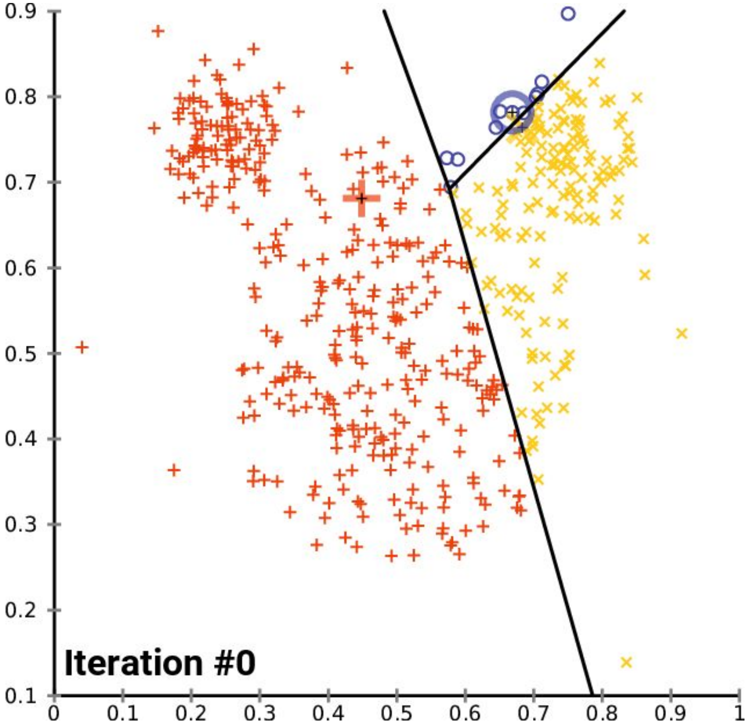
DPGMM Gibbs Sampler

Dirichlet Process Gaussian Mixture Models

Li Sun, Yimeng Xu, Jiahao Xu

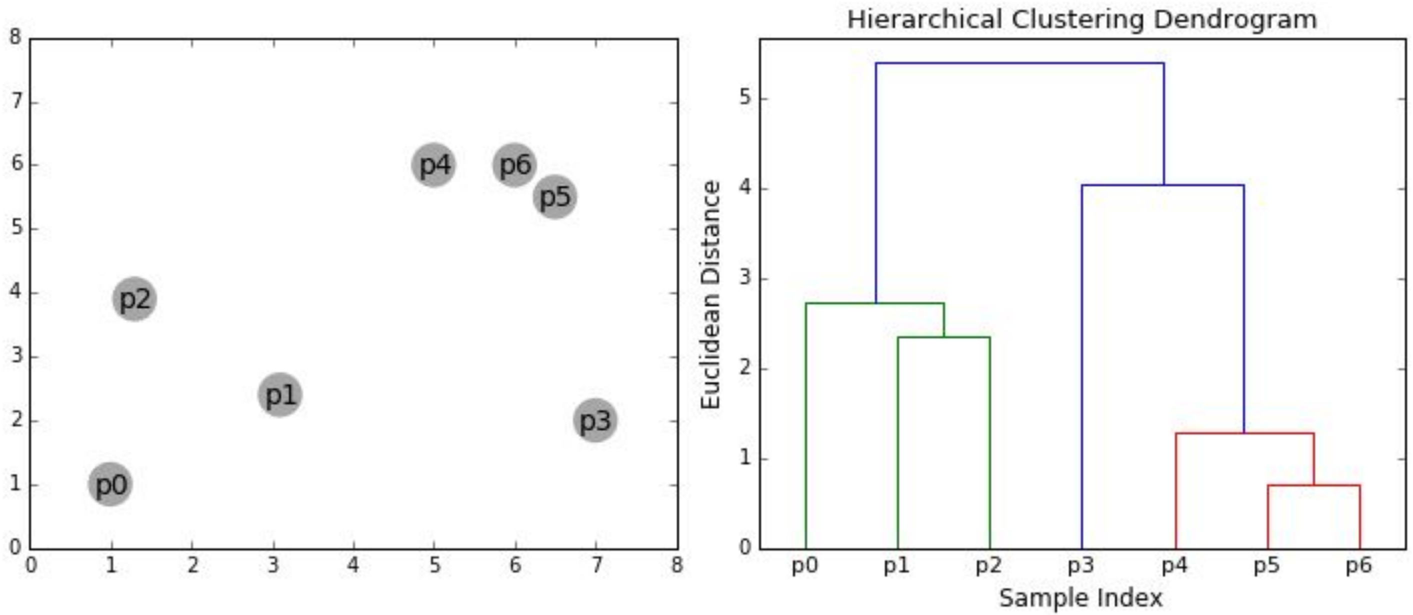
MATH 640: Bayesian Statistics

Background -- kmeans



https://en.wikipedia.org/wiki/K-means_clustering#/media/File:K-means_convergence.gif

Background -- Hierarchical Clustering



