# Variational Bayes Algorithm

## Lecture Notes

Ahmet Ademoglu, *PhD*
Bogazici University
Institute of Biomedical Engineering

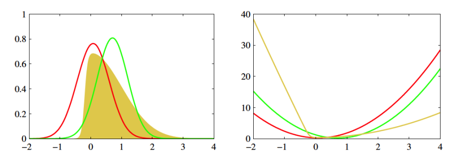
Some concepts and illustrations in this lecture are adapted from the textbooks,
**Pattern Recognition and Machine Learning**, C. M. Bishop, *Springer*, 2006.

We have a probabilistic model for $p(\mathbf{y}, \theta)$ and our objective is to find an approximation for the posterior distribution of hidden variables $\theta$, $p(\theta|\mathbf{y})$ as well as a distribution for the model evidence $p(\mathbf{y})$.

## Laplace approximation

A family of approximation techniques called *Variational Bayes* is a local Gaussian approximation to a mode (*i.e.*, a maximum) of the distribution.

$p(\theta|\mathbf{y})$ $\qquad$ $\log p(\theta|\mathbf{y})$




Yellow: Original, Red: Laplace Approximation, Green: Variational Approximation

## Kullback-Leibler divergence

$D_{KL}(q(\theta)||p(\theta|\mathbf{y},\lambda)) = -\int q(\theta)\log\frac{p(\theta|\mathbf{y},\lambda)}{q(\theta)}d\theta$

## $\mathcal{L}$ functional

$$\mathcal{L}(q,\lambda) = \int q(\theta)\log\frac{p(\mathbf{y},\theta|\lambda)}{q(\theta)}d\theta$$

## Model Evidence $p(\mathbf{y})$ and Free Energy $\mathcal{L}$

Marginal log-likelihood of $p(\mathbf{y})$ can be written as

$$\log p(\mathbf{y}|\lambda) = \mathcal{L}(q,\lambda) + D_{KL}(q(\theta)||p(\theta|\mathbf{y},\lambda))$$

and it is easier to optimize $\log p(\mathbf{y},\theta|\lambda)$ than $\log p(\mathbf{y}|\lambda)$ which can be done using *Expectation-Maximization (EM) Algorithm*.

### $1^{st}$ Stage Optimization: Expectation

$\mathcal{L}(q,\lambda)$ is maximized by fixing $\lambda = \lambda^m$ and choosing $q(\theta) = p(\theta|\mathbf{y},\lambda^m)$. This makes $D_{KL}(q(\theta)||p(\theta|\mathbf{y},\lambda))$ vanish and the lower bound $\mathcal{L}$ equal to $\log p(\mathbf{y}|\theta)$.

### $2^{nd}$ Stage Optimization: Maximization

$q(\theta)$ is fixed and $\mathcal{L}(q,\lambda)$ is maximized for $\lambda$ to obtain $\lambda^{m+1}$.
Since $q(\theta)$ will be no more equal to $p(\theta|y,\lambda^{m+1})$,
$D_{KL}(q(\theta)||p(\theta|\mathbf{y},\lambda^{m+1})) \neq 0$.
The new lower bound will be increased and expressed as

$$\mathcal{L}(q,\lambda) = \int p(\theta|\mathbf{y},\lambda^m)\log p(\theta,\mathbf{y}|\lambda)d\theta - \int p(\theta|\mathbf{y},\lambda^m)\log p(\theta|\mathbf{y},\lambda^m)d\theta$$
$$= Q(\lambda|\lambda^m) + \ Const$$

which is to be maximized for $\lambda$ to find $\lambda^{m+1}$.

E-Step: Compute $p(\theta|\mathbf{y},\lambda^m)$

M-Step: Evaluate $\lambda^{m+1} = \arg\max_{\lambda} Q(\lambda|\lambda^m)$

The *EM algorithm* requires us know the $p(\theta|\mathbf{y}, \lambda)$ or to compute $\int p(\theta|\mathbf{y}, \lambda^m) \log p(\theta, \mathbf{y}|\lambda) d\theta$.
In some cases, this is not possible which makes *EM algorithm* inapplicable.

## Variational Framework

Alternatively, $p(\theta|\mathbf{y}, \lambda)$ can be replaced by an assumed distribution $q$ such that it maximizes $\mathcal{L}(q(\theta), \lambda)$ keeping $\lambda$ fixed.
The lower bound $\mathcal{L}$ now becomes a functional because $q(\theta)$ is a variable now.
This requires to use *variational calculus* or to determine the change of the functional $\mathcal{L}$ with respect to the change in $q(\theta)$.

### Mean Field Approximation from Statistical Physics

For Bayesian inference, $q(\theta)$ over which we make an optimization can be assumed as a function to be factorized as

$$q(\theta) = \prod_{i=1}^{M} q_i(\theta_i)$$

## Ordinary Calculus

$$y(x+\epsilon) = y(x) + \frac{dy(x)}{dx}\epsilon + O(\epsilon^2)$$

For a function of several variables $y(x_1, \ldots, x_D)$

$$y(x_1+\epsilon_1, \ldots, x_D+\epsilon_D) = y(x_1, \ldots, x_D) + \sum_{i=1}^{D} \frac{\partial y}{\partial x_i}\epsilon_i + O(\epsilon^2)$$

## Variational Calculus

Using first order approximation

$$F[y(x)+\epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)}\eta(x)dx + O(\epsilon^2)$$

Around a maximum or a minimum region, $F[y(x)]$ will be very close to $F[y(x)+\epsilon\eta(x)]$ which will imply that $\displaystyle\int \frac{\partial F}{\delta y(x)}\eta(x)dx = 0$.

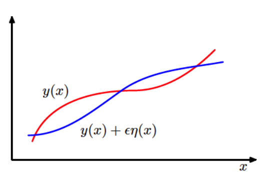
If we assume that $F[y(x)] = \displaystyle\int G\left[y(x), y'(x), x\right] dx$,

$$F[y(x)+\epsilon\eta(x)] = F[y(x)] + \epsilon \int \left\{\frac{\partial G}{\partial y}\eta(x) + \frac{\partial G}{\partial y'}\eta'(x)\right\} dx + O(\epsilon^2)$$

Using integration by parts, $\int \frac{\partial G}{\partial y'}\eta'(x)dx = \frac{\partial G}{\partial y'}\eta(x)|_{\mathcal{B}} - \int \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right)\eta(x)dx$

Since $\eta(x)$ is chosen to be as zero for the end points $\mathcal{B}$ as shown below;

$$F[y(x)+\epsilon\eta(x)] = F[y(x)]$$




$$+\epsilon \int \left\{\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right)\right\}\eta(x)dx + O(\epsilon^2)$$

which yields *Euler-Lagrange* equations;

$$\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) = 0$$

for the functional derivative to vanish.

$$
\begin{aligned}
KL(p||q) &= -\int p(\mathbf{Z}) \left[ \sum_{i=1}^{M} \log q_i(\mathbf{Z}_i) d\mathbf{Z} \right] + const \\
&= -\int \left( p(\mathbf{Z}) \log q_j(\mathbf{Z}_j) + p(\mathbf{Z}) \sum_{i \neq j} \log q_i(\mathbf{Z}_i) \right) d\mathbf{Z} + const \\
&= -\int p(\mathbf{Z}) \log q_j(\mathbf{Z}_j) d\mathbf{Z} + const \\
&= -\int \log q_j(\mathbf{Z}_j) \left[ \int p(\mathbf{Z}) \prod_{i \neq j} d\mathbf{Z}_i \right] d\mathbf{Z}_j + const \\
&= -\int \log q_j(\mathbf{Z}_j) H_j(\mathbf{Z}_j) d\mathbf{Z}_j + const \;\; \text{where } H_j(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} d\mathbf{Z}_i
\end{aligned}
$$

