NPFL108 - Bayesian inference

Filip Jurcicek

UFAL, MFF, Charles University in Prague
Malostranske namesti 25, Prague, 14300, Czech republic

May 16, 2014

Abstract

This work follows the NPFL108 course on Bayesian inference. In NPFL108, a student is introduced into basics of Bayesian inference. The presented techniques include analytic solutions for the simplest cases, the Variational Inference and Expectation Propagation algorithms, Markov Chain Monte Carlo (MCMC) techniques, represented by Gibbs sampling.

Contents

1 Introduction

In this work, the Bayesian models will be described using graphical models. A graphical model represents variables as nodes and dependencies between them as oriented edges. In the graphical representation in this article, the observed variables are filled with light blue colour and the unobserved (hidden) variables are filled with white colour. In addition, the fixed parameters of the priors are filled with the light red colour.

An important problem in graphical models is the process of finding the probability distributions of unobserved (some times called latent or hidden) variables given the observed variables. This process is called inference. This work introduces basic methods of Bayesian inference on a set of simple examples. The presented techniques include analytic solutions for the simplest cases, the Variational Inference and Expectation Propagation algorithms, Markov Chain Monte Carlo (MCMC) techniques, represented by Gibbs sampling.

This work is based on the lecture notes for the NPFL108 - Bayesian inference course. Some derivations and/or results were obtained or corrected with the help of the students of the course. Namely: Ondřej Dušek, Matěj Korvas, ...

1.1 Bayes rule, prior, likelihood, joint distribution, and posterior

2 Tractable Bayesian Inference

2.1 Example: The binomial distribution

Graphical model Figure 1 depict a graphical model representing inference of a parameters a Bernoulli distribution. There are N observations $D = \{x_1, \ldots, x_N\}$ and we aim to compute the posterior distribution for the proportion - hidden variable θ given the observation and prior. Note that the observations have values 0 or 1, where 0 usually stands for failure and 1 for success.

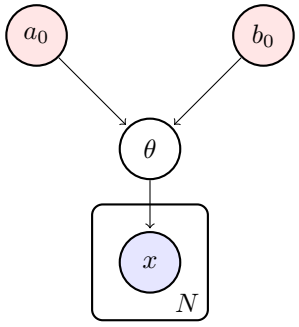


Figure 1: The graphical model representing the joint distribution for the observations $x_1, \ldots, x_N$ and the proportion θ.

Generative process The model depicted on Figure 1 assumes the following generative process:

1. $\theta \sim Beta(\cdot|a_0, b_0)$
2. $x_i \sim Bern(\cdot|\theta) \quad \forall i \in \{1, \ldots, N\}$

where the parameters a_0 and b_0 are priors set manually and x_i is an observed variable.

Before continuing, please recall that:

$$Beta(\theta|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1},$$
$$Bern(x|\theta) = \theta^x(1-\theta)^{1-x}.$$

Please note that:

$$Bern(x = 1|\theta) = \theta.$$

Prior As noted above, the prior is formed by the Beta distribution $Beta(\theta|a, b)$.

Likelihood The likelihood is defined as follow:

$$\begin{aligned} p(D|\boldsymbol{w}) &= p(\boldsymbol{x}|\theta) \\ &= \prod_i p(x_i|\theta) \\ &= \prod_i Bern(x_i|\theta) \\ &= \prod_i \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_i x_i}(1-\theta)^{\sum_i (1-x_i)}. \end{aligned}$$

Joint distribution The joint distribution is defined as follows:

$$\begin{aligned}
p(\boldsymbol{w}, D) &= p(\boldsymbol{x}, \theta) \\
&= p(\theta) \prod_i p(x_i|\theta) \\
&= Beta(\theta|a_0, b_0) \prod_i Bern(x_i|\theta) \\
&= \frac{\Gamma(a_0 + b_0)}{\Gamma(a_0)\Gamma(b_0)} \theta^{a_0 - 1}(1-\theta)^{b_0 - 1} \prod_i \theta^{x_i}(1-\theta)^{1-x_i} \\
&= \frac{\Gamma(a_0 + b_0)}{\Gamma(a_0)\Gamma(b_0)} \theta^{a_0 - 1}(1-\theta)^{b_0 - 1} \theta^{\sum_i x_i}(1-\theta)^{\sum_i (1-x_i)} \\
&= \frac{\Gamma(a_0 + b_0)}{\Gamma(a_0)\Gamma(b_0)} \theta^{a_0 - 1 + \sum_i x_i}(1-\theta)^{b_0 - 1 + \sum_i (1-x_i)}.
\end{aligned}$$

Posterior The posterior distribution is defined as follows:

$$p(\boldsymbol{w}|D) = \frac{1}{p(D)} p(\boldsymbol{w}, D),$$

where

$$p(D) = \int p(\boldsymbol{w}, D) d\boldsymbol{w}.$$

However instead of solving the normalisation constant $p(D)$, we will inspect the likelihood whether it has a form of some convenient distribution.

$$\begin{aligned}
p(\boldsymbol{w}|D) &\propto p(\boldsymbol{w}, D) \\
&\propto p(\boldsymbol{x}, \theta) \\
&\propto \frac{\Gamma(a_0 + b_0)}{\Gamma(a_0)\Gamma(b_0)} \theta^{a_0 - 1 + \sum_i x_i}(1-\theta)^{b_0 - 1 + \sum_i (1-x_i)}
\end{aligned}$$

Since the Bernoulli and Beta distributions are conjugate, one can observe trough introspection that the posterior has the form of a Beta distribution:

$$\begin{aligned}
p(\boldsymbol{w}|D) &= p(\theta|a_N, b_N) \\
&= Beta(\theta|a_N, b_N) \\
a_N &= a_0 - 1 + \sum_i x_i \\
b_N &= b_0 - 1 + \sum_i (1 - x_i)
\end{aligned}$$

To simplify the result, one can denote $\sum_i x_i$ as number of successes n_1 and $\sum_i (1 - x_i)$ as number of failures n_0. Then one gets:

$$\begin{aligned}
p(\boldsymbol{w}|D) &= Beta(\theta|a_N, b_N) \\
a_N &= a_0 + n_1 - 1 \\
b_N &= b_0 + n_0 - 1
\end{aligned}$$

2.2 Example: The multinomial distribution

Graphical model Figure 2 depict a graphical model representing inference of a parameters a Multinomial distribution. There are N observations $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ and we aim to compute the posterior distribution for the proportion - hidden variable $\boldsymbol{\theta}$ given the observation and prior. Note that the observations $\boldsymbol{x}_i$ are d dimensional vectors where all values are 0 except one being set to 1. The position of 1 indicates the cardinal value of the observation, e.g. $[0, 0, 0, 1, 0]$ represents value 3 from $0, \ldots, 4$.

Generative process The model depicted on Figure 2 assumes the following generative process:

1. $\boldsymbol{\theta} \sim Dir(\cdot|\boldsymbol{\alpha}_0)$
2. $\boldsymbol{x}_i \sim Multi(\cdot|\boldsymbol{\theta}) \quad \forall i \in \{1, \ldots, N\}$

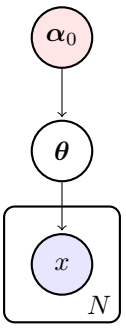


Figure 2: The graphical model representing the joint distribution for the observations $x_1, \ldots, x_N$ and the θ parameter.

where the parameters $\boldsymbol{\alpha}$ is a vector of priors set manually and $\boldsymbol{x}_i$ are observed variables.

Before continuing, please recall that:

$$Dir(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{j=1}^{d}\Gamma(\alpha_j)}\prod_{j=1}^{d}\theta_j^{\alpha_j-1}.$$

Where:

$$\alpha_j \geq 0,$$
$$\alpha_0 = \sum_{j=1}^{d}\alpha_j.$$

$$Multi(\boldsymbol{x}|\boldsymbol{\theta}) = \prod j = 1^d \theta_j^{x_j},$$

where:

$$\theta_j \geq 0,$$
$$\sum_j \theta_j = 1.$$

Please note that:

$$P(x = j|\boldsymbol{\theta}) = Multi(x = [\ldots, 1, \ldots]|\theta) = \boldsymbol{\theta}_j$$

Prior As noted above, the prior is formed by the Dirichlet distribution $Dir(\boldsymbol{\theta}|\boldsymbol{\alpha})$.

Likelihood The likelihood is defined as follow:

$$\begin{aligned} p(D|\boldsymbol{w}) &= p(\boldsymbol{x}|\boldsymbol{\theta}) \\ &= \prod_i p(\boldsymbol{x}_i|\boldsymbol{\theta}) \\ &= \prod_i Multi(\boldsymbol{x}_i|\boldsymbol{\theta}) \\ &= \prod_i \prod_{j=1}^{d} \theta_j^{x_{ij}} \\ &= \prod_{j=1}^{d} \theta_j^{\sum_i x_{ij}} \end{aligned}$$

Joint distribution The joint distribution is defined as follows:

$$\begin{aligned}
p(\boldsymbol{w}, D) &= p(\boldsymbol{x}, \theta) \\
&= p(\theta) \prod_i p(x_i|\theta) \\
&= Dir(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_i Multi(\boldsymbol{x}_i|\boldsymbol{\theta}) \\
&= \frac{\Gamma(\alpha_0)}{\prod_{j=1}^d \Gamma(\alpha_j)} \prod_{j=1}^d \theta_j^{\alpha_j - 1} \prod_{j=1}^d \theta_j^{\sum_i x_{ij}} \\
&= \frac{\Gamma(\alpha_0)}{\prod_{j=1}^d \Gamma(\alpha_j)} \prod_{j=1}^d \theta_j^{\alpha_j + \sum_i x_{ij} - 1}
\end{aligned}$$

Posterior The posterior distribution is defined as follows:

$$\begin{aligned}
p(\boldsymbol{w}|D) &= \frac{1}{p(D)} p(\boldsymbol{w}, D), \\
&\propto p(\boldsymbol{w}, D) \\
&\propto p(\boldsymbol{x}, \boldsymbol{\theta}) \\
&\propto \frac{\Gamma(\alpha_0)}{\prod_{j=1}^d \Gamma(\alpha_j)} \prod_{j=1}^d \theta_j^{\alpha_j + \sum_i x_{ij} - 1}
\end{aligned}$$

Since the Multinomial and Dirichlet distributions are conjugate, one can observe trough introspection that the posterior has the form of a Dirichlet distribution:

$$\begin{aligned}
p(\boldsymbol{w}|D) &= p(\boldsymbol{\theta}|\boldsymbol{\alpha}_N) \\
&= Dir(\boldsymbol{\theta}|\boldsymbol{\alpha}_N) \\
\alpha_{Nj} &= \alpha_{0j} + \sum_i x_i - 1
\end{aligned}$$

To simplify the result, one can denote $\sum_i x_{ij}$ as n_j . Then one gets:

$$\begin{aligned}
p(\boldsymbol{w}|D) &= Dir(\boldsymbol{\theta}|\boldsymbol{\alpha}_N) \\
\alpha_{Nj} &= \alpha_{0j} + n_j - 1 \\
\alpha_{N0} &= \sum_{j=1}^d \alpha_{Nj}
\end{aligned}$$

2.3 Example: The normal distribution

Let us start with a simple problem of Bayesian inference for the normal distribution. We will study four situations:

1. Inference of the mean while the variance is known
2. Inference of the variance while the mean is known
3. Inference of the mean and variance using conjugate prior
4. Inference of the mean and variance using non-conjugate prior

In all cases, we aim to compute the posterior distributions for the unknown variables, e.g. the mean and variance.