https://dashee87.github.io/data%20science/general/

DPGMM

$$X_i|\mu_i \sim \mathcal{N}(\mu_i, \sigma^2)$$
$$\mu_i|G \sim G$$
$$G = DP(\alpha, \mathcal{N}(\mu_0, \sigma_0^2))$$

Same μ → Same Cluster

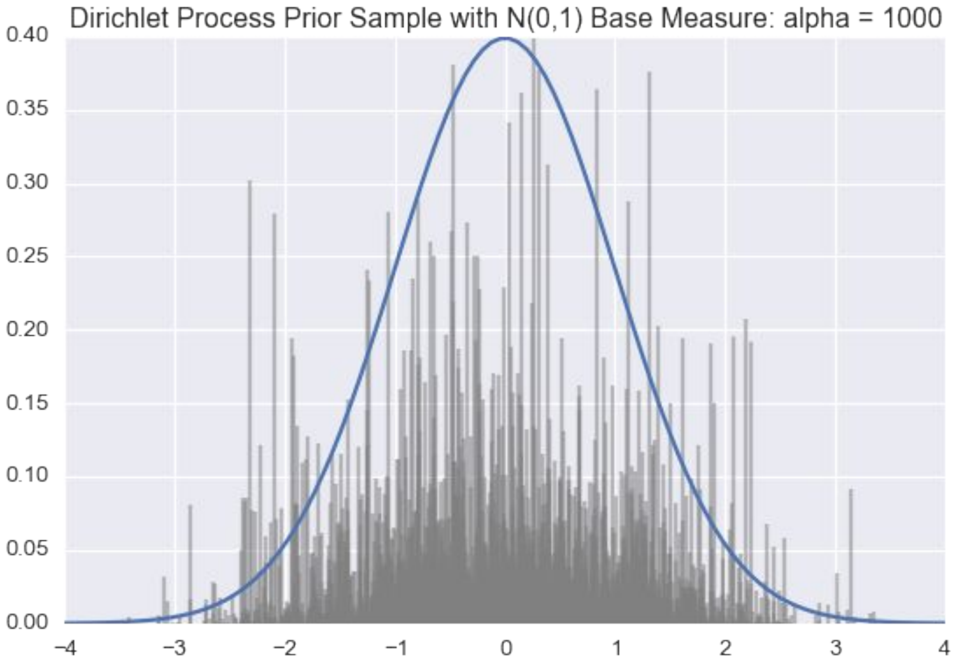
$\mu \sim F$ If F is continuous

$$Pr(\mu_i = \mu_j) = 0$$

We need a **discrete approximation**!

