 - V_j)$$

where for each j, $V_j \sim Beta(1, \alpha)$.

The Dirichlet Process

```
set.seed(3)
alpha<-2.5;M<-5000
nreps<-2000
pi.mat<-matrix(0,nrow=nreps,ncol=M)
for(i in 1:nreps){
    V<-rbeta(M,1,alpha)
    pi.mat[i,]<-c(V[1],cumprod(1-V[-M])*V[-1])
}
round(pi.mat[1,1:6],5)

+ [1] 0.38992 0.12755 0.35586 0.02867 0.02705 0.01836

round(pi.mat[2,1:6],5)

+ [1] 0.06067 0.18180 0.14346 0.25954 0.09613 0.17481

round(pi.mat[3,1:6],5)

+ [1] 0.59511 0.10145 0.06439 0.08052 0.00782 0.06107
```

The Dirichlet Process

π sequence is eventually zero to machine accuracy

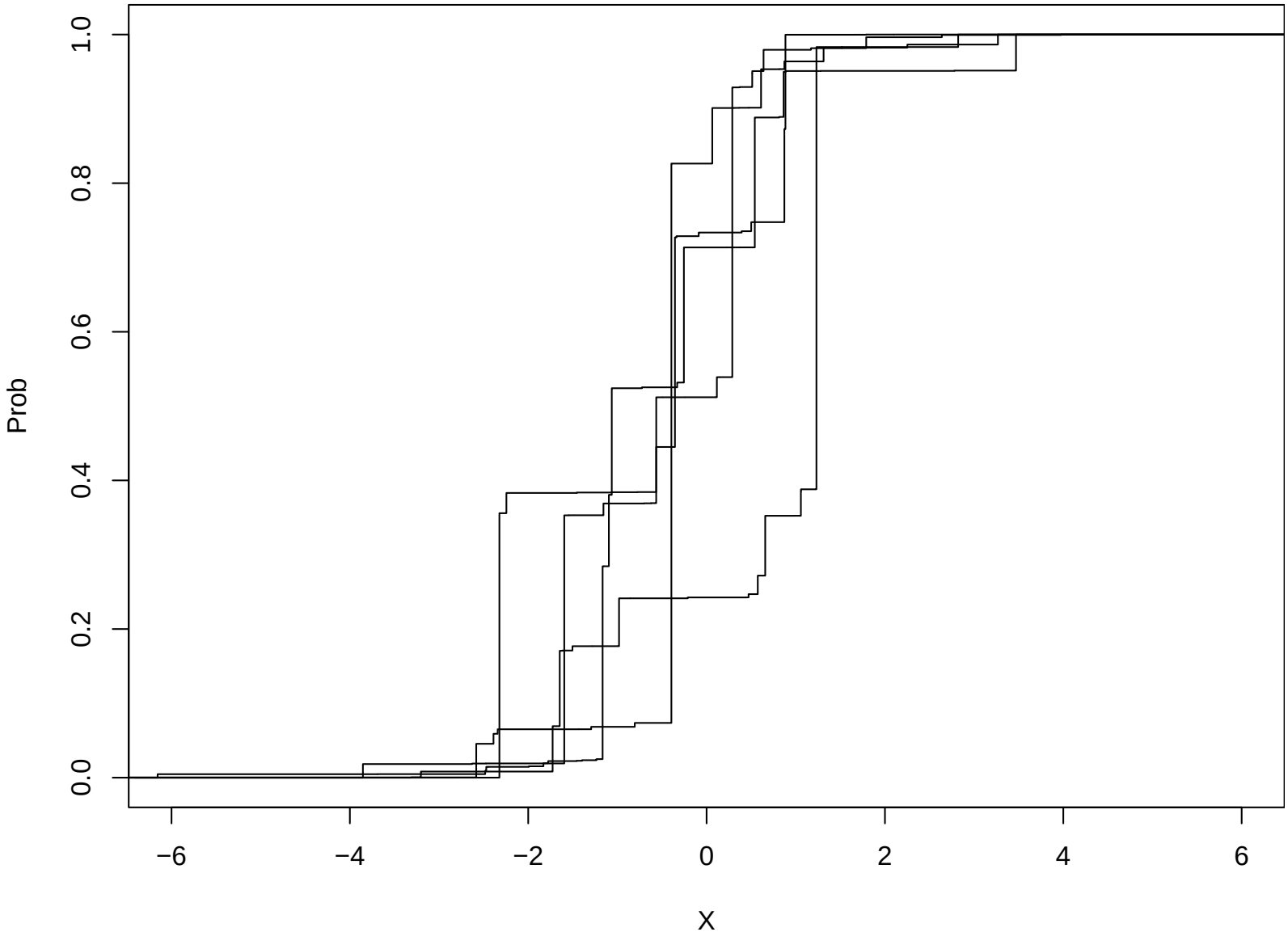
```
pi.trunc<-apply(pi.mat,1,which.min)
pi.trunc[1:10]

+  [1] 1854 1877 1881 1848 1880 1878 1834 1883 1892 1908

pi.mat[1,c(pi.trunc[1]-1,pi.trunc[1])]

+ [1] 1.976263e-323  0.000000e+00
```

