

# Research on *Variational Message Passing*

—CSE692 Final Report

Wei Hu  
SUNY SB-CS

whu@cs.sunysb.edu

December 16, 2007

## Abstract

In this report, I will introduce a powerful variational inference method named *Variational Message Passing (VMP)* and show some results using this method. First, I will talk about the fundamentals of variational inference. Then I will introduce the bases of VMP—*exponential family* and *conjugate-exponential model* and arrive at the VMP algorithm. Next, a specific graphical model—*Latent Dirichlet Allocation (LDA)* will be explored to show how to use VMP. Finally, VMP will be run on some toy examples and results reported.

## 1 Introduction

Graphical model is a powerful probabilistic tool in many research areas such as computer vision [7], natural language processing [1] and data mining [6]. Among many issues related to graphical models, *inference* is the most important one. In some simple cases, exact inference algorithms [2] such as junction tree and sum-product can be used. But in many real applications, exact inference is intractable so that we have to turn to approximation inference. All the approximation inference algorithms fall into one of the two categories [5]: sampling algorithms and variational algorithms. In recent years, variational inference attracts more and more interests because of its strong theoretical basis and high potential on developing efficient algorithms.

So for my final project, I wanted to learn some knowledge about graphical model, especially about variational inference on graphical models. In [8], a very powerful variational inference method called *Variational Message Passing (VMP)* was proposed. It can widely used on many different kinds of graphical models and its efficiency is high. So in this final project, I first learned VMP and related topics deeply, and then derived VMP on a popular generative model called *Latent Dirichlet Allocation* and point out that the algorithm procedure proposed in an ICCV07 paper [3] is totally wrong. Notice that this part involved a great deal of hard work and was entirely finished by myself. Finally, I run VMP on Gaussian Mixture Models and compared VMP with the classical Expectation Maximization method.

## 2 Fundamentals of variational inference

### 2.1 Notations

Most of the notations in this report are borrowed from [8]. A directed graph  $G$  can be expressed as  $G = (\mathbf{X}, \mathbf{E})$ , where  $\mathbf{X} = \{X_i\}$  is the set of nodes in the graph and  $\mathbf{E}$  is the set of edges. Normally, every node in the graph corresponds to a random variable (univariate or multivariate), so  $\mathbf{X}$  can also be used to denote the set of random variables. By the definition of conditional independence in graphical model, we have

$$P(\mathbf{X}) = \prod_i P(X_i | \text{pa}_i) \quad (1)$$

Among all the random variables in the graph, some are *visible variables* (also called *observed variables*), denoted as  $\mathbf{V}$ ; and the others are *hidden variables*, denoted as  $\mathbf{H}$ . So  $\mathbf{X} = \{\mathbf{V}, \mathbf{H}\}$ . Inference is in general to obtain the posterior distribution  $P(\mathbf{H}|\mathbf{V})$ .

### 2.2 Basic Ideas of Variational Inference

In many cases,  $P(\mathbf{H}|\mathbf{V})$  cannot be calculated exactly, but we can try to approximate it. So the goal of variational inference is to find a tractable variational distribution  $Q(\mathbf{H})$  that closely approximates the true posterior distribution  $P(\mathbf{H}|\mathbf{V})$ . The “difference” of the two distribution can be measured with Kullback-Leibler divergence:  $KL(Q \parallel P) = \int_{\mathbf{H}} Q(\mathbf{H}) \log \frac{Q(\mathbf{H})}{P(\mathbf{H}|\mathbf{V})}$ . So now the goal is to minimize this KL divergence in terms of  $Q$ .

$$\begin{aligned} KL(Q \parallel P) &= \int_{\mathbf{H}} Q(\mathbf{H}) \log \frac{Q(\mathbf{H})}{P(\mathbf{H}|\mathbf{V})} \\ &= \int_{\mathbf{H}} Q(\mathbf{H}) \log \frac{Q(\mathbf{H})}{P(\mathbf{H}, \mathbf{V})} + \int_{\mathbf{H}} Q(\mathbf{H}) \log P(\mathbf{V}) \\ &= \mathcal{L}(Q) + \log(P(\mathbf{V})) \end{aligned}$$

where

$$\mathcal{L}(Q) = \int_{\mathbf{H}} Q(\mathbf{H}) \log \frac{Q(\mathbf{H})}{P(\mathbf{H}, \mathbf{V})} \quad (2)$$

is called Helmholtz free energy.

Obviously,  $\arg_Q \min KL(Q \parallel P) = \arg_Q \min(\mathcal{L}(Q) + \log(P(\mathbf{V}))) = \arg_Q \min \mathcal{L}(Q)$ , so minimizing  $KL(Q \parallel P)$  equals to minimizing Helmholtz free energy  $\mathcal{L}(Q)$ . In the following sections, we are going to focus on minimizing  $\mathcal{L}(Q)$ .