Dirichlet Process (DP)

$$G \sim DP(\alpha, H)$$
$$[G(A_1), \ldots, G(A_K)] \sim Dirichlet(\alpha H(A_1), \ldots, \alpha H(A_K))$$



DPGMM

Notation:

Cluster assignment

$$Z_i = k \Leftrightarrow \mu_i \ in\ kth\ partition$$

$$p_k = P(Z = k)$$

$$p = \{p_1, \ldots, p_K\}$$

$$n_{k,-i} = count\ of \mu_j (j \neq i)\ in\ kth\ partition$$

$$-i = \{1, \ldots, i-1, i+1, \ldots, N\}$$

$$X_i | \mu \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$Z_i | p \sim Discrete(p_1, \ldots, p_K)$$

$$\mu_i \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

$$p \sim Dir(\alpha/K, \ldots, \alpha/K)$$

$$K \to \infty$$

Predictive Distribution

$$P(Z_i = k) = p_k$$

$$L(Z) = \prod_{i=1}^{N} P(Z_i) = \prod_{k=1}^{K} p_k^{n_k}$$

$$p_1, \ldots, p_K \sim DIR(\tfrac{\alpha}{K}, \ldots, \tfrac{\alpha}{K})$$

$$P(Z_i = m | Z_{-i}) = \frac{P(Z_i = m, Z_{-i})}{P(Z_{-i})}$$

$$= \frac{\int_{p_1,\ldots,p_K} P(Z_i = m, Z_{-i} | p_1, \ldots, p_K) P(p_1, \ldots, p_K) dp_1 \ldots dp_K}{\int_{p_1,\ldots,p_K} P(Z_{-i} | p_1, \ldots, p_K) P(p_1, \ldots, p_K) dp_1 \ldots dp_K}$$

Predictive Distribution

$$
\begin{aligned}
denominator &= \int_{p_1,\dots,p_K} P(Z_{-i}|p_1,\dots,p_K)P(p_1,\dots,p_K)dp_1\dots dp_K \\
&= \int_{p_1,\dots,p_K} [\prod_{k=1}^{K} p_k^{n_{k,-i}}] \frac{\Gamma(\sum_{k=1}^{K}\frac{\alpha}{K})}{\prod_{k=1}^{K}\Gamma(\frac{\alpha}{K})} \prod_{k=1}^{K} p_k^{\frac{\alpha}{K}-1} dp_1\dots dp_K \\
&= \frac{\Gamma(\sum_{k=1}^{K}\frac{\alpha}{K})}{\prod_{k=1}^{K}\Gamma(\frac{\alpha}{K})} \int_{p_1,\dots,p_K} \prod_{k=1}^{K} p_k^{\frac{\alpha}{K}+n_{k,-i}-1} dp_1\dots dp_K \\
&= \frac{\Gamma(\sum_{k=1}^{K}\frac{\alpha}{K})}{\prod_{k=1}^{K}\Gamma(\frac{\alpha}{K})} \frac{\prod_{k=1}^{K}\Gamma(\frac{\alpha}{K}+n_{k,-i})}{\Gamma(\sum_{k=1}^{K}(\frac{\alpha}{K}+n_{k,-i}))} \int_{p_1,\dots,p_K} \frac{\Gamma(\sum_{k=1}^{K}(\frac{\alpha}{K}+n_{k,-i}))}{\prod_{k=1}^{K}\Gamma(\frac{\alpha}{K}+n_{k,-i})} \prod_{k=1}^{K} p_k^{\frac{\alpha}{K}+n_{k,-i}-1} dp_1\dots dp_K
\end{aligned}
$$

$$
numerator = \frac{\Gamma(\sum_{k=1}^{K}\frac{\alpha}{K})}{\prod_{k=1}^{K}\Gamma(\frac{\alpha}{K})} \frac{\Gamma(\frac{\alpha}{K}+n_{m,-i}+1)\cdot\prod_{k=1}^{K}\Gamma(\frac{\alpha}{K}+n_{k,-i})}{\Gamma(\frac{\alpha}{K}+n_{m,-i}+1+\sum_{k=1,k\neq m}^{K}(\frac{\alpha}{K}+n_{k,-i}))}
$$

$$
denominator = \frac{\Gamma(\sum_{k=1}^{K}\frac{\alpha}{K})}{\prod_{k=1}^{K}\Gamma(\frac{\alpha}{K})} \frac{\prod_{k=1}^{K}\Gamma(\frac{\alpha}{K}+n_{k,-i})}{\Gamma(\sum_{k=1}^{K}(\frac{\alpha}{K}+n_{k,-i}))}
$$

