Weekly Report I

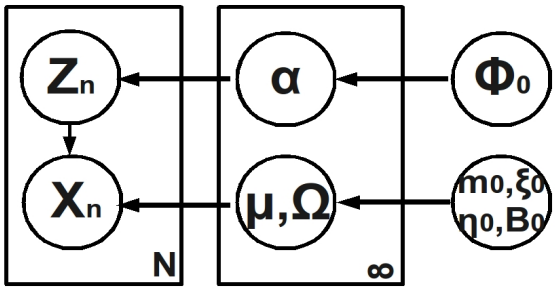
Donglai Wei

2010.4.5

1. Comparison Among EE,collapsed EE and ME for DPmixture

1.1 Settings

Graphical Model for DP mixture



0) Notation:

x_n:observation;
z_n:assignment;
θ:parameter (Hidden Variables besides z_n : α, μ, Ω);
hyper: hyper-parameters (Prior:$\phi_0, m_0, B_0, \eta_0, \xi_0$)

1) Conditional Probability:

EE(stick-breaking):
$p(x_n, z_n, \theta|hyper) = \mathcal{M}(z_n|\alpha')\mathcal{B}(1, \phi_0)\mathcal{N}(x_n|z_n, \mu, \Omega)\mathcal{N}(\mu|m_0, \xi_0\Omega)\mathcal{W}(\Omega|\eta_0, B_0)$

ME(limit of finite mixture):
$p(x_n, z_n, \theta|hyper) = \mathcal{M}(z_n|\alpha)\mathcal{D}(\alpha|\phi_0)\mathcal{N}(x_n|z_n, \mu, \Omega)\mathcal{N}(\mu|m_0, \xi_0\Omega)\mathcal{W}(\Omega|\eta_0, B_0)$

2) Free Energy:

$\mathcal{F}([z], K_+) = \sum_{c=1}^{K_+}[\frac{DN_c}{2}log\pi + \frac{D}{2}log\frac{\xi_c}{\xi_0}logdet(B_c) - \frac{\eta_0}{2}logdet(B_0) - log\frac{\Gamma_D(\frac{\eta_c}{2})}{\Gamma_D(\frac{\eta_0}{2})}$
$+\frac{1}{K_+}log\frac{\Gamma(N+\phi_0)}{\Gamma(\phi_0)} - log(\Gamma(N_c) - log\phi_0]$

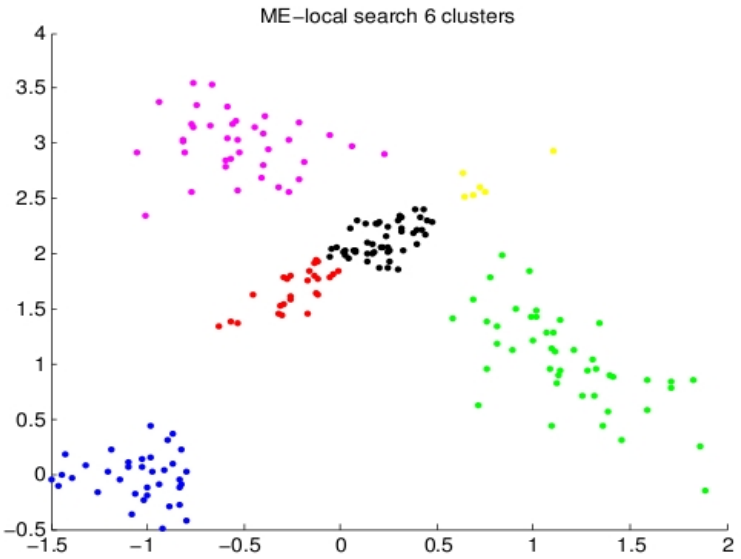
Table 1: Prior 1 (with mean and std)