Constrained minimization of

$$-\int \log q_j(\mathbf{Z}_j) H_j(\mathbf{Z}_j) d\mathbf{Z}_j + \lambda \left( \int q_j(\mathbf{Z}_j) d\mathbf{Z}_j - 1 \right)$$

using $G[q_j(\mathbf{Z}_j)] = -\log q_j(\mathbf{Z}_j) H_j(\mathbf{Z}_j) + \lambda q_j(\mathbf{Z}_j)$

Euler-Lagrange equations yields $\frac{\partial G}{\partial q_j} = -\frac{H_j(\mathbf{Z}_j)}{q_j(\mathbf{Z}_j)} + \lambda = 0.$

Since $\lambda = \int H_j(\mathbf{Z}_j) d\mathbf{Z}_j = \int [p(\mathbf{Z}) \prod_{i \neq j} d\mathbf{Z}_i] d\mathbf{Z}_j = 1$

$$q_j(\mathbf{Z}_j) = H_j(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} d\mathbf{Z}_i$$

$$\mathcal{L}(q,\theta) = \int q(\theta) \log \frac{p(\theta, \mathbf{y}|\lambda)}{q(\theta)} d\theta = \int \prod_i q_i \left[ \log p(\theta, \mathbf{y}|\lambda) - \sum_j \log q_j \right] d\theta$$

$$= \int \prod_i q_i \log p(\theta, \mathbf{y}|\lambda) \prod_j d\theta_j - \sum_j \int \prod_i q_i \log q_j d\theta_i$$

$$= \int q_j \left[ \int \log p(\theta, \mathbf{y}|\lambda) \prod_{i \neq j} q_i d\theta_i \right] d\theta_j - \int q_j \log q_j d\theta_j - \sum_{i \neq j} \int q_i \log q_i d\theta_i$$

$$= \int q_j \log \tilde{p}(\theta_j, \mathbf{y}|\lambda) d\theta_j - \int q_j \log q_j d\theta_j - \sum_{i \neq j} \int q_i \log q_i d\theta_i$$

$$= -KL(q_j || \tilde{p}) - \sum_{i \neq j} \int q_i \log q_i d\theta_i$$

where we define a new distribution $\log \tilde{p}(\theta_j, \mathbf{y}|\lambda)$ (with a constant for normalization)

$$\log \tilde{p}(\theta_j, \mathbf{y}|\lambda) = \int \log p(\theta, \mathbf{y}|\lambda) \prod_{i \neq j} q_i d\theta_i + const = \mathop{\mathrm{E}}_{i \neq j}[\log p(\theta, \mathbf{y}|\lambda)] + const$$

$\mathcal{L}(q,\theta)$ will be minimized when $KL(q_j || \tilde{p}) = 0$ or $q_j(\theta_j) = \tilde{p}(\theta_j, \mathbf{y}|\lambda)$.
The general optimal solution with a normalization constant will be

$$q_j^\star(\theta_j) = \exp\left( \mathop{\mathrm{E}}_{i \neq j}[\log p(\theta, \mathbf{y}|\lambda)] \right) / \int \exp\left( \mathop{\mathrm{E}}_{i \neq j}[\log p(\theta, \mathbf{y}|\lambda)] \right) d\theta_j$$

## The Univariate Gaussian

Given a set of independent observations $\mathbf{y} = \{y_1, \ldots, y_N\}$ drawn from a Gaussian distribution, we will determine the posterior for the mean $\mu$ and the precision $\tau$;
The likelihood function is

$$p(\mathbf{y}|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(y_n-\mu)^2\right\}.$$

The conjugate prior distributions for $\mu$ and $\tau$ are

$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1})$$
$$p(\tau) \quad = Gam(\tau|a_0,b_0) = \tau^{a_0-1}e^{-b_0\tau}b_0^{a_0}/\Gamma(a_0)$$

Factorized variational approximation for the posterior distribution is

$$q(\mu,\tau) = q_\mu(\mu)q_\tau(\tau)$$

Applying the variational approach,

$$\log q_\mu^\star(\mu) = \mathrm{E}_\tau\left[\log p(\mathbf{y}|\mu,\tau) + \log p(\mu|\tau)\right] + const \ (p(\tau))$$

$$= -\frac{\mathrm{E}_\tau[\tau]}{2}\left\{\lambda_0(\mu-\mu_0)^2 + \sum_{n=1}^{N}(y_n-\mu)^2\right\} + const$$

Assuming that $q_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \lambda_N^{-1})$, we can obtain

$$\mu(\mu_N\lambda_N) = \mu\left(\frac{\mathrm{E}_\tau[\tau]}{2}\right)\left(2\lambda_0\mu_0 + 2\sum_{n=1}^{N}y_n\right) \qquad \longrightarrow \mu_N = \frac{\lambda_0\mu_0 + \sum_{n=1}^{N}y_n}{\lambda_0+N}$$

$$\mu^2\left(-\frac{1}{2}\lambda_N\right) = \mu^2\left(-\frac{\mathrm{E}_\tau[\tau]}{2}\right)(\lambda_0+N) \qquad \longrightarrow \lambda_N = \mathrm{E}_\tau[\tau](\lambda_0+N)$$

Similarly, for $q_\tau(\tau)$, we have

$$\log q^\star_\tau(\tau) = \mathrm{E}_\mu\left[\log p(\mathbf{y}|\mu,\tau) + \log p(\mu|\tau)\right] + \log p(\tau) + const$$
$$= (a_0 - 1)\log(\tau) - b_0\tau + \frac{N}{2}\log(\tau)$$
$$-\frac{\tau}{2}\mathrm{E}_\mu\left[\sum_{n=1}^{N}(y_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + const$$

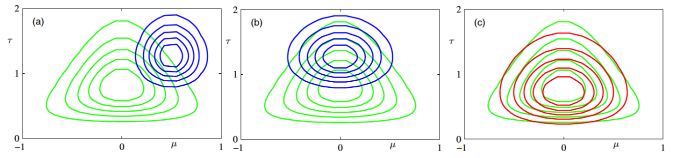
Assuming that $q_\tau(\tau)$ will be a gamma distribution with $Gam(\tau|a_N, b_N)$, we can obtain

$$(a_N - 1)\log(\tau) = (a_0 - 1 + \frac{N}{2})\log(\tau)$$
$$(-b_N)\tau = -(b_0 + \frac{1}{2}\mathrm{E}_\mu\left[\sum_{n=1}^{N}(y_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right])\tau$$

or

$$a_N = a_0 + \frac{N}{2} \qquad b_N = b_0 + \frac{1}{2}\mathrm{E}_\mu\left[\sum_{n=1}^{N}(y_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right]$$

Posterior distribution $p(\theta|\mathbf{y},\lambda) = p(\mu,\tau|\mathbf{y})$ approximated by $q(\mu,\tau)$ after iterations.



## Linear Regression

$$\mathbf{y} = f(\Theta) + \epsilon$$

where $\mathbf{y} = [y_1, \ldots, y_N]^T$ are the noisy observations, $\Theta$ is the vector of known/hidden variables and $\epsilon$ is the *i.i.d.* additive noise vector with $p(\epsilon) = \mathcal{N}(\epsilon|\mathbf{0}, \beta^{-1}\mathbf{I})$.

If we model the observations as a linear combination of $M$ basis functions $\phi_m(\mathbf{x})$, then

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{w} + \epsilon$$

where $\Theta = \{\mathbf{x}, \mathbf{w}\}$, $\mathbf{x} = [x_1, \ldots, x_M]$ are known variables, $\mathbf{w} = [w_1, \ldots, w_M]^T$ are hidden variables and

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \cdots & \phi_M(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \cdots & \phi_M(\mathbf{x}_N) \end{bmatrix}$$

our likelihood function becomes $p(\mathbf{y}|\mathbf{w}, \beta) = \mathcal{N}(\mathbf{y}|\boldsymbol{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I})$.