### 2.3 Factorized Variational Distribution

Choose a variational distribution from factorized distribution family, i.e. let  $Q(\mathbf{H})$  have the following form

$$Q(\mathbf{H}) = \prod_i Q_i(\mathbf{H}_i) \quad (3)$$

where  $\{\mathbf{H}_i\}$  are the disjoint groups of variables, and every  $Q_i(\mathbf{H}_i)$  is a probability density function. Substituting (3) into (2), we can get

$$\begin{aligned}
\mathcal{L}(\mathbf{Q}) &= \int_{\mathbf{H}} \left( \prod_j Q_j(\mathbf{H}_j) \right) \sum_i \log Q_i(\mathbf{H}_i) - \int_{\mathbf{H}} \left( \prod_j Q_j(\mathbf{H}_j) \right) \log P(\mathbf{H}, \mathbf{V}) \\
&= \int \cdots \int_{\mathbf{H}_1 \cdots \mathbf{H}_{i-1} \mathbf{H}_{i+1} \cdots \mathbf{H}_I} \left( \prod_{j \neq i} Q_j(\mathbf{H}_j) \right) \int_{\mathbf{H}_i} \sum_i Q_i(\mathbf{H}_i) \log Q_i(\mathbf{H}_i) \\
&\quad - \int_{\mathbf{H}_i} Q_i(\mathbf{H}_i) \left( \int \cdots \int_{\mathbf{H}_1 \cdots \mathbf{H}_{i-1} \mathbf{H}_{i+1} \cdots \mathbf{H}_I} \left( \prod_{j \neq i} Q_j(\mathbf{H}_j) \right) \log P(\mathbf{H}, \mathbf{V}) \right) \\
&= \sum_i \int_{\mathbf{H}_i} Q_i(\mathbf{H}_i) \log Q_i(\mathbf{H}_i) - \int_{\mathbf{H}_j} Q_j(\mathbf{H}_j) \langle \log P(\mathbf{H}, \mathbf{V}) \rangle_{\sim Q_j(\mathbf{H}_j)} \\
&= \int_{\mathbf{H}_j} Q_j(\mathbf{H}_j) \log Q_j(\mathbf{H}_j) - \int_{\mathbf{H}_j} Q_j(\mathbf{H}_j) \langle \log P(\mathbf{H}, \mathbf{V}) \rangle_{\sim Q_j(\mathbf{H}_j)} + \text{terms not in } Q_j \\
&= \int_{\mathbf{H}_j} Q_j(\mathbf{H}_j) \log \frac{Q_j(\mathbf{H}_j)}{Q_j^*(\mathbf{H}_j)} + \text{terms not in } Q_j \\
&= \text{KL}(Q_j \parallel Q_j^*) + \text{terms not in } Q_j
\end{aligned}$$

where  $\sim Q_j(\mathbf{H}_j) = \prod_{i \neq j} Q_i(\mathbf{H}_i)$  and  $\log Q_j^*(\mathbf{H}_j) = \langle \log P(\mathbf{H}, \mathbf{V}) \rangle_{\sim Q_j(\mathbf{H}_j)}$

So by fixing  $Q_1^{\text{old}}(\mathbf{H}_1), \dots, Q_{j-1}^{\text{old}}(\mathbf{H}_{j-1}), Q_{j+1}^{\text{old}}(\mathbf{H}_{j+1}), \dots, Q_I^{\text{old}}(\mathbf{H}_I)$ , we can minimize  $\mathcal{L}(\mathbf{Q})$  in terms of  $Q_j$ . And by the above formula, we can see when  $Q_j = Q_j^* + \text{terms not in } Q_j$ ,  $\mathcal{L}(\mathbf{Q})$  gets its minimum.

Then how to compute  $\log Q_j^*(\mathbf{H}_j)$ ? For simplicity, let every hidden variable group only contains one hidden variable, i.e.  $\mathbf{H}_j = \{H_j\}$ . From (1), we have

$$\begin{aligned}
\log Q_j^*(\mathbf{H}_j) &= \langle \log P(\mathbf{H}, \mathbf{V}) \rangle_{\sim Q_j(\mathbf{H}_j)} \\
&= \left\langle \sum_i \log P(X_i | \text{pa}_i) \right\rangle_{\sim Q_j(\mathbf{H}_j)} \\
&= \langle \log P(H_j | \text{pa}_j) \rangle_{\sim Q_j(H_j)} + \sum_{k \in \text{ch}_j} \langle \log P(X_k | \text{pa}_k) \rangle_{\sim Q_j(H_j)} \\
&\quad + \text{terms not in } Q_j
\end{aligned}$$

