Imperial College
London

NOTES

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

495 Advanced Statistical Machine Learning and Pattern Recognition

Author:
Thomas Teh (CID: 0124 3008)

Date: January 31, 2017

1 Expectation Maximization

1.1 General Approach for Expectation Maximization

1.1.1 Notes

1. The goal of the Expectation-Maximization is to find maximum likelihood solutions for models having latent variables
2. The general concept is that since our knowledge of the latent variables in $\boldsymbol{Z}$ is given by the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{Z},\boldsymbol{\theta})$, we use the expectation of the latent variables instead of the actual values.
3. EM algorithm can be used to find MAP solutions models in which a prior is defined over the parameters.

1.1.2 Algorithm

Given a joint distribution $p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$ over observed variables $\boldsymbol{X}$ and latent variables $\boldsymbol{Z}$, governed by parameter $\boldsymbol{\theta}$ the goal is to maximize the likelihood function $p(\boldsymbol{X}|\boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$.

1. Choose an initial for the parameters $\boldsymbol{\theta}^{old}$.
2. E Step: Evaluate $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old})$
3. M Step: Evaluate $\boldsymbol{\theta}^{new}$ given by

   $$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$$

   where

   $$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$$

4. Check for convergence of either the log-likelihood or the parameter values. If convergence is not satisfied

   $$\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$$

   and return to step 2

1.2 Gaussian Mixture Models

The Gaussian mixture distribution can be written as a linear superposition of Gaussians

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k,\Sigma_k)$$

1.2.1 Formulation of the Gaussian Mixture Models

Let $\boldsymbol{z}$ be a K-dimensional binary random variable with 1-of-K representation.

$$p(z_k = 1) = \pi_k \Rightarrow p(\boldsymbol{z}) = \prod_{i=1}^{K} \pi_k^{z_k}$$

Conditional probability of $\boldsymbol{x}$ given a particular value for latent variable $\boldsymbol{z}$:

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)^{z_k}$$

Using Bayes theorem and marginalize the latent variable $\boldsymbol{z}$

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

Similarly, the posterior probability of $\boldsymbol{z}$ is given by

$$\gamma(z_k) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{x})} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)}$$

1.2.2 Maximum Likelihood

The log-likelihood function is given by

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k) \right)$$

The maximum likelihood method will yield the following:

$$\begin{aligned}
\boldsymbol{\mu}_k &= \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \\
\Sigma_k &= \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top}{\sum_{n=1}^{N} \gamma(z_{nk})} \\
\pi_k &= \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}
\end{aligned}$$

Issues with maximum likelihood:

1. Presence of singularities: When we have at least two components in the mixture, one of them can have a finite variance and assign finite probability to all the data points, while the other component can shrink onto one specific data point and therefore contribute to an ever increasing additive value to the log likelihood.

2. Identifiability: Solutions may not be unique, hence it may be hard to interpret the parameter values discovered by a model.

3. The log likelihood equation is difficult to optimize over.

1.2.3 Expectation Maximization Formulation

Supposed that in addition to $\boldsymbol{X}$, we were also given the values of the latent variables $\boldsymbol{Z}$, the likelihood and the log likelihood function are given by

$$p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)^{z_{nk}}$$
$$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\Big(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k\Big)$$

1. **Expectation Step**:

   Taking the expectation on the log likelihood

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \Sigma, \boldsymbol{\pi})] &= \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}[z_{nk}]\Big(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k\Big) \\ &= G(\boldsymbol{\theta})\end{aligned}$$

$$p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K} [\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)]^{z_{nk}}$$

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}[z_{nk}] &= \frac{\sum_{z_{nk}} z_{nk}[\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)]^{z_{nk}}}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \Sigma_j)} \\ &= \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \Sigma_j)} \\ &= \gamma(z_{nk})\end{aligned}$$

2. **Maximization Step**: By taking the derivative of $(\boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$ and set them to 0, the parameters can be found to be

$$\begin{aligned}\boldsymbol{\mu}_k &= \frac{\sum_{n=1}^{N} \gamma(z_{nk})\boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \\ \Sigma_k &= \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top}{\sum_{n=1}^{N} \gamma(z_{nk})} \\ \pi_k &= \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}\end{aligned}$$

1.2.4 Expectation Maximization Algorithm

Given a Gaussian mixture model, the goal is to maximize the likelihood functions w.r.t to the parameters:

1. Initialize the means $\boldsymbol{\mu}_k$, covariances Σ_k and mixing coefficients π_k and evaluate the initial of the log likelihood.