2.3.1 Inference of the mean while the variance is known

Figure 3 depict a graphical model representing inference of a mean of an univariate normal distribution assuming that the variance σ^2 is known. There is one observation x and we aim to compute the posterior distribution for the hidden variable μ given the observation and prior.

The model depicted on Figure 3 assumes the following generative process:

1. $\mu \sim N(\cdot|\mu_0, \sigma_0^2)$
2. $x \sim N(\cdot|\mu, \sigma^2)$

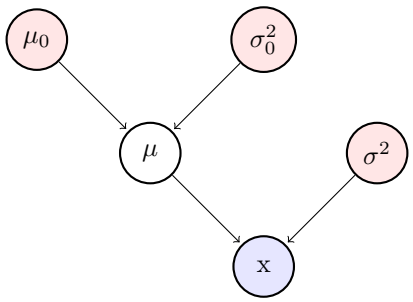


Figure 3: The graphical model for prediction of the observation x given the mean μ where the mean μ is unknown.

where the parameters μ_0, σ_0^2, and σ^2 are priors set manually and x is an observed variable.

To compute the posterior for μ, we must define the joint distribution for x and μ given the manually set prior parameters μ_0, σ_0^2 and known variance σ^2 of the observations:

$$p(x, \mu|\sigma^2, \mu_0, \sigma_0^2) = p(x|\mu, \sigma^2)p(\mu|\mu_0, \sigma_0^2), \tag{1}$$

where

$$p(x|\mu, \sigma^2) = N(x|\mu, \sigma^2) \tag{2}$$
$$p(\mu|\mu_0, \sigma_0^2) = N(\mu|\mu_0, \sigma_0^2) \tag{3}$$

Note that we use the normal distribution as a prior for the mean. The important property of this prior is that it is conjugate to the normal distribution used to model the probability of the observation.

Further, we have to compute the posterior of the μ. Using the Bayes rule, we get:

$$p(x, \mu|\sigma^2, \mu_0, \sigma_0^2) = p(\mu|x, \sigma^2, \mu_0, \sigma_0^2)p(x|\sigma^2, \mu_0, \sigma_0^2) \tag{4}$$

Therefore, the posterior for the mean μ is:

$$p(\mu|x, \sigma^2, \mu_0, \sigma_0^2) = \frac{p(x, \mu|\sigma^2, \mu_0, \sigma_0^2)}{p(x|\sigma^2, \mu_0, \sigma_0^2)} \tag{5}$$
$$= \frac{p(x, \mu|\sigma^2, \mu_0, \sigma_0^2)}{\int_\mu p(x, \mu|\sigma^2, \mu_0, \sigma_0^2)d\mu} \tag{6}$$

Substituting (1) into (6), we get:

$$p(\mu|x, \sigma^2, \mu_0, \sigma_0^2) = \frac{p(x|\mu, \sigma^2)p(\mu|\mu_0, \sigma_0^2)}{\int_\mu p(x|\mu, \sigma^2)p(\mu|\mu_0, \sigma_0^2)d\mu} \tag{7}$$

Since all factors in the dividend and divisor in (7) contain μ, the fraction cannot be further simplified. To compute the posterior of μ, lets substitute (2) and (3) into (7):

$$p(\mu|x, \sigma^2, \mu_0, \sigma_0^2) = \frac{N(x|\mu, \sigma^2)N(\mu|\mu_0, \sigma_0^2)}{\int_\mu N(x|\mu, \sigma^2)N(\mu|\mu_0, \sigma_0^2)d\mu} = N(\mu|\mu_x, \sigma_x^2) \tag{8}$$

Given that the prior of μ is a normal distribution and therefore conjugate with a normal distribution used for modelling the observation x, the posterior is again a normal distribution which will be denoted as $N(\mu|\mu_x, \sigma_x^2)$. However, to avoid the integration in the divisor, it is easier to compute only the dividend and then by completing the squares compute the full posterior. Before continuing, lets us recall the definition of the normal distribution:

$$N(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}. \tag{9}$$

Using this definition and substituting it into (8), we get:

$$
\begin{aligned}
p(\mu|x,\sigma^2,\mu_0,\sigma_0^2) &\propto N(x|\mu,\sigma^2)N(\mu|\mu_0,\sigma_0^2) \\
&\propto \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\frac{1}{(2\pi\sigma_0^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\} \\
&\propto \frac{1}{(2\pi\sigma^2)^{1/2}}\frac{1}{(2\pi\sigma_0^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(x^2-2x\mu+\mu^2)-\frac{1}{2\sigma_0^2}(\mu^2-2\mu\mu_0+\mu_0^2)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}x^2-\frac{1}{\sigma^2}2x\mu+\frac{1}{\sigma^2}\mu^2+\frac{1}{\sigma_0^2}\mu^2-\frac{1}{\sigma_0^2}2\mu\mu_0+\frac{1}{\sigma_0^2}\mu_0^2\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\left[\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right]\mu^2-2\mu\left[\frac{1}{\sigma^2}x+\frac{1}{\sigma_0^2}\mu_0\right]+\frac{1}{\sigma^2}x^2+\frac{1}{\sigma_0^2}\mu_0^2\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\mu^2-2\mu\frac{1}{\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}}\left[\frac{1}{\sigma^2}x+\frac{1}{\sigma_0^2}\mu_0\right]+\frac{1}{\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}}\left[\frac{1}{\sigma^2}x^2+\frac{1}{\sigma_0^2}\mu_0^2\right]\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\mu^2-2\mu\left[\frac{\sigma_0^2}{\sigma_0^2+\sigma^2}x+\frac{\sigma^2}{\sigma^2+\sigma_0^2}\mu_0\right]+\frac{\sigma_0^2}{\sigma_0^2+\sigma^2}x^2+\frac{\sigma^2}{\sigma^2+\sigma_0^2}\mu_0^2\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\mu^2-2\mu\left[\frac{\sigma_0^2}{\sigma_0^2+\sigma^2}x+\frac{\sigma^2}{\sigma^2+\sigma_0^2}\mu_0\right]\right)\right\} \qquad (10)
\end{aligned}
$$

Note that $\frac{\sigma_0^2}{\sigma_0^2+\sigma^2}x^2+\frac{\sigma^2}{\sigma^2+\sigma_0^2}\mu_0^2$ is independent of μ and therefore a constant. Consequently, it can be omitted. As described in (8), the posterior $p(\mu|x,\sigma^2,\mu_0,\sigma_0^2)$ has the form of $N(\mu|\mu_x,\sigma_x^2)$ and it is proportionate to (10).

$$
p(\mu|x,\sigma^2,\mu_0,\sigma_0^2)=N(\mu|\mu_x,\sigma_x^2)\propto\exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\mu^2-2\mu\left[\frac{\sigma_0^2}{\sigma_0^2+\sigma^2}x+\frac{\sigma^2}{\sigma^2+\sigma_0^2}\mu_0\right]\right)\right\} \qquad (11)
$$

Completing the squares of the exponent in (11), a careful reader can notice that:

$$
\mu_x=\frac{\sigma_0^2}{\sigma_0^2+\sigma^2}x+\frac{\sigma^2}{\sigma_0^2+\sigma^2}\mu_0 \qquad (12)
$$

$$
\frac{1}{\sigma_x^2}=\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2} \qquad (13)
$$

So far, we assumed that there is only one observation x. However, the introduced approach can be used in a similar way even if there is $x_1,\ldots,x_N$ observations. Figure 4 depicts a graphical model which explicitly expresses the multiple observations. One can imagine that if there are more observations then the graphical representation can become cluttered, therefore multiple repeated nodes are expressed more compactly in a form of one node in *a plate* labelled with a number indicating the number of times the node should be replicated.

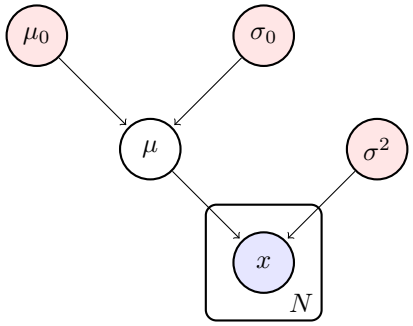


Figure 4: The graphical model representing the joint distribution for the observations $x_1,\ldots,x_N$ and the mean μ.

In case of multiple observations $x_1,\ldots,x_N$, the joint distribution is defined as:

$$
\begin{aligned}
p(\boldsymbol{x},\mu|\sigma^2,\mu_0,\sigma_0^2) &= p(\mu|\mu_0,\sigma_0^2)\prod_{i=1}^{N}p(x_i|\mu,\sigma^2) \\
&= N(\mu|\mu_0,\sigma_0^2)\prod_{i=1}^{N}N(x_i|\mu,\sigma^2) \qquad (14)
\end{aligned}
$$

where $\boldsymbol{x} = \{x_1, \ldots, x_N\}$. Using similar technique as in (10), one can derive that posterior of the mean μ is:

$$
\begin{aligned}
p(\mu|\boldsymbol{x}, \sigma^2, \mu_0, \sigma_0^2) &= N(\mu|\mu_N, \sigma_N^2) \\
\mu_N &= \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 \\
\frac{1}{\sigma_N^2} &= \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}
\end{aligned}
\tag{15}
$$

where $\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is empirical mean.

Sometimes, it is convenient to use the precision instead of the variance since it can significantly simplify the calculations and the result. Since the precision is defined as

$$
\lambda = \frac{1}{\sigma^2}, \tag{16}
$$

the results using the precision is

$$
\begin{aligned}
p(\mu|\boldsymbol{x}, \lambda, \mu_0, \lambda_0) &= N(\mu|\mu_N, \lambda_N^{-1}) \\
\mu_N &= \frac{N\lambda\mu_{ML} + \lambda_0\mu_0}{N\lambda + \lambda_0} \\
\lambda_N &= N\lambda + \lambda_0
\end{aligned}
\tag{17}
$$

2.3.2 Inference of the variance while the mean is known

Figure 5 depict a graphical model representing inference of a variance of an univariate normal distribution assuming that the variance σ^2 is known. This is the opposite situation when compared to the case presented in the previous section. There are N observations $x_1, \ldots, x_N$ and we aim to compute the posterior distribution for the hidden variable σ^2 given the observations and prior.

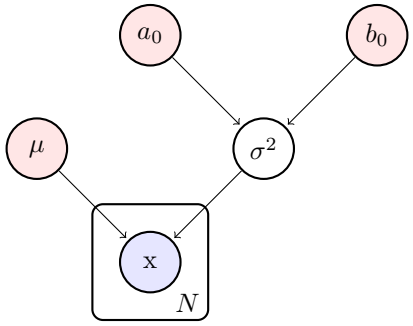


Figure 5: The graphical model for prediction of the observations $x_1, \ldots, x_N$ given the mean μ where the variance σ^2 is unknown.