Learning Algorithm	Initial Cluster Numbers	Number of clusters	RandIndex
EE(sampled initialization)	10	4.1(0.3)	0.99(0.00)
	20	4.1(0.3)	0.99(0.00)
EE(random initialization)	10	4.4(0.48)	0.97(0.01)
	20	4.5(0.83)	0.96(0.01)
Collapsed EE	10	5.0(1.0)	0.97(0.22)
	20	7.1(1.7)	0.85(0.07)
ME	1(Top-down)	4	1
	200(Bottom-up)	4	1
	200(Local search,1-200 update)	6	0.91
	200(..+Merge)	4	1
	200(Local search,random update)	5.9(1.37)	0.94(0.05)
	200(..+Merge)	4(0)	1(0)

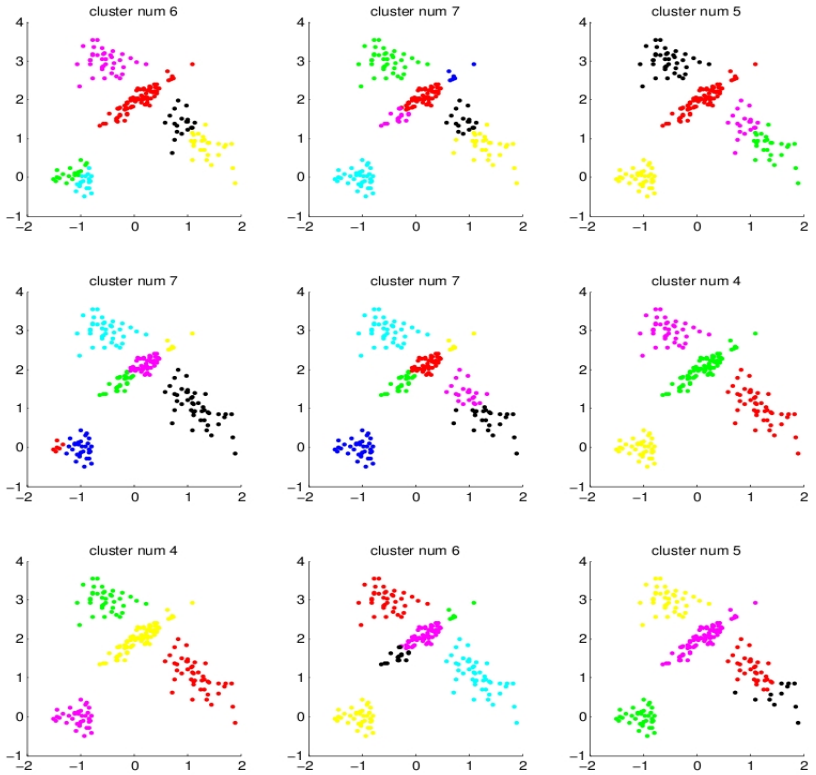
1.2 Prior 1 (default ME(top-down))

ϕ_0=2
$m_0 = \bar{x}$=(0.0217,1.6134)
B_0=b0r*D*S=[0.1505,0.0113;0.0113,0.2518]
η_0=2
ξ_0=0.1
(b0r=0.1151(random),D=2,S=[0.6536,0.0493;0.0493,1.0936](cov matrix))
(Algorithms are implemented by Kurihara,free energy threshold for EE,c-EE is 1e-10)