The Dirichlet Process

The probability masses are placed at points drawn from G_X: suppose $G_X \equiv Normal(0, 2^2)$.

```
sig<-2
X.mat<-matrix(rnorm(M*nreps,0,sig),nrow=nreps)
par(mar=c(4,4,1,0))
plot(X.mat[1,],rep(0,M),type='n',
     xlim=range(-6,6),ylim=range(0,1),xlab='X',ylab='Prob')
for(i in 1:5){
     X.vec<-X.mat[i,]
     pi.vec<-pi.mat[i,order(X.vec)]
     X.vec<-sort(X.vec)
     lines(X.vec,cumsum(pi.vec),type='s')
}
```

The Dirichlet Process



The Dirichlet Process

For *any partition*

$$B_1, \ldots, B_K, B_{K+1}$$

of $\mathbb{R}$, the *amounts of probability*

$$P_1, \ldots, P_K, P_{K+1}$$

assigned to the elements of the partition form a

$$Dirichlet(K+1; \alpha_1, \alpha_2, \ldots, \alpha_K, \alpha_{K+1})$$

random vector, where

$$\alpha_k = \alpha G_X(B_k) \qquad k = 1, \ldots, K+1$$

The Dirichlet Process

The *Dirichlet distribution* is a model for a collection of probabilities

$$f(\pi_1, \ldots, \pi_K) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_{K+1})}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{K+1})} \pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1} (1 - \pi_1 - \cdots - \pi_K)^{\alpha_{K+1} - 1}$$

where

$$0 < \sum_{k=1}^{K} \pi_k < 1$$

Two ways to simulate:

(i) *Direct:* For $k = 1, \ldots, K+1$, $W_k \sim Gamma(\alpha_k, 1)$ independent, then

$$\pi_k = \frac{W_k}{W_1 + \cdots + W_K} \qquad k = 1, \ldots, K$$

The Dirichlet Process

(ii) *Sequential:*

- $\pi_1 \sim Beta(\alpha_1, \alpha_2 + \cdots + \alpha_{K+1})$
- for $k = 2, 3, \ldots, K$

$$\pi_k = (1 - \pi_1 - \pi_2 - \cdots - \pi_{k-1})V_k$$

where

$$V_k \sim Beta(\alpha_k, \alpha_{k+1} + \cdots + \alpha_{K+1})$$

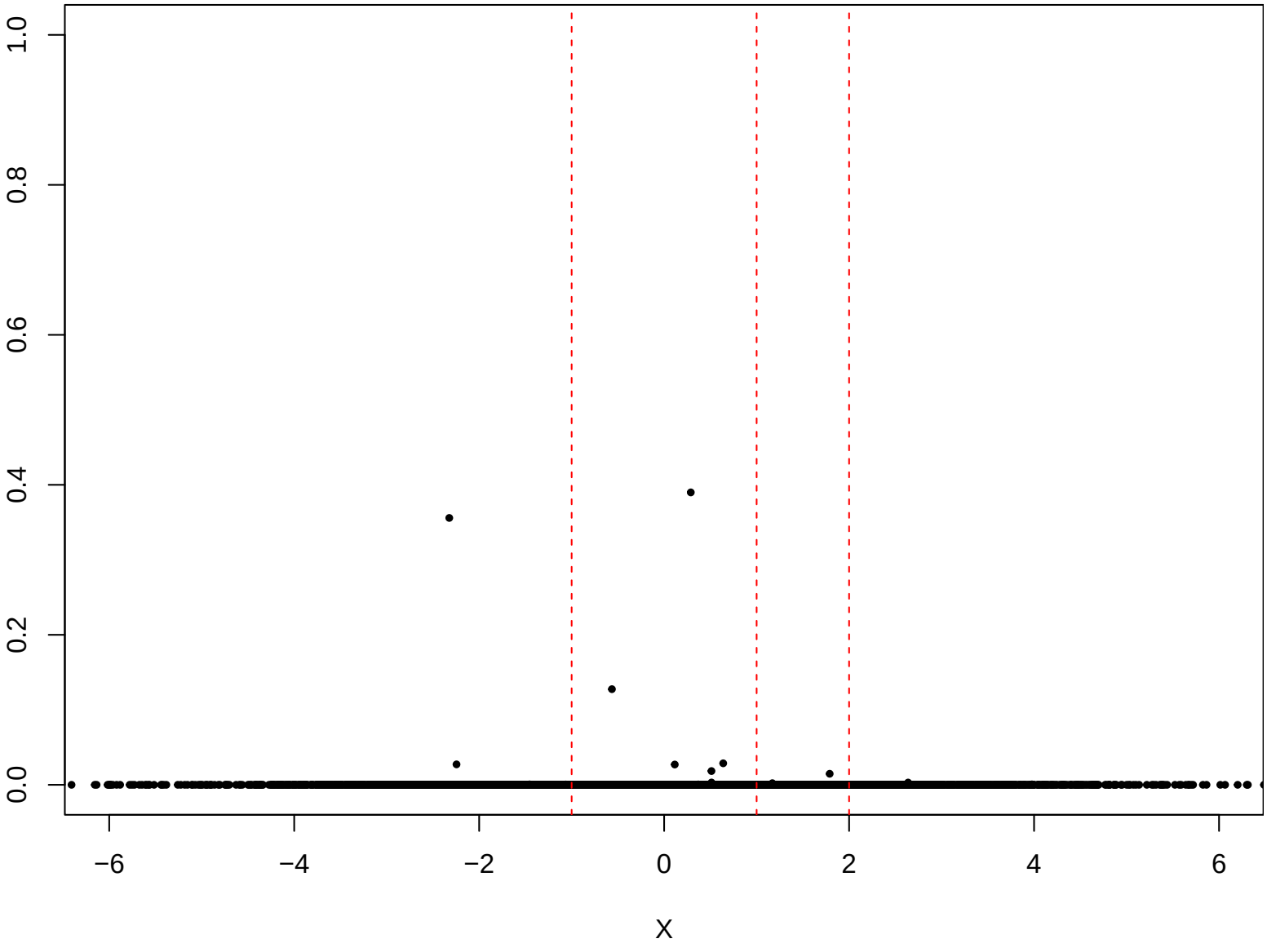
Form a partition:

```
K<-3
b.v<-c(-Inf,-1,1,2,Inf)
G.v<-diff(pnorm(b.v,0,sig))
round(G.v,5)

+ [1] 0.30854 0.38292 0.14988 0.15866

al.k<-alpha*G.v
al.k

+ [1] 0.7713438 0.9573123 0.3747057 0.3966381

al.k/sum(al.k)

+ [1] 0.3085375 0.3829249 0.1498823 0.1586553
```