To get an analytical solution, we require the prior for the σ^2 parameter to be conjugate with the normal distribution. The calculations will be greatly simplified if we work with precision λ instead of the variance σ^2. The graphical model using precision is at Figure 6.

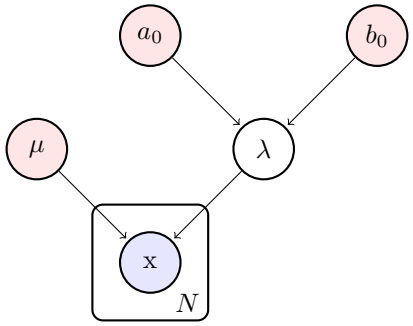


Figure 6: The graphical model for prediction of the observations $x_1, \ldots, x_N$ given the mean μ where the precision λ is unknown.

The model depicted on Figure 6 assumes the following generative process:

1. $\lambda \sim Gam(\cdot|a_0, b_0)$

2. $x_i \sim N(\cdot|\mu, \lambda^{-1}) \quad \forall i \in \{1, \ldots, N\}$

where the parameters μ_0, a_0, and b_0 are priors set manually and x is an observed variable.

Substitution (16) into (9), the normal distribution using the precision is as follows:

$$N(x|\mu, \sigma^2) = N(x|\mu, \lambda^{-1}) = \frac{\lambda^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\lambda}{2}(x-\mu)^2\right\} \propto \lambda^{1/2} \exp\left\{-\frac{\lambda}{2}(x-\mu)^2\right\} \tag{18}$$

The conjugate prior for the precision λ is the *gamma* distribution defined by:

$$Gam(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp\{-b\lambda\} \propto \lambda^{a-1} \exp\{-b\lambda\}. \tag{19}$$

The joint distribution represented by the graphical model at Figure 6 is given by:

$$\begin{aligned} p(\boldsymbol{x}, \lambda|\mu, a_0, b_0) &= p(\lambda|a_0, b_0) \prod_{i=1}^{N} p(x_i|\mu, \lambda^{-1}) \\ &= Gam(\lambda|a_0, b_0) \prod_{i=1}^{N} N(x_i|\mu, \lambda^{-1}) \end{aligned}$$

where $\boldsymbol{x} = \{x_1, \ldots, x_N\}$. The posterior of the precision λ is therefore derived as follows:

$$\begin{aligned} p(\lambda|\boldsymbol{x}, \mu, a_0, b_0) &= \frac{Gam(\lambda|a_0, b_0) \prod_{i=1}^{N} N(x_i|\mu, \lambda^{-1})}{\int_\lambda Gam(\lambda|a_0, b_0) \prod_{i=1}^{N} N(x_i|\mu, \lambda^{-1}) d\lambda} \\ &\propto Gam(\lambda|a_0, b_0) \prod_{i=1}^{N} N(x_i|\mu, \lambda^{-1}) \\ &\propto \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda^{a_0-1} \exp\{-b_0\lambda\} \prod_{i=1}^{N} \lambda^{1/2} \exp\left\{-\frac{\lambda}{2}(x_i-\mu)^2\right\} \\ &\propto \lambda^{a_0-1} \exp\{-b_0\lambda\} \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right\} \\ &\propto \lambda^{a_0+N/2-1} \exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right\} \\ &\propto \lambda^{[a_0+N/2]-1} \exp\left\{-\left[b_0 + \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right]\lambda\right\} \end{aligned}$$

Since the normal and gamma distributions are conjugate, the posterior for the precision λ has the form of the gamma distribution and its parameters can be computed as follows:

$$\begin{aligned} p(\lambda|\boldsymbol{x}, \mu, a_0, b_0) &= Gam(\lambda|a_N, b_N) \\ a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2 = b_0 + \frac{N}{2\lambda_{ML}}, \end{aligned} \tag{20}$$

where $\lambda_{ML} = N/\sum_{i=1}^{N}(x_i-\mu)^2$.

2.3.3 Inference of the mean and variance using a conjugate prior

Figure 7 depicts a graphical model representing inference of a mean and precision of an univariate normal distribution. In this section, we will consider a conjugate prior for the mean and precision.

The model depicted on Figure 7 assumes the following generative process:

1. $\lambda \sim Gam(\cdot|a_0, b_0)$
2. $\mu \sim N(\cdot|\mu_0, (\beta_0\lambda)^{-1})$
3. $x_i \sim N(\cdot|\mu, \lambda^{-1}) \quad \forall i \in \{1, \ldots, N\}$

where the parameters μ_0, β_0^2, a_0, and b_0 are priors set manually and x is an observed variable.

Again, we will use the precision instead of the variance since it will greatly simplify the derivation of the solution. A conjugate prior for the mean μ and the precision λ is the normal-gamma distribution defined as:

$$p(\mu, \lambda|\mu_0, \beta_0, a_0, b_0) = N(\mu|\mu_0, (\beta_0\lambda)^{-1}) Gam(\lambda|a_0, b_0).$$

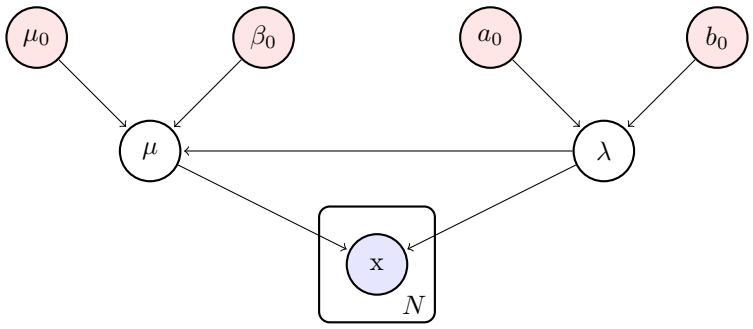


Figure 7: The graphical model for prediction of the observations $x_1, \ldots, x_N$ given the mean μ and the precision λ. Where both the mean and variance is unknown.

Therefore, the joint distribution represented by the graphical model is:

$$\begin{aligned} p(\boldsymbol{x}, \mu, \lambda | \mu_0, \beta_0, a_0, b_0) &= p(\mu, \lambda | \mu_0, \beta_0, a_0, b_0) \prod_{i=1}^{N} p(x_i | \mu, \lambda^{-1}) \\ &= N(\mu | \mu_0, (\beta_0 \lambda)^{-1}) Gam(\lambda | a_0, b_0) \prod_{i=1}^{N} N(x_i | \mu, \lambda^{-1}) \end{aligned}$$

where $\boldsymbol{x} = \{x_1, \ldots, x_N\}$. And the posterior can be computed as follows:

$$p(\mu, \lambda | \boldsymbol{x}, \mu_0, \beta_0, a_0, b_0) = \frac{N(\mu | \mu_0, (\beta_0 \lambda)^{-1}) Gam(\lambda | a_0, b_0) \prod_{i=1}^{N} N(x_i | \mu, \lambda^{-1})}{\int_{\mu,\lambda} N(\mu | \mu_0, (\beta_0 \lambda)^{-1}) Gam(\lambda | a_0, b_0) \prod_{i=1}^{N} N(x_i | \mu, \lambda^{-1}) d\mu \lambda}$$

$$\propto N(\mu | \mu_0, (\beta_0 \lambda)^{-1}) Gam(\lambda | a_0, b_0) \prod_{i=1}^{N} N(x_i | \mu, \lambda^{-1})$$

$$\propto \frac{(\beta_0 \lambda)^{1/2}}{(2\pi)^{1/2}} \exp\left\{ -\frac{\beta_0 \lambda}{2} (\mu - \mu_0)^2 \right\} \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda^{a_0 - 1} \exp\{-b_0 \lambda\} \prod_{i=1}^{N} \lambda^{1/2} \exp\left\{ -\frac{\lambda}{2} (x_i - \mu)^2 \right\}$$

$$\propto \lambda^{1/2} \lambda^{[a_0 + N/2] - 1} \exp\left\{ -\frac{\beta_0 \lambda}{2} (\mu - \mu_0)^2 \right\} \exp\{-b_0 \lambda\} \exp\left\{ -\frac{\lambda}{2} \sum_{i=1}^{N} (x_i - \mu)^2 \right\}$$

$$\propto \lambda^{1/2} \lambda^{[a_0 + N/2] - 1} \exp\left\{ -\frac{\beta_0 \lambda}{2} (\mu - \mu_0)^2 - b_0 \lambda - \frac{\lambda}{2} \sum_{i=1}^{N} (x_i - \mu)^2 \right\}$$

$$\propto \lambda^{1/2} \lambda^{[a_0 + N/2] - 1} \exp\left\{ -\frac{\beta_0 \lambda}{2} (\mu^2 - 2\mu\mu_0 + \mu_0^2) - b_0 \lambda - \frac{\lambda}{2} \sum_{i=1}^{N} (x_i^2 - 2x_i \mu + \mu^2) \right\}$$

$$\propto \lambda^{1/2} \lambda^{[a_0 + N/2] - 1} \exp\left\{ -\frac{\beta_0 \lambda}{2} \left(\mu^2 - 2\mu\mu_0 + \mu_0^2 + \frac{2b_0}{\beta_0} + \frac{1}{\beta_0} \sum_{i=1}^{N} x_i^2 - \frac{1}{\beta_0} \sum_{i=1}^{N} 2x_i \mu + \frac{1}{\beta_0} \sum_{i=1}^{N} \mu^2 \right) \right\}$$

$$\propto \lambda^{1/2} \lambda^{[a_0 + N/2] - 1} \exp\left\{ -\frac{\beta_0 \lambda}{2} \left(\mu^2 - 2\mu\mu_0 + \mu_0^2 + \frac{2b_0}{\beta_0} + \frac{1}{\beta_0} \sum_{i=1}^{N} x_i^2 - \frac{2\mu}{\beta_0} \sum_{i=1}^{N} x_i + \frac{N\mu^2}{\beta_0} \right) \right\}$$

$$\propto \lambda^{1/2} \lambda^{[a_0 + N/2] - 1} \exp\left\{ -\frac{\beta_0 \lambda}{2} \left(\left(1 + \frac{N}{\beta_0} \right) \mu^2 - 2\mu \left(\mu_0 + \frac{1}{\beta_0} \sum_{i=1}^{N} x_i \right) + \mu_0^2 + \frac{2b_0}{\beta_0} + \frac{1}{\beta_0} \sum_{i=1}^{N} x_i^2 \right) \right\}$$

$$\propto \lambda^{1/2} \lambda^{[a_0 + N/2] - 1} \exp\left\{ -\frac{\beta_0 (1 + \frac{N}{\beta_0}) \lambda}{2} \left(\mu^2 - 2\mu \left(1 + \frac{N}{\beta_0} \right)^{-1} \left(\mu_0 + \frac{1}{\beta_0} \sum_{i=1}^{N} x_i \right) \right. \right.$$

$$\left. \left. + \left(1 + \frac{N}{\beta_0} \right)^{-1} \left(\mu_0^2 + \frac{2b_0}{\beta_0} + \frac{1}{\beta_0} \sum_{i=1}^{N} x_i^2 \right) \right) \right\}$$