Prior 1: ME,1-200 update



Prior 1: ME,random update



Prior 1: c-EE,K=20

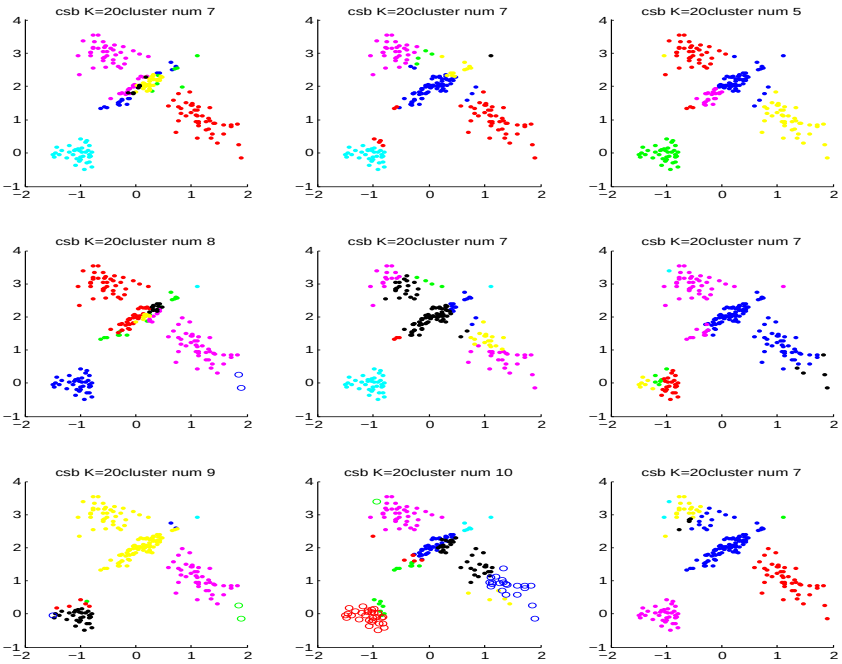


Table 2: Prior 2 (with mean and std)

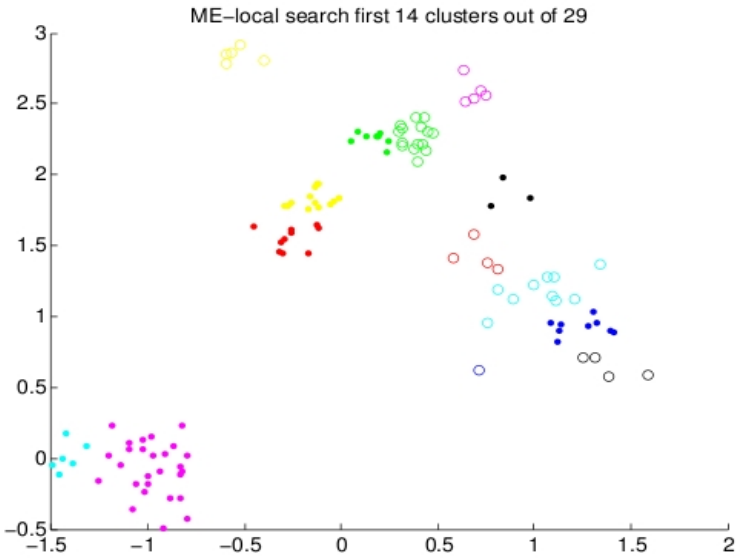
Learning Algorithm	Initial Cluster Numbers	Number of clusters	RandIndex
EE(sampled initialization)	10 20	4.1(0.3) 4.1(0.3)	0.99(0.00) 0.99(0.00)
EE(random initialization)	10 20	6.9(1.2) 7.6(1.8)	0.97(0.01) 0.96(0.03)
Collapsed EE	10 20	5.6(1.6) 8.1(1.6)	0.90(0.22) 0.87(0.08)
ME	1(Top-down) 200(Bottom-up) 200(Local search,1-200 update) 200(..+Merge) 200(Local search,random update) 200(..+Merge)	4 4 29 4 29.7(1.49) 4(0)	1 1 0.77 1 0.77(0.01) 1(0)