Here is the intuition. Due to conditional independency, if we want to update the PDF of a node  $\mathbf{H}_j$  (i.e.  $Q_j(H_j)$ ), we only have to consider its *Markov Blanket*, i.e., its parents  $\text{pa}_j$ , its children  $\{X_k\}, k \in \text{ch}_j$  and its co-parents  $\{\text{pa}_k\}, k \in \text{ch}_j$ . All the other nodes will result in terms not related to  $Q_j$ . So we can simply let:

$$\log Q_j^*(\mathbf{H}_j) = \langle \log P(H_j | \text{pa}_j) \rangle_{\sim Q_j(H_j)} + \sum_{k \in \text{ch}_j} \langle \log P(X_k | \text{pa}_k) \rangle_{\sim Q_j(H_j)} \quad (4)$$

The two terms in the right hand side are still not easy to compute in general cases. In the next two sections, we will investigate a special conditional distribution family called *exponential family*

and a special model called *conjugate-exponential model*. They have several great properties which can be used to compute the two terms in a efficient manner.

## 3 Exponential Family and Conjugate-Exponential Model

### 3.1 Exponential Family

A conditional distribution is in the exponential family if it can be written in this form:

$$P_{\mathbf{X}|\Theta}(\mathbf{X} = \mathbf{x}|\Theta) = \exp(\vec{\eta}(\Theta)^T \vec{\mathbf{u}}(\mathbf{x}) + h(\mathbf{x}) + g(\Theta)) \quad (5)$$

where  $\vec{\eta}(\Theta)$  is called the *natural parameter vector* and  $\vec{\mathbf{u}}(\mathbf{x})$  is called the *natural statistic vector*. Notice that random variable  $\mathbf{X}$  can be univariate or multivariate and the dimension of  $\vec{\mathbf{u}}(\mathbf{x})$  is not necessarily equal to  $\mathbf{X}$ . See Example 1.

#### Example 1. Gaussian distribution

The probability density function of Gaussian distribution is:

$$\begin{aligned} P(X = x|\mu, \sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \\ &= \exp\left(\left[\begin{array}{c} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{array}\right]^T \begin{bmatrix} x \\ x^2 \end{bmatrix} - \frac{1}{2} \left(\frac{\mu^2}{\sigma^2} + 2 \log \sigma\right) - \frac{1}{2} \log 2\pi\right) \end{aligned}$$

Here we can see

$$\vec{\eta}(\mu, \sigma) = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}, \vec{\mathbf{u}}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}, h(x) = -\frac{1}{2} \log 2\pi, g(\mu, \sigma) = -\frac{1}{2} \left(\frac{\mu^2}{\sigma^2} + 2 \log \sigma\right)$$

□

Many common distributions are in the exponential family, such as Dirichlet distribution (see Example 2), Multinomial distribution (see Example 3), Gamma distribution, Poisson distribution and so on.

#### Example 2. Dirichlet distribution

With parameters  $\alpha_1, \dots, \alpha_K > 0 (K \geq 2)$ , and constrains:  $x_1, \dots, x_K > 0, \sum_i x_i = 1$ , the probability density function of the Dirichlet distribution is:

$$\begin{aligned} P(\vec{\mathbf{X}} = (x_1, \dots, x_K) | \alpha_1, \dots, \alpha_K) &= \frac{1}{\mathbf{B}(\vec{\alpha})} \prod_{i=1}^K x_i^{\alpha_i - 1} \\ &= \exp\left(\sum_{i=1}^K \alpha_i \log x_i - \sum_{i=1}^K \log x_i - \log \mathbf{B}(\vec{\alpha})\right) \end{aligned}$$

where the normalizing constant  $\mathbf{B}(\vec{\alpha})$  is the multinomial beta function, which can be expressed in terms of the gamma function:

$$\mathbf{B}(\vec{\alpha}) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}$$

Here we can see

$$\vec{\eta}(\vec{\alpha}) = \vec{\alpha}, \vec{\mathbf{u}}(\vec{\mathbf{x}}) = \log \vec{\mathbf{x}}, h(\vec{\mathbf{x}}) = - \sum_{i=1}^K \log x_i, g(\vec{\alpha}) = - \log \mathbf{B}(\vec{\alpha})$$