Add up the probabilities in each partition element:

```
par(mar=c(4,3,1,0))
plot(X.mat[1,],pi.mat[1,],pch=19,cex=0.5,xlab='X',
        xlim=range(-6,6),ylim=range(0,1))
abline(v=b.v,lty=2,col='red')
```

The Dirichlet Process



The Dirichlet Process

```
#Classify each X to a partition element
class.mat<-t(apply(X.mat,1,cut,breaks=b.v,
                   include.lowest=TRUE,labels=FALSE))
table(class.mat[1,])

+
+    1    2    3    4
+ 1522 1920  762  796
```

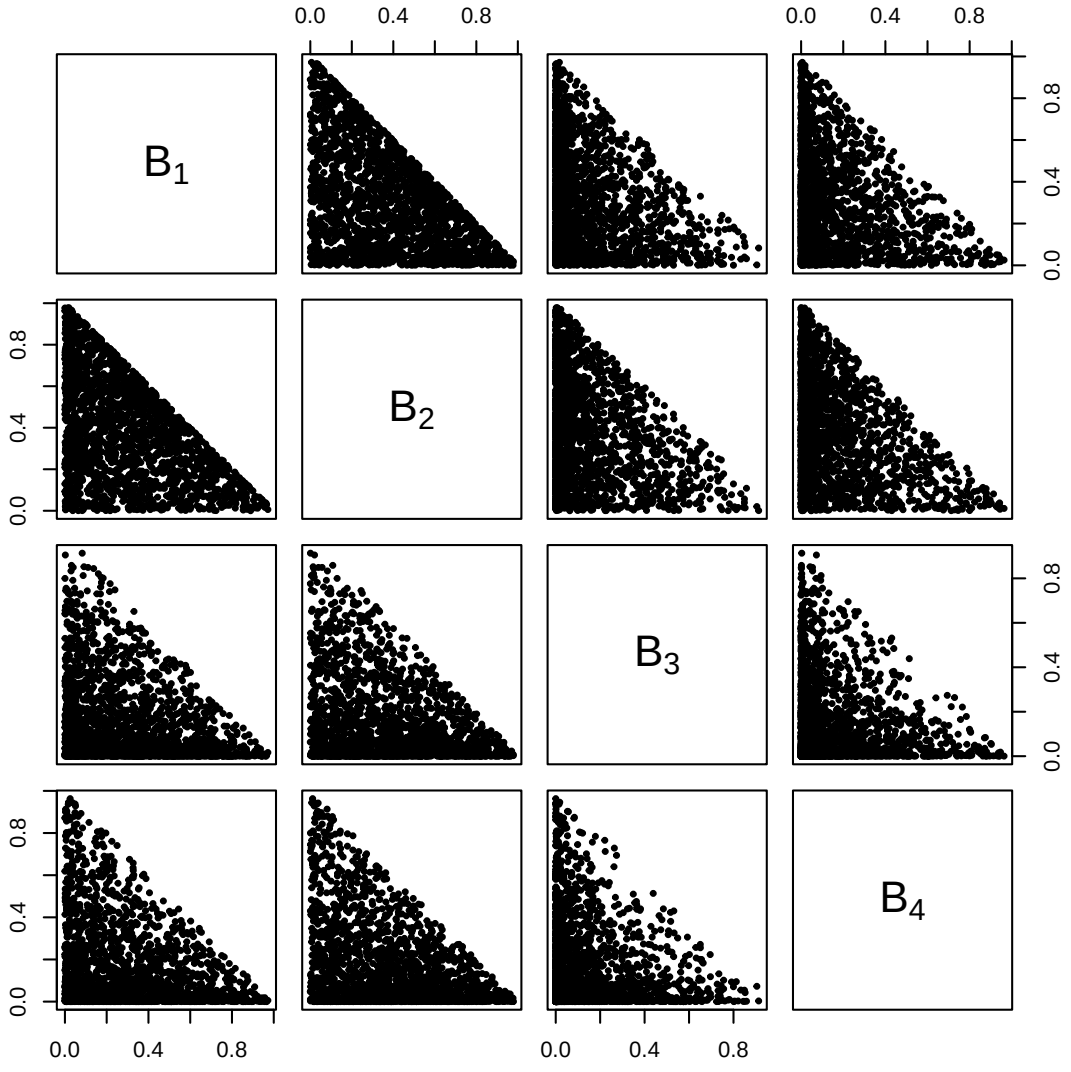
The Dirichlet Process

```
#Sum the probabilities
prob.totals<-matrix(0,nrow=nreps,ncol=K+1)
for(i in 1:nreps){
    for(k in 1:(K+1)){
        prob.totals[i,k]<-sum(pi.mat[i,class.mat[i,]==k])
    }
}
apply(prob.totals,2,mean)  #Mean probability over 1000 replicates

+ [1] 0.3075493 0.3783651 0.1498851 0.1642005

al.k/sum(al.k)             #Expected value

+ [1] 0.3085375 0.3829249 0.1498823 0.1586553
```