1.3 Prior 2 (default EE)

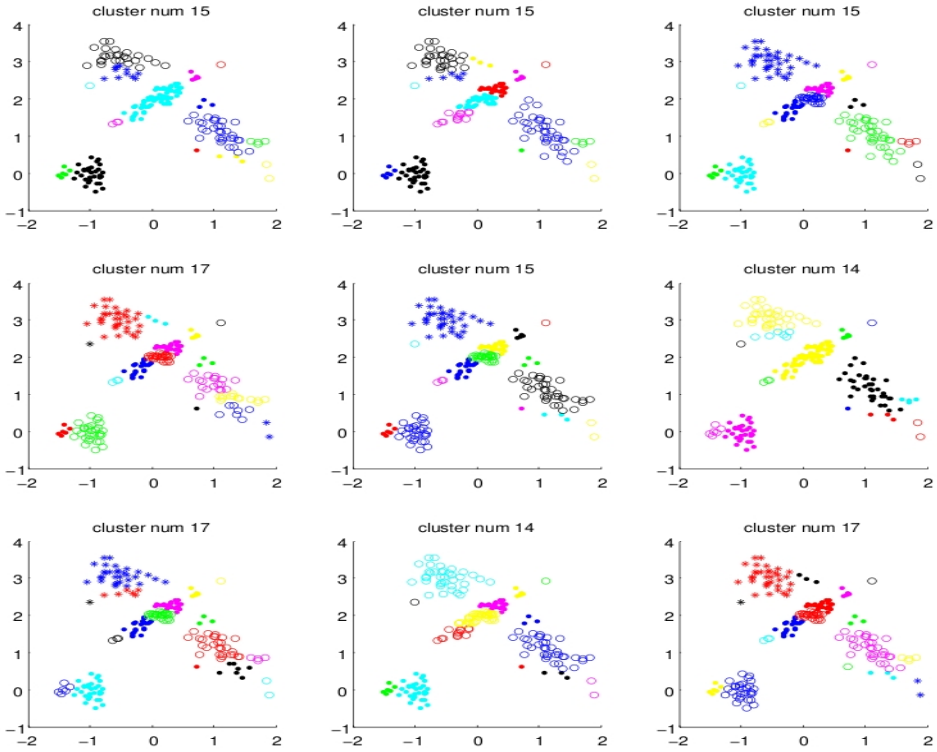
Differences:
ϕ_0=1
η_0=3
ξ_0=0.01
$B_0 = \eta_0\xi_0$max(eig(S))*D*eye(D)=[0.033,0;0,0.033]

The change of B_0, ξ_0 that control the shape of the cluster matter significantly.