□

**Example 3. Multinomial distribution**

With parameters  $\theta_1, \dots, \theta_K \geq 0, \sum_i \theta_i = 1 (K \geq 2)$ , and constrains:  $\sum_i x_i = N$ , the probability mass function of the Multinomial distribution is:

$$\begin{aligned} \mathbf{P}(\vec{\mathbf{X}} = (x_1, \dots, x_K) | N, \theta_1, \dots, \theta_K) &= \frac{N!}{\prod_{i=1}^K x_i!} \prod_{i=1}^K \theta_i^{x_i} \\ &= \exp\left(\sum_{i=1}^K x_i \log \theta_i + \log N! - \sum_{i=1}^K \log x_i!\right) \\ &= \exp\left(\sum_{i=1}^{K-1} x_i \log \theta_i + (N - \sum_{i=1}^{K-1} x_i) \log(\theta_K) + \log N! - \sum_{i=1}^K \log x_i!\right) \\ &= \exp\left(\sum_{i=1}^{K-1} x_i \log(\theta_i/\theta_K) + N \log(\theta_K) + \log N! - \sum_{i=1}^K \log x_i!\right) \end{aligned}$$

Here we can see

$$\vec{\eta}(N, \vec{\theta}) = \begin{bmatrix} \log(\frac{\theta_1}{\theta_K}) \\ \vdots \\ \log(\frac{\theta_{K-1}}{\theta_K}) \end{bmatrix}, \vec{\mathbf{u}}(\vec{\mathbf{x}}) = \begin{bmatrix} x_1 \\ \vdots \\ x_{K-1} \end{bmatrix}, h(\vec{\mathbf{x}}) = - \sum_{i=1}^K \log x_i!, g(N, \vec{\theta}) = N \log(\theta_K) + \log N!$$

Notice that univariate discrete distribution is a special case of Multinomial distribution, where  $N = 1$

□

The exponential family has several good properties. In the following, I will show some that are helpful in inference.

**Property 1. Expectation of natural statistic vector**

It is very easy and efficient to calculate the expectation of natural statistic vector. Reparameterize  $\Theta$  to  $\vec{\eta}$  in (5), we can get:

$$\mathbf{P}_{\mathbf{X}|\vec{\eta}}(\mathbf{X} = \mathbf{x}|\vec{\eta}) = \exp(\vec{\eta}^T \vec{\mathbf{u}}(\mathbf{x}) + h(\mathbf{x}) + \tilde{g}(\vec{\eta}))$$

Integrate the above formula with respect to  $\mathbf{X}$ , we have

$$\int_{\mathbf{X}} \exp(\vec{\eta}^T \vec{\mathbf{u}}(\mathbf{X}) + h(\mathbf{X}) + \tilde{g}(\vec{\eta})) d\mathbf{X} = \int_{\mathbf{X}} \mathbf{P}_{\mathbf{X}|\vec{\eta}}(\mathbf{X}|\vec{\eta}) d\mathbf{X} = 1$$

then differentiate with respect to  $\vec{\eta}$

$$\int_{\mathbf{X}} \frac{d}{d\vec{\eta}} \exp(\vec{\eta}^T \vec{\mathbf{u}}(\mathbf{X}) + h(\mathbf{X}) + \tilde{g}(\vec{\eta})) d\mathbf{X} = 0$$

$$\int_{\mathbf{X}} \mathbf{P}_{\mathbf{X}|\vec{\eta}}(\mathbf{X}|\vec{\eta}) \left[ \vec{\mathbf{u}}(\mathbf{X}) + \frac{d\tilde{g}(\vec{\eta})}{d\vec{\eta}} \right] d\mathbf{X} = 0$$

And so the expectation of the natural statistic vector is given by

$$\langle \vec{\mathbf{u}}(\mathbf{X}) \rangle_{\mathbf{P}_{\mathbf{X}|\vec{\eta}}(\mathbf{X}|\vec{\eta})} = -\frac{d\tilde{g}(\vec{\eta})}{d\vec{\eta}} \quad (6)$$

□

**Example 4.** *Expectation of natural statistic vector in Dirichlet distribution*

In Example 2, we have  $\vec{\eta}(\vec{\alpha}) = \vec{\alpha}$ , so  $\tilde{g}(\vec{\eta}) = g(\vec{\alpha}) = -\log \mathbf{B}(\vec{\alpha})$ , then

$$-\frac{d\tilde{g}(\vec{\eta})}{d\vec{\eta}} = \frac{d \log \mathbf{B}(\vec{\alpha})}{d\vec{\alpha}} = \frac{d}{d\vec{\alpha}} \left( \sum_{i=1}^K \log \Gamma(\alpha_i) - \log \Gamma\left(\sum_{i=1}^K \alpha_i\right) \right)$$

$$= \begin{bmatrix} \Psi(\alpha_1) - \Psi(\alpha_0) \\ \vdots \\ \Psi(\alpha_K) - \Psi(\alpha_0) \end{bmatrix}$$

where  $\alpha_0 = \sum_{i=1}^K \alpha_i$ .