Plot the sampled partition probabilities:

```
pairs(prob.totals,pch=19,cex=0.5,
        labels=c(expression(B[1]),expression(B[2]),
                 expression(B[3]),expression(B[4])))
```

The Dirichlet Process

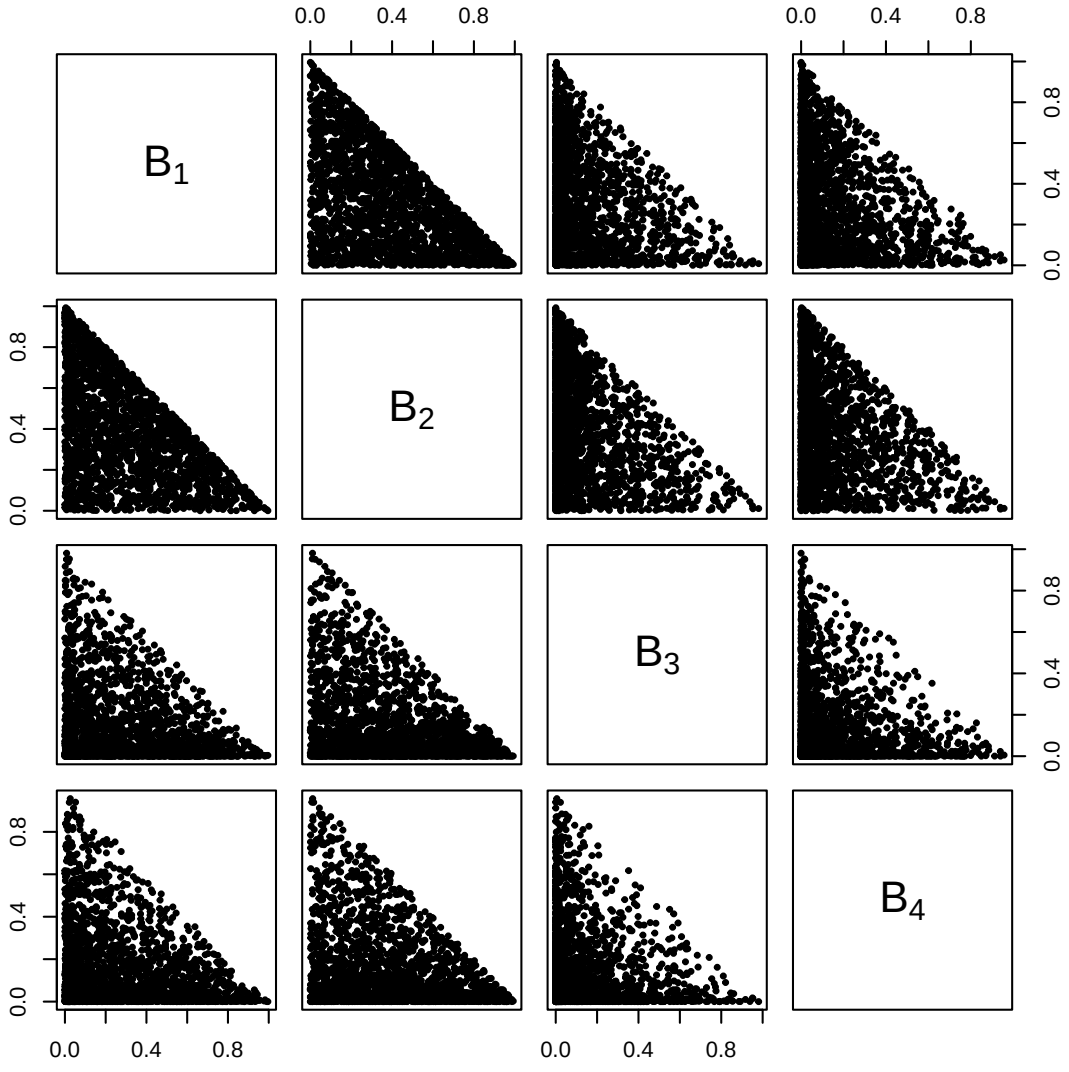


The Dirichlet Process

Sampling the Dirichlet distribution for comparison:

```
library(extraDistr)
dirprob.mat<-rdirichlet(nreps, al.k)
pairs(dirprob.mat,pch=19,cex=0.5,
        labels=c(expression(B[1]),expression(B[2]),
                    expression(B[3]),expression(B[4])))
```