A nonstationary Gaussian prior distribution with a distinct inverse variance $\alpha_m$ for each weight $w_m$ is

$$p(\mathbf{w}|\alpha) = \prod_{m=1}^{M} \mathcal{N}\left(w_m|0, \alpha_m^{-1}\right).$$

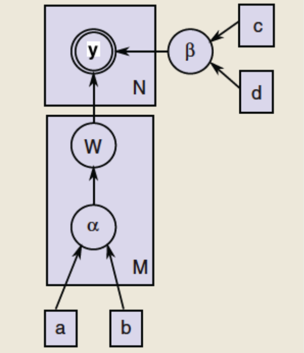
In order to constrain the precision parameters $\alpha_m$, we model them as random variables with conjugate distributions of Gamma prior

$$p(\alpha|a, b) = \prod_{m=1}^{M} Gamma(\alpha_m|a, b)$$

Prior distribution of noise precision is also modeled as

$$p(\beta|c, d) = Gamma(\beta|c, d)$$

Graphical model for linear regression problem.




Bayesian inference requires to determine the posterior distribution

$$p(\mathbf{w}, \alpha, \beta|\mathbf{y}) = p(\mathbf{y}|\mathbf{w}, \beta)p(\mathbf{w}, \alpha)p(\alpha)p(\beta)/p(\mathbf{y})$$

whose normalization constant $p(\mathbf{y})$ cannot be computed analytically.

## Variational Approximation

In this problem, $\theta = \{\mathbf{w}, \alpha, \beta\}$ and $\lambda = \{a, b, c, d\}$

Factorized variational approximation for the posterior distribution $p(\theta|\mathbf{y}, \lambda)$ is

$$p(\mathbf{w}, \alpha, \beta|\mathbf{y}, a, b, c, d) \approx q(\mathbf{w}, \alpha, \beta) = q(\mathbf{w})q(\alpha)q(\beta)$$

$$\log q^{\star}(\mathbf{w}) = \mathrm{E}_{q(\alpha)q(\beta)}\left[\log p(\mathbf{y}, \mathbf{w}, \alpha, \beta)\right] + const$$

$$= \mathrm{E}_{q(\alpha)q(\beta)}\left[\log p(\mathbf{y}|\mathbf{w}, \beta) + \log p(\mathbf{w}|\alpha)\right] + const$$

$$= \mathrm{E}_{q(\alpha)q(\beta)}\left[-\frac{\beta}{2}(\mathbf{y} - \mathbf{\Phi w})^T(\mathbf{y} - \mathbf{\Phi w}) - \frac{1}{2}\sum_{m=1}^{M}\alpha_m w_m^2\right] + const$$

$$= -\frac{1}{2}\mathrm{E}_{q(\beta)}[\beta]\left[\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{\Phi w} + \mathbf{w}^T\mathbf{\Phi}^T\mathbf{\Phi w}\right] - \frac{1}{2}\sum_{m=1}^{M}\mathrm{E}_{q(\alpha)}\left[\alpha_m\right] w_m^2 + const$$

$$= -\frac{1}{2}\mathbf{w}^T\underbrace{\left[\mathrm{E}_{q(\beta)}[\beta]\mathbf{\Phi}^T\mathbf{\Phi} + \mathrm{E}_{q(\alpha)} diag[\alpha_1, \ldots, \alpha_m]\right]}_{\Sigma^{-1}}\mathbf{w} + \mathbf{w}^T\underbrace{\left[\mathrm{E}_{q(\beta)}[\beta]\mathbf{\Phi}^T\right]}_{\Sigma^{-1}\mu}\mathbf{y} + const$$

Therefore, $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mu, \Sigma)$.

## Self-Study Question

Show that the posterior distribution for $\alpha$ can be given as

$$q(\alpha) = \prod_{m=1}^{M} Gamma(\alpha_m | \tilde{a}, \tilde{b}_m)$$

where

$$\tilde{a} = a + 1/2$$
$$\tilde{b}_m = b + \frac{1}{2}\mathrm{E}_{q(\mathbf{w})}[w_m^2]$$

and the posterior distribution for $\beta$ as

$$q(\beta) = Gamma(\beta | \tilde{c}, \tilde{d}_m)$$

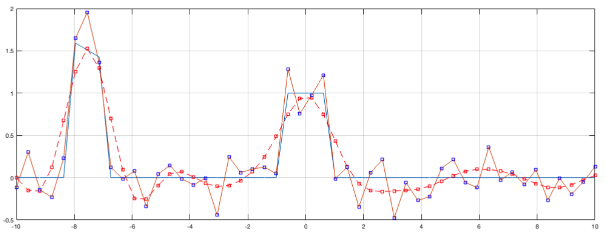
where

$$\tilde{c} = c + N/2$$
$$\tilde{d} = d + \frac{1}{2}\mathrm{E}_{q(\mathbf{w})}[||\mathbf{y} - \Phi\mathbf{w}||^2]$$

## Numerical Example for VB Linear regression

```
N=50; %no of samples and no of basis functions
sigma_2 = 4e-2; %noise variance
a=0; b=0; % hyperparameters for alpha
SNR = 6.6; % SNR in decibels
sigma_phi_2 = 1; % variance for gaussian kernel
x= linspace(-10,10,N)';
y0 = zeros(size(x));
y0(find((x>-1)&(x<1)))=1;
y0(find((x>-8)&(x<-7))) = -0.2*x(find((x>-8)&(x<-7))) ;
n = randn(size(y0));
n=(n-mean(n))/std(n)*sqrt(sigma_2);
y=y0+n; % noisy observations
plot(x,y0,x,y);
10*log10(std(y0).^2/std(n).^2) % SNR empirical
Phi=exp(-0.5/sigma_phi_2*(repmat(x,1,N)-repmat(x',N,1) ).^2);
% Initialize hyperparameters
a = randn(N,1)*1e-3;
b = randn(N,1)*1e-3;
c = randn(1,1)*1e-3;
d = randn(1,1)*1e-3;
```

```
% VB iterations
for i = 1:200,
% estimation of hyperparameters for w
 Sigma_ = pinv(c/d*Phi'*Phi + diag(a./b));
 mu= Sigma_ * (c/d *Phi')*y;
% estimation of hyperparameters for alpha
 a = a + 0.5;
 b = b + 0.5*(diag(mu*mu' + Sigma_));
% estimation of hyperparameters for beta
 c = c + N/2;
 d = d+0.5*(y'*y-2*y'*Phi*mu+trace(Phi'*Phi *Sigma_)+mu'*Phi'*Phi*mu);
end
plot(x,y0,x,  Phi*mu,'--rs' , x,y,'-.bs',x,y);grid
```



When the approximate posterior distributions $q(\mathbf{w})$, $q(\alpha)$ and $q(\beta)$ are iteratively updated until convergence, the posterior $p(\mathbf{w}, \alpha, \beta | \mathbf{y}, a, b, c, d)$ can be approximately determined by $q(\mathbf{w}, \alpha, \beta)$.
The true prior distribution for the weights can also be found by

$$p(\mathbf{w}|a, b) = \int p(\mathbf{w}, \alpha | a, b) d\alpha = \int p(\mathbf{w}|\alpha) p(\alpha | a, b) d\alpha$$

**Self-Study Question**

Show that true prior distribution for the weights is a Student-t distribution

$$p(\mathbf{w}|a, b) = \int \prod_{m=1}^{M} \mathcal{N}(w_m | 0, \alpha_m^{-1}) Gamma(\alpha_m | a, b) d\alpha_m = \prod_{m=1}^{M} Student(w_m | \mu, \lambda, \nu)$$

where
$\mu = 0$, $\lambda = a/b$ and $\nu = 2a$.

