Presentation Antoniak's paper

Olga Lazerova, Umberto Junior Mele

Mixtures of Dirichlet process with applications to Bayesian Nonparametric problems.

To make Bayesian analysis be fruitful als in treating non-parametric problems, we need to set a prior distribution on the set of probability distributions that could be acceptable for a given problem on a given sample space. So there are some desiderable properties of a prior distribution for non-parametric problems, that need to be satisfied:

- The support of the prior distribution should be large, with respect to some suitable topology on the space of probabilty distributions on the sample space.
- Posterior distributions given a sample of observation from the true probability distribution should be manageable analythically and easy to determine.
- It should be possible to express conveniently the expectation of some simple loss function;
- The class **D** of random prior distribution on the sample space, should be closed, in the sense that if the prior is a member of **D**, then the posterior is a member also of **D**.
- The class **D** should be "*rich*", so that there will exist a member of **D** capable of expressing any prior information or belief.
- Furthermore, should be parametrized in a manner which can be readily interpreted in relation to prior information and belief.

Ferguson[1] has defined a process called Dirichlet process which is particulary strong in satisfying almost all the previouse conditions, but is slightly deficient with respect to the codition related to the "*richness*" of the prior distributions. Although Ferguson was succesful in using the Dirichlet processes for Bayesian analyses of several non-parametric problems, there are some statistical models for which the closure property does not hold. Since we want to fix also this latest problem, Antoniak[2] suggested the Mixture of Dirichlet Process that is an improvement of the Dirichlet Process proposed by Ferguson, which, roughly, is a Dirichlet Process where the parameter α is itself random.

Mixtures of Dirichlet Processes

To understand better the MDP structure, we need first to define a slight generalization of the usual definition of a transition probability:

Definition 1 let $(\Omega, \mathcal{A})$ and $(U, \mathcal{B})$ be two measurable spaces. A *transition measure* on $U \times \mathcal{A}$ is a mapping α of $U \times \mathcal{A}$ into $[0, \infty)$ such that:

- for every $u \in U$, $\alpha(u, \cdot)$ is a finite, nonnegative, nonnull measure on $(\Omega, \mathcal{A})$.
- for every $A \in \mathcal{A}$, $\alpha(\cdot, A)$ is measurable on $(U, \mathcal{B})$

Note that this differs from the usual transition probability kernel, since now we want to use $\alpha(u, \cdot)$ as parameter for the Dirichlet process, therefore it doesn't need to be $\alpha(u, \Omega) = 1$.

[1] Thomas S. Ferguson. *A Bayesian Analysis of some nonparametric problems*
[2] *Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems*

So we can define the MDP as:

Definition 2 Let $(\Omega, \mathcal{A})$ be a measurable space, let $(U, \mathcal{B}, H)$ be a probability space, called the index space, and let α be a transition measure on $U \times \mathcal{A}$. We say P is a MPD on $(\Omega, \mathcal{A})$ with mixing distribution H on the index space $(U, \mathcal{B})$, and transition measure α, if for all k=1,... and any measurable partition A_1,A_2,...,A_k of Ω we have

$$\mathcal{P}(P(A_1) \leq y_1, ..., P(A_k) \leq y_k) = \int_U D\Big(y_1, y_2, ..., y_k | \alpha(u, A_1), \alpha(u, A_2), ..., \alpha(u, A_k)\Big) dH(u)$$

Where $D\Big(\theta_1, ..., \theta_k | \alpha(u, A_1), \alpha(u, A_2), ..., \alpha(u, A_k)\Big)$ is a Dirichlet distribution function with parameters $(\alpha_1, \alpha_2, ..., \alpha_k)$

Roughly, we may consider the index u as a random variable with distribution H and conditional given u, P is a Dirichlet process with parameter $\alpha(u, \cdot)$.

The next definition and the next proposition are been used to compute the likelihood function of the values sampled from the posterior distribution sampled from the Posterior Dirichlet Process.

Definition 3 Let P be a mixture of Dirichlet process on $(\Omega, \mathcal{A})$ with mixing distribution H on the index space $(U, \mathcal{B})$ and transition measure α on $U \times \mathcal{A}$. We say that $\theta_1, \theta_2, ..., \theta_n$ is a sample of size n from P if for any $m = 1, 2, ...$ and measurable sets $A_1, A_2, ..., A_m, C_1, C_2, ..., C_n$ we have that:

$$\mathcal{P}\{\theta_1 \in C_1, ..., \theta_n \in C_n | u, P(A_1), ..., P(A_m), P(C_1), ..., P(C_n)\} = \prod_{i=1}^{n} P(C_i) \quad a.s.$$

Proposition 1 If $P \sim \int_U \mathcal{D}\big(\alpha(u, \cdot)\big) dH(u)$ and θ is a sample of size one from P, then for any meeasurable set A,

$$\mathcal{P}(\theta \in A) = \int_U \frac{\alpha(u, A)}{\alpha(u, \Omega)} dH(u)$$

PROOF. Since $\mathcal{P}(\theta \in A | u, P(A)) = P(A) \quad a.s.$, hence:

$$\mathcal{P}(\theta \in A | u) = \mathcal{E}_{P(A)|u}\{\mathcal{P}\big(\theta \in A | u, P(A)\big) | u\} \quad a.s.$$

$$= \mathcal{E}\{P(A) | u\} = \frac{\alpha(u, \cdot)}{\alpha(u, \Omega)} \quad a.s.$$

Finally:

$$\mathcal{P}(\theta \in A) = \mathcal{E}_U \left[\frac{\alpha(u, \cdot)}{\alpha(u, \Omega)} \right] = \int_U \frac{\alpha(u, \cdot)}{\alpha(u, \Omega)} dH(u)$$

THEOREM 1 Let P be a Dirichlet process on $(\Omega, \mathcal{A})$ with parameter α. Let θ be a sample size 1 from P, and $A \in \mathcal{A}$ be any measurable set such that $\alpha(A) > 0$. Then the conditional distribution of P, given $\theta \in A$, is a mixture of Dirichlet processes on $(\Omega, \mathcal{A})$, with index space $(A, \mathcal{A} \cap A)$, and transition measure α on $A \times (\mathcal{A} \cap A)$, where $H_A(\cdot) = \alpha(\cdot)/\alpha(A)$ on A and $\alpha(u, \cdot) = \alpha + \delta_u$ for $u \in A$.