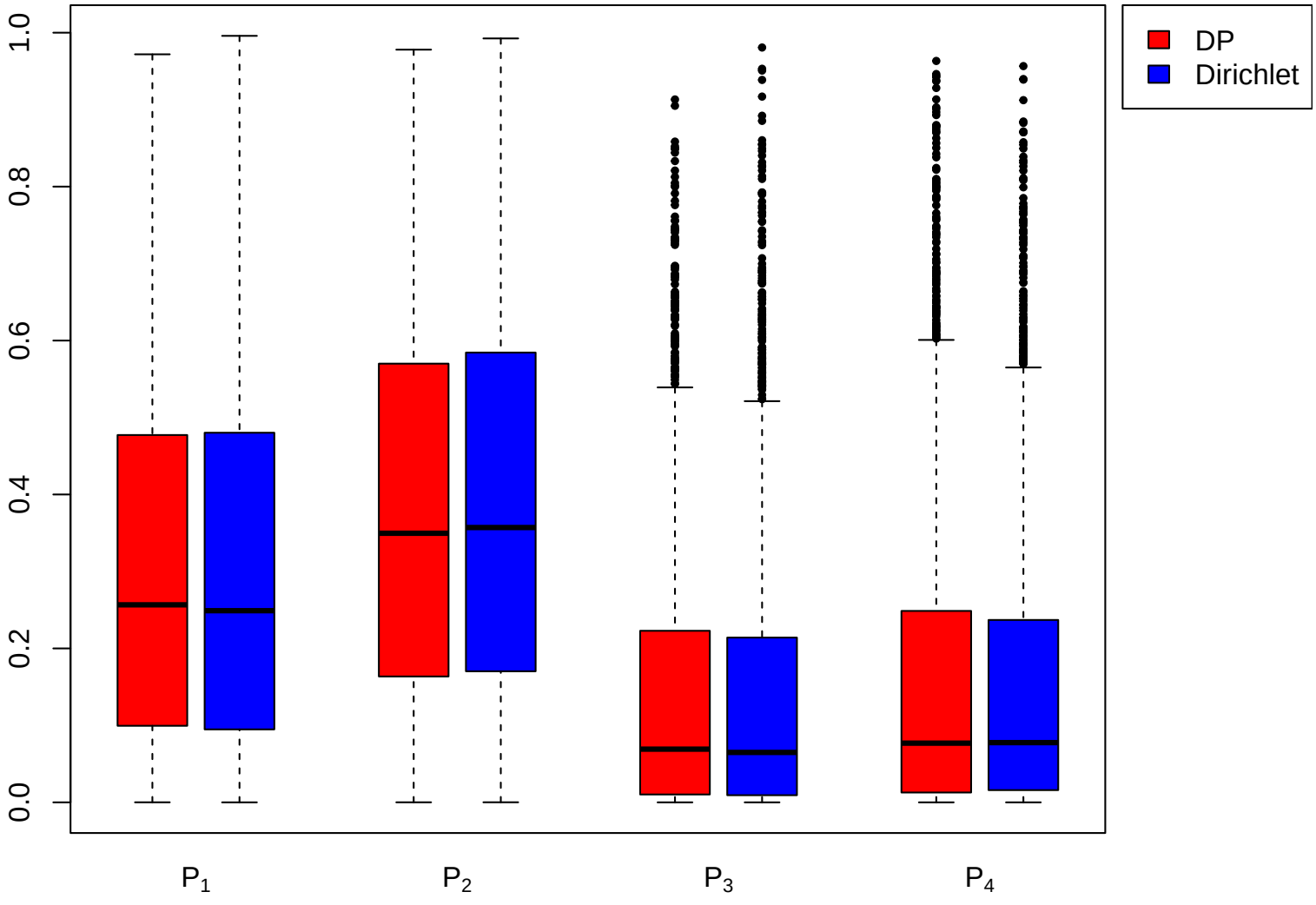
The Dirichlet Process



Comparison:

```
all.probs<-cbind(prob.totals,dirprob.mat)[,c(1,5,2,6,3,7,4,8)]
par(mar=c(5,4,1,6),xpd=TRUE,xaxt='n')
avec<-c(1,2,4,5,7,8,10,11)
boxplot(all.probs,pch=19,cex=0.5,col=c('red','blue'),at=avec)
legend("topright", c("DP", "Dirichlet"),
            fill = c("red", "blue"),inset=c(-0.2,0))
nvec<-c(expression(P[1]),expression(P[2]),
            expression(P[3]),expression(P[4]))
text(c(1.5,4.5,7.5,10.5),rep(-0.1,4),nvec)
```

The Dirichlet Process



The Dirichlet Process

The *random mass function* $\widetilde{f}(x)$ for new random variable Y

$$\widetilde{f}(x) = \sum_{i=1}^{\infty} \pi_i \delta_{X_i}(x)$$

has the random probability specification

$$\Pr[Y = X_i] = \pi_i \qquad i = 1, 2, \ldots$$

We can easily draw a sample of Y values

```
n<-10
id<-sample(1:M,size=n,prob=pi.mat[1,],rep=T)
Y<-X.mat[1,id]
round(Y,5)

+   [1] -2.32324  0.28831 -2.32324  0.11514 -0.56453  0.28831 -0.56453
+   [9]  0.28831 -0.56453
```