## Gaussian Mixture Models (GMM)

GMM is based on a number of observations that are assumed to be generated by $K$ Gaussians whose means, covariances and the probability (weight) that a point comes from each of the Gaussians are to be determined.

For the sake of convenience, the Gaussian functions are given with their means and precisions as

$$k^{th} \text{ Gaussian function : } \phi_k(\mathbf{y}|\mu_k, \Lambda_k) \triangleq \frac{exp\left(-\frac{1}{2}(\mathbf{y}-\mu_k)^T\Lambda_k(\mathbf{y}-\mu_k)\right)}{(2\pi)^{D/2}|\Lambda_k^{-1}|^{1/2}}$$

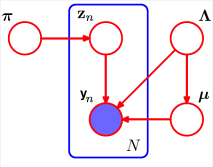
$$\text{Prior probability of mixture components : } \sum_{k=1}^{K} \pi_k = 1 \text{ and } \pi_k \geq 0.$$

Given $N$ *i.i.d.* samples $\mathbf{y}_1, \ldots \mathbf{y}_n \in \mathcal{R}^D$ from a GMM with $K$ components, we estimate its parameter set $\lambda = \{(\pi_k, \mu_k, \Lambda_k)\}_{k=1}^{K}$.

It is assumed that the data $\mathbf{y}$ are sampled using the following procedure;

i) Randomly sample one component $k$ using the probability vector $\pi = [\pi_1, \ldots, \pi_K]$.

ii) Generate an observation by sampling from the density $\phi_k(\mathbf{y})$ of component $k$.




For each observation $\mathbf{y}_n$, we have a hidden variable $\mathbf{z}_n$ which is a $K-$dimensional binary vector whose elements but one is 0.
For a set of observations $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ and $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$, the conditional distribution of $\mathbf{Z}$ is

$$p(\mathbf{Z}|\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

and the likelihood of $\mathbf{Y}$ is

$$p(\mathbf{Y}|\mathbf{Z}, \mu, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\mathbf{y}_n|\mu_k, \Lambda_k^{-1})^{z_{nk}}$$

## Variational Bayes for GMM Training

The conjugate priors we use for $\lambda = \{(\pi_k, \mu_k, \Lambda_k)\}_{k=1}^{K}$ are Dirichlet and Gauss-Wishart *i.e.*

$$p(\pi) = Dir(\pi|\alpha_1, \ldots, \alpha_K) = \frac{\Gamma\left(\sum_{m=1}^{K} \alpha_m\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

usually it is assumed that $\alpha_k = \alpha_0$ for all $k$.

$$\mathcal{W}(\mu, \Lambda) = \prod_{k=1}^{K} p(\mu_k, \Lambda_k) = \prod_{k=1}^{K} p(\mu_k|\Lambda_k) p(\Lambda_k)$$

where

$$p(\mu_k|\Lambda_k) = \mathcal{N}(\mu_k|\mathbf{m}_0, (\beta_0\Lambda_k)^{-1})$$

and $p(\Lambda_k)$ is the Wishart distribution

$$p(\Lambda_k) = \mathcal{W}_0(\Lambda_k|\mathbf{W}_0, \nu_0) = \frac{|\Lambda_k|^{(\nu - D - 1)}|e^{-\frac{1}{2} Trace(\mathbf{W}_0^{-1}\Lambda_k)})}{|\mathbf{W}_0|^{\nu_0/2}\left(2^{\nu_0 D/2}\pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma\left(\frac{\nu_0 + 1 - i}{2}\right)\right)}$$

The hyperparameter set in full Bayesian GMM is $\{\alpha, \mathbf{m}_0, \beta_0, \nu_0, \mathbf{W}_0\}$.
The set of hidden variables and parameters is $\mathbf{\Theta} = \{\mathbf{Z}, \lambda\}$.
Using mean-field approximation, $p(\Theta) = q(\mathbf{Z})q(\pi)q(\mu, \Lambda)$
and performing the calculations, we can obtain

$$q(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$$

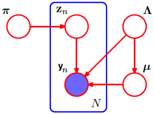
$$q(\pi) = Dir(\pi|\alpha_1, \ldots, \alpha_K)$$

$$q(\mu, \Lambda) = \prod_{k=1}^{K} q(\mu_k|\Lambda_k)q(\Lambda_k)$$

$$q(\mu_k|\Lambda_k) = \prod_{k=1}^{K} \mathcal{N}(\mu_k|\mathbf{m}_k, \beta_k\Lambda_k)$$

$$q(\Lambda_k) = \prod_{k=1}^{K} \mathcal{W}(\Lambda_k|\mathbf{W}_k, \nu_k)$$

Iteration equations for $\{r_{nk}, \mathbf{m}_k, \beta_k, \nu_k, \mathbf{W}_k\}$ are determined by computing the conditional expectations of log posterior distribution of $p(\mathbf{Y}, \mathbf{Z}, \lambda|\alpha, \mathbf{m}_0, \beta_0, \nu_0, \mathbf{W}_0)$.




## Joint Distribution of Random Variables

$$p(\mathbf{Y}, \mathbf{Z}, \pi, \mu, \Lambda) = p(\mathbf{Y}|\mathbf{Z}, \mu, \Lambda)p(\mathbf{Z}|\pi)p(\pi)p(\mu|\Lambda)p(\Lambda)$$

## Variational Distribution of Random Variables

$$q(\mathbf{Z}, \pi, \mu, \Lambda) = q(\mathbf{Z})q(\pi)q(\mu)q(\Lambda)$$

Optimization of $\log q^{\star}(\mathbf{Z}) = \mathrm{E}_{\pi,\mu,\Lambda}\left[\log p(\mathbf{Y}, \mathbf{Z}, \pi, \mu, \Lambda)\right]$ yields

$$\log q^{\star}(\mathbf{Z}) = \mathrm{E}_{\mu,\Lambda}\left[\log p(\mathbf{Y}|\mathbf{Z}, \mu, \Lambda)\right] + \mathrm{E}_{\pi}\left[\log p(\mathbf{Z}|\pi)\right] + const$$

$$\log q^{\star}(\mathbf{Z}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z^{nk} \log \rho_{nk} + const$$

where

$$\log \rho_{nk} = \mathrm{E}_{\pi}\left[\log p(\pi_k)\right] + \frac{1}{2}\mathrm{E}_{\Lambda_k}\left[\log |\Lambda_k|\right] - \frac{D}{2}\log(2\pi) - \frac{1}{2}\mathrm{E}_{\mu_k,\Lambda_k}\left[(\mathbf{y}_n - \mu_k)^T \Lambda_k (\mathbf{y}_n - \mu_k)\right]$$

$$q^\star(\mathbf{Z}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$$

In order to obtain a normalized distribution for $q_\Theta$, we know that

$q(\mathbf{Z}) = q(\mathbf{z}_1) \ldots q(\mathbf{z}_N)$ for any $n$, $q(\mathbf{z}_n) = \prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$ and $\sum_{j=1}^{K} z_{nj} = 1$ .

As a simple example, when we sum $q(z_n)$ over all possibilities of $z_{nk}$ like for instance for $K = 3$, $\theta_n$ can be $\{[1\ 0\ 0], [0\ 1\ 0], [0\ 0\ 1]\}$ and

$$\sum_{z_n} \prod_{k=1}^{3} \rho_{nk}^{\theta_{nk}} = \rho_{n1} + \rho_{n2} + \rho_{n3} = \sum_{j=1}^{3} \rho_{nj}$$

The normalized form of $q(z_n)$ can be expressed as $q^\star(z_n) = \prod_{k=1}^{K} r_{nk}^{z_{nk}}$

where $r_{nk} = \rho_{nk} / \sum_{j=1}^{K} \rho_{nj}$.