PROOF Let $(B_1, B_2, ..., B_n)$ be any measurable partition of Ω. Since A is measurable we obtain a refined measurable partition by letting $B_i^{'} = A \cap B_i$, and $B_i^0 = A^c \cap B_i$ so that $A = \bigcup_{i=1}^n B_i^{'}$. It follows that the conditional distribution of $P(B_1^{'}), ..., P(B_n^{'}), P(B_1^0), ..., P(B_n^0)$ given $\theta \in B_i^{'}$ is a simple Dirichlet with parameters $\left(\alpha(B_1^{'}), ..., \alpha(B_i^{'}) + 1, ..., \alpha(B_n^0)\right)$. Integrating this respect to the conditional probability that $\theta \in B_i^{'}$, given $\theta \in A$, we obtain the conditional distribution of $P(B_1^{'}), ..., P(B_n^0)$, given $\theta \in A$, is :

$$\sum_{i=1}^{n} \frac{\alpha(B_i^{'})}{\alpha(A)} \mathcal{D}(\alpha(B_1^{'}), ...\alpha(B_i^{'}) + 1, ..., \alpha(B_n^0)$$

Where we recognize $H(\cdot) = \alpha(\cdot)/\alpha(A)$ on A.

Corollary 1.1 Let P be a mixture of Dirichlet processes on $(\Omega, \mathcal{A})$ with index space also $(\Omega, \mathcal{A})$, and transition measure $\alpha_u = \alpha + \delta_u$. If the distribution H on the index space $(\Omega, \mathcal{A})$ is given by $H(A) = \alpha(A)/\alpha(\Omega)$, then P is in fact a simple Dirichlet process on $(\Omega, \mathcal{A})$ with parameter α. In symbols

$$\int_\theta \mathcal{D}(\alpha + \delta_u) \frac{\alpha(du)}{\alpha(\Omega)} = \mathcal{D}(\alpha)$$

So , if we combine the results of Theorem 1, Corollary 1.1 and Proposition 1, we get:

THEOREM 2. Let P be a Dirichlet process on $(\Omega, \mathcal{A})$ with parameter α, and let θ be a sample from P. Let $A \in \mathcal{A}$. Then the conditional distribution of P given $P(A)$ and $\theta \in A$ is the same as the conditional distribution of P given $P(A)$.

Roughly speaking, if $P(A)$ is known, then the event $\theta \in A$ tells us nothing more about the process.

Before we can procede to the most interesting theorems about mixtures of Dirichlert processes, we must stop to examine the measure theoretic structure we have created and insure the existence of certain conditional distributions by adding appropriate regularity conditions to the underlying spaces.

Definition 5. A standard Borel space is a measurable space $(\Omega, \mathcal{A})$, in which $\mathcal{A}$ is a countably generated, and for which there exist a bi-measurable mapping (i.e., the inverse mapping is also measurable) between $(\Omega, \mathcal{A})$ and some complete separable metric space $(Y, \mathcal{C})$.

Let's try to clarify some intuition behind standart Borel spaces. Borel spaces were introduced by G.W. Mackey as a tool in the study of representations of algebraic structures as sets of linear mappings on vector spaces. For example, representations of locally compact groups as unitary operators on a Hilbert space; or representations of C???-algebras as operators on a Hilbert space. It is natural in representation theory to ???identify??? representations that are in some sense ???equivalent???; whence a heightened interest in quotient spaces modulo equivalence relations.

A goal of representation theory is to associate with an algebraic object $\mathcal{A}$ a ???dual object??? $\mathcal{A}*$, defined in terms of representations that are ???primitive??? in an appropriate sense; the hope is that $\mathcal{A}$??? is amenable to analysis and that properties of $\mathcal{A}$ and $\mathcal{A}$??? are reflected in each other.

In the best of all worlds (locally compact abelian groups) the dual object is an algebraic object of the same sort, and one recovers the original object by taking the dual of its dual (in the case of locally compact abelian groups, this is the Pontrjagin duality theorem). More commonly, $\mathcal{A}???$ is a set without algebraic structure; it may, grudgingly, have one or more topologies, usually with a meager supply of open sets (rarely enough for the Hausdorff separation property).

Example: Let (E_i, B_i) be a family of Borel spaces, E the direct union (or ???sum-set???) of the sets E_i . Regard $(E_i)_i???I$ as a partition of E. (Caution: It is conceivable that the original sets Ei are all equal; but after identifying E_i with a subset of E , the E_i are pairwise disjoint.) Write $f_i : E_i???E$ for the insertion mappings. If B is the final Borel structure for the family $(f_i)_i???I$, one calls (E, B) the direct union (or ???sum???) of the Borel spaces (E_i, B_i). For a set $S???E$, $f_i^{???1}(S) = S???E_i$; thus

$$B = \{S???E : S???E_i???B_i \text{ for all } i???I\}$$

Evidently $B???E_i = B_i$, thus (E_i, B_i) is a sub-Borel space of (E, B), and $B_i = \{B???B : B???E_i\}$.

With these definitions we can now state a really important theorem which says, roughly, that if we sample from a mixture of Dirichlet processes, and the sample is distorted by random error, the posterior distribution of the process is again a mixture of Dirichlet processes.