$$\propto \lambda^{1/2}\lambda^{[a_0+N/2]-1}\exp\left\{-\frac{(\beta_0+N)\lambda}{2}\left(\mu^2-2\mu\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0+\sum_{i=1}^{N}x_i\right)\right.\right.$$
$$\left.\left.+\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0^2+2b_0+\sum_{i=1}^{N}x_i^2\right)\right)\right\}$$

$$\propto \lambda^{1/2}\lambda^{[a_0+N/2]-1}\exp\left\{-\frac{(\beta_0+N)\lambda}{2}\left(\mu^2-2\mu\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0+\sum_{i=1}^{N}x_i\right)+\left[\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0+\sum_{i=1}^{N}x_i\right)\right]^2\right.\right.$$
$$\left.\left.-\left[\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0+\sum_{i=1}^{N}x_i\right)\right]^2+\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0^2+2b_0+\sum_{i=1}^{N}x_i^2\right)\right)\right\}$$

$$\propto \lambda^{1/2}\lambda^{[a_0+N/2]-1}\exp\left\{-\frac{(\beta_0+N)\lambda}{2}\left(\left(\mu-\left[\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0+\sum_{i=1}^{N}x_i\right)\right]\right)^2\right.\right.$$
$$\left.\left.-\left[\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0+\sum_{i=1}^{N}x_i\right)\right]^2+\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0^2+2b_0+\sum_{i=1}^{N}x_i^2\right)\right)\right\}$$

$$\propto \lambda^{1/2}\exp\left\{-\frac{(\beta_0+N)\lambda}{2}\left(\mu-\left[\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0+\sum_{i=1}^{N}x_i\right)\right]\right)^2\right\}$$
$$\lambda^{[a_0+N/2]-1}\exp\left\{-\frac{(\beta_0+N)\lambda}{2}\left(-\left[\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0+\sum_{i=1}^{N}x_i\right)\right]^2+\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0^2+2b_0+\sum_{i=1}^{N}x_i^2\right)\right)\right\}$$

$$\propto \lambda^{1/2}\exp\left\{-\frac{(\beta_0+N)\lambda}{2}\left(\mu-\left[\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0+\sum_{i=1}^{N}x_i\right)\right]\right)^2\right\}$$
$$\lambda^{[a_0+N/2]-1}\exp\left\{-\frac{\lambda}{2}\left(-\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0+\sum_{i=1}^{N}x_i\right)^2+\beta_0\mu_0^2+2b_0+\sum_{i=1}^{N}x_i^2\right)\right\}$$

$$\propto \lambda^{1/2}\exp\left\{-\frac{(\beta_0+N)\lambda}{2}\left(\mu-\left[\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0+\sum_{i=1}^{N}x_i\right)\right]\right)^2\right\}$$
$$\lambda^{[a_0+N/2]-1}\exp\left\{-\left[b_0+\frac{1}{2}\left(-\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0+\sum_{i=1}^{N}x_i\right)^2+\beta_0\mu_0^2+\sum_{i=1}^{N}x_i^2\right)\right]\lambda\right\}$$

Since we used a conjugate prior, the posterior has the same form as the prior and its parameters can be identified from the equation above as follows:

$$p(\mu,\lambda|\mu_N,\beta_N,a_N,b_N)=N(\mu|\mu_N,(\beta_N\lambda)^{-1})Gam(\lambda|a_N,b_N)$$
$$\mu_N=\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0+\sum_{i=1}^{N}x_i\right)$$
$$\beta_N=\beta_0+N$$
$$a_N=a_0+\frac{N}{2}$$
$$b_N=b_0+\frac{1}{2}\left(-\left(\beta_0+N\right)^{-1}\left(\beta_0\mu_0+\sum_{i=1}^{N}x_i\right)^2+\beta_0\mu_0^2+\sum_{i=1}^{N}x_i^2\right)$$

2.3.4 Inference of the mean and variance using a non-conjugate prior

In the previous section, we considered a conjugate prior for the mean and precision. After a complex manipulation with the posterior, we derived a closed form solution. However, we cannot always design a conjugate prior or compute the posterior in a closed form. In this section, we will try to compute the posterior distribution for the normal distribution when the prior is not conjugate. Figure 8 depict a graphical model representing inference of the mean and precision of an univariate normal distribution using non-conjugate prior.

The model depicted on Figure 8 assumes the following generative process:

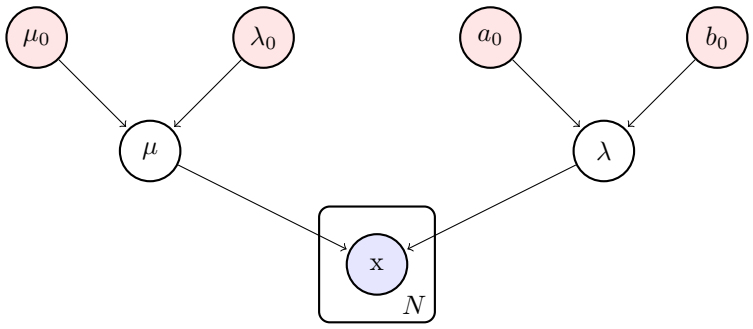


Figure 8: The graphical model for prediction of the observation x given the mean μ and the precision λ. Where both the mean and precision are unknown and the prior is not conjugate.

1. $\lambda \sim Gam(\cdot|a_0, b_0)$
2. $\mu \sim N(\cdot|\mu_0, \lambda_0^{-1})$
3. $x_i \sim N(\cdot|\mu, \lambda^{-1}) \quad \forall i \in \{1, \ldots, N\}$

where the parameters μ_0, λ_0, a_0, and b_0 are priors set manually and x is an observed variable.

The prior distribution for the hidden parameters according to the graphical model is defined as follows:

$$p(\mu, \lambda|\mu_0, \lambda_0, a_0, b_0) = N(\mu|\mu_0, \lambda_0^{-1})Gam(\lambda|a_0, b_0). \tag{21}$$

Consequently, the joint distribution is:

$$\begin{aligned} p(\boldsymbol{x}, \mu, \lambda|\mu_0, \lambda_0, a_0, b_0) &= p(\mu, \lambda|\mu_0, \lambda_0, a_0, b_0) \prod_{i=1}^{N} p(x_i|\mu, \lambda^{-1}) \\ &= N(\mu|\mu_0, \lambda_0^{-1})Gam(\lambda|a_0, b_0) \prod_{i=1}^{N} N(x_i|\mu, \lambda^{-1}) \end{aligned}$$

where $\boldsymbol{x} = \{x_1, \ldots, x_N\}$. The posterior is then defined as:

$$p(\mu, \lambda|\boldsymbol{x}, \mu_0, \lambda_0, a_0, b_0) \propto N(\mu|\mu_0, \lambda_0^{-1})Gam(\lambda|a_0, b_0) \prod_{i=1}^{N} N(x_i|\mu, \lambda^{-1}) \tag{22}$$

However, the posterior (22) does not have a form of the prior defined in (24). The unavailability of a simple analytical solution to the posterior greatly complicates inference in such models and therefore approximation techniques must be used.

There are several techniques dealing with inference in intractable models. The most popular are based on Markov Chain Monte Carlo (MCMC) method, e.g. Gibbs sampling, Variational Inference, or Expectation Propagation.

3 Inference in discrete Bayesian networks

4 The Laplace Approximation

5 Variational Inference

Variational Inference (VI) is based on the calculus of variations, i.e., a generalisation of standard calculus. VI deals with functionals, functions and derivatives of functionals rather than functions, variables and derivatives. In variational calculus similar rules apply. VI can be applied to models of either continuous or discrete random variables. VI approximates both the posterior distribution: $p(\boldsymbol{w}|D)$, and its normalisation constant (model evidence): $p(D)$, where D is the evidence – data, and $\boldsymbol{w}$ are unknown parameters.

Variational inference is based on decomposition of model evidence

$$p(D) = \int p(\boldsymbol{w}, D)dw = \int p(D|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$$

as follows

$$\log p(D) = L(q) + KL(q||p)$$
$$\log p(D) = L(q(\boldsymbol{w})) + KL(q(\boldsymbol{w})||p(\boldsymbol{w}|D))$$

where $p(\boldsymbol{w}|D)$ is our true distribution and $q(\boldsymbol{w})$ is its approximation. $L(q)$ approximates $\log p(D)$ and we want to maximise it. The Kullback-Leibler (KL) divergence measures the "distance" from $q(\boldsymbol{w})$ to $p(\boldsymbol{w}|D)$ and we want to minimise it.

$$L(q(\boldsymbol{w}))) = \int q(\boldsymbol{w}) \log \left\{ \frac{p(\boldsymbol{w}, D)}{q(\boldsymbol{w})} \right\} d\boldsymbol{w}$$

is lower bound and

$$KL(q(\boldsymbol{w})||p(\boldsymbol{w}|D)) = \int q(\boldsymbol{w}) \log \left\{ \frac{q(\boldsymbol{w})}{p(\boldsymbol{w}|D)} \right\} d\boldsymbol{w}$$

is the Kullback-Leibler divergence. The KL divergence is also known as relative-entropy and has these properties:

1. $KL(p||p) = 0$,
2. $KL(q||p) = 0$ if and only if $q = p$,
3. $KL(q||p) \geq 0$ for all q and p.

Decomposition of the $p(D)$ evidence can be verified as follows:

$$\begin{aligned}
\log p(D) &= L(q) + KL(q||p) \\
&= \int q(\boldsymbol{w}) \log \left\{ \frac{p(\boldsymbol{w}, D)}{q(\boldsymbol{w})} \right\} d\boldsymbol{w} + \int q(\boldsymbol{w}) \log \left\{ \frac{q(\boldsymbol{w})}{p(\boldsymbol{w}|D)} \right\} d\boldsymbol{w} \\
&= \int q(\boldsymbol{w}) \left\{ \log \left\{ \frac{p(\boldsymbol{w}, D)}{q(\boldsymbol{w})} \right\} + \log \left\{ \frac{q(\boldsymbol{w})}{p(\boldsymbol{w}|D)} \right\} \right\} d\boldsymbol{w} \\
&= \int q(\boldsymbol{w}) \log \left\{ \frac{p(\boldsymbol{w}, D)}{q(\boldsymbol{w})} \frac{q(\boldsymbol{w})}{p(\boldsymbol{w}|D)} \right\} d\boldsymbol{w} \\
&= \int q(\boldsymbol{w}) \log \left\{ \frac{p(\boldsymbol{w}, D)}{p(\boldsymbol{w}|D)} \right\} d\boldsymbol{w} \\
&= \int q(\boldsymbol{w}) \log \left\{ \frac{p(\boldsymbol{w}|D)p(D)}{p(\boldsymbol{w}|D)} \right\} d\boldsymbol{w} \\
&= \int q(\boldsymbol{w}) \log p(D) d\boldsymbol{w} \\
&= \log p(D) \int q(\boldsymbol{w}) d\boldsymbol{w} \\
&= \log p(D) \cdot 1 \\
&= \log p(D)
\end{aligned}$$

Theoretical properties of the Variational inference are very favourable. Although it is an approximation, it is guaranteed to converge to a local optimum. Since the KL divergence satisfies $KL(q||p) \geq 0$, one can see that the quantity $L(q)$ is a lower bound on the log likelihood function $\log p(D)$. The goal of Variational inference is the variational lower bound $L(q)$ with respect to the approximate $q(\boldsymbol{w})$ distribution, or to minimise the $KL(q||p)$ divergence.