Prior 2: ME,1-200 update, first 14 clusters



Prior 2: ME,random update



Prior 2: EE,sis=0,K=20

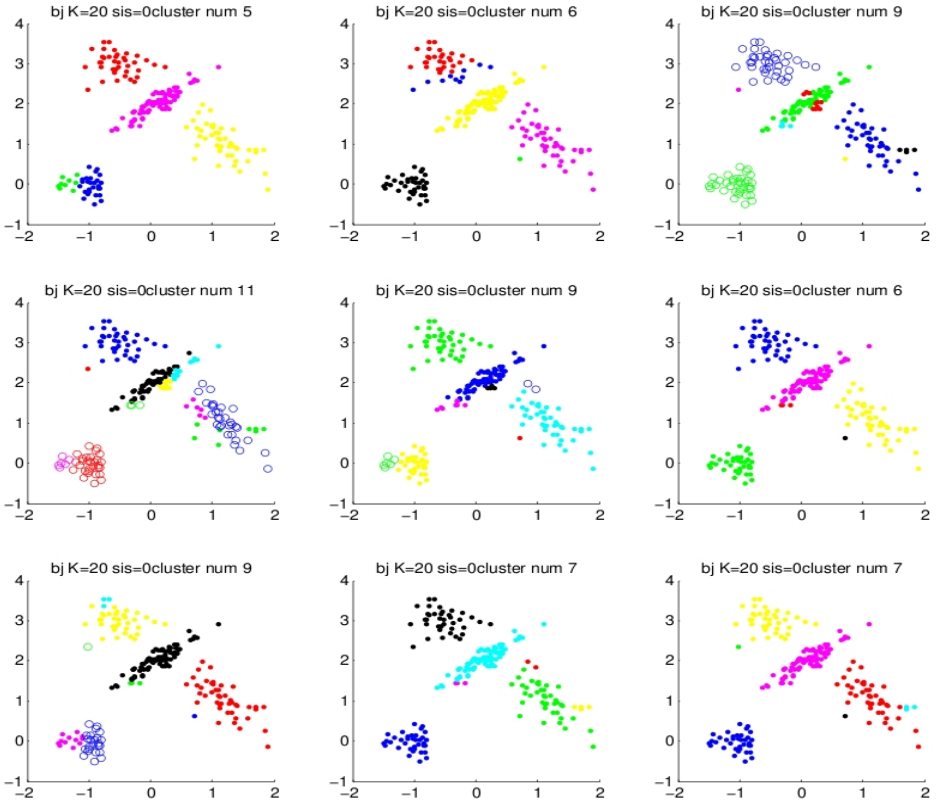


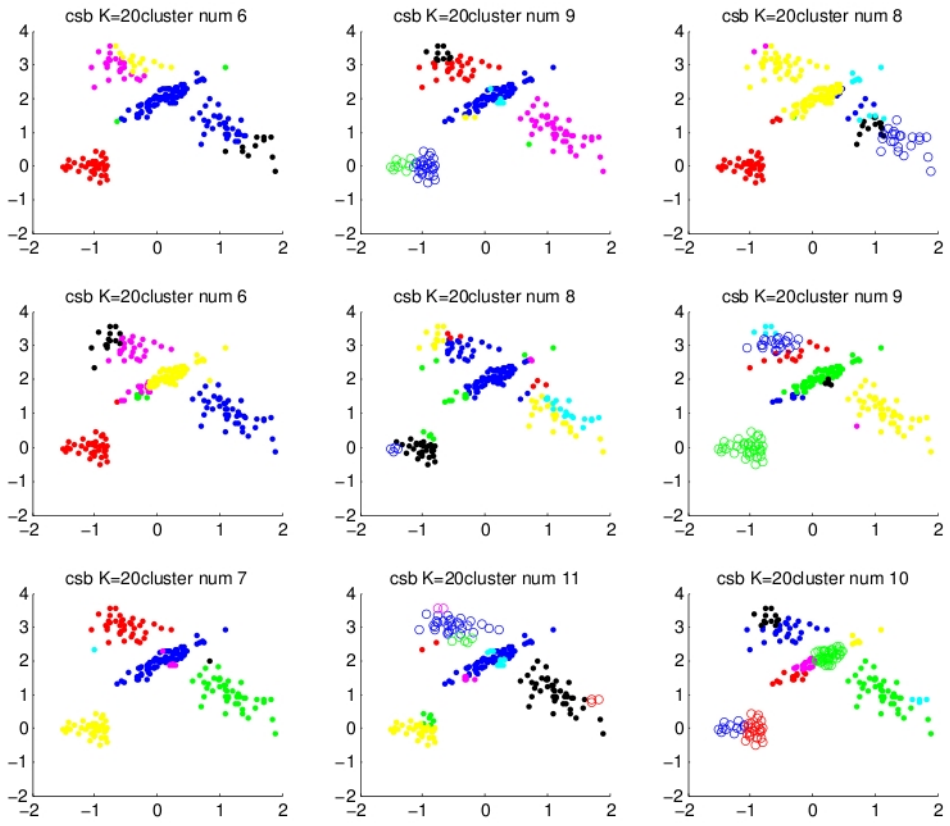
Table 3: Change from Prior 1 to 2

parameter	♯ clusters	RandIndex
ξ_0	8	0.88
η_0	10	0.87
ϕ_0	6	0.91
B_0	17	0.81

Table 4: Change from Prior 2 to 1

parameter	♯ clusters	RandIndex
ξ_0	18	0.80
η_0	25	0.79
ϕ_0	29	0.77
B_0	10	0.87

Prior 2: c-EE,K=20

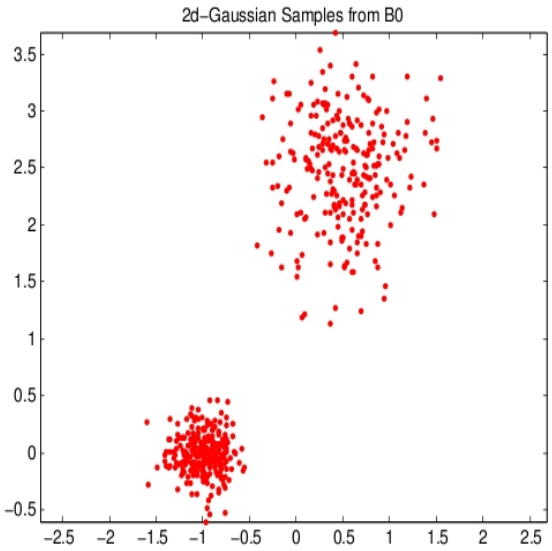


1.3 So, what's the matter

0) Empirical:(table 3 and 4)
i) Start from Prior 1, we change the parameter to Prior 2 one at a time:(update 1-200)
ii) Start from Prior 2, we change the different parameter back to Prior 1 one at a time:(update 1-200)