THEOREM 3 Let P be a mixture of Dirichlet processes on a standard Borel space $(\Omega, \mathcal{A})$ with standard Borel space $(U, \mathcal{B})$, distribution H on $(U, \mathcal{B})$, and transition measure α on $U \times \mathcal{A}$. Let $(X, \mathcal{C})$ be a standard Borel sample space, and F a transition probability from $\Omega \times \mathcal{C}$ to $[0, 1]$. If θ is a sample from P, i.e., $\theta|P, u \sim P$ and $X|P, u, \theta \sim F(\theta, \cdot)$, then the distribution of P given $X = x$ is a mixing of Dirichlet processes onn $(\Omega, \mathcal{A})$, with index space $(\Omega \times U, \mathcal{A} \times \mathcal{B})$, transition measure $\alpha_u + \delta_\theta$ on $(\Omega \times U) \times \mathcal{A}$, and mixing distribution H_x on the index space $(\Omega \times U, \mathcal{A} \times \mathcal{B})$ where H_x is the conditional distribution of (θ, u) given $X = x$. In symbols, if:

$$u \sim H, \qquad P|u \sim \mathcal{D}(\alpha_u), \qquad P \sim \int_U \mathcal{D}(\alpha_u)\, dH(u),$$

$$\theta|P, u \sim P, \qquad \textit{and} \qquad X|P, u, \theta \sim F(\theta, \cdot)$$

$$\text{then } (P|X = x) \sim \textstyle\int_{\Omega \times U} \mathcal{D}(\alpha_u + \delta_\theta)\, dH_x(\theta, u).$$

PROOF. The distribution of P given (θ, u, x) is a $\mathcal{D}(\alpha_u + \delta_\theta)$, independent of x, the distribution of X given (θ, u) is $F(\theta, \cdot)$ independent of u; the distribution of θ given u is $\alpha(u, \cdot)/\alpha(u, \Omega)$; and the distribution of u is H. The last three define the joint distribution of (θ, u, X) as given in the theorem, and conditional distribution of P given (θ, u, X), namely $\mathcal{D}(\alpha_u + \delta_\theta)$, with respect to the conditional distribution of (θ, u) give X, that lead us to the formula given in the theorem, which we recognize as a mixture of Dirichlet processes.

So now we state two corollaries of Theorem 3 which treat cases that occurs frequently in applications. No proof is needed since they are simply special cases of Therem 3.

Corollary 3.1 let P be a Dirichlet process on a standard Borel space $(\Omega, \mathcal{A})$, with parameter α and let θ be a sample from P. Let be $(X, \mathcal{C})$ be a standard Borel sample space and F a transition probability from $\Omega \times \mathcal{C}$ to $[0, 1]$. If the conditional distribution of X given P and θ is $F(\theta, \cdot)$, then the conditional distribution of P given $X = x$ is a mixture of Dirichlet processes on $(\Omega, \mathcal{A})$ with mixing distribution H on the index space $(\Omega, \mathcal{A})$ and transition measure $\alpha(\theta, \cdot) = \alpha(\cdot) + \delta_\theta(\cdot)$, where the mixing distribution H on $(\Omega, \mathcal{A})$ considered as the index space is the conditional distribution of θ given $X = x$; in concise notation:

$$P \sim \mathcal{D}(\alpha),\ \theta \sim P,\ X \sim F(\theta, \cdot) \Rightarrow P|X \sim \int \mathcal{D}(\alpha + \delta_\theta)\, dH_x(\theta)$$

This corollary is very useful since this property is used to write analytically the posterior of the DPM model, since it claims that the posterior distribution of the process given the distorted sample is a mixture of Dirichlet processes.

Corollary 3.2 Let P be a mixture of Dirichlet processes on a standard Borel space $(\Omega, \mathcal{A})$, with standard Borel index space $(U, \mathcal{B})$, distribution H in $(U, \mathcal{B})$, and transition measure α on $U \times \mathcal{A}$. If θ is a sample from P, then P given θ is a mixture of Dirichlet processes on $(\Omega, \mathcal{A})$, with transition measure $\alpha_u + \delta_\theta$, and distribution H_θ on $(U, \mathcal{B})$, where H_θ is the conditional distribution of u given θ. In symbols, if $P \sim \int_U \mathcal{D}(\alpha_u)\, dH(u)$ and $\theta \sim P$ then:

$$P|\theta \sim \int_U \mathcal{D}(\alpha_u + \delta_\theta)\, dH_\theta(u)$$

Properties of samples from Dirichlet processes.

The preceding theorems were stated rather formally to reveal the underlying measure theoretic structure, but only for a sample of size one, to avoid cumbersome notation. In what follows we will develop more useful formulas for sample of size n, with less regard for elaborate formalism. We will see that the the join distribution of multiple samples possesses some peculiar properties caused by a virtual 'memory' of the process.

The reason for this is that although α may be nonatomic, the conditional distribution of P given θ, is a Dirichlet process with parameter $\alpha + \delta_\theta$, which is already atomic with an atom of measure 1 at θ. This property of the Dirichlet process is very similar to what happens in Polya urn schemes and in one sense characterizes Dirichlet processes.

On the other hand, one can consider the probability that θ_n is a new value, distinct from any previous observation $\theta_1, \theta_2, ..., \theta_{n-1}$, so we define W_i as a random variable which equals 1 if θ is new and zero otherwise, then $\mathcal{P}(W_i = 1) = \alpha(\Omega)/(\alpha(\Omega) + i - 1)$. If we further define $Z_n = \sum_{i=1}^n W_i$ then this random variableis simply the number of distinct values of θ which have occurred in the first n observations.