2. E Step: Evaluate the responsibilities using the current parameter values

$$\gamma(z_k) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{x})} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)}$$

3. M Step: Re-estimate the parameters using the current responsibilities

$$\begin{aligned}
\boldsymbol{\mu}_k^{new} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n \\
\Sigma_k^{new} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \\
\pi_k^{new} &= \frac{N_k}{N}
\end{aligned}$$

   where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

4. Evaluate the log likelihood

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k) \right)$$

   and check for convergence of either parameter of the log likelihood. If the convergence criterion is not satisfied, return to step 2

1.3 Bernoulli Mixture Models

$$\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\mu}) &= \prod_{i=1}^{D} \mu_i^{x_i} (1 - \mu_i)^{1 - x_i} \\
\mathbb{E}[\boldsymbol{x}] &= \boldsymbol{\mu} \\
\mathbb{V}[\boldsymbol{x}] &= diag\{\mu_i(1 - \mu_i)\}
\end{aligned}$$

The mixture of the Bernoulli distributions is given by

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k)$$
$$p(\boldsymbol{x}|\boldsymbol{\mu}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki}^{x_i})^{1-x_i}$$
$$\mathbb{E}[\boldsymbol{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k$$
$$\mathbb{V}[\boldsymbol{x}] = \sum_{k=1}^{K} \pi_k \left(\Sigma_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top\right) - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\top$$
$$\Sigma_k = diag\{\mu_{ki}(1-\mu_{ki})\}$$

Let $\boldsymbol{z}$ be the one hot representation, we have

$$p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\mu}) = \prod_{k=1}^{K} p(\boldsymbol{x}|\boldsymbol{\mu}_k)^{z_k}$$
$$p(\boldsymbol{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

The log likelihood function is given by

$$\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\mu},\boldsymbol{\pi}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left\{\ln \pi_k + \sum_{i=1}^{D}[x_{ni}\ln \mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})]\right\}$$

The E step: we then take the expectation of the log likelihood above

$$\mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\mu},\boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})\left\{\ln \pi_k + \sum_{i=1}^{D}[x_{ni}\ln \mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})]\right\}$$

$$\gamma(z_{nk}) = \frac{\sum_{z_{nk}} z_{nk}[\pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k)]^{z_{nk}}}{\sum_{z_{nj}}[\pi_j p(\boldsymbol{x}_n|\boldsymbol{\mu}_j)]^{z_{nj}}} = \frac{\pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k)}{\sum_{j=1}^{K}\pi_j p(\boldsymbol{x})_{\boldsymbol{n}}|\boldsymbol{\mu}_j)}$$

The M step: we then maximize the log likelihood wrt to the parameters

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})$$
$$\pi_k = \frac{N_k}{N}$$
$$N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

1.4 Convergence of the EM Algorithm

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) \Rightarrow \ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + KL(q \parallel p)$$

$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln\left[\frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right]$$

$$KL(q \parallel p) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln\left[\frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right]$$

In the E step, the lower bound $\mathcal{L}(q,\boldsymbol{\theta}^{old})$ is maximized with respect to $q(\boldsymbol{Z})$. The solution to this maximization problem can be done by setting $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old})$ such that $KL(q \parallel p) = 0$.

In the M step, the lower bound $\mathcal{L}(q,\boldsymbol{\theta}^{old})$ is maximized with respect to $\boldsymbol{\theta}$. However, the KL divergence will no longer be zero. Hence, the increase in the upper bound of the log likelihood function is greater than the increase in the lower bound.

By substituting $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old})$, we get the lower bound to be the following after the E step:

$$\begin{aligned}\mathcal{L}(q,\boldsymbol{\theta}) &= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) - \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \\ &= \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) + const\end{aligned}$$

2 Probabilistic PCA

$$\begin{aligned}\boldsymbol{x} &= \boldsymbol{W}\boldsymbol{y} + \boldsymbol{\mu} + \boldsymbol{e} \\ \boldsymbol{e} &\sim \mathcal{N}(\boldsymbol{e}|\boldsymbol{0},\sigma^2\boldsymbol{I}) \\ \boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{y}|\boldsymbol{0},\boldsymbol{I})\end{aligned}$$

2.1 Maximum Likelihood

1. We are interested in the distribution of $\boldsymbol{x}$ and the parameters $\boldsymbol{\theta}$. The marginal distribution of $\boldsymbol{x}$ can be obtained by integrating $\boldsymbol{y}$ out of the joint distribution.