The Dirichlet Process

There are *ties* in this sample

```
n<-100
id<-sample(1:M,size=n,prob=pi.mat[1,],rep=T)
table(round(X.mat[1,id],5))

+
+ -2.32324 -2.24477 -0.56453  0.11514  0.28831  0.51173  0.63875  1.78
+       49        2       10        1       33        2        1
+  2.63767
+        1

round(cbind(X.mat[1,1:8],pi.mat[1,1:8]),5)

+           [,1]    [,2]
+ [1,]  0.28831 0.38992
+ [2,] -0.56453 0.12755
+ [3,] -2.32324 0.35586
+ [4,]  0.63875 0.02867
+ [5,]  0.11514 0.02705
+ [6,]  0.51173 0.01836
+ [7,]  2.63767 0.00302
+ [8,]  1.17036 0.00217
```

The Dirichlet Process

The *Polya Urn* approach for simulating Y:

1. simulate $Y_1 \sim G_X$;
2. for $i = 2, 3, \ldots, n$ generate Y_i from the *mixture distribution*

$$\frac{\alpha}{\alpha + i - 1} G_X(.) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{Y_j}(.)$$

that is

- with probability $\alpha/(\alpha + i - 1)$, simulate $Y_i \sim G_X(.)$;
- with probability $1/(\alpha + i - 1)$, simulate Y_i *uniformly* on $\{Y_1, \ldots, Y_{i-1}\}$.