Although $P(W_n = 1)$ is a monotone decreasing in n nevertheless note that
$\mathcal{E}\{Z_n\} = \alpha(\Omega) \sum_{i=1}^n 1/(\alpha(\Omega) + i - 1) \approx \alpha(\Omega)[log\big((n + \alpha(\Omega))/\alpha(\Omega)\big)]$. Hence $\mathcal{E}\{Z_n\} \to \infty, \quad n \to \infty.$

Moreover, since the distribution of the distinct values is simply $\alpha(\cdot)/\alpha(\Omega)$, this property can be used in the usual way to obtain information about the shape of α if it unknown.

On the other hand, the rate at which new distinct values appear depends only on the magnitude of $\alpha(\Omega)$, and not in the shape of $\alpha(\cdot)$, and this property should enable us to obtain information about the magnitude of $\alpha(\Omega)$ if it is unknown. Since the probability of observing exactly k distinct values in a sample of size n, is proportional to $\mathcal{P}(Z_n = k) \propto \frac{\alpha(\Omega)^k}{\alpha(\Omega)^{(n)}}$.

The significance of the foregoing is that if one knows he is sampling from some Dirichlet process with unknown parameter α, then we can obtain, independently, consistent estimates for $\alpha(\cdot)/\alpha(\Omega)$ and $\alpha(\Omega)$ by making use of the properties given above. If, however, there were some doubt that the process was in fact a Dirichlet process, then the only discriminating feature lest is the actual pattern of multiplicities observed. As a first step in such discrimination we show that we can, in fact, determine the fine structure of the probability of various patterns of multimplicities in a sample from a Dirichlet process.

DEFINITION 6 Let $\theta_1, \theta_2, ..., \theta_n$ be a sample from a Dirichlet process P. We will say that the sample belongs to the class $C(m_1, m_2, ..., m_n)$, and write $(\theta_1, \theta_2, ..., \theta_n) \in C(m_1, m_2, ..., m_n)$, if there are m_1 distinct values of θ that occurs only once, m_2 that occurs twice,. . . ,m_n that occur exactly n times.

Two immediate consequences of this definition are that $n = \sum_{i=1}^n i \cdot m_i$, and the total number of distinct values of θ that occur is $Z_n = \sum_{i=1}^n m_i$.

Proposition 3 Let P be a Dirichlet process on a standard Borel space $(\Omega, \mathcal{A})$, with parameter α, and let the parameter be nonatomic. Let $(\theta_1, \theta_2, ..., \theta_n)$ is a sample of size n from P, then:

$$\mathcal{P}\{(\theta_1, \theta_2, ..., \theta_n) \in C(m_1, m_2, ..., m_n)\} = \frac{n!}{\prod_{i=1}^n i^{m_i} \cdot (m_i!)} \cdot \frac{\alpha(\Omega)^{\sum_{i=1}^n m_i}}{\alpha(\Omega)^{(n)}}\}$$

Proposition 4 Let $P \sim \int_U \mathcal{D}(\alpha_u)\ dH(u)$, where α_u is nonatomic for all $u \in U$, and let $(\theta_1, \theta_2, ..., \theta_n)$ be a sample of size n from P. Then the posterior distribution of u, given $(\theta_1, \theta_2, ..., \theta_n) \in C(m_1, m_2, ..., m_n) \in \mathcal{A}$ is determined by

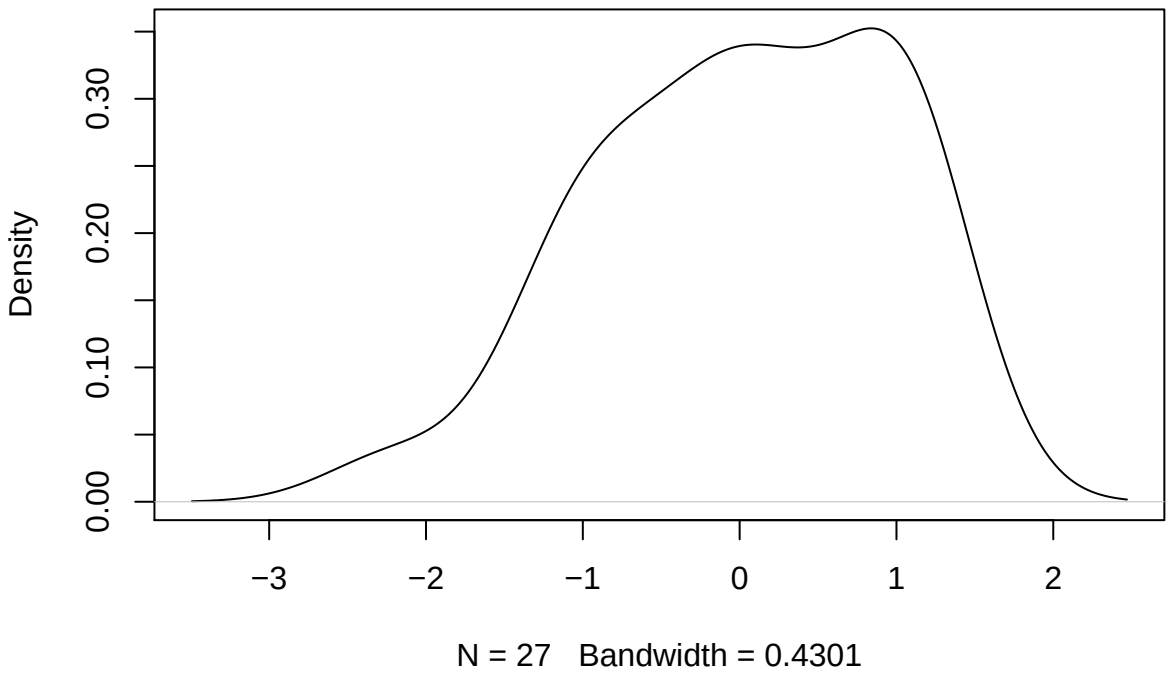
$$\mathcal{P}\big(u \in B | \underline{\theta} \in C(\underline{m})\big) = \frac{\int_B \frac{\alpha(u,\Omega)^{z_n}}{\alpha(u,\Omega)^{(n)}}\ dH(u)}{\int_U \frac{\alpha(u,\Omega)^{z_n}}{\alpha(u,\Omega)^{(n)}}\ dH(u)}$$