Alternative derivation of the lower bound $L(q)$ is based on the Jensen's inequality:

$$\begin{aligned}
\log p(D) &= \log \int p(D, \boldsymbol{w}) d\boldsymbol{w} = \log \int q(\boldsymbol{w}) \frac{p(D, \boldsymbol{w})}{q(\boldsymbol{w})} d\boldsymbol{w} \\
&\geq \int q(\boldsymbol{w}) \log \frac{p(D, \boldsymbol{w})}{q(\boldsymbol{w})} d\boldsymbol{w} \\
&\geq L(q)
\end{aligned}$$

5.1 Variational Mean Field

In Variational Mean Field, one assumes that q factors with respect to a partition of w into M disjoint groups w_i, with $i = \{1, \ldots, M\}$. No further assumptions are made about q.

$$q(\boldsymbol{w}) = \prod_i^M q_i(w_i) \tag{23}$$

which can be written with explicit parameters, $\boldsymbol{\theta}$, for the the approximation as

$$q(\boldsymbol{w}; \boldsymbol{\theta}) = \prod_i^M q_i(w_i; \theta_i)$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_i]$.

The presented version of the Variational inference is sometimes called "global" since it tries to optimise the full joint probability. Since even this can be found intractable, one can derive a local version of the Variational inference, where only individual factors are independently optimised using the Variational inference.

Substituting q in $KL(q||p)$ and looking for the dependence with respect to q_j is similar to coordinate ascend when optimising $KL(q(\boldsymbol{w}; \boldsymbol{\theta})||p(\boldsymbol{w}|D))$.

$$q(\boldsymbol{w}) = \prod_i^M q_i(w_i) = q_1(w_1) q_2(w_2) \ldots q_M(w_M)$$

We iteratively optimise $q(\boldsymbol{w})$ with respect to $q_i(w_i|\theta_i)$ for $i \in \{1, \ldots, M\}$.

Derivation:

$$\begin{aligned}
KL(q||p) &= \int \prod_{i=1}^M q_i(w_i) \log \left\{ \frac{\prod_{k=1}^M q_k(w_k)}{p(\boldsymbol{w}|D)} \right\} d\boldsymbol{w} \\
&= \int \prod_{i=1}^M q_i(w_i) \left\{ \sum_{k=1}^M \log q_k(w_k) - \log p(\boldsymbol{w}|D) \right\} d\boldsymbol{w} \\
&= \int \prod_{i=1}^M q_i(w_i) \left\{ \sum_{k=1}^M \log q_k(w_k) - \log p(\boldsymbol{w}, D) + \log p(D) \right\} d\boldsymbol{w} \\
&= \int \prod_{i=1}^M q_i(w_i) \left\{ \sum_{k=1}^M \log q_k(w_k) - \log p(\boldsymbol{w}, D) \right\} d\boldsymbol{w} + C_1 \\
&= \int \prod_{i=1}^M q_i(w_i) \left\{ \sum_{k=1}^M \log q_k(w_k) \right\} d\boldsymbol{w} - \int \prod_{i=1}^M q_i(w_i) \left\{ \log p(\boldsymbol{w}, D) \right\} d\boldsymbol{w} + C_1 \\
&= \sum_{k=1}^M \int \prod_{i=1}^M q_i(w_i) \log q_k(w_k) d\boldsymbol{w} - \int \prod_{i=1}^M q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w} + C_1 \\
&= \int \prod_{i=1}^M q_i(w_i) \log q_j(w_j) d\boldsymbol{w} + \sum_{k=1; k \neq j}^M \int \prod_{i=1}^M q_i(w_i) \log q_k(w_k) d\boldsymbol{w} - \int \prod_{i=1}^M q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w} + C_1 \\
&= \int \prod_{i=1}^M q_i(w_i) \log q_j(w_j) d\boldsymbol{w} + \sum_{k=1; k \neq j}^M \int q_k(w_k) \log q_k(w_k) \prod_{i=1; i \neq k}^M q_i(w_i) d\boldsymbol{w} - \int \prod_{i=1}^M q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w} + C_1 \\
&= \int \prod_{i=1}^M q_i(w_i) \log q_j(w_j) d\boldsymbol{w} + \sum_{k=1; k \neq j}^M \int q_k(w_k) \log q_k(w_k) \int \prod_{i=1; i \neq k}^M q_i(w_i) d\boldsymbol{w}_{\setminus k} dw_k - \int \prod_{i=1}^M q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w} + C_1 \\
&= \int \prod_{i=1}^M q_i(w_i) \log q_j(w_j) d\boldsymbol{w} + \sum_{k=1; k \neq j}^M \int q_k(w_k) \log q_k(w_k) dw_k - \int \prod_{i=1}^M q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w} + C_1 \\
&= \int \prod_{i=1}^M q_i(w_i) \log q_j(w_j) d\boldsymbol{w} - \int \prod_{i=1}^M q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w} + C_2
\end{aligned}$$

Derivation continuation:

$$
\begin{aligned}
KL(q||p) &= \int \prod_{i=1}^{M} q_i(w_i) \log q_j(w_j) d\boldsymbol{w} - \int \prod_{i=1}^{M} q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w} + C_2 \\
&= \int q_j(w_j) \log q_j(w_j) \prod_{i=1;i\neq j}^{M} q_i(w_i) d\boldsymbol{w} - \int \prod_{i=1}^{M} q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w} + C_2 \\
&= \int q_j(w_j) \log q_j(w_j) dw_j - \int \prod_{i=1}^{M} q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w} + C_2 \\
&= \int q_j(w_j) \log q_j(w_j) dw_j - \int q_j(w_j) \int \prod_{i=1;i\neq j}^{M} q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w} + C_2 \\
&= \int q_j(w_j) \log q_j(w_j) dw_j - \int q_j(w_j) \log \left(\exp \left\{ \int \prod_{i=1;i\neq j}^{M} q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w}^{\setminus j} \right\} \right) dw_j + C_2 \\
&= \int q_j(w_j) \log q_j(w_j) dw_j - \int q_j(w_j) \log \left(\exp \left\{ \int \prod_{i=1;i\neq j}^{M} q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w}^{\setminus j} \right\} \right) dw_j + C_2 \cdot 1 \\
&= \int q_j(w_j) \log q_j(w_j) dw_j - \int q_j(w_j) \log \left(\exp \left\{ \int \prod_{i=1;i\neq j}^{M} q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w}^{\setminus j} \right\} \right) dw_j - C_2 \int q_j(w_j) dw_j \\
&= \int q_j(w_j) \log q_j(w_j) dw_j - \int q_j(w_j) \log \left(\exp \left\{ \int \prod_{i=1;i\neq j}^{M} q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w}^{\setminus j} \right\} \right) dw_j - \int q_j(w_j) \log \exp(-C_2) dw_j \\
&= \int q_j(w_j) \log q_j(w_j) dw_j - \int q_j(w_j) \log \left(\exp \left\{ \int \prod_{i=1;i\neq j}^{M} q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w}^{\setminus j} - C_2 \right\} \right) dw_j \\
&= \int q_j(w_j) \log \frac{q_j(w_j)}{\exp \left\{ \int \prod_{i=1;i\neq j}^{M} q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w}^{\setminus j} + C_3 \right\}} dw_j \\
&= KL \left(q_j(w_j) || \exp \left\{ \int \prod_{i=1;i\neq j}^{M} q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w}^{\setminus j} + C_3 \right\} \right)
\end{aligned}
$$

In general, $KL(q||p)$ is minimised when both $q = p$. The optimal q_j given that the other factors are kept fixed is:

$$
\begin{aligned}
q_j(w_j; \theta_j) &\propto \exp \left\{ \int \prod_{i=1;i\neq j}^{M} q_i(w_i) \log p(\boldsymbol{w}, D) d\boldsymbol{w}^{\setminus j} \right\} \\
&\propto \exp E_{q^{\setminus j}} \left[\log p(\boldsymbol{w}, D) \right]
\end{aligned}
$$

Please note that equality does not apply here because of the constant C_3. More often, we work with the log version and the normalisation constant is found by introspection.

$$\log q_j(w_j; \theta_j) = E_{q^{\setminus j}} \left[\log p(\boldsymbol{w}, D) \right] + C_4$$

5.2 Example: VMF - Unknown Mean and Variance of a normal distribution, with improper priors

In this section, we will try to compute the posterior distribution for the normal distribution when the prior is improper. Figure 9 depict a graphical model representing inference of the mean and precision of an univariate normal distribution using non-conjugate prior.

As we set the priors for μ and λ to be improper priors:

$$
\begin{aligned}
p(\mu) &= \mu_0 \\
p(\lambda) &= 1/\lambda
\end{aligned}
$$

the prior distribution for the hidden parameters factors as follows:

$$p(\mu, \lambda | \mu_0) = \mu_0 \frac{1}{\lambda}.$$

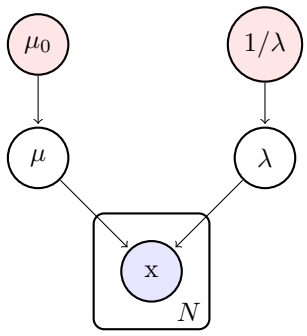


Figure 9: The graphical model for prediction of the observation x given the mean μ and the precision λ. Where both the mean and precision are unknown and the prior is improper.

While these priors are computationally convenient, they are not conjugate. Therefore, the posterior will have a different form compared to the priors.
We enforce that the posterior approximation factors

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

and solve for the optimal factors

$$\log q_\mu(\mu) = E_{q_\lambda} [\log p(D, \mu, \lambda)]$$
$$\log q_\lambda(\lambda) = E_{q_\mu} [\log p(D, \mu, \lambda)]$$

Goal: infer the posterior distribution of the mean μ and precision λ of a normal distribution given independent observations $D = \{x_1, \ldots, x_N\}$.