By (6) and  $\vec{\mathbf{u}}(\vec{\mathbf{X}}) = \log \vec{\mathbf{X}}$ , we have

$$\langle \log X_i \rangle = \Psi(\alpha_i) - \Psi(\alpha_0), i = 1, \dots, K$$

□

**Example 5.** *Expectation of natural statistic vector in Multinomial distribution*

In Example 3, we have  $\vec{\eta}(N, \vec{\theta}) = \left[ \log\left(\frac{\theta_1}{\theta_K}\right), \dots, \log\left(\frac{\theta_{K-1}}{\theta_K}\right) \right]^T$ . By letting  $\eta_i = \log\left(\frac{\theta_i}{\theta_K}\right)$ ,  $i = 1, \dots, K-1$ , we can get

$$\theta_i = \frac{\exp(\eta_i)}{1 + \sum_{j=1}^{K-1} \exp(\eta_j)}, i = 1, \dots, K-1$$

This is the reparameterization.

$$g(N, \vec{\theta}) = N \log(\theta_K) + \log N! \Rightarrow \tilde{g}(\vec{\eta}) = -N \log\left(1 + \sum_{j=1}^{K-1} \exp(\eta_j)\right) + \log N!$$

So

$$-\frac{d\tilde{g}(\vec{\eta})}{d\vec{\eta}} = \left[ N \frac{\exp(\eta_1)}{1 + \sum_{j=1}^{K-1} \exp(\eta_j)}, \dots, N \frac{\exp(\eta_{K-1})}{1 + \sum_{j=1}^{K-1} \exp(\eta_j)} \right]^T = [N\theta_1, \dots, N\theta_{K-1}]^T$$

By (6) and  $\vec{\mathbf{u}}(\vec{\mathbf{X}}) = [X_1, \dots, X_{K-1}]^T$ , we have

$$\langle X_i \rangle = N\theta_i, i = 1, \dots, K-1$$

and  $\langle X_K \rangle = \langle N - \sum_{j=1}^{K-1} X_j \rangle = N - N \sum_{j=1}^{K-1} \theta_j = N\theta_K$ .

□

## 3.2 Conjugate-Exponential Model

A parent distribution  $P(\mathbf{Y}|\text{pa}_{\mathbf{Y}})$  is said to be conjugate to a child distribution  $P(\mathbf{X}|\mathbf{Y}, \text{cp}_{\mathbf{Y}})$  if  $P(\mathbf{Y}|\text{pa}_{\mathbf{Y}})$  has the same functional form, with respect to  $\mathbf{Y}$ , as  $P(\mathbf{X}|\mathbf{Y}, \text{cp}_{\mathbf{Y}})$ . And if the parent distribution and the child distribution are both in exponential family, we say that they form a *conjugate-exponential model*. By (5), we have:

$$\begin{aligned}\log P(\mathbf{Y}|\text{pa}_{\mathbf{Y}}) &= \vec{\eta}_{\mathbf{Y}}(\text{pa}_{\mathbf{Y}})^{\top} \vec{\mathbf{u}}_{\mathbf{Y}}(\mathbf{Y}) + h_{\mathbf{Y}}(\mathbf{Y}) + g_{\mathbf{Y}}(\text{pa}_{\mathbf{Y}}) \\ \log P(\mathbf{X}|\mathbf{Y}, \text{cp}_{\mathbf{Y}}) &= \vec{\eta}_{\mathbf{X}}(\mathbf{Y}, \text{cp}_{\mathbf{Y}})^{\top} \vec{\mathbf{u}}_{\mathbf{X}}(\mathbf{X}) + h_{\mathbf{X}}(\mathbf{X}) + g_{\mathbf{X}}(\mathbf{Y}, \text{cp}_{\mathbf{Y}})\end{aligned}$$

By the definition of conjugate, we can rewrite the second formula as:

$$\log P(\mathbf{X}|\mathbf{Y}, \text{cp}_{\mathbf{Y}}) = \vec{\eta}_{\mathbf{X}\mathbf{Y}}(\mathbf{X}, \text{cp}_{\mathbf{Y}})^{\top} \vec{\mathbf{u}}_{\mathbf{Y}}(\mathbf{Y}) + g_{\mathbf{X}\mathbf{Y}}(\mathbf{X}, \text{cp}_{\mathbf{Y}})$$

**Example 6.** *Dirichlet distribution is conjugate to Multinomial distribution*

By Example 2 and 3, we can easily see that Dirichlet distribution is conjugate to Multinomial distribution.  $\square$