We made a little example to make this last part clear:

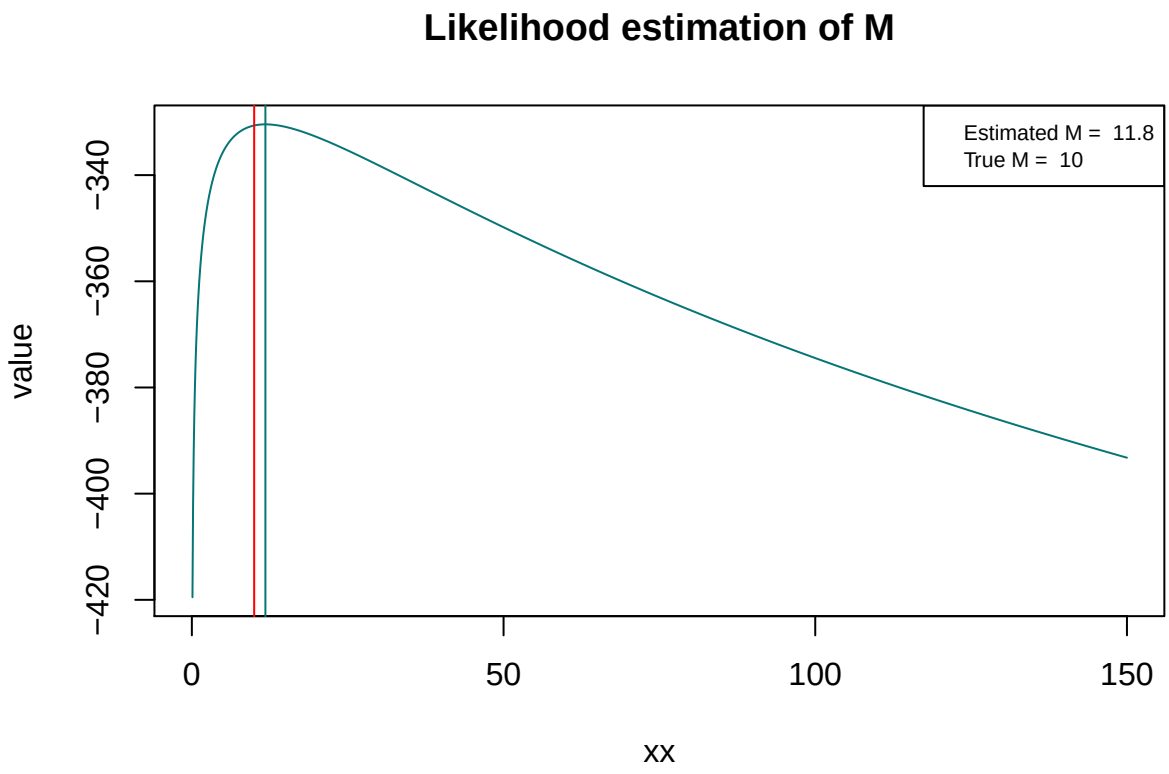
```
sethuraman.cost <- function(number.obs, M){
  n <- 5000
  y <- rnorm(n)
  thet <- rbeta(n,shape1 = 1, shape2 = M)
  prob <- rep(0,n)
  prob[1] <- thet[1]
  for(i in 2:n){
    prob[i]<- thet[i]*prod(1 - thet[1:i-1])
  }
  dat <- sample(y,size= number.obs, prob=prob,replace=T)
  return(dat)
}


function.M <- function(obs, M){
  n = length(obs)
  Z = length(unique(obs))
  num = Z*log(M)
  den = log(M)
  for(i in 1:(n-1)){
    den = den + log(M + i)
  }
  return(num - den)
}

par(mfrow=c(1,1))
M = 10
number.obs=100
observed <- sethuraman.cost(number.obs,M=M)
uniq.obs <- unique(observed)
plot(density(uniq.obs), main = 'Estimation of alpha()')
```




```
xx <- seq(from = 0 , to = 150, by = 0.1)
value <- function.M(observed, xx)
plot(xx, value,type = 'l',
     col= rgb(1,114,116,250,maxColorValue = 255),
     main = 'Likelihood estimation of M')
m.bar <- xx[which.max(value)]
segments(m.bar,-1e10,m.bar,1e10,col= rgb(1,114,116,250,maxColorValue = 255))
segments(M,-1e10,M,1e10,col = 'red')
legend('topright',
       legend = c(paste('Estimated M = ',round(m.bar,1)),
                  paste('True M = ',M)),
       cex = .7)
```




```
compute.prob <- function(obs, M){
  lo <- function.M(obs, M)
  un.obs <- unique(obs)
  num = 0
  den = 0
  m.vec = rep(0, length(obs))
  for(el in un.obs){
    m.vec[sum(el == obs)] = m.vec[sum(el == obs)] + 1
  }
  for(i in (1:length(obs))){
    num = num + log(i)
    den = den + m.vec[i]*log(i) + log(factorial(m.vec[i]))
  }
  return(lo + num - den)
}

exp(compute.prob(observed, m.bar))
```