The normalized overall distribution is

$q^\star(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$ with $\sum_{k=1}^{K} r_{nk} = 1$ and $r_{nk} \geq 0$ as required from mixture probabilities.

$$\log q^\star(\pi,\mu,\Lambda) = \mathrm{E}_{\mathbf{Z}}\Big[\log p(\mathbf{Y},\mathbf{Z},\pi,\mu,\Lambda)\Big] + const$$

$$= \mathrm{E}_{\mathbf{Z}}\Big[\log p(\mathbf{Y}|\mathbf{Z},\mu,\Lambda)\Big] + \mathrm{E}_{\mathbf{Z}}\Big[\log p(\mathbf{Z}|\pi)\Big]$$
$$+\mathrm{E}_{\mathbf{Z}}\Big[\log p(\pi)\Big] + \mathrm{E}_{\mathbf{Z}}\Big[\log p(\mu,\Lambda)\Big] + const$$

$$= \log p(\pi) + \sum_{k=1}^{K}\log p(\mu_k,\Lambda_k) + \mathrm{E}_{\mathbf{Z}}\left[\log p(\mathbf{Z}|\pi)\right]$$
$$+\sum_{k=1}^{K}\sum_{n=1}^{N}\mathrm{E}_{\mathbf{Z}}[z_{nk}]\log\mathcal{N}(\mathbf{y}_n|\mu_k,\Lambda_k^{-1}) + const$$

which can be partitioned into $q^\star(\pi,\mu,\Lambda) = q(\pi)\prod_{k=1}^{K} q(\mu_k,\Lambda_k)$

where $\log q^\star(\pi) = (\alpha_0 - 1)\sum_{k=1}^{K}\log\pi_k + \sum_{k=1}^{K}\sum_{n=1}^{N}\mathrm{E}_{\mathbf{Z}}[z_{nk}]\log\pi_k + const$

or $q^\star(\pi) = Dir(\pi|\alpha)$ with $\alpha_k = \alpha_0 + \sum_{n=1}^{N}\mathrm{E}_{\mathbf{Z}}[z_{nk}] = \alpha_0 + \sum_{n=1}^{N} r_{nk}$.

## Self-Study Question

Using $q^\star(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}}$, show that $\mathrm{E}_{\mathbf{Z}}[z_{nk}] = r_{nk}$.

$$q^\star(\mu_k,\Lambda_k) = q(\mu_k|\Lambda_k)q(\Lambda_k) = \mathcal{N}(\mu_k|\mathbf{m}_k,(\beta_k\Lambda_k)^{-1})\mathcal{W}_k(\Lambda_k|\mathbf{W}_k,\nu_k)$$

**Self-Study Question**

Show that $q^\star(\mu_k,\Lambda_k)$ can be expressed as a Gauss-Wishart distribution
$q^\star(\mu_k,\Lambda_k) = \mathcal{N}(\mu_k|\mathbf{m}_k,(\beta_k\Lambda_k)^{-1})\mathcal{W}(\Lambda_k|\mathbf{W}_k,\nu_k)$
where

$$\begin{aligned}
\beta_k &= \beta_0 + N_k \\
\mathbf{m}_k &= \tfrac{1}{\beta_k}(\beta_0\mathbf{m}_0 + N_k\bar{\mathbf{y}}_k) \\
\mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \tfrac{\beta_0 N_k}{\beta_0+N_k}(\bar{\mathbf{y}}_k-\mathbf{m}_0)(\bar{\mathbf{y}}_k-\mathbf{m}_0)^T \\
\nu_k &= \nu_0 + N_k
\end{aligned}$$

with

$$N_k = \sum_{n=1}^{N} r_{nk}, \quad \bar{\mathbf{y}}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{y}_n \text{ and } \mathbf{S}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{y}_n-\bar{\mathbf{y}}_k)(\mathbf{y}_n-\bar{\mathbf{y}}_k)^T$$

Since we have determined the $q^\star(\pi)$, $q^\star(\mu,\Lambda)$ and $q^\star(\Lambda)$, we can evaluate the expectations in $\log^\star(\mathbf{Z})$

$$\log \rho_{nk} = \mathrm{E}_{\pi}\left[\log p(\pi_k)\right] + \frac{1}{2}\mathrm{E}_{\Lambda_k}\left[\log|\Lambda_k|\right] - \frac{D}{2}\log(2\pi) \\ -\frac{1}{2}\mathrm{E}_{\mu_k,\Lambda_k}\left[(\mathbf{y}_n-\mu_k)^T\Lambda_k(\mathbf{y}_n-\mu_k)\right].$$

Show that

$$\mathrm{E}_{\mu_k,\Lambda_k}\left[(\mathbf{y}_n-\mu_k)^T\Lambda_k(\mathbf{y}_n-\mu_k)\right] = D\beta_k^{-1} + \nu_k(\mathbf{y}_n-\mathbf{m}_k)^T\mathbf{W}_k(\mathbf{y}_n-\mathbf{m}_k)$$

$$\log\tilde{\Lambda}_k = \mathrm{E}_{\Lambda_k}\left[\log|\Lambda_k|\right] = \sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}\right) + D\log(2) + \log(|\mathbf{W}_k|)$$

$$\log\tilde{\pi} = \mathrm{E}_{\pi_k}\left[\log\pi_k\right] = \psi(\alpha_k) - \psi(\hat{\alpha})$$

where $\hat{\alpha} = \sum_k \alpha_k$ and
$\psi(\alpha) = \frac{d}{d\alpha}\log\Gamma(\alpha)$ is the *Digamma* function.

## Example for GMM Model

$N = 600$ data points are generated using a GMM Model with $K = 2$, with

$$\mu = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ -3 \end{bmatrix} \right\}, \Lambda^{-1} = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right\} \text{ and } \pi = \{0.3, 0.7\}.$$

### Generation of Simulated Data

```
N=300; % no of data points
K=2; % no of clusters
D=2; % dimension of data vectors
Pi = [0.3 0.7]; % mixture probabilities
p1=cumsum(Pi);
for i=1:N,
P(i) =sum( (rand()>=p1))+1;
end;
Mu_0 = [0 3 3 -3 -3 ;0 -3 3 3 -3]; % mean vectors of clusters
% covariance matrices of clusters
Lambda{1} = inv([1 0 ;0 1]);
Lambda{2} = inv([1 0.5 ;0.5 1 ]);
% compute the coloring matrix
for i=1:K,
L{i} = inv(chol(Lambda{i}));
end;
% sample the gaussian deviates
for i=1:N,
Y(:,i) = L{P(i)}*randn(D,1)+ Mu_0(:,P(i));
end;
plot(Y(1,:),Y(2,:),'*');
```