The likelihood of μ and λ is

$$
\begin{aligned}
p(D|\mu,\lambda) &= \prod_{i=1}^{N} p(x_i|\mu,\lambda) = \prod_{i=1}^{N} N(x_i|\mu,\lambda^{-1}) \\
\log p(D|\mu,\lambda) &= \sum_{i=1}^{N} \log N(x_i|\mu,\lambda^{-1}) \\
&= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\lambda^{-1}}} \exp\left\{-\frac{\lambda}{2}(x_i-\mu)^2\right\} \\
&= -\frac{N}{2}\log 2\pi\lambda^{-1} - \frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2 \\
&= \frac{N}{2}\log\lambda - \frac{N}{2}\log 2\pi - \frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2 \\
&= \frac{N}{2}\log\lambda - \frac{N}{2}\log 2\pi - \frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu_{ML}+\mu_{ML}-\mu)^2 \\
&= \frac{N}{2}\log\lambda - \frac{N}{2}\log 2\pi - \frac{\lambda}{2}\sum_{i=1}^{N}((x_i-\mu_{ML})-(\mu-\mu_{ML}))^2 \\
&= \frac{N}{2}\log\lambda - \frac{N}{2}\log 2\pi - \frac{\lambda}{2}\sum_{i=1}^{N}((x_i-\mu_{ML})^2 - 2(x_i-\mu_{ML})(\mu-\mu_{ML}) + (\mu-\mu_{ML})^2) \\
&= \frac{N}{2}\log\lambda - \frac{N}{2}\log 2\pi - \frac{\lambda}{2}\left[N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2 - \sum_{i=1}^{N}2(x_i-\mu_{ML})(\mu-\mu_{ML})\right] \\
&= \frac{N}{2}\log\lambda - \frac{N}{2}\log 2\pi - \frac{\lambda}{2}\left[N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2 - 2(\mu-\mu_{ML})\sum_{i=1}^{N}(x_i-\mu_{ML})\right] \\
&= \frac{N}{2}\log\lambda - \frac{N}{2}\log 2\pi - \frac{\lambda}{2}\left[N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2 - 2(\mu-\mu_{ML})\left(\sum_{i=1}^{N}x_i - \sum_{i=1}^{N}\mu_{ML}\right)\right] \\
&= \frac{N}{2}\log\lambda - \frac{N}{2}\log 2\pi - \frac{\lambda}{2}\left[N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2 - 2(\mu-\mu_{ML})\left(\sum_{i=1}^{N}x_i - \sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{N}x_j\right)\right] \\
&= \frac{N}{2}\log\lambda - \frac{N}{2}\log 2\pi - \frac{\lambda}{2}\left[N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2 - 2(\mu-\mu_{ML})\left(\sum_{i=1}^{N}x_i - \sum_{j=1}^{N}x_j\right)\right] \\
&= \frac{N}{2}\log\lambda - \frac{N}{2}\log 2\pi - \frac{\lambda}{2}\left[N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2 - 2(\mu-\mu_{ML})\cdot 0\right] \\
&= \frac{N}{2}\log\lambda - \frac{\lambda}{2}\left[N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2\right] + \text{const}
\end{aligned}
$$

where $\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is empirical mean.

5.2.1 $\log q_\mu(\mu)$

Derivation:

$$\begin{aligned}
\log q_\mu(\mu) &= E_{q_\lambda}\left[\log p(D,\mu,\lambda)\right] \\
&= E_{q_\lambda}\left[\log p(D|\mu,\lambda)p(\mu)p(\lambda)\right] \\
&= E_{q_\lambda}\left[\log p(D|\mu,\lambda) + \log p(\mu) + \log p(\lambda)\right] \\
&= \int q_\lambda(\lambda)\left[\log p(D|\mu,\lambda) + \log p(\mu) + \log p(\lambda)\right] d\lambda \\
&= \int q_\lambda(\lambda)\log p(D|\mu,\lambda)d\lambda + \int q_\lambda(\lambda)\log p(\mu)d\lambda + \int q_\lambda(\lambda)\log p(\lambda)d\lambda \\
&= \int q_\lambda(\lambda)\log p(D|\mu,\lambda)d\lambda + \log p(\mu)\int q_\lambda(\lambda)d\lambda + C_1 \\
&= \int q_\lambda(\lambda)\log p(D|\mu,\lambda)d\lambda + \log(\mu_0)\cdot 1 + C_1 \\
&= \int q_\lambda(\lambda)\log p(D|\mu,\lambda)d\lambda + C_2 \\
&= \int q_\lambda(\lambda)\left(\frac{N}{2}\log\lambda - \frac{\lambda}{2}\left[N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2\right]\right)d\lambda + C_3 \\
&= \int q_\lambda(\lambda)\left(-\frac{N\lambda}{2}(\mu-\mu_{ML})^2\right)d\lambda + C_4 \\
&= \left(-\frac{N}{2}(\mu-\mu_{ML})^2\right)\int q_\lambda(\lambda)\lambda d\lambda + C_4 \\
&= -\frac{N}{2}(\mu-\mu_{ML})^2 E_{q_\lambda}[\lambda] + C_4
\end{aligned}$$

By introspection, one can observe that

$$\begin{aligned}
q_\mu(\mu) &\propto \exp\left\{-\frac{NE_{q_\lambda}[\lambda]}{2}(\mu-\mu_{ML})^2\right\} \\
&= N(\mu;\mu_{ML},\lambda_N^{-1})
\end{aligned}$$

where

$$\lambda_N = NE_{q_\lambda}[\lambda].$$

Recall that

$$N(x|\mu,\lambda^{-1}) = \frac{\lambda^{1/2}}{(2\pi)^{1/2}}\exp\left\{-\frac{\lambda}{2}(x-\mu)^2\right\}$$

5.2.2 $\log q_\lambda(\lambda)$

Derivation:

$$\begin{aligned}
\log q_\lambda(\lambda) &= E_{q_\mu}\left[\log p(D,\mu,\lambda)\right] \\
&= E_{q_\mu}[\log p(D|\mu,\lambda) + \log p(\mu) + \log p(\lambda)] \\
&= E_{q_\mu}[\log p(D|\mu,\lambda)] + E_{q_\mu}[\log p(\mu)] + E_{q_\mu}[\log p(\lambda)] \\
&= E_{q_\mu}[\log p(D|\mu,\lambda)] + E_{q_\mu}[\log p(\lambda)] + C_1 \\
&= E_{q_\mu}[\log p(D|\mu,\lambda)] + \log p(\lambda) + C_1 \\
&= \log p(\lambda) + E_{q_\mu}[\log p(D|\mu,\lambda)] + C_1 \\
&= -\log\lambda + E_{q_\mu}\left[\frac{N}{2}\log\lambda - \frac{\lambda}{2}\left[N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2\right]\right] + C_2 \\
&= \frac{N-2}{2}\log\lambda - \frac{\lambda}{2}E_{q_\mu}\left[N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2\right] + C_2 \\
&= \frac{N-2}{2}\log\lambda - \frac{N\lambda}{2}E_{q_\mu}\left[\mu^2 - 2\mu\mu_{ML} + \mu_{ML}^2 + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right] + C_2 \\
&= \log\lambda^{\frac{N}{2}-1} - \frac{N}{2}\left(E_{q_\mu}[\mu^2] - 2E_{q_\mu}[\mu]\mu_{ML} + \mu_{ML}^2 + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)\lambda + C_2 \\
&= \log\lambda^{\frac{N}{2}-1} - \frac{N}{2}\left(\lambda_N^{-1} + \mu_{ML}^2 - 2\mu_{ML}\mu_{ML} + \mu_{ML}^2 + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)\lambda + C_2 \\
&= \log\lambda^{\frac{N}{2}-1} - \frac{N}{2}\left(\lambda_N^{-1} + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)\lambda + C_2
\end{aligned}$$

By introspection, one can observe that

$$\begin{aligned}
q_\lambda(\lambda) &\propto \lambda^{\frac{N}{2}-1}\exp\left\{-\frac{N}{2}\left(\lambda_N^{-1} + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)\lambda\right\} \\
&= Gam(\lambda; a_N, b_N)
\end{aligned}$$

where

$$\begin{aligned}
a_N &= \frac{N}{2} \\
b_N &= \frac{N}{2}\left(\lambda_N^{-1} + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)
\end{aligned}$$

Recall that

$$Gam(\lambda; a, b) = b^a\frac{1}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda)$$

5.2.3 Summary

This gives the following optimal factors given that the other factor is fixed

$$\begin{aligned}
q_\mu(\mu) &= N(\mu|\mu_{ML}, \lambda_N^{-1}) \\
q_\lambda(\lambda) &= Gam(\lambda|a_N, b_N)
\end{aligned}$$

where

$$\begin{aligned}
\lambda_N &= NE_{q_\lambda}[\lambda] = N\frac{a_N}{b_N} \\
a_N &= \frac{N}{2} \\
b_N &= \frac{N}{2}\left(\lambda_N^{-1} + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)
\end{aligned}$$

We iteratively optimise q_μ and q_λ until convergence.

5.3 Example: VMF - Unknown Mean and Variance of a normal distribution, with non-conjugate priors

In this section, we will try to compute the posterior distribution for the normal distribution when the prior is not conjugate. Figure 10 depict a graphical model representing inference of the mean and precision of an univariate normal distribution using non-conjugate prior.

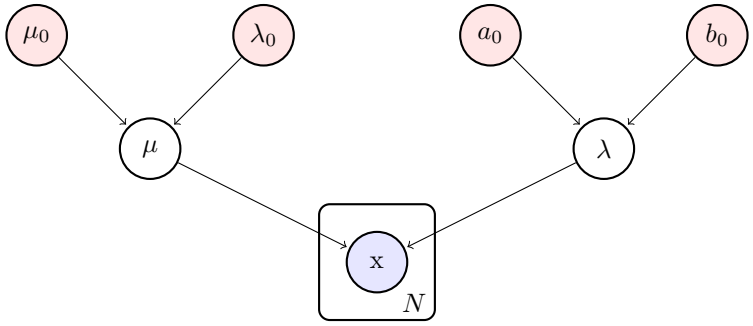


Figure 10: The graphical model for prediction of the observation x given the mean μ and the precision λ. Where both the mean and precision are unknown and the prior is not conjugate.

As we set the priors for μ and λ to be non-conjugate priors:

$$p(\mu) = N(\mu|\mu_0, \lambda_0^{-1}),$$
$$p(\lambda) = Gam(\lambda|a_0, b_0),$$

the prior distribution for the hidden parameters factors as follows:

$$p(\mu, \lambda|\mu_0, \lambda_0, a_0, b_0) = N(\mu|\mu_0, \lambda_0^{-1})Gam(\lambda|a_0, b_0). \tag{24}$$

We enforce that the posterior approximation factors

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

and solve for the optimal factors

$$\log q_\mu(\mu) = E_{q_\lambda}\left[\log p(D, \mu, \lambda)\right]$$
$$\log q_\lambda(\lambda) = E_{q_\mu}\left[\log p(D, \mu, \lambda)\right]$$