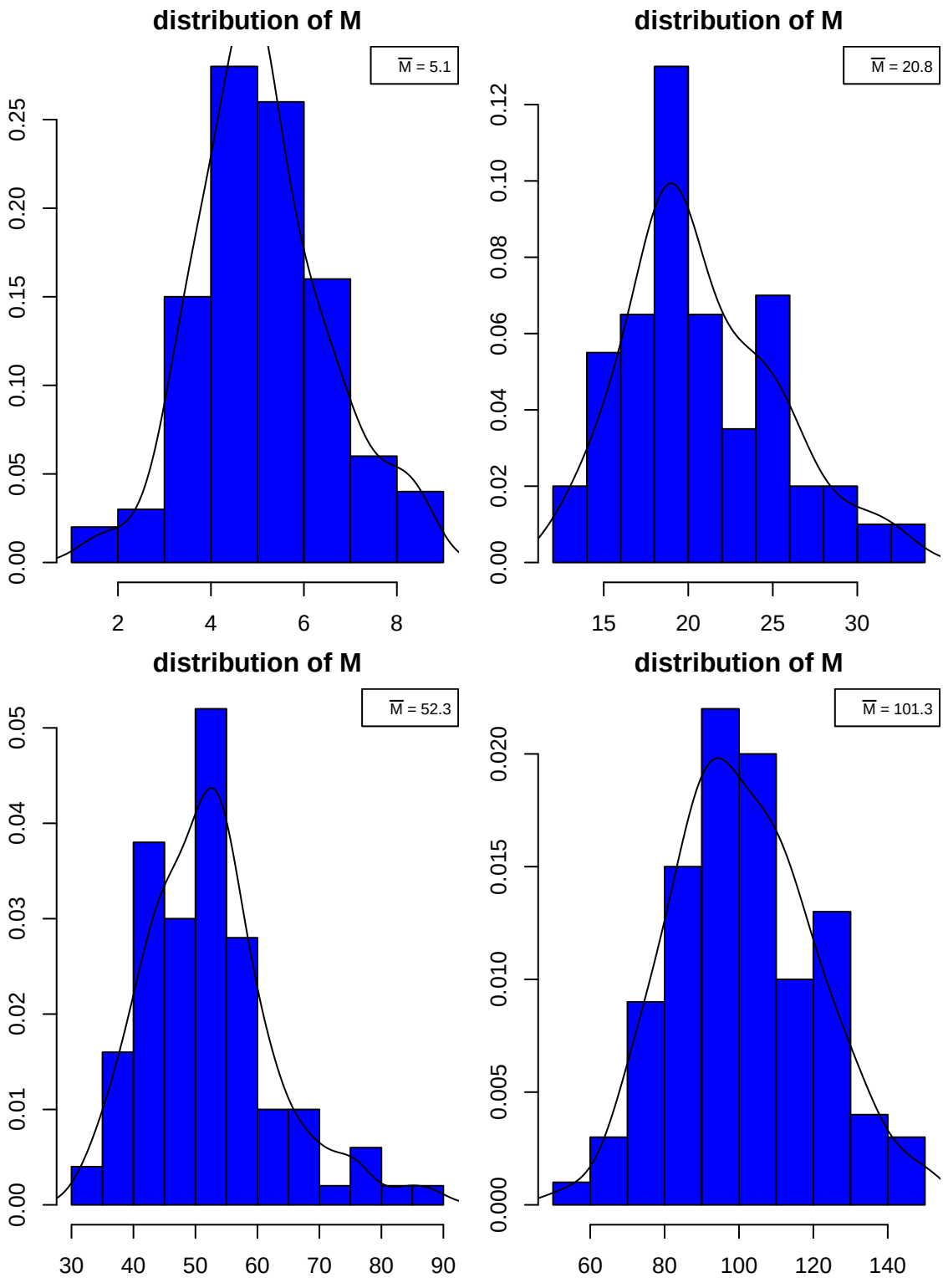
```
## [1] 4.315549e-08
```

```
exp(compute.prob(observed, 10))
```

```
## [1] 3.479967e-08
```

```
M <- c(5, 20 , 50, 100)
number.obs=100
sim= 100
m.estimated <- list()
for(i in 1:4){
  m.estimated[[i]] <- rep(0, sim)
  for(z in 1:sim){
    observed <- sethuraman.cost(number.obs,M=M[i])
    value <- function.M(observed, xx)
    m.estimated[[i]][z] <- xx[which.max(value)]
  }
}
```

```
par(mfrow = c(2,2), mar =c(2,2,2,1))
for(i in 1:4){
  hist(m.estimated[[i]], main = 'distribution of M', probability = T, col = 'blue')
  lines(density(m.estimated[[i]]))
  m <- mean(m.estimated[[i]])
  legend('topright',
         legend = substitute(paste(bar(M),' = ',m), list(m = round(m,1))),
         cex = 0.7)
}
```


distribution of M
M = 5.1
0.00
0.05
0.10
0.15
0.20
0.25
2
4
6
8
distribution of M
M = 20.8
0.00
0.02
0.04
0.06
0.08
0.10
0.12
15
20
25
30
distribution of M
M = 52.3
0.00
0.01
0.02
0.03
0.04
0.05
30
40
50
60
70
80
90
distribution of M
M = 101.3
0.000
0.005
0.010
0.015
0.020
60
80
100
120
140

We can conclude from Propposition 4 that is α is nonatomic and $\alpha_u(\Omega) = M$ is a costant independent of u, then the event $\underline{\theta} \in C(\underline{m})$ is independent of the event $u \in B$, and hence $\underline{\theta} \in C(\underline{m})$ provides no new information about u. This is a consequence of the assumption that α_u be nonatomic. If α_u is atomic the event $\underline{\theta} \in C(\underline{m})$ may still provide information about u even when $\alpha(u, \Omega)$ is independent of u.