## 4 Variational Message Passing

### 4.1 Optimization of Q

Now we arrive at a right point that (4) can be calculated. Substituting a conjugate-exponential model in terms of  $Y$  into (4), we have:

$$\begin{aligned}\log Q_Y^*(Y) &= \langle \vec{\eta}_{\mathbf{Y}}(\text{pa}_{\mathbf{Y}})^{\top} \vec{\mathbf{u}}_{\mathbf{Y}}(Y) + h_{\mathbf{Y}}(Y) + g_{\mathbf{Y}}(\text{pa}_{\mathbf{Y}}) \rangle_{\sim Q_Y(Y)} \\ &\quad + \sum_{k \in \text{ch}_{\mathbf{Y}}} \langle \vec{\eta}_{\mathbf{X}\mathbf{Y}}(X_k, \text{cp}_k)^{\top} \vec{\mathbf{u}}_{\mathbf{Y}}(Y) + g_{\mathbf{X}\mathbf{Y}}(X_k, \text{cp}_k) \rangle_{\sim Q_Y(Y)} \\ &= \left[ \langle \vec{\eta}_{\mathbf{Y}}(\text{pa}_{\mathbf{Y}}) \rangle_{\sim Q_Y(Y)} + \sum_{k \in \text{ch}_{\mathbf{Y}}} \langle \vec{\eta}_{\mathbf{X}\mathbf{Y}}(X_k, \text{cp}_k) \rangle_{\sim Q_Y(Y)} \right]^{\top} \vec{\mathbf{u}}_{\mathbf{Y}}(Y) + h_{\mathbf{Y}}(Y) + \text{terms not in } Q_j\end{aligned}$$

It follows that  $Q_Y^*$  is an exponential family distribution of the same form as  $P(\mathbf{Y}|\text{pa}_{\mathbf{Y}})$  but with a natural parameter vector

$$\vec{\eta}_{\mathbf{Y}}^* = \langle \vec{\eta}_{\mathbf{Y}}(\text{pa}_{\mathbf{Y}}) \rangle_{\sim Q_Y(Y)} + \sum_{k \in \text{ch}_{\mathbf{Y}}} \langle \vec{\eta}_{\mathbf{X}\mathbf{Y}}(X_k, \text{cp}_k) \rangle_{\sim Q_Y(Y)} \quad (7)$$

Moreover, we can reparameterise  $\langle \vec{\eta}_{\mathbf{Y}}(\text{pa}_{\mathbf{Y}}) \rangle_{\sim Q_Y(Y)}$  and  $\sum_{k \in \text{ch}_{\mathbf{Y}}} \langle \vec{\eta}_{\mathbf{X}\mathbf{Y}}(X_k, \text{cp}_k) \rangle_{\sim Q_Y(Y)}$  in terms of their corresponding dependent variables, i.e.:

$$\langle \vec{\eta}_{\mathbf{Y}}(\text{pa}_{\mathbf{Y}}) \rangle_{\sim Q_Y(Y)} = \widetilde{\vec{\eta}}_{\mathbf{Y}}(\{\langle \vec{\mathbf{u}}_i \rangle_{Q_i(i)}\}_{i \in \text{pa}_{\mathbf{Y}}}) \quad (8)$$

$$\langle \vec{\eta}_{\mathbf{X}\mathbf{Y}}(X_k, \text{cp}_k) \rangle_{\sim Q_Y(Y)} = \widetilde{\vec{\eta}}_{\mathbf{X}\mathbf{Y}}(\langle \vec{\mathbf{u}}_k \rangle_{Q_k(k)}, \{\langle \vec{\mathbf{u}}_j \rangle_{Q_j(j)}\}_{j \in \text{cp}_k}) \quad (9)$$

Now updating every  $Q_Y(Y)$  will be easy and efficient.

## 4.2 VMP Algorithm

Based on (8) and (9), we can define two different kind of *message* on the conjugate-exponential model.

1) The message from a parent node  $Y$  to a child node  $X$  is given by:

$$\mathbf{m}_{Y \rightarrow X} = \langle \vec{\mathbf{u}}_Y \rangle_{Q_Y(Y)} \quad (10)$$

2) The message from a child node  $X$  to a parent node  $Y$  is given by:

$$\mathbf{m}_{X \rightarrow Y} = \widetilde{\eta}_{XY} \left( \langle \vec{\mathbf{u}}_X \rangle_{Q_X(X)}, \{\mathbf{m}_{i \rightarrow X}\}_{i \in \text{cp}_Y} \right) \quad (11)$$

Then by (7), we have:

$$\vec{\eta}_Y^* = \widetilde{\eta}_Y \left( \{\mathbf{m}_{i \rightarrow Y}\}_{i \in \text{pa}_Y} \right) + \sum_{j \in \text{ch}_Y} \mathbf{m}_{j \rightarrow Y} \quad (12)$$

We have now reached the point where VMP algorithm can be define, see Algorithm 1.