5.3.1 $\log q_\mu(\mu)$

Derivation:

$$\begin{aligned}
\log q_\mu(\mu) &= E_{q_\lambda}\left[\log p(D,\mu,\lambda)\right] \\
&= E_{q_\lambda}\left[\log p(D|\mu,\lambda)p(\mu)p(\lambda)\right] \\
&= E_{q_\lambda}\left[\log p(D|\mu,\lambda) + \log p(\mu) + \log p(\lambda)\right] \\
&= E_{q_\lambda}\left[\log p(D|\mu,\lambda)\right] + E_{q_\lambda}\left[\log p(\mu)\right] + E_{q_\lambda}\left[\log p(\lambda)\right] \\
&= E_{q_\lambda}\left[\log p(D|\mu,\lambda)\right] + E_{q_\lambda}\left[\log p(\mu)\right] + C_1 \\
&= E_{q_\lambda}\left[\frac{N}{2}\log\lambda - \frac{\lambda}{2}\left(N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)\right] + E_{q_\lambda}\left[-\frac{\lambda_0}{2}(\mu-\mu_0)^2\right] + C_2 \\
&= E_{q_\lambda}\left[\frac{N}{2}\log\lambda - \frac{\lambda}{2}\left(N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)\right] - \frac{\lambda_0}{2}(\mu-\mu_0)^2 + C_2 \\
&= E_{q_\lambda}\left[-\frac{\lambda}{2}N(\mu-\mu_{ML})^2\right] - \frac{\lambda_0}{2}(\mu-\mu_0)^2 + C_3 \\
&= -\frac{N}{2}(\mu-\mu_{ML})^2 E_{q_\lambda}\left[\lambda\right] - \frac{\lambda_0}{2}(\mu-\mu_0)^2 + C_3 \\
&= -\frac{N}{2}(\mu^2 - 2\mu\mu_{ML} + \mu_{ML}^2)E_{q_\lambda}\left[\lambda\right] - \frac{\lambda_0}{2}(\mu^2 - 2\mu\mu_0 + \mu_0^2) + C_3 \\
&= -\frac{N}{2}\mu^2 E_{q_\lambda}\left[\lambda\right] + N\mu\mu_{ML}E_{q_\lambda}\left[\lambda\right] - \frac{\lambda_0}{2}\mu^2 + \lambda_0\mu\mu_0 + C_4 \\
&= -\frac{N}{2}\mu^2 E_{q_\lambda}\left[\lambda\right] - \frac{\lambda_0}{2}\mu^2 + N\mu\mu_{ML}E_{q_\lambda}\left[\lambda\right] + \lambda_0\mu\mu_0 + C_4 \\
&= -\frac{NE_{q_\lambda}\left[\lambda\right] + \lambda_0}{2}\mu^2 + (NE_{q_\lambda}\left[\lambda\right]\mu_{ML} + \lambda_0\mu_0)\mu + C_4 \\
&= -\frac{NE_{q_\lambda}\left[\lambda\right] + \lambda_0}{2}\left(\mu^2 - 2\frac{NE_{q_\lambda}\left[\lambda\right]\mu_{ML} + \lambda_0\mu_0}{NE_{q_\lambda}\left[\lambda\right] + \lambda_0}\mu + C_5\right)
\end{aligned}$$

By completing the squares, one can derive that

$$\begin{aligned}
q_\mu(\mu) &\propto \exp\left\{-\frac{NE_{q_\lambda}[\lambda] + \lambda_0}{2}\left(\mu - \frac{NE_{q_\lambda}\left[\lambda\right]\mu_{ML} + \lambda_0\mu_0}{NE_\lambda[\lambda] + \lambda_0}\right)^2\right\} \\
&= N(\mu;\mu_N,\lambda_N^{-1})
\end{aligned}$$

where

$$\begin{aligned}
\mu_N &= \frac{NE_{q_\lambda}\left[\lambda\right]\mu_{ML} + \lambda_0\mu_0}{NE_{q_\lambda}[\lambda] + \lambda_0} \\
\lambda_N &= NE_{q_\lambda}[\lambda] + \lambda_0
\end{aligned}$$

5.3.2 $\log q_\lambda(\lambda)$

Derivation:

$$
\begin{aligned}
\log q_\lambda(\lambda) &= E_{q_\mu}\left[\log p(D,\mu,\lambda)\right] \\
&= E_{q_\mu}\left[\log p(D|\mu,\lambda)p(\mu)p(\lambda)\right] \\
&= E_{q_\mu}\left[\log p(D|\mu,\lambda) + \log p(\mu) + \log p(\lambda)\right] \\
&= E_{q_\mu}\left[\log p(D|\mu,\lambda)\right] + E_{q_\mu}\left[\log p(\mu)\right] + E_{q_\mu}\left[\log p(\lambda)\right] \\
&= E_{q_\mu}\left[\log p(D|\mu,\lambda)\right] + E_{q_\mu}\left[\log p(\lambda)\right] + C_1 \\
&= E_{q_\mu}\left[\frac{N}{2}\log\lambda - \frac{\lambda}{2}\left(N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)\right] + E_{q_\mu}\left[(a_0-1)\log\lambda - b_0\lambda\right] + C_2 \\
&= E_{q_\mu}\left[\frac{N}{2}\log\lambda - \frac{\lambda}{2}\left(N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)\right] + (a_0-1)\log\lambda - b_0\lambda + C_2 \\
&= E_{q_\mu}\left[\frac{N}{2}\log\lambda\right] + E_{q_\mu}\left[-\frac{\lambda}{2}\left(N(\mu-\mu_{ML})^2 + \sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)\right] + (a_0-1)\log\lambda - b_0\lambda + C_3 \\
&= \frac{N}{2}\log\lambda - \frac{N}{2}E_{q_\mu}\left[(\mu-\mu_{ML})^2 + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right]\lambda + (a_0-1)\log\lambda - b_0\lambda + C_3 \\
&= \left(a_0 + \frac{N}{2} - 1\right)\log\lambda - \left(b_0 + E_{q_\mu}\left[(\mu-\mu_{ML})^2 + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right]\right)\lambda + C_3 \\
&= \left(a_0 + \frac{N}{2} - 1\right)\log\lambda - \left(b_0 + \frac{N}{2}E_{q_\mu}\left[\mu^2 - 2\mu\mu_{ML} + \mu_{ML}^2 + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right]\right)\lambda + C_3 \\
&= \left(a_0 + \frac{N}{2} - 1\right)\log\lambda - \left(b_0 + \frac{N}{2}\left(E_{q_\mu}[\mu^2] - 2E_{q_\mu}[\mu]\mu_{ML} + \mu_{ML}^2 + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)\right)\lambda + C_3 \\
&= \left(a_0 + \frac{N}{2} - 1\right)\log\lambda - \left(b_0 + \frac{N}{2}\left(\lambda_N^{-1} + \mu_N^2 - 2\mu_N\mu_{ML} + \mu_{ML}^2 + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)\right)\lambda + C_3 \\
&= \log\lambda^{a_0+\frac{N}{2}-1} - \left(b_0 + \frac{N}{2}\left(\lambda_N^{-1} + (\mu_N-\mu_{ML})^2 + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)\right)\lambda + C_3
\end{aligned}
$$

By introspection, one can observe that

$$
\begin{aligned}
q_\lambda(\lambda) &\propto \lambda^{a_0+\frac{N}{2}-1}\exp\left\{-\left(a_0 + \frac{N}{2}\left(\lambda_N^{-1} + (\mu_N-\mu_{ML})^2 + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)\right)\lambda\right\} \\
&= Gam(\lambda; a_N, b_N)
\end{aligned}
$$

where

$$
\begin{aligned}
a_N &= a_0 + \frac{N}{2} \\
b_N &= b_0 + \frac{N}{2}\left(\lambda_N^{-1} + (\mu_N-\mu_{ML})^2 + \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2\right)
\end{aligned}
$$

5.3.3 Summary

This gives the following optimal factors given that the other factor is fixed

$$
\begin{aligned}
q_\mu(\mu) &= N(\mu|\mu_N, \lambda_N^{-1}) \\
q_\lambda(\lambda) &= Gam(\lambda|a_N, b_N)
\end{aligned}
$$

where

$$\mu_N = \frac{NE_{q_\lambda}[\lambda]\,\mu_{ML} + \lambda_0\mu_0}{N(a_N/b_N) + \lambda_0} = \frac{N(a_N/b_N)\mu_{ML} + \lambda_0\mu_0}{N(a_N/b_N) + \lambda_0}$$
$$\lambda_N = NE_{q_\lambda}[\lambda] + \lambda_0 = N(a_N/b_N) + \lambda_0$$
$$a_N = a_0 + \frac{N}{2}$$
$$b_N = b_0 + \frac{N}{2}\left(\lambda_N^{-1} + (\mu_N - \mu_{ML})^2 + \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})^2\right)$$
$$= b_0 + \frac{N}{2}\left(\lambda_N^{-1} + \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_N)^2\right)$$

We iteratively optimise q_μ and q_λ until convergence.

5.4 Gradient ascend

First, one selects q to be a parametric distribution: $q(\boldsymbol{w}|\boldsymbol{\theta})$ for which $L(q)$ can be computed analytically. Then, one can use a gradient ascend (hill climbing) to maximise $L(q)$ with respect of parameters, $\boldsymbol{\theta}$, of $q(\boldsymbol{w};\boldsymbol{\theta})$. The lower bound then becomes a function of $\boldsymbol{\theta}$ and can be optimised. This can be very tedious.

5.5 Example: GA - Inference of the mean and variance using a non-conjugate prior

Using gradient ascend, the goal is to approximate the posterior (22) with the with a product of marginals in the form of the prior as defined in (24). The interesting point here is that we know what the posterior will look like.

The approximating distribution (23) is defined as follows:

$$q(\mu, \lambda|\mu_N, \lambda_N, a_N, b_N) = N(\mu|\mu_N, \lambda_N^{-1})Gam(\lambda|a_N, b_N), \tag{25}$$

where $w_1 = \mu$ and $w_2 = \lambda$, and the factors q_i are

$$q_1(w_1|\mu_N, \lambda_N) \equiv q_\mu(\mu|\mu_N, \lambda_N) = N(\mu|\mu_N, \lambda_N^{-1})$$
$$q_2(w_2|a_N, b_N) \equiv q_\lambda(\lambda|a_N, b_N) = Gam(\lambda|a_N, b_N)$$

The true distribution is defined as follows:

$$p(\mu, \lambda|\boldsymbol{x}, \mu_0, \lambda_0, a_0, b_0) \propto N(\mu|\mu_0, \lambda_0^{-1})Gam(\lambda|a_0, b_0)\prod_{i=1}^{N} N(x_i|\mu, \lambda^{-1}). \tag{26}$$

One can notice that we know the joint posterior distribution (26) only up to the normalisation constant. However, this is not a significant problem since the minimum of the KL divergence does not depend on the normalisation constant. Therefore, the KL divergence can be computed as:

$$KL(q||p) = \int_{\mu,\lambda} N(\mu|\mu_N, \lambda_N^{-1})Gam(\lambda|a_N, b_N)\log\frac{N(\mu|\mu_N, \lambda_N^{-1})Gam(\lambda|a_N, b_N)}{1/Z_{\boldsymbol{x},\mu_0,\lambda_0,a_0,b_0}\cdot N(\mu|\mu_0, \lambda_0^{-1})Gam(\lambda|a_0, b_0)\prod_{i=1}^{N} N(x_i|\mu, \lambda^{-1})}d\mu d\lambda,$$

where the $Z_{\boldsymbol{x},\mu_0,\lambda_0,a_0,b_0}$ is the unknown normalisation constant of the true posterior.