LEMMA 1 Let $P \sim \int_U D(\alpha_u) dH(u)$ as in theorem 3, let $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_n)$ be a sample of size n from P, and suppose there exists a σ-finite, σ-additive measure μ on $(\Omega, \mathcal{A})$ such that for each $u \in U$,

1. α_u is σ-additive and absolutely continuous with respect to μ, and
2. the measure μ has mass one to at each atom of α_u. Then

$$dH_{\underline{\theta}}(u) = \frac{\frac{1}{M_u^{(n)}} \cdot \prod_{i=1}^{n} \alpha_u^{'}(\theta^{'})(m_u + 1)^{(n(\theta^{'})-1)} dH(u)}{\int_U \frac{1}{M_u^{(n)}} \cdot \prod_{i=1}^{n} \alpha_u^{'}(\theta^{'})(m_u + 1)^{(n(\theta^{'})-1)} dH(u)}$$

where $\alpha_u^{'}(\cdot)$ denotes the Radon-Nikodyn derivate of $\alpha_u(\cdot)$ with respect to μ; $\theta_i^{'}$ is the ith distinct value of θ in $\underline{\theta}$; $n(\theta_i^{'})$ is the number of times the value $\theta_i^{'}$ occurs in $\underline{\theta}$; $M_u = \alpha_u(\Omega)$ and $m_u(\theta_i^{'}) = \alpha_u^{'}(\theta_i^{'})$ if $\theta_i^{'}$ is an atom of α_u, zero otherwise.

PROOF We obtain the joint distribution of $(u, \theta_1, \theta_2, ..., \theta_n)$ by calculating the likelihood function of $\underline{\theta}$ and making the appropriate normalization. We know that the likelihood of θ_{k+1} given $u, \underline{\theta}$ is

$$\frac{\alpha_u^{'}(\theta_{k+1}) d\mu}{M_u + k}$$

for a value of θ_{k+1} which has not occur previously, and

$$\frac{[m_u(\theta_{k+1}) + j] d\mu}{M_u + k}$$

for a value of θ_{k+1} which has occured previously j times.

Hence the Likelihood of $\underline{\theta}|u$ is :

$$\frac{1}{M_u^{(n)}} \cdot \prod_{i=1}^{n} \alpha_u^{'}(\theta^{'})(m_u + 1)^{(n(\theta^{'})-1)} d\mu^n$$

we obtain $dH_{\underline{\theta}}(u)$ by multiplying the previously by $dH(u)$ and dividing by the marginal of $\underline{\theta}$.

A discrimination problem

Let's assume we are given samples $X_{i1}, ...X_{ik_i}$ coming from $D(p_1)...D(p_k)$ and $Y \sim D(p_j)$ where $1 \leq j \leq k$. We are assuming that $\alpha_1, ..\alpha_k \sim Pois(\lambda_k)$ with the same support. We'd like to know the distrubution of Y i.e. classify Y to one of k groups. So we are interested in minimization of the misclassification risk r_i of Y_j which is proportional to $P(Y|i) = \prod_{j=1}^{n} \frac{\alpha_{i,j}^{'(k_j)}}{M^{(n)}}$ which is strickly related to the *Proposition 3* where we defined the probability of a particular sequence in in the class $C(m_1, .., m_n)$.

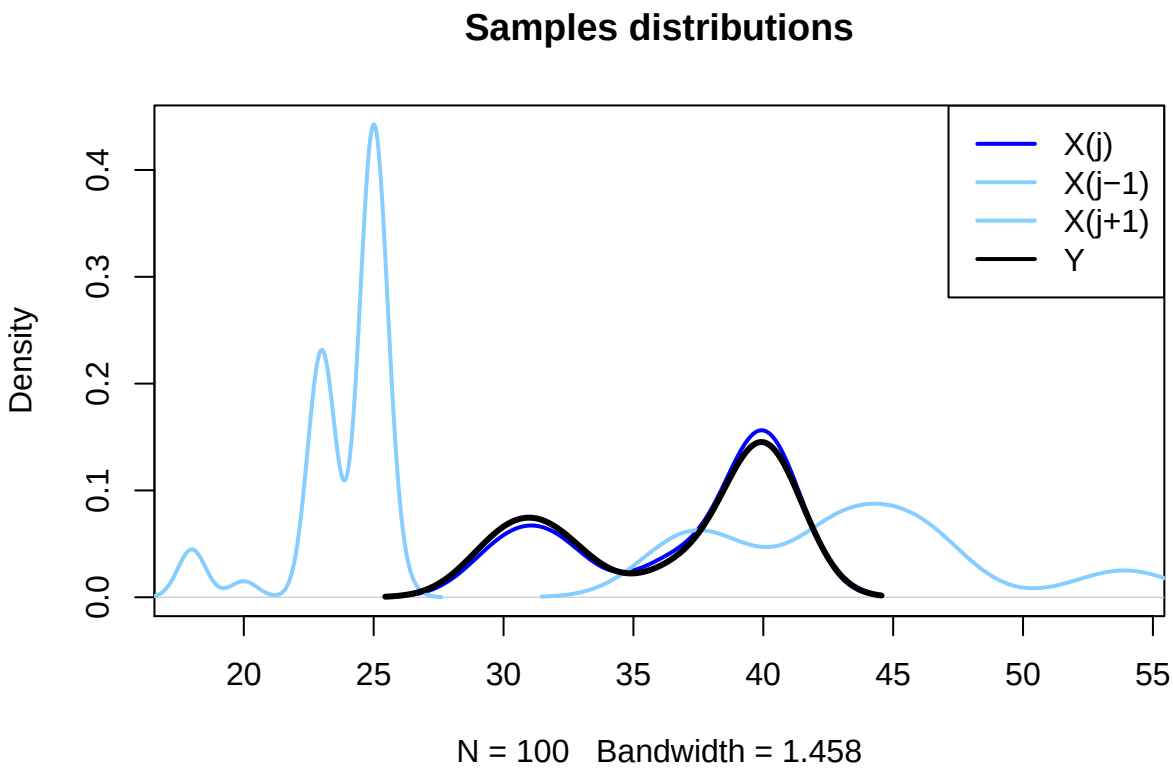
```
k = 10 #number of samples X from different distributions we consider
lambda =seq(1,100,length.out = k)
M = 1
```

```r
theta = seq(1,100,length= 100) #support
n = length(theta)
fun1 = function(n,lambda)(rpois(n,lambda))
s.const <- function(fun,mu, M){
  n <- 5000
  y <- fun(n,mu)
  thet <- rbeta(n,shape1 = 1, shape2 = M)
  prob <- rep(0,n)
  prob[1] <- thet[1]
  prob[2:n] = sapply(2:n, function(i) thet[i] * prod(1 - thet[1:(i-1)]))
  return(list('alpha'=y,'pi'=prob))
}

X = matrix(NA, n,k)
j = sample(seq(2,k-2),1) #random choice of distribution for y
for (i in 1:k){
  obj= s.const(fun1,mu = lambda[i],M = M)
  X[,i] <- sample(obj$alpha,size= n, prob=obj$pi,replace=T)
  if (i == j)(obj_y = obj)
}
y <- sample(obj_y$alpha,size= n, prob=obj_y$pi,replace=T)
```