$$
P(Z_i = m|Z_{-i}) = \frac{\frac{\alpha}{K}+n_{m,-i}}{\alpha+N-1}
$$

Chinese Restaurant Process (CRP)

$$P(Z_i = m | Z_{-i}) = \frac{n_{m,-i}}{\alpha + N - 1} \quad \forall i \in [1, K]$$

$$P(Z_i = K+1 | Z_{-i}) = \frac{\alpha}{\alpha + N - 1}$$

https://commons.wikimedia.org/wiki/File:Chinese_Restaurant_Process_for_DP(0.5,H).gif

Predictive Probability

in cluster m

Likelihood if in cluster m

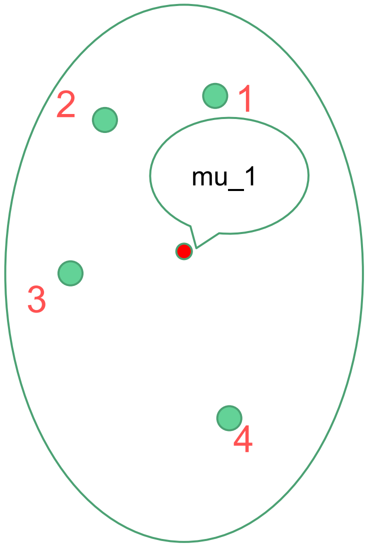
$$P(Z_i = m | Z_{-i}, X) \propto P(Z_i = m | Z_{-i}, \alpha) \cdot P(X_i | Z_i = m, \mu_i)$$

$$\propto \begin{cases} n_{m,-i} \cdot \mathcal{N}(x_i; \mu_{[m]}, \sigma^2) & \textit{existing cluster } m \\ \alpha \int_\mu Pr(X_i|\mu) \cdot Pr(\mu|\mu_0) = \mathcal{N}(x_i; \mu_0, \sigma^2 + \sigma_0^2) & \textit{new cluster} \end{cases}$$

Gibbs Sampler

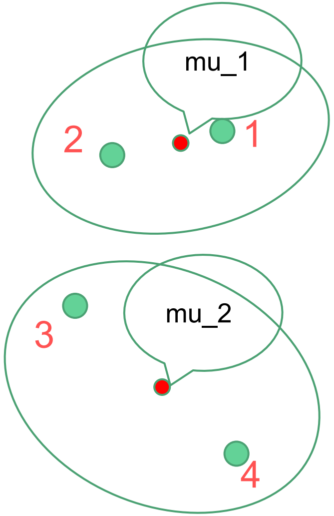
Data	Initial Z0	1st run Z1	p.m.f	2nd run Z2	p.m.f	3rd run Z3	p.m.f
x1	Cluster 1	Cluster 1 New cluster	$\propto N(x_1;\mu_1,\sigma^2)$ $\propto \alpha N(x_1;\mu_0,\sigma^2+\sigma_0^2)$	Cluster 1 Cluster 2 New cluster	$\propto N(x_1;\mu_1,\sigma^2)$ $\propto N(x_2;\mu_2,\sigma^2)$	Cluster 1 Cluster 3 New cluster	$\propto N(x_1;\mu_1,\sigma^2)$ $\propto N(x_3;\mu_3,\sigma^2)$
x2	Cluster 1	…		…			
x3	Cluster 1	…		…			
Realization	Z01=1 Z02=1 Z03=1	Z11=1 u1 Z12=2 u2 Z13=2 u2		Z21=1 u1 Z22=1 u1 Z23=3 u3		Z31=1 u1 Z32=2 u2 Z33=1 u1	
Likelihood	L(x1,x2,3|u1)	L(x1|u1) L(x2,x3|u2)		L(x1,x2|u1) L(x3|u3)		L(x1,x3|u1) L(x2|u2)	
Prior	$u \sim N(\mu_0,\sigma_0^2)$						
Sample from posterior	u1	u1,u2		u1, u3		u1, u2	

Gibbs Sampler -- Run 0



Initialized as one cluster and the first cluster center is sampled from posterior distribution.