## VB Algorithm for GMM with $K = 2$ clusters

```
% VB EM for GMM
% Set hyperparameters for priors
alpha_0 = ones(1,1)/K;  % Dirichlet for p(pi)
W_0 = eye(D,D)*100;
W_0_inv = inv(W_0);
nu_0 =10+D; % Wishart for p(mu,Lambda)
beta_0 = 1;
m0 = zeros(D,1); % Gaussian for p(mu | Lambda)
% Initialize hidden variables and parameters
m = randn(D,K);
beta =beta_0*ones(1,K);
nu = nu_0*ones(1,K);
alpha = alpha_0*ones(1,K);
alpha_hat = sum(alpha);
E_pi= (psi(alpha)- psi(alpha_hat));
for k=1:K,
W{k} = W_0;
ps = sum(psi((nu(k)+1-[1:2])/2));
E_L(k) =  ps+ D*log(2) + log(det(W{k}));
end;
% iterate over {r,m,Lambda, beta,nu,W}
for iter = 1:500,
%E-STEP
for k=1:K,for n=1:N,
log_rho(n,k)=E_pi(k)+0.5*E_L(k)-0.5*D/beta(k) -0.5*nu(k)*(Y(:,n)-m(:,k))'*(W{k})*(Y(:,n)-m(:,k));end;
end;
for n=1:N,Z(n) =  logsumexp(log_rho(n,:),2); end;
for k=1:K,r(:,k) = exp( log_rho(:,k) - Z'); end;
```

```
%M_STEP
N_k = sum(r)+1e-10;
Y_bar = Y*r./(ones(2,1)*N_k);
for k=1:K,s = Y-Y_bar(:,k)*ones(1,N);
s1 = 0;
for n=1:N,  s1 = s1 + r(n,k)*s(:,n)*s(:,n)'   ; end;
S{k} = s1/N_k(k);
end;
alpha = alpha_0 + N_k;
alpha_hat = sum(alpha);
beta = beta_0 + N_k;
nu = nu_0 + N_k +1;
Pi_k = (alpha_0 + N_k)./(K*alpha_0+N);
m    = (beta_0*m0 + (ones(D,1)*N_k).*Y_bar)./(ones(D,1)*beta);
for k=1:K,
W_inv{k} = W_0_inv+ N_k(k)*S{k} + ((beta_0*N_k(k))/(beta_0 +...
N_k(k)))*(Y_bar(:,k)-m0*ones(1,K))*(Y_bar(:,k)-m0*ones(1,K))';
W{k} = inv(W_inv{k});
end;
E_pi= psi(alpha)- psi(alpha_hat);
for k=1:K,
ps = sum(psi((nu(k)+1-[1:2])/2));
E_L(k) =  ps+ D*log(2) + log(det(W{k}));
end;
end; % iter
for k=1:K,
MU(:,k) =m(:,k); LA{k} =inv(nu(k)*W{k});
PI(k) = Pi_k(k);
end;
```

## Variational Lower Bound for Model Evaluation

Remember that the lower bound is

$$\mathcal{L} = \sum_{\mathbf{Z}} \int\int\int q(\mathbf{Z},\pi,\mu,\Lambda) \log\left\{\frac{p(\mathbf{Y},\mathbf{Z},\pi,\mu,\Lambda)}{q(\mathbf{Z},\pi,\mu,\Lambda)}\right\} d\pi\, d\mu\, d\Lambda$$

$$= \mathbb{E}_{\mathbf{Z},\pi,\mu,\Lambda}\left[\log\ p(\mathbf{Y},\mathbf{Z},\pi,\mu,\Lambda)\right] - \mathbb{E}_{\mathbf{Z},\pi,\mu,\Lambda}\left[\log\ q(\mathbf{Z},\pi,\mu,\Lambda)\right]$$

$$= \mathbb{E}_{\mathbf{Z},\mu,\Lambda}\left[\log\ p(\mathbf{Y}|\mathbf{Z},\mu,\Lambda)\right] + \mathbb{E}_{\mathbf{Z},\pi}\left[\log\ p(\mathbf{Z}|\pi)\right] + \underbrace{\mathbb{E}_{\pi}\left[\log\ p(\pi)\right]}_{\log\tilde{\pi}}$$

$$+ \mathbb{E}_{\mu,\Lambda}\left[\log\ p(\mu,\Lambda)\right] - \mathbb{E}_{\mathbf{Z}}\left[\log\ q(\mathbf{Z})\right] - \mathbb{E}_{\pi}\left[\log\ q(\pi)\right] - \mathbb{E}_{\mu,\Lambda}\left[\log\ q(\mu,\Lambda)\right]$$

We can evaluate terms as

$$\mathrm{E}_{\mathbf{Z},\mu,\Lambda}\left[\log\ p(\mathbf{Y}|\mathbf{Z},\mu,\Lambda)\right] = \mathrm{E}_{\mathbf{Z},\mu,\Lambda}\left[\sum_{n=1}^{N}\sum_{k=1}^{K}\log\mathcal{N}(\mathbf{y}_k|\mu_k,\Lambda_k)^{z_{nk}}\right]$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K}\mathrm{E}_{\mathbf{Z},\mu,\Lambda}\left[z_{nk}[\log|\Lambda_k|^{1/2} + \log(2\pi)^{-D/2} - \tfrac{1}{2}(\mathbf{y}_n-\mu_k)^T\Lambda_k(\mathbf{y}_n-\mu_k)]\right]$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K}\underbrace{\mathrm{E}_{\mathbf{Z}}[z_{nk}]}_{1}\left[\tfrac{1}{2}\underbrace{\mathrm{E}_{\Lambda_k}[\log|\Lambda_k|]}_{2} + \log(2\pi)^{-D/2} - \tfrac{1}{2}\underbrace{E_{\mu_k,\Lambda_k}[(\mathbf{y}_n-\mu_k)^T\Lambda_k(\mathbf{y}_n-\mu_k)]}_{3}\right]$$

We use the following relations

1. $\mathrm{E}_{\mathbf{Z}}[z_{nk}] = r_{nk}$ ,
2. $\mathrm{E}_{\Lambda_k}[\log|\Lambda_k|] = \log\tilde{\Lambda}_k$
3. $\mathrm{E}_{\mu_k,\Lambda_k}\left[(\mathbf{y}_n-\mu_k)^T\Lambda_k(\mathbf{y}_n-\mu_k)\right] = D\beta_k^{-1} + \nu_k(\mathbf{y}_n-\mathbf{m}_k)^T\mathbf{W}_k(\mathbf{y}_n-\mathbf{m}_k)$
4. $\sum_{n=1}^{N} r_{nk} = N_k,$
5. $\bar{\mathbf{y}}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{y}_n$
6. $\mathbf{S}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{y}_n-\bar{\mathbf{y}}_k)$
7. $\sum_{n=1}^{N} r_{nk}\mathbf{y}_n\mathbf{y}_n^T = \sum_{n=1}^{N} r_{nk}(\mathbf{y}_n-\bar{\mathbf{y}}_k)(\mathbf{y}_n-\bar{\mathbf{y}}_k)^T + N_k\bar{\mathbf{y}}_k\bar{\mathbf{y}}_k^T$

Using identity 3

$\nu_k tr[\mathbf{W}_k \sum_{n=1}^{N} r_{nk}(\mathbf{y}_n - \mathbf{m}_k)(\mathbf{y}_n - \mathbf{m}_k)^T] =$

$\nu_k tr[\mathbf{W_k} \sum_{n=1}^{N} r_{nk} \left(\mathbf{y}_n\mathbf{y}_n^T - \mathbf{y}_n\mathbf{m}_k^T - \mathbf{m}_k\mathbf{y}_n^T + \mathbf{m}_k\mathbf{m}_k^T\right)]$