1) Theoretical:
Different Prior B_0, ξ_0 prefers different shape of the cluster. Shown in the figure below, B_0

from Prior 1(right) is more sparse and elliptical than that from Prior 2(left)
Easily imagined,small ξ_0 will generate small clusters since μ_c is evaluated by $\mathcal{N}(\mu|m_0, \xi_0\Omega)$



2 CRF for HDP

2.1 CRP for DPmixture

One step back to derive the Allocation term $p(z_1, ..., z_N|\phi_0)$ in the free energy with CRP.
Recall that **Free Energy:**
$\mathcal{F}([z], K_+) = \sum_{c=1}^{K_+}[\frac{DN_c}{2}log\pi + \frac{D}{2}log\frac{\xi_c}{\xi_0}logdet(B_c) - \frac{\eta_0}{2}logdet(B_0) - log\frac{\Gamma_D(\frac{\eta_c}{2})}{\Gamma_D(\frac{\eta_0}{2})}$
$+\frac{1}{K_+}log\frac{\Gamma(N+\phi_0)}{\Gamma(\phi_0)} - log(\Gamma(N_c)) - log\phi_0]$

Known: N datas,K clusters, N_i datas in each cluster,ϕ_0 for new cluster

Unknown: Cluster Assignment z_n

Allocation Term:
$p(z_1, ..., z_N|\phi_0) = \Pi_{j=1}^{j=N} p(z_j|z_1, ..., z_{j-1}, \phi_0)$

$p(z_j|z_1, ..., z_{j-1}, \phi_0) = \sum_{k=1}^{K(j)} \frac{m_{k(j)}}{j-1+\phi_0}\delta_{z_j=k} + \frac{\phi_0}{j-1+\phi_0}\delta_{z_j=K(j)+1}$

1) Partition:$\Pi_{j=1}^{j=N} \frac{1}{\phi_0+j-1} = \frac{\Gamma(\phi_0)}{\Gamma(N+\phi_0)}$

2) Forming new clusters:ϕ_0^K

3) Accumulating for all clusters:$\Pi_{j=1}^{j=N}(N_j-1)! = \Pi_{j=1}^{j=N}\Gamma(N_j)$

So,$-log(p(z_1,...,z_N|\phi_0)) = \sum_{c=1}^{K} \underline{+\frac{1}{K}log\frac{\Gamma(N+\phi_0)}{\Gamma(\phi_0)} - log(\Gamma(N_c)) - log\phi_0}$

2.2 CRF for HDP

Here, for the Gaussian case:
$\lambda = (m_0, B_0, \eta_0, \xi_0)$
$\theta_k = (\mu_k, \Omega_k)$

Known:
1) Global: J Restaurants, K global dishes, T tables, N datas,
t_{ji} : the table that customer i in Restaurant j sits;
k_{jt} : the dish that table t in Restaurant j serves;
$\vec{k}$: new dish different from what has been served;
$\vec{t}$: the table different from what has been occupied;
2) Counting so far (during the process)
m_{jk} : number of tables in Restaurant j serving dish k;
$m_{\cdot k}$: number of tables serving dish k;
$m_{j\cdot}$: number of tables in Restaurant j;
$m_{\cdot\cdot}$: number of tables;
n_{jtk} : number of customers in Restaurant j at table t eating dish k;
$n_{jt\cdot}$: number of customers in Restaurant j at table t;
$n_{j\cdot\cdot}$: number of customers in Restaurant j;