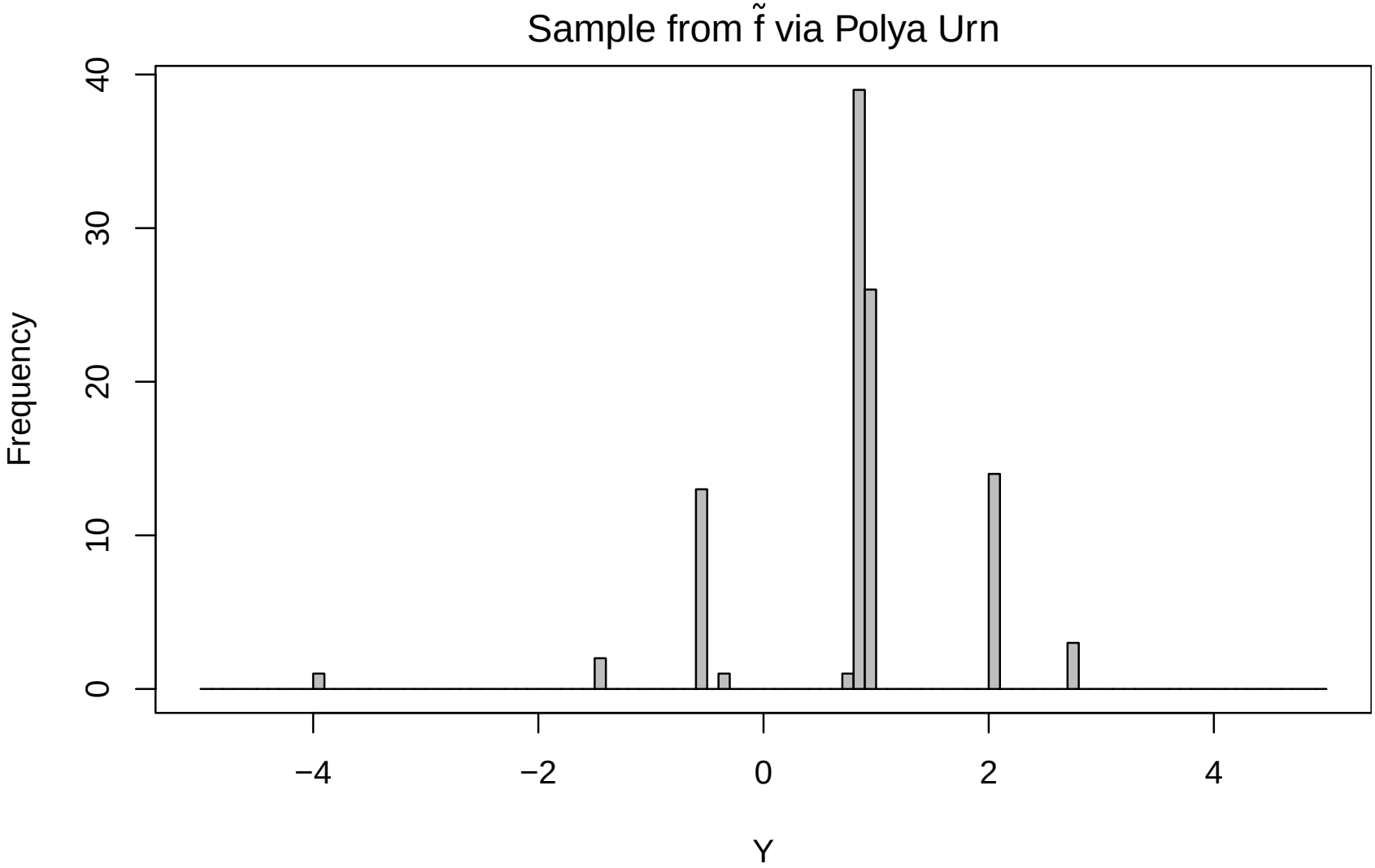
The Dirichlet Process

```
#Polya urn
n<-100
Y<-numeric(n)
Y[1]<-rnorm(1,0,sig)
for(i in 2:n){
        u<-runif(1)
        if(u < alpha/(alpha+i-1)){
                Y[i]<-rnorm(1,0,sig)
        }else{
                Y[i]<-sample(Y[1:(i-1)],size=1)
        }
}
table(round(Y,5))

+
+ -3.97199 -1.43244 -0.53967 -0.36081  0.79516   0.8305   0.9805  2.01
+         1         2        13         1         1        39        26
+  2.78123
+         3
```

The Dirichlet Process

```
par(mar=c(4,4,2,0))
tval<-expression(paste('Sample from ', widetilde(f),' via Polya Urn'))
hist(Y,col='gray',breaks=seq(-5,5,by=0.1), main=tval)
box()
```

The Dirichlet Process



Posterior calculation:

- Data: $y_1, y_2, \ldots, y_n$ iid from $\widetilde{f}$
- Prior: $\widetilde{f} \sim DP(\alpha, G_X)$;
- Posterior: $\widetilde{f} \sim DP(\alpha^\star, G_X^\star)$: for set B,

$$\alpha^\star = \alpha + n$$

$$G_X^\star(B) = \frac{\alpha}{\alpha + n} G_X(B) + \frac{1}{\alpha + n} \sum_{j=1}^{n} \delta_{y_j}(B)$$

In prior and posterior, the unknown $\widetilde{f}$ is almost surely *discrete*.