---

**Algorithm 1** The variational message passing algorithm

---

- 1: Initialize each factor distribution  $Q_{X_j}$  by initializing the corresponding moment vector  $\langle \mathbf{u}_{X_j}(X_j) \rangle_{Q_{X_j}(X_j)}$
  - 2: For each node  $X_j$  in turn,
    - Retrieve messages from all parent and child nodes, as defined in (10) and (11).
    - Compute updated natural parameter vector  $\vec{\eta}_{X_j}^*$  using (12).
    - Compute updated moment vector  $\langle \mathbf{u}_{X_j}(X_j) \rangle_{Q_{X_j}(X_j)}$  given the new setting of the parameter vector.
- 

## 5 Latent Dirichlet Allocation—A Real Application

In this section, we will explore how VMP algorithm can be used in LDA. LDA is a generative model widely used in text analysis and computer vision and I will introduce it first.

### 5.1 Introduction of LDA

Latent Dirichlet Allocation (LDA) was first proposed in [4]. In LDA model, the observations are  $D$  documents and  $R_d$  words in the  $d$ -th ( $d = 1, \dots, D$ ) document. It assumes the following generative process for each word  $w^{d,r}$  ( $d = 1, \dots, D, r = 1, \dots, R_d$ ).

1. For each document  $d$ , choose topic distribution parameters from a Dirichlet distribution:  $\vec{\theta}_d \sim \text{Dirichlet}(\vec{\alpha})$ .
2. Choose topic-word distribution parameters from a Dirichlet distribution:  $\beta_{i,j} \sim \text{Dirichlet}(\lambda_{i,j})$ .

3. For each word  $w^{d,r}$  in document  $d$ :

- (a) Choose a topic from the topic distribution:  $z^{d,r} \sim \text{Multinomial}(\vec{\theta}_d)$ .
- (b) Choose the word  $w^{d,r}$  from the topic-word distribution given the corresponding topic  $z^{d,r}$ :  $w^{d,r} \sim \text{P}(w^{d,r} | z^{d,r}, \beta) = \text{Multinomial}(\beta_{:,z^{d,r}})$ .

where  $\beta_{:,z^{d,r}}$  is the  $z^{d,r}$ -th column of the matrix  $\beta_{W \times K}$ .  $W$  is the number of possible words,  $K$  is the number of possible topics.

## 5.2 VMP on LDA

Let us see how VMP is used on LDA.

1. Initialize  $\langle \mathbf{u}_{X_j}(\vec{X}_j) \rangle_{Q_{X_j}(X_j)}$ .

- (a) For  $\vec{\theta}^d$ . From 4,  $\mathbf{m}_{\vec{\alpha} \rightarrow \vec{\theta}^d} = \langle \vec{\mathbf{u}}_{\vec{\theta}^d}(\vec{\theta}^d) \rangle = \langle \log \vec{\theta}^d \rangle = [\Psi(\alpha_1) - \Psi(\alpha_0), \dots, \Psi(\alpha_K) - \Psi(\alpha_0)]^T$ .
- (b) For  $z^{d,r}$ . From 5,  $\mathbf{m}_{\vec{\theta}^d \rightarrow z^{d,r}} = \langle \vec{\mathbf{u}}_{z^{d,r}}(z^{d,r}) \rangle = [\theta_1^d, \dots, \theta_K^d]$ .
- (c) For  $\beta$ . Similar to  $\vec{\theta}^d$ , we have,  $[\mathbf{m}_{\lambda \rightarrow \beta}]_{i,j} = \langle \log \beta_{i,j} \rangle = \Psi(\lambda_{i,j}) - \Psi(\lambda_0)$ .

2. Update natural parameter vector.

- (a) For  $\vec{\theta}^d$ . Because it doesn't have co-parents, so the update is simple:

$$\alpha_k \leftarrow \alpha_k + D * R_d * \theta_k^d, k = 1, \dots, K$$

- (b) For  $z^{d,r}$ . It only has one parent  $\vec{\theta}^d$  and  $\widetilde{\eta}_{z^{d,r}} = \eta_{z^{d,r}}$ , so the first part of updating formula (7) will be right  $\mathbf{m}_{\vec{\alpha} \rightarrow \vec{\theta}^d}$ .  $z^{d,r}$  has a co-parent  $\beta$ , and  $\log \text{P}(w^{d,r} | z^{d,r}, \beta) = \log^T \beta_{w^{d,r},:} * [\delta(z^{d,r} = 1), \dots, \delta(z^{d,r} = K)]^T$ , so the second part of (7) will be  $[\mathbf{m}_{\lambda \rightarrow \beta}]_{w^{d,r},:}$ . So the update is given by:

$$\theta_k \leftarrow [\Psi(\alpha_k) - \Psi(\alpha_0)] + [\Psi(\lambda_{w^{d,r},k}) - \Psi(\lambda_0)]$$