Unknown: Table Assignment t_{ji},Dish Assignment k_{jt}
Allocation Term:

$$p(k_{11},...,k_{JT_J}|\gamma) = \Pi_{j=1}^{J}(\Pi_{i=1}^{T_j}p(k_{ji}|\vec{k}_1,...\vec{k}_{j-1},k_{j1},...,k_{j(i-1)},\gamma))$$

$$p(t_{11},...,t_{JN_J}|\alpha) = \Pi_{j=1}^{J}(\Pi_{i=1}^{N_j}p(t_{ji}|t_{j1},...,t_{j(i-1)},\alpha))$$

$$p(k_{jt}|\vec{k}_1,...\vec{k}_{j-1},k_{j1},...,k_{j(t-1)},\gamma) = \sum_{k=1}^{K}\frac{m_{\cdot k}}{m_{\cdot\cdot}-1+\gamma}\delta_{k_{jt}=k} + \frac{\gamma}{m_{\cdot\cdot}-1+\gamma}\delta_{k_{ji}=\vec{k}}$$

$$p(t_{ji}|t_{j1},...,t_{j(i-1)},\alpha) = \sum_{t=1}^{m_{j\cdot}}\frac{n_{jt\cdot}}{i-1+\alpha}\delta_{t_{ji}=t} + \frac{\alpha}{i-1+\alpha}\delta_{t_{ji}=\vec{t}}$$

1) Partition:
k: $\Pi_{w=1}^{\sum_{j=1}^{J}m_{j\cdot}}\frac{1}{\gamma+w-1} = \Pi_{w=1}^{m_{\cdot\cdot}}\frac{1}{\gamma+w-1} = \frac{\Gamma(\gamma)}{\Gamma(T+\gamma)}$

t: $\Pi_{j=1}^{J}\Pi_{i=1}^{i=n_{j..}}\frac{1}{\alpha+i-1}=\Pi_{j=1}^{j=J}\frac{\Gamma(\alpha)}{\Gamma(n_{j..}+\alpha)}$

2) Forming new clusters (1st point in the cluster):

k: γ^K

t: $\Pi_{j=1}^{J}\alpha^{m_{j.}}$

3) Accumulating for each clusters (other points in the cluster):

k: $\Pi_{k=1}^{k=K}(m_{.k}-1)!=\Pi_{k=1}^{k=K}\Gamma(m_{.k})$

t: $\Pi_{j=1}^{j=J}\Pi_{t=1}^{t=m_{j.}}(n_{jt.}-1)!=\Pi_{j=1}^{j=J}\Pi_{t=1}^{t=m_{j.}}\Gamma(n_{jt.})$

So, **Free Energy:**

$\mathcal{F}([z])$

$=$

(Likelihood)$\sum_{k=1}^{K}[\frac{Dn_{..k}}{2}log\pi+\frac{D}{2}log\frac{\xi_k}{\xi_0}+\frac{\eta_k}{2}logdet(B_k)-\frac{\eta_0}{2}logdet(B_0)-log\frac{\Gamma_D(\frac{\eta_k}{2})}{\Gamma_D(\frac{\eta_0}{2})}]$

$+$

(Allocation:)$\underline{\sum_{j=1}^{J}\sum_{t=1}^{m_{j.}}[\frac{1}{m_{j.}}log\frac{\Gamma(n_{j..}+\alpha)}{\Gamma(\alpha)}-log(\Gamma(n_{jt.})-log\alpha]+\sum_{k=1}^{K}[\frac{1}{K}log\frac{\Gamma(T+\gamma)}{\Gamma(\gamma)}-log(\Gamma(n_{..k})-log\gamma]}$

$=$

(t-term)$\sum_{j=1}^{J}\sum_{t=1}^{m_{j.}}[\underline{\frac{1}{m_{j.}}log\frac{\Gamma(n_{j..}+\alpha)}{\Gamma(\alpha)}-log(\Gamma(n_{jt.})-log\alpha+\frac{1}{Jm_{j.}}log\frac{\Gamma(T+\gamma)}{\Gamma(\gamma)}]}$