Replacing the first term with identity 7, and using identities 6, 5 and 4 we obtain

$$\underbrace{\sum_{n=1}^{N} r_{nk}(\mathbf{y}_n - \bar{\mathbf{y}}_k)(\mathbf{y}_n - \bar{\mathbf{y}}_k)^T}_{N_k\mathbf{S}_k} + N_k\bar{\mathbf{y}}_k\bar{\mathbf{y}}_k^T + \sum_{n=1}^{N} r_{nk}\left(-\underbrace{\mathbf{y}_n\mathbf{m}_k^T}_{N_k\bar{\mathbf{y}}_k\mathbf{m}_k^T} - \mathbf{m}_k\mathbf{y}_n^T + \mathbf{m}_k\mathbf{m}_k^T\right)$$

$$\nu_k tr[\mathbf{W}_k \sum_{n=1}^{N} r_{nk}(\mathbf{y}_n - \mathbf{m}_k)(\mathbf{y}_n - \mathbf{m}_k)^T] = \nu_k tr[\mathbf{W_k}(N_k\mathbf{S}_k + N_k(\bar{\mathbf{y}}_k - \mathbf{m}_k))(\bar{\mathbf{y}}_k - \mathbf{m}_k)^T]$$

$$\mathrm{E}_{\mathbf{Z},\mu,\Lambda}\left[\log\ p(\mathbf{Y}|\mathbf{Z},\mu,\Lambda)\right] = \frac{1}{2}\sum_{k=1}^{K} N_k\Big\{\log\ \tilde{\Lambda}_k - D\beta_k^{-1} - \nu_k\ \mathrm{tr}(\mathbf{S}_k\mathbf{W}_k) - \nu_k(\bar{\mathbf{y}}_k - \mathbf{m}_k)^T\mathbf{W}_k(\bar{\mathbf{y}}_k - \mathbf{m}_k) - D\ \log(2\pi)\Big\}$$

Remembering the priors:

$$p(\pi) = Dir(\pi|\alpha_1, \dots, \alpha_K) = \frac{\Gamma\left(\sum_{m=1}^{K} \alpha_m\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$ and $\alpha_k = \alpha_0$ for all $k$.

$$\mathbb{E}_\pi\left[\log\ p(\pi)\right] = \log C(\alpha_0) + (\alpha_0 - 1)\sum_{k=1}^{K} \log \tilde{\pi}_k$$

$$\mathbb{E}_{\mathbf{Z},\pi}\left[\log\ p(\mathbf{Z}|\pi)\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\ \log \tilde{\pi}_k \qquad \log p(\mathbf{Z}|\pi) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \log \pi_k$$

$$\mathcal{W}(\mu, \Lambda) = \prod_{k=1}^{K} p(\mu_k, \Lambda_k) = \prod_{k=1}^{K} p(\mu_k|\Lambda_k)p(\Lambda_k)$$

where

$$p(\mu_k|\Lambda_k) = \mathcal{N}(\mu_k|\mathbf{m}_0, (\beta_0\Lambda_k)^{-1})$$

and $p(\Lambda_k)$ is the Wishart distribution

$$p(\Lambda_k) = \mathcal{W}_0(\Lambda_k|\mathbf{W}_0, \nu_0) = \frac{|\Lambda_k|^{(\nu - D - 1)}|e^{-\frac{1}{2}Trace(\mathbf{W}_0^{-1}\Lambda_k)})}{|\mathbf{W}_0|^{\nu_0/2}\left(2^{\nu_0 D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu_0+1-i}{2}\right)\right)}$$

$$q(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$$
$$q(\pi) = Dir(\pi|\alpha_1, \ldots, \alpha_K)$$
$$q(\mu, \Lambda) = \prod_{k=1}^{K} q(\mu_k|\Lambda_k)q(\Lambda_k)$$
$$q(\mu_k|\Lambda_k) = \prod_{k=1}^{K} \mathcal{N}(\mu_k|\mathbf{m}_k, \beta_k\Lambda_k)$$
$$q(\Lambda_k) = \prod_{k=1}^{K} \mathcal{W}(\Lambda_k|\mathbf{W}_k, \nu_k)$$

$$\mathbb{E}_{\mu,\Lambda}\Big[\log p(\mu,\Lambda)\Big] = \mathbb{E}_{\mu,\Lambda}\Big[\log p(\mu_k|\Lambda_k) + \log p(\Lambda_k)\Big]$$

$$= \mathbb{E}_{\mu,\Lambda}\Big[\log(\tfrac{1}{2\pi})^{D/2} + \log|\beta_0\Lambda_k|^{1/2} - \tfrac{1}{2}(\mu_k - \mathbf{m}_0)^T\beta_0\Lambda_k(\mu_k - \mathbf{m}_0)^T + \log B(\mathbf{W}_0,\nu_0) + \tfrac{\nu_0 - D - 1}{2}\log|\Lambda_k| - \tfrac{1}{2}tr[\mathbf{W}_0^{-1}\Lambda_k]\Big]$$

$$= \tfrac{D}{2}\log\tfrac{\beta_0}{2\pi} + \log B(\mathbf{W}_0,\nu_0) + \tfrac{1}{2}\mathbb{E}_{\Lambda_k}[\log|\Lambda_k|] - \tfrac{1}{2}\beta_0\mathbb{E}_{\mu_k,\Lambda_k}[(\mu_k - \mathbf{m}_0)^T\Lambda_k(\mu_k - \mathbf{m}_0)] + \tfrac{\nu_0 - D - 1}{2}\mathbb{E}_{\Lambda_k}[\log|\Lambda_k|] - \tfrac{1}{2}tr[\mathbf{W}_0^{-1}\mathbb{E}_{\Lambda_k}[\Lambda_k]]$$

The third term in the above expression is evaluated as

$$\mathbb{E}_{\mu_k,\Lambda_k}[(\mu_k - \mathbf{m}_0)^T\Lambda_k(\mu_k - \mathbf{m}_0)] = \int \mathbb{E}_{\Lambda_k}[tr[\Lambda_k(\mu_k - \mathbf{m}_0)(\mu_k - \mathbf{m}_0)^T]]\mathcal{N}(\mu_k|\mathbf{m}_k,(\beta_k\Lambda_k)^{-1})d\mu_k$$

$$\mathbb{E}_{\mu_k}[\mu_k] = \mathbf{m}_k,\ \mathbb{E}_{\mu_k}[\mu_k\mu_k^T] = \mathbf{m}_k\mathbf{m}_k^T + (\beta_k\Lambda_k)^{-1}$$

$$\mathbb{E}_{\Lambda_k}[tr[\Lambda_k\beta_k^{-1}\Lambda_k^{-1} + \Lambda_k(\mathbf{m}_k - \mathbf{m}_0)(\mathbf{m}_k - \mathbf{m}_0)^T]] = D/\beta_k + tr[\underbrace{\nu_k\mathbf{W}_k}_{\mathbb{E}_{q(\Lambda_k)}[\Lambda_k]}(\mathbf{m}_k - \mathbf{m}_0)(\mathbf{m}_k - \mathbf{m}_0)^T]$$

The last term is $-\frac{1}{2}tr[\mathbf{W}_0^{-1}\mathbb{E}_{\Lambda_k}[\Lambda_k]] = -\frac{1}{2}tr[\nu_k\mathbf{W}_0^{-1}\mathbf{W}_k]$

$$p(\mu, \Lambda) = \mathcal{W}(\mu, \Lambda) = \prod_{k=1}^{K} p(\mu_k, \Lambda_k) = \prod_{k=1}^{K} p(\mu_k|\Lambda_k)p(\Lambda_k)$$

where

$$p(\mu_k|\Lambda_k) = \mathcal{N}(\mu_k|\mathbf{m}_0, (\beta_0\Lambda_k)^{-1})$$

and $p(\Lambda_k)$ is the Wishart distribution

$$p(\Lambda_k) = \mathcal{W}_0(\Lambda_k|\mathbf{W}_0, \nu_0) = \frac{|\Lambda_k|^{(\nu_0 - D - 1)}|e^{-\frac{1}{2}\,Trace(\mathbf{W}_0^{-1}\Lambda_k)})}{|\mathbf{W}_0|^{\nu_0/2}\left(2^{\nu_0 D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu_0+1-i}{2}\right)\right)}$$

$$\mathbb{E}_{\mu,\Lambda}\Big[\log p(\mu, \Lambda)\Big]$$

$$= \tfrac{1}{2}\textstyle\sum_{k=1}^{K}\left\{D\ \log(\beta_0/2\pi) + \log\tilde{\Lambda}_k - \frac{D\beta_0}{\beta_k} - \beta_0\nu_k(\mathbf{m}_k - \mathbf{m}_0)^T\mathbf{W}_k(\mathbf{m}_k - \mathbf{m}_0)\right\} +$$
$$K\ \log B(\mathbf{W}_0, \nu_0) + \tfrac{(\nu_0 - D - 1)}{2}\textstyle\sum_{k=1}^{K}\log\tilde{\Lambda}_k - \tfrac{1}{2}\sum_{k=1}^{K}\nu_k \mathrm{tr}(\mathbf{W}_0^{-1}\mathbf{W}_k)$$