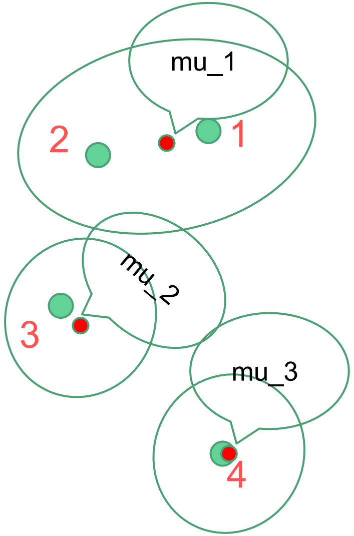
Gibbs Sampler -- Run 1



Point 3 and point 4 are assigned to a new cluster with center mu_2,

Resample mu_1 from the posterior distribution based only on point 1 and point 2. Sample mu_2 from the posterior distribution based on the point 3 and point 4

Gibbs Sampler -- Run 2

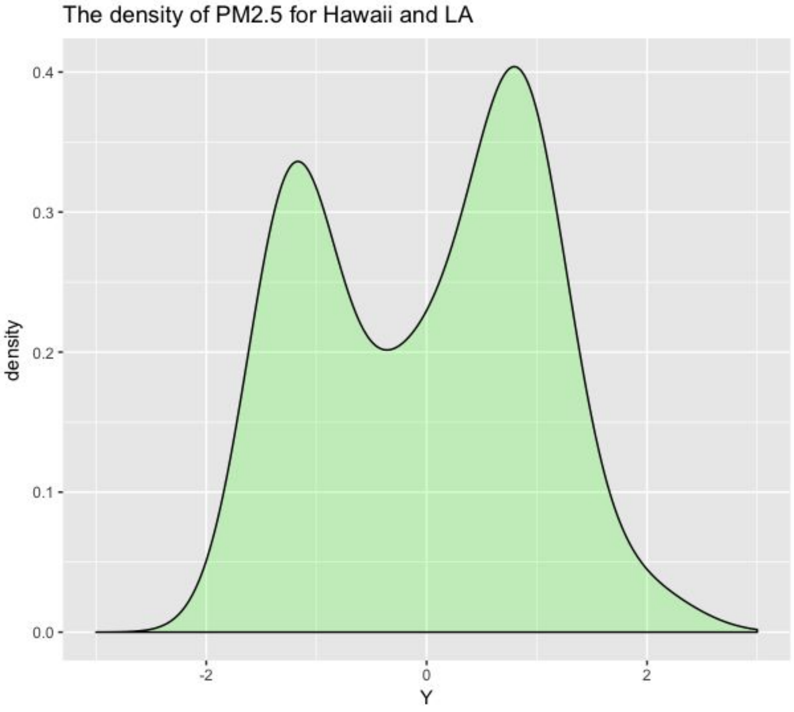


Same process

Point 4 is assigned to a new cluster with center mu_3

Sample mu_1 from the posterior based on point 1 and point 2. Sample mu_2 from the posterior based on point 3. Sample mu_3 from the posterior based on point 4

Data Overview



Summary after normalization:

Min. :-1.5101

1st Qu.:-1.0983

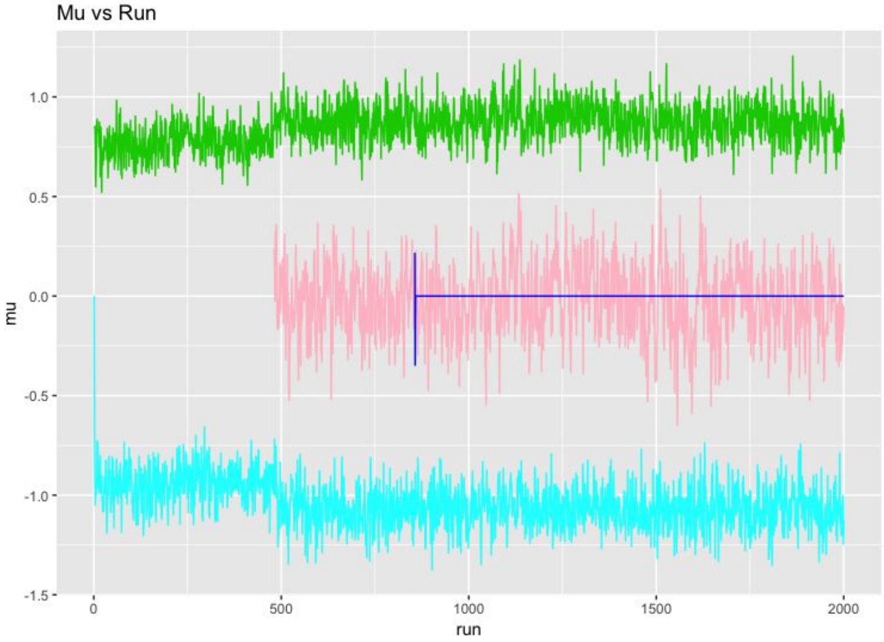
Median : 0.2021

Mean : 0.0000

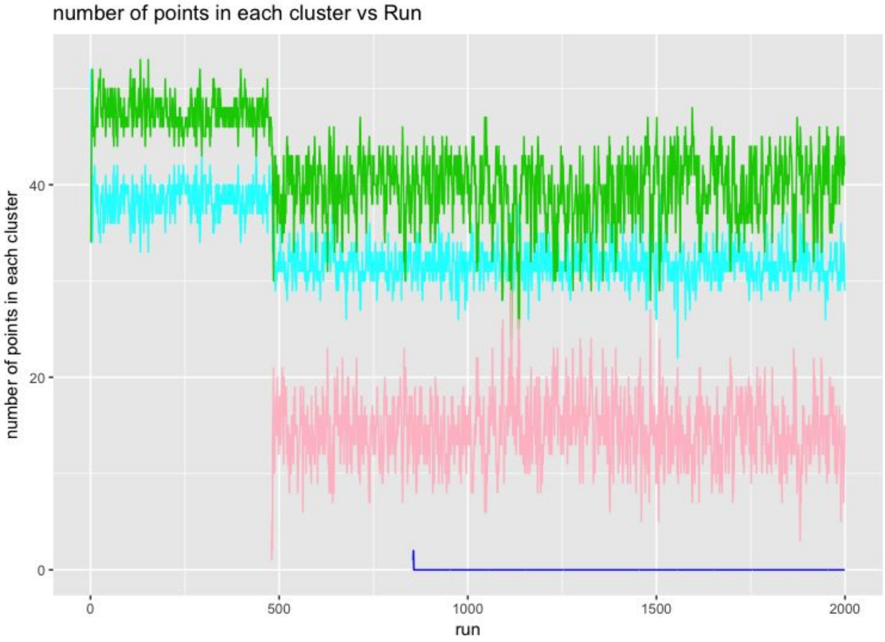
3rd Qu.: 0.8632

Max. : 2.2612

Result

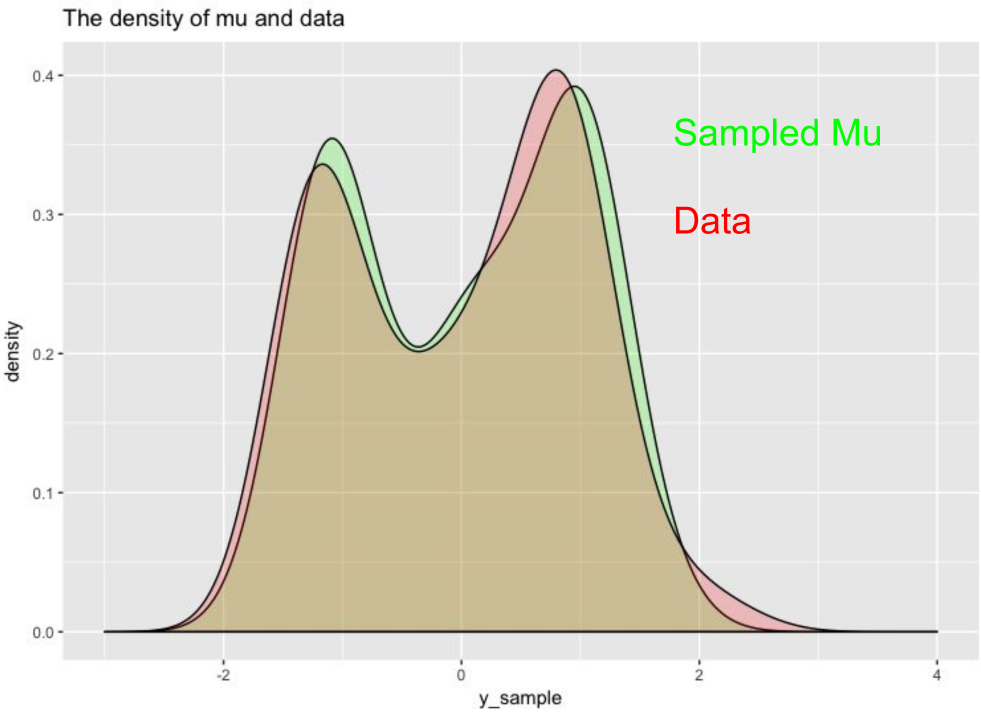


Mu vs Run
mu
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
500
1000
1500
2000
run

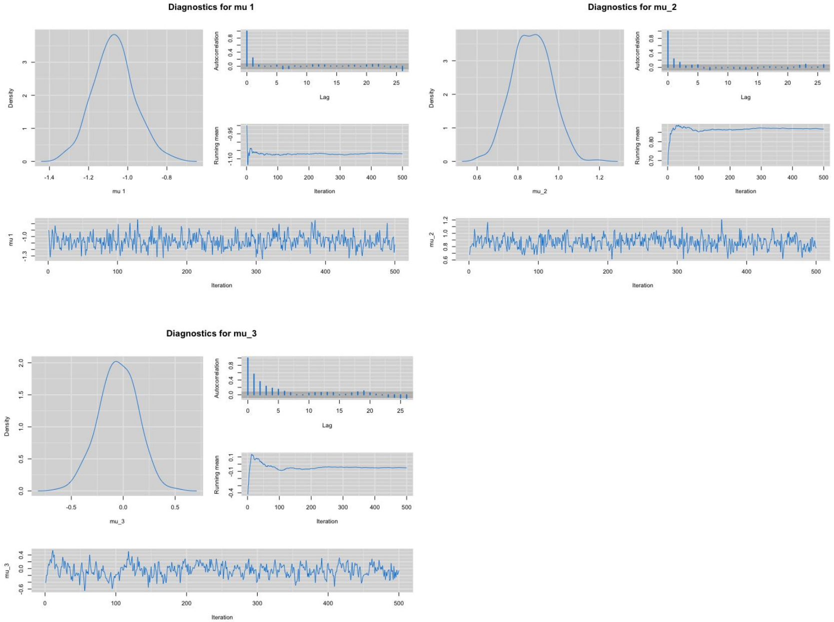


number of points in each cluster vs Run
number of points in each cluster
40
20
0
0
500
1000
1500
2000
run

Result



The density of mu and data
density
y_sample
Sampled Mu
Data



Diagnostics for mu 1
Diagnostics for mu_2
Diagnostics for mu_3

Discussion

Pros:

- Do not need to pre specify the number of clusters
- Can generate a distribution of a centroid in each cluster
- Ease to explain

Cons:

- Need more computing power than other methods
- Need to turn many parameters

Future Work:
Expand on 2-D data and multivariate normal prior.

Also sample variance in the Gibbs sampler.

Try different prior for different data.

Thank you!