Let's compare disribution of X_{j-1}, X_j, X_{j+1} and Y:

```r
plot(density(X[,j]),col = "blue",lwd =2,
     xlim=range(min(X[,j-1]),max(X[,j+1])),
     ylim=range(0,max(c(max(density(X[,j-1])$y),max(density(X[,j])$y),max(density(X[,j+1])$y)))),
                main = "Samples distributions")
lines(density(X[,j-1]),col = "skyblue1",lwd =2)
lines(density(X[,j+1]),col = "skyblue1",lwd =2)
lines(density(y),lwd = 3)
legend("topright",
       c("X(j)","X(j-1)","X(j+1)","Y"), col = c("blue", "skyblue1", "skyblue1","black"),
       lwd = 2)
```

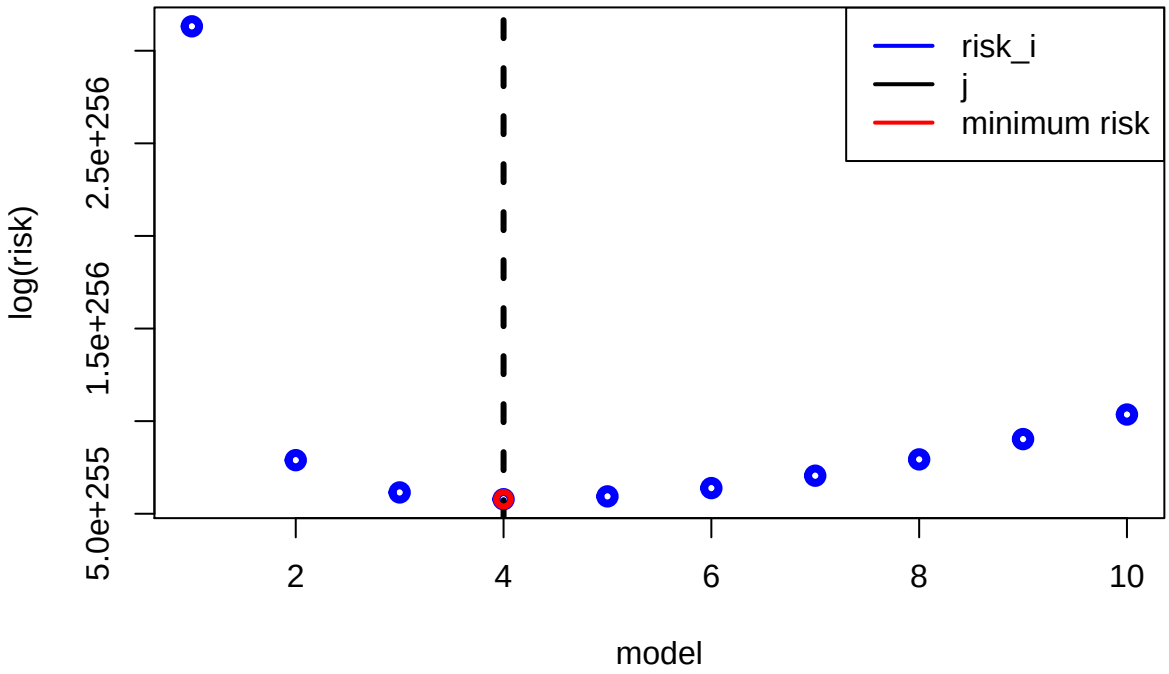


Now we can compute risk r_i for each of the distributions:

```
C = matrix(NA,n,k)
k_ = rep(NA,length.out = length(theta))
k_ = sapply(1:length(theta), function(i) sum(y==i)) #number of Y_i = i
for (i in 1:k){
  C[,i] = sapply(1:length(theta), function(m) sum(X[,i] == m))
}
alpha = matrix(NA,length(theta),k)
for (l in 1:k){
  alpha[,l] = dpois(theta,lambda[l])*M
  alpha[,l]+C[,l] #alpha'
}

poli = function(x,n){
  p = x
  for (i in 1:(n-1)){
    p = prod(p,x+i)
  }
  if (n ==0)(p = 1)
  return(p)
}
probs = matrix(NA,length(theta),length(lambda)) #risk
for (i in theta){
  for (l in 1:k){
    probs[i,l] = log(poli(alpha[i,l],k_[i])/poli(M,length(theta)))
  }
```

```
}
probs = sapply(1:length(lambda),function(i)prod(probs[,i]))
plot(probs,ylab = "log(risk)", xlab = "model",col = "blue", lwd = 4)
abline(v = j, lwd=3, lty=2)
points(x = which.min(probs),y = probs[which.min(probs)], col = "red", lwd = 3)
legend("topright",
       c("risk_i","j","minimum risk"), col = c("blue", "black","red"),
       lwd = 2)
```



As we see the classification problem was solved correctly.