The Dirichlet Process

Posterior sampling: for the data generating model, we specify that

$$Y \sim Gamma(3, 1/2)$$

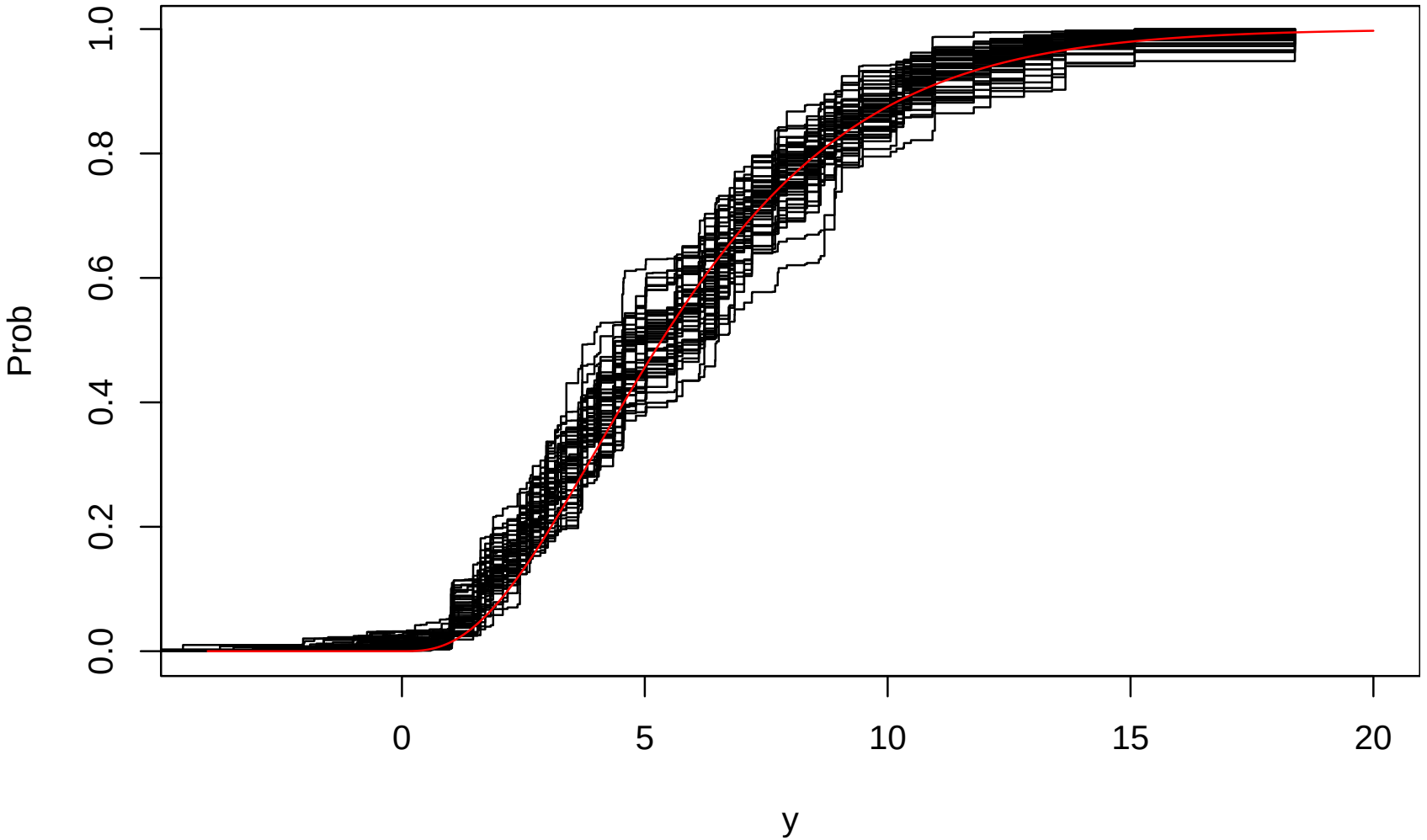
but retain the prior base measure $G_X \equiv Normal(0, 2^2)$.

```
n<-100
Y<-rgamma(n,3,0.5)
al.star<-alpha+n
nsamp<-2000
pi.samp<-X.samp<-matrix(0,nrow=nsamp,ncol=M)
for(i in 1:nsamp){
    V<-rbeta(M,1,al.star)
    pi.samp[i,]<-c(V[1],cumprod(1-V[-M])*V[-1])
    tmp.samp1<-rnorm(M,0,sig)
    tmp.samp2<-sample(Y,size=M,rep=T)
    u<-runif(M)
    X.samp[i,]<-(u < alpha/al.star)*tmp.samp1+
                (u > alpha/al.star)*tmp.samp2
}
```

The Dirichlet Process

```
par(mar=c(4,4,2,0))
xv<-seq(-4,20,by=0.01)
yv<-pgamma(xv,3,0.5)
plot(xv,yv,type='n',ylab='Prob',xlab='y')
for(i in 1:50){
    X.vec<-X.samp[i,]
    pi.vec<-pi.samp[i,order(X.vec)]
    X.vec<-sort(X.vec)
    lines(X.vec,cumsum(pi.vec),type='s')
}
lines(xv,yv,col='red')
```

The Dirichlet Process



The Dirichlet Process

- The true data generating distribution is *continuous*, but the *discrete* distributions sampled in the posterior provide a good approximation;
- The prior and posterior place non-zero probability on mass functions with support on *negative values*.

The Dirichlet Process

- for large n,

$$G_X^\star(B) \backsimeq \frac{1}{n}\sum_{j=1}^{n}\delta_{y_j}(B)$$

so G_X concentrates on the *empirical cdf*, and the posterior can be sampled by generating probabilities

$$\pi \sim Dirichlet(n; 1, \ldots, 1)$$

and placing them on the *observed data* $\{y_1, \ldots, y_n\}$.