$+$(k-term)$\sum_{k=1}^{K}[\frac{n_{..k}D}{2}log\pi+\frac{D}{2}log\frac{\xi_k}{\xi_0}+\frac{\eta_k}{2}logdet(B_k)-\frac{\eta_0}{2}logdet(B_0)-log\frac{\Gamma_D(\frac{\eta_k}{2})}{\Gamma_D(\frac{\eta_0}{2})}-\underline{log(\Gamma(n_{..k})-log\gamma]}$

where:

$\xi_k=\xi_0+n_{..k}$

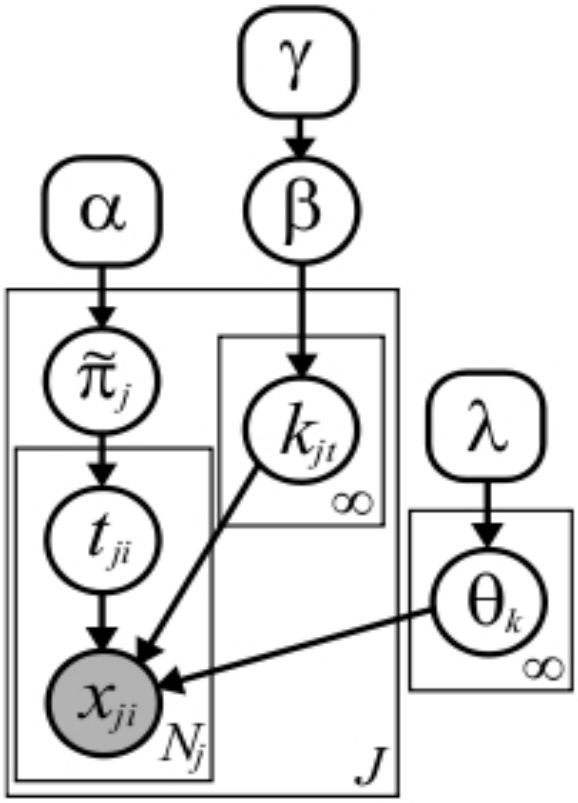
$m_k=\frac{n_{..k}\vec{x}_{(k_{jt_{ji}}=k)}+\xi_0 m_0}{\xi_k}$

$\eta_k=\eta_0=n_{..k}$

$B_k=B_0+n_{..k}S_k+\frac{n_{..k}\xi_0}{\xi k}(\vec{x}_{(k_{jt_{ji}}=k)}-m_0)(\vec{x}_{(k_{jt_{ji}}=k)}-m_0)^T$

$S_k: samplecovariance$

Graphical Model for HDP



3. Search?

The Key for ME algorithm is to search the optimanl assignment to minimize free energy.

3.1 Local Search

It's kind of a black box since we put no prior on which datas could be in the same cluster. We just randomly throw one (or more) nodes into the black box and find the best assignment for it conditioning on others.

3.2 Search with Prior knowledge

0) One step aside, considering the case of Linear Programming. Given a set of linear constraints of $x_1, ...x_n$, we want to find their optimal "assignment" to maximize a linear object function. Without prior knowledge, the search is hopeless in R^n. But when we know that the constraints actually form a convex simplex with the optimal solution on some of the vertexes, the search becomes promising.

1) In our case, for mixture of Gaussian or image(later), there is the assumption of "Spatial Smoothness".
i) The Split and Merge technique implemented by Kurihara for ME DPmixture actually makes use of "Spatial Smoothness" from a top-down approach.
Below is the matlab code for the heuristic split:

$[D, N] = size\,(data)\,;$
$[v, l] = eig\,(cov\,(data))\,;$
$[max\_val, max\_index] = max\,(diag\,(l))\,;$
$diff = v\,(:, max\_index) * sqrt\,(l\,(max\_index, max\_index))\,;$
$center = mean\,(data, 2)\,;$
$center1 = center + diff;$
$center2 = center - diff;$
ii) From a bottom-up approach, we can make use K-nearest Neighbour or Density-based Clustering as an auxilary. In each Group j, we can change datas that are "close" to each other as an α expansion.