$$\begin{aligned}p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N|\boldsymbol{\theta}) &= \int_{\boldsymbol{y}_i} p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N|\boldsymbol{\theta}) d\boldsymbol{y}_1 \ldots d\boldsymbol{y}_N \\ &= \int_{\boldsymbol{y}_i} \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{W},\boldsymbol{\mu},\sigma) p(\boldsymbol{y}_i) d\boldsymbol{y}_1 \ldots d\boldsymbol{y}_N \\ &= \prod_{i=1}^{N} \int_{\boldsymbol{y}_i} p(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{W},\boldsymbol{\mu},\sigma) p(\boldsymbol{y}_i) d\boldsymbol{y}_i\end{aligned}$$

2. The joint distribution can be obtained by applying Bayes' Rule.

$$p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{W},\boldsymbol{\mu},\sigma)p(\boldsymbol{y}_i)$$
$$p(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{W},\boldsymbol{\mu},\sigma)p(\boldsymbol{y}_i) = \left[(2\pi\sigma^2)^{-\frac{F}{2}} e^{-\frac{1}{2\sigma^2}(\boldsymbol{x}_i-\boldsymbol{\mu}-\boldsymbol{W}\boldsymbol{y}_i)^\top(\boldsymbol{x}_i-\boldsymbol{\mu}-\boldsymbol{W}\boldsymbol{y}_i)}\right]\left[(2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\boldsymbol{y}_i{}^\top\boldsymbol{y}_i}\right]$$

3. By collecting the $\boldsymbol{y}$ in the exponents, completing the squares and then applying the Woodbury identity (more details please refer to notes in 496):

$$p(\boldsymbol{x}_i|\boldsymbol{W},\boldsymbol{\mu},\sigma^2) = \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu},\boldsymbol{D})$$
$$\boldsymbol{D} = \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}$$
$$p(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{W},\boldsymbol{\mu},\sigma^2) = \mathcal{N}(\boldsymbol{y}_i|\boldsymbol{M}^{-1}\boldsymbol{W}^\top(\boldsymbol{x}_i-\boldsymbol{\mu}),\sigma^2\boldsymbol{M}^{-1})$$
$$\boldsymbol{M} = \sigma^2\boldsymbol{I} + \boldsymbol{W}^\top\boldsymbol{W}$$

4. Maximizing the likelihood and solve for the parameters

$$\boldsymbol{S}_t = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top \text{ ($\boldsymbol{S}_t$ is the covariance matrix)}$$
$$\sigma^2 = \frac{1}{F-d}\sum_{j=d+1}^{F}\lambda_j$$
$$\boldsymbol{W}_d = \boldsymbol{U}_d(\boldsymbol{\Lambda}-\sigma^2\boldsymbol{I})^{\frac{1}{2}}\boldsymbol{V}^\top$$

5. Hence we no longer have a projection but:

$$\mathbb{E}_{p(\boldsymbol{y}_i|\boldsymbol{x}_i)}[\boldsymbol{y}_i] = \boldsymbol{M}^{-1}\boldsymbol{W}^\top(\boldsymbol{x}_i-\boldsymbol{\mu})$$
$$\widehat{\boldsymbol{x}_i} = \boldsymbol{W}\mathbb{E}_{p(\boldsymbol{y}_i|\boldsymbol{x}_i)}[\boldsymbol{y}_i] + \boldsymbol{\mu}$$

2.2 EM PPCA

1. Write out the log likelihood of the joint distribution of the observed variable and the latent variable

$$p(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{\theta})p(\boldsymbol{y}_i)$$
$$\ln p(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta}) = \sum_{i=1}^{N}(\ln p(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{\theta}) + \ln p(\boldsymbol{y}_i))$$
$$\ln p(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{\theta}) = -\frac{F}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{x}_i-\boldsymbol{W}\boldsymbol{y}_i-\boldsymbol{\mu})^\top(\boldsymbol{x}_i-\boldsymbol{W}\boldsymbol{y}_i-\boldsymbol{\mu})$$
$$\ln p(\boldsymbol{y}_i) = -\frac{1}{2}\boldsymbol{y}_i^\top\boldsymbol{y}_i - \frac{D}{2}\ln 2\pi$$

2. Take the expectation on the log likelihood on the joint distribution:

$$\ln p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \left[-\frac{F}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} (\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{y}_i - \boldsymbol{\mu})^\top (\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{y}_i - \boldsymbol{\mu}) - \frac{1}{2} \boldsymbol{y}_i^\top \boldsymbol{y}_i - \frac{D}{2} \ln 2\pi \right.$$

$\mathbb{E}_{p(\boldsymbol{Y}|\boldsymbol{X})}[\ln p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta})] =$