- (c) For  $\beta$ . Similar to updating  $\vec{\theta}^d$ , we have:

$$\lambda_{i,k} \leftarrow \lambda_{i,k} + \sum_d \sum_r^{R_d} \delta(w^{d,r} = i) \theta_k$$

**Based on these, we can see that the procedure described in [3] is wrong. I doubt how the authors got the good results and why such a paper got published.**

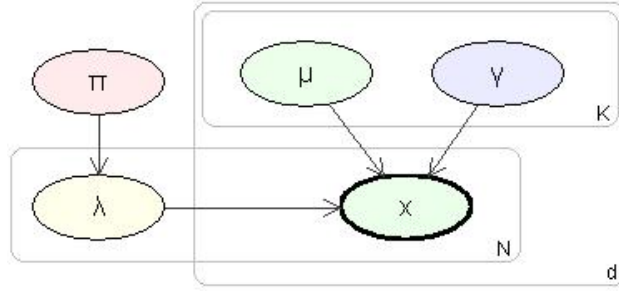


Figure 1: Gaussian Mixture Model under the framework of Variational Message Passing

## 6 Experiments on Gaussian Mixture Models

The derivation of VMP on Gaussian Mixture Models (GMM) is simpler compared to the derivation on LDA shown above, so we omit it here.

To use VMP, we have to convert the classical GMM into a conjugate-exponential model. The new version of GMM is shown in Figure 1, where  $x$ 's are the observation data points,  $\mu$ 's are the means and  $\gamma$ 's are the precisions,  $\lambda$ 's are hidden variables indicating which component the data point is drawn from and  $\pi$  is the parameters of a Dirichlet distribution. And  $d$  is the dimension of sample space,  $N$  is the number of data points,  $K$  is the number of components.

I run 500 trails, in all of which  $d = 2$ ,  $N = 1000$ ,  $K = 10$ . For each trail, the true parameters were randomly generated and fixed afterwards. With the true parameters, 1000 points were then drawn from the GMM as the observations. VMP and classical Expectation Maximization (EM) were run on them respectively. Some results are shown in Figure ???.

It is not easy to compare the results obtained by these two algorithms. I came up a measurement called *average relative error rate* of the means (ARER\_mean). It is defined as follows:

1. For each true mean of each trail, find out all the estimated means near it, and then calculate the Euclidean distances between them.
2. For each estimated mean of any trail, find out all the true means near it, and then calculate the Euclidean distances between them. Redundancy should be first removed here.
3. Sum up all the above distances, divide the result by the number of true means, and we get a relative error rate for each trail.
4. Average all the relative error rates to get the average relative error rate of the means.

The ARER\_mean of VMP is 0.12, while the ARER\_mean of EM is 0.19. That means VMP can get good estimation of means than EM does.

I couldn't come up any measurement to compare the results in terms of precisions. But finally, in terms of efficiency, VMP is also better than EM. Because the average running time of VMP is only 5.7 seconds while the average running time of EM is 10.2 seconds.

In conclusion, VMP is better than EM on GMM in terms of accuracy of means and efficiency.

## 7 Conclusions

In this report, I reported how I did my final project in detail. First, I studied hard and learned well the Variational Message Passing algorithm and related topics. Then I explored how VMP can be used in LDA, a popular but complex model. I did all the math by myself and made a proof that there are mistakes in an ICCV07 paper [3]. Finally, I run VMP in some toy examples and compare it with EM and showed that VMP is better than EM.

## References

- [1] J. A. Bilmes. References for: Graphical model research in audio, speech and language processing. Technical report, Department of Electrical Engineering, University of Washington, 2003. [1](#)
- [2] C. M. Bishop. *Pattern Recognition and Machine Learning*, chapter 8. 2002. [1](#)
- [3] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *Proceedings of IEEE international conference on computer vision*, 2007. [1](#), [5.2](#), [7](#)
- [4] M. I. J. David M. Blei, Andrew Y. Ng. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. [5.1](#)
- [5] M. I. Jordan. *Graphical models*, 2003. [1](#)
- [6] R. Kruse and C. Borgelt. *Data Mining with Graphical Models*, volume 2534/2002, pages 259–287. 2002. [1](#)
- [7] J. M. Rehg, V. Pavlovic, T. S. Huang, and W. T. Freeman. Guest editors’ introduction to the special section on graphical models in computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):785–786, 2003. [1](#)
- [8] J. Winn and C. M. Bishop. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005. [1](#), [2.1](#)