Remembering that

$$q(\mu_k|\Lambda_k) = \prod_{k=1}^{K} \mathcal{N}(\mu_k|\mathbf{m}_k, \beta_k\Lambda_k) \text{ and } q(\Lambda_k) = \prod_{k=1}^{K} \mathcal{W}(\Lambda_k|\mathbf{W}_k, \nu_k)$$

$$E_{q(\mu_k,\Lambda_k)}\left[(\mu_k - \mathbf{m}_k)^T \beta_k\Lambda_k(\mu_k - \mathbf{m}_k)\right] = E_{q(\mu_k,\Lambda_k)} Tr[(\mu_k - \mathbf{m}_k)(\mu_k - \mathbf{m}_k)^T(\beta_k\Lambda_k)]$$
$$= E_{q(\mu_k,\Lambda_k)}[Tr(\mu_k\mu_k^T - 2\mu_k\mathbf{m}_k^T + \mathbf{m}_k\mathbf{m}_k^T)(\beta_k\Lambda_k)] = E_{q(\Lambda_k)} Tr[Cov(\mu_k)(\beta_k\Lambda_k)] = D$$

$$E_{q(\mu_k,\Lambda_k)}[\log(|\beta_k\Lambda_k|)] = E_{q(\mu_k,\Lambda_k)}[\log(|\beta_k\Lambda_k|)] = E_{q(\mu_k,\Lambda_k)}[D\log\beta_k + \log|\Lambda_k|]$$

$$\mathbb{E}_{\mu,\Lambda}\Big[\log q(\mu,\Lambda)\Big] = \textstyle\sum_{k=1}^{K}\left\{\frac{1}{2}\log\tilde{\Lambda}_k + \frac{D}{2}\log\left(\frac{\beta_k}{2\pi}\right) - \frac{D}{2} - H\big[q(\Lambda_k)\big]\right\}$$

$$\mathbb{E}_{\mathbf{Z}}\Big[\log q(\mathbf{Z})\Big] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\ \log r_{nk} \qquad q(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}}$$

$$\mathbb{E}_{\pi}\Big[\log q(\pi)\Big] = \log C(\alpha) + \sum_{k=1}^{K}(\alpha_k - 1)\ \log\tilde{\pi}_k$$

where $H\big[q(\Lambda_k)\big]$ , $C(\alpha)$ and $B(\mathbf{W},\nu)$ are defined as

$$H[q(\Lambda)] = -\int \log(q(\Lambda))q(\Lambda)d\Lambda$$

$$C(\alpha) = \frac{\Gamma(\hat{\alpha})}{\Gamma(\alpha_1)\ldots\Gamma(\hat{\alpha_K})} \text{ and } B(\mathbf{W},\nu) = |\mathbf{W}|^{-\nu/2}\left(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1}$$

## Variational Lower Bound $\mathcal{L}$

$$
\frac{1}{2}\sum_{k=1}^{K} N_k\Big\{ \log\ \tilde{\Lambda}_k - D\beta_k^{-1} - \nu_k\ \mathrm{tr}(\mathbf{S}_k\mathbf{W}_k) - \nu_k(\bar{\mathbf{y}}_k - \mathbf{m}_k)^T\mathbf{W}_k(\bar{\mathbf{y}}_k - \mathbf{m}_k)\Big\}
$$
$$
+(\alpha_0 - 1)\sum_{k=1}^{K} \log\tilde{\pi}_k + \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\ \log\tilde{\pi}_k
$$
$$
+\frac{1}{2}\sum_{k=1}^{K}\Big\{ \log\tilde{\Lambda}_k - \frac{D\beta_0}{\beta_k} - \beta_0\nu_k(\mathbf{m}_k - \mathbf{m}_0)^T\mathbf{W}_k(\mathbf{m}_k - \mathbf{m}_0)\Big\}
$$
$$
+\frac{(\nu_0 - D - 1)}{2}\sum_{k=1}^{K}\log\tilde{\Lambda}_k - \frac{1}{2}\sum_{k=1}^{K}\nu_k\mathrm{tr}(\mathbf{W}_0^{-1}\mathbf{W}_k)
$$
$$
+\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\ \log r_{nk} + \sum_{k=1}^{K}(\alpha_k - 1)\ \log\tilde{\pi}_k
$$
$$
+\sum_{k=1}^{K}\Big\{\frac{1}{2}\log\tilde{\Lambda}_k + \frac{D}{2}\log\Big(\frac{\beta_k}{2\pi}\Big) - H\big[q(\Lambda_k)\big]\Big\}
$$

$$
\log\rho_{nk} = \log\tilde{\pi}_k + \frac{1}{2}\log\tilde{\Lambda}_k - \frac{D}{2}\log(2\pi) \qquad r_{nk} = \rho_{nk}\Big/\sum_{j=1}^{K}\rho_{nj}
$$

$$
\nu_k = \nu_0 + N_k \qquad \alpha_k = \alpha_0 + \sum_{n=1}^{N} r_{nk} \qquad \beta_k = \beta_0 + N_k
$$

$$
\mathbf{m}_k = \frac{1}{\beta_k}(\beta_0\mathbf{m}_0 + N_k\bar{\mathbf{y}}_k)
$$
$$
\mathbf{W}_k^{-1} = \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{\mathbf{y}}_k - \mathbf{m}_0)(\bar{\mathbf{y}}_k - \mathbf{m}_0)^T
$$

$$
N_k = \sum_{n=1}^{N} r_{nk}, \quad \bar{\mathbf{y}}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{y}_n \text{ and } \mathbf{S}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{y}_n - \bar{\mathbf{y}}_k)(\mathbf{y}_n - \bar{\mathbf{y}}_k)^T
$$

## Self-Study Question

How that the entropy of $H[q(\lambda_k)]$ is

$$H[q(\lambda_k)] = -\int q(\lambda_k) \log q(\Lambda_k) d\Lambda_k$$

$$= -(\nu_k - D - 1)\log \tilde{\Lambda}_k + \frac{D\nu_k}{2} + \frac{\nu_k}{2}\log|W_k| + \frac{\nu_k D}{2}\log 2 + \frac{D(D-1)}{4}\log \pi - \sum_{i=1}^{D} \Gamma(\frac{\nu_k + 1 - i}{2})$$

## Self-Study Question

Generate $N = 600$ data points using a GMM model with $K = 5$, where

$\mu = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ -3 \end{bmatrix}, \begin{bmatrix} 3 \\ 3 \end{bmatrix} \begin{bmatrix} -3 \\ 3 \end{bmatrix}, \begin{bmatrix} -3 \\ -3 \end{bmatrix} \right\}$,

$\Lambda^{-1} = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \right\}$.

and $\pi = \{0.2, 0.3, 0.3, 0.1, 0.1\}$ .

Using the variational Bayes, estimate the distribution parameters and the lower bound $\mathcal{L}$ by running the algorithm for $k = 2, 3, \ldots, 10$ and show that the $\mathcal{L}$ is minimized for the model with $K = 5$.