This can be further expanded as

$$
\begin{aligned}
KL(q||p) = \int_{\mu,\lambda} & N(\mu|\mu_N, \lambda_N^{-1}) Gam(\lambda|a_N, b_N) \Big(\log N(\mu|\mu_N, \lambda_N^{-1}) + \log Gam(\lambda|a_N, b_N) \\
& - \log N(\mu|\mu_0, \lambda_0^{-1}) - \log Gam(\lambda|a_0, b_0) - \sum_{i=1}^{N} \log N(x_i|\mu, \lambda^{-1}) + \log Z_{\boldsymbol{x},\mu_0,\lambda_0,a_0,b_0} \Big) d\mu d\lambda \\
= & \int_{\mu,\lambda} N(\mu|\mu_N, \lambda_N^{-1}) Gam(\lambda|a_N, b_N) \log N(\mu|\mu_N, \lambda_N^{-1}) d\mu d\lambda \\
& + \int_{\mu,\lambda} N(\mu|\mu_N, \lambda_N^{-1}) Gam(\lambda|a_N, b_N) \log Gam(\lambda|a_N, b_N) d\mu d\lambda \\
& - \int_{\mu,\lambda} N(\mu|\mu_N, \lambda_N^{-1}) Gam(\lambda|a_N, b_N) \log N(\mu|\mu_0, \lambda_0^{-1}) d\mu d\lambda \\
& - \int_{\mu,\lambda} N(\mu|\mu_N, \lambda_N^{-1}) Gam(\lambda|a_N, b_N) \log Gam(\lambda|a_0, b_0) d\mu d\lambda \\
& - \int_{\mu,\lambda} N(\mu|\mu_N, \lambda_N^{-1}) Gam(\lambda|a_N, b_N) \sum_{i=1}^{N} \log N(x_i|\mu, \lambda^{-1}) d\mu d\lambda \\
& + \int_{\mu,\lambda} N(\mu|\mu_N, \lambda_N^{-1}) Gam(\lambda|a_N, b_N) \log Z_{\boldsymbol{x},\mu_0,\lambda_0,a_0,b_0} d\mu d\lambda
\end{aligned}
$$

Now, some of the factors can be integrated out:

$$
\begin{aligned}
KL(q||p) = & \int_{\mu} N(\mu|\mu_N, \lambda_N^{-1}) \log N(\mu|\mu_N, \lambda_N^{-1}) d\mu \\
& + \int_{\lambda} Gam(\lambda|a_N, b_N) \log Gam(\lambda|a_N, b_N) d\lambda \\
& - \int_{\mu} N(\mu|\mu_N, \lambda_N^{-1}) \log N(\mu|\mu_0, \lambda_0^{-1}) d\mu \\
& - \int_{\lambda} Gam(\lambda|a_N, b_N) \log Gam(\lambda|a_0, b_0) d\lambda \\
& - \int_{\mu,\lambda} N(\mu|\mu_N, \lambda_N^{-1}) Gam(\lambda|a_N, b_N) \sum_{i=1}^{N} \log N(x_i|\mu, \lambda^{-1}) d\mu d\lambda \\
& + \log Z_{\boldsymbol{x},\mu_0,\lambda_0,a_0,b_0}
\end{aligned}
$$

In our case, the minimisation of $KL(q||p)$ is performed with respect to the $\mu_N, \lambda_N^{,} a_N, b_N$ parameters of the approximating distribution 25. The simplest solution is to compute partial derivatives of the KL divergence with respect to these parameters, and then perform gradient descend.

Let first derive the partial derivative for μ_N:

$$
\begin{aligned}
\frac{\partial KL(q||p)}{\partial \mu_N} = & \frac{\partial}{\partial \mu_N} \int_{\mu} N(\mu|\mu_N, \lambda_N^{-1}) \log N(\mu|\mu_N, \lambda_N^{-1}) d\mu \\
& - \frac{\partial}{\partial \mu_N} \int_{\mu} N(\mu|\mu_N, \lambda_N^{-1}) \log N(\mu|\mu_0, \lambda_0^{-1}) d\mu \\
& - \frac{\partial}{\partial \mu_N} \int_{\mu,\lambda} N(\mu|\mu_N, \lambda_N^{-1}) Gam(\lambda|a_N, b_N) \sum_{i=1}^{N} \log N(x_i|\mu, \lambda^{-1}) d\mu d\lambda
\end{aligned} \tag{27}
$$

Before continuing, lets us recall the definition of the normal (18) and the gamma (19) distributions:

$$
N(x|\mu, \lambda^{-1}) = \frac{\lambda^{1/2}}{(2\pi)^{1/2}} \exp\left\{ -\frac{\lambda}{2}(x-\mu)^2 \right\} \tag{28}
$$

$$
Gam(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp\{-b\lambda\}. \tag{29}
$$

Now substitute (28) and (29) into (27).

$$\begin{aligned}
\frac{\partial KL(q||p)}{\partial \mu_N} = & \frac{\partial}{\partial \mu_N} \int_\mu \frac{\lambda_N^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\lambda_N}{2}(\mu-\mu_N)^2\right\} \log\left\{\frac{\lambda_N^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\lambda_N}{2}(\mu-\mu_N)^2\right\}\right\} d\mu \\
& - \frac{\partial}{\partial \mu_N} \int_\mu \frac{\lambda_N^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\lambda_N}{2}(\mu-\mu_N)^2\right\} \log\left\{\frac{\lambda_0^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\lambda_0}{2}(\mu-\mu_0)^2\right\}\right\} d\mu \\
& - \frac{\partial}{\partial \mu_N} \int_{\mu,\lambda} \frac{\lambda_N^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\lambda_N}{2}(\mu-\mu_N)^2\right\} \frac{1}{\Gamma(a_N)} b^{a_N} \lambda^{a_N-1} \exp\{-b_N\lambda\} \cdot \\
& \quad \cdot \sum_{i=1}^N \log \frac{\lambda^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\lambda}{2}(x_i-\mu)^2\right\} d\mu d\lambda
\end{aligned}$$

However, this is typically difficult to solve this way. Therefore, another approach building on functional analysis, where one tries to compute derivatives with respect to functions instead of parameters, is be easier to grasp:

$$\begin{aligned}
\frac{\partial KL(q||p)}{\partial N(\mu|\mu_N,\lambda_N^{-1})} = & \frac{\partial}{\partial N(\mu|\mu_N,\lambda_N^{-1})} \int_\mu N(\mu|\mu_N,\lambda_N^{-1}) \log N(\mu|\mu_N,\lambda_N^{-1}) d\mu \\
& - \frac{\partial}{\partial N(\mu|\mu_N,\lambda_N^{-1})} \int_\mu N(\mu|\mu_N,\lambda_N^{-1}) \log N(\mu|\mu_0,\lambda_0^{-1}) d\mu \\
& - \frac{\partial}{\partial N(\mu|\mu_N,\lambda_N^{-1})} \int_{\mu,\lambda} N(\mu|\mu_N,\lambda_N^{-1}) Gam(\lambda|a_N,b_N) \sum_{i=1}^N \log N(x_i|\mu,\lambda^{-1}) d\mu d\lambda \\
= & (\log N(\mu|\mu_N,\lambda_N^{-1}) + 1) - \log N(\mu|\mu_0,\lambda_0^{-1}) - \int_\lambda Gam(\lambda|a_N,b_N) \sum_{i=1}^N \log N(x_i|\mu,\lambda^{-1}) d\lambda
\end{aligned}$$

Setting the derivative equal to zero, one can compute the approximation:

$$\begin{aligned}
0 &= (\log N(\mu|\mu_N,\lambda_N^{-1}) + 1) - \log N(\mu|\mu_0,\lambda_0^{-1}) - \int_\lambda Gam(\lambda|a_N,b_N) \sum_{i=1}^N \log N(x_i|\mu,\lambda^{-1}) d\lambda \\
\log N(\mu|\mu_N,\lambda_N^{-1}) &= \log N(\mu|\mu_0,\lambda_0^{-1}) + \int_\lambda Gam(\lambda|a_N,b_N) \sum_{i=1}^N \log N(x_i|\mu,\lambda^{-1}) d\lambda - 1 \\
N(\mu|\mu_N,\lambda_N^{-1}) &\propto \exp\left\{\log N(\mu|\mu_0,\lambda_0^{-1}) + \int_\lambda Gam(\lambda|a_N,b_N) \sum_{i=1}^N \log N(x_i|\mu,\lambda^{-1}) d\lambda\right\} \\
N(\mu|\mu_N,\lambda_N^{-1}) &\propto \exp\left\{\int_\lambda Gam(\lambda|a_N,b_N) \log\left(N(\mu|\mu_0,\lambda_0^{-1}) \prod_{i=1}^N N(x_i|\mu,\lambda^{-1})\right) d\lambda\right\} \qquad (30)
\end{aligned}$$

Using the results (17), (30) can be written as:

$$
\begin{aligned}
N(\mu|\mu_N, \lambda_N^{-1}) &\propto \exp\left\{\int_\lambda Gam(\lambda|a_N, b_N) \log N(\mu|\mu_X, \lambda_X^{-1}) d\lambda\right\} \\
\mu_X &= \frac{N\lambda\mu_{ML} + \lambda_0\mu_0}{N\lambda + \lambda_0} \\
\lambda_X &= N\lambda + \lambda_0 \\
N(\mu|\mu_N, \lambda_N^{-1}) &\propto \exp\left\{E_{Gam(\lambda|a_N,b_N)} \log N(\mu|\mu_X, \lambda_X^{-1})\right\} \\
N(\mu|\mu_N, \lambda_N^{-1}) &\propto \exp\left\{E_{Gam(\lambda|a_N,b_N)}\left[\frac{1}{2}\log(N\lambda + \lambda_0) - \frac{(N\lambda + \lambda_0)}{2}\left(\mu - \frac{N\lambda\mu_{ML} + \lambda_0\mu_0}{N\lambda + \lambda_0}\right)^2\right]\right\} \\
N(\mu|\mu_N, \lambda_N^{-1}) &\propto \exp\left\{E_{Gam(\lambda|a_N,b_N)}\left[-\frac{(N\lambda + \lambda_0)}{2}\left(\mu - \frac{N\lambda\mu_{ML} + \lambda_0\mu_0}{N\lambda + \lambda_0}\right)^2\right]\right\} \\
N(\mu|\mu_N, \lambda_N^{-1}) &\propto \exp\left\{E_{Gam(\lambda|a_N,b_N)}\left[-\frac{(N\lambda + \lambda_0)}{2}\left(\mu^2 - 2\mu\frac{N\lambda\mu_{ML} + \lambda_0\mu_0}{N\lambda + \lambda_0} + \left(\frac{N\lambda\mu_{ML} + \lambda_0\mu_0}{N\lambda + \lambda_0}\right)^2\right)\right]\right\} \\
N(\mu|\mu_N, \lambda_N^{-1}) &\propto \exp\left\{E_{Gam(\lambda|a_N,b_N)}\left[-\frac{(N\lambda + \lambda_0)}{2}\mu^2 + \frac{(N\lambda + \lambda_0)}{2}2\mu\frac{N\lambda\mu_{ML} + \lambda_0\mu_0}{N\lambda + \lambda_0}\right]\right\} \\
N(\mu|\mu_N, \lambda_N^{-1}) &\propto \exp\left\{E_{Gam(\lambda|a_N,b_N)}\left[-\frac{(N\lambda + \lambda_0)}{2}\mu^2 + (N\lambda\mu_{ML} + \lambda_0\mu_0)\mu\right]\right\} \\
N(\mu|\mu_N, \lambda_N^{-1}) &\propto \exp\left\{E_{Gam(\lambda|a_N,b_N)}\left[-\frac{N\lambda}{2}\mu^2 - \frac{\lambda_0}{2}\mu^2 + N\lambda\mu_{ML}\mu + \lambda_0\mu_0\mu\right]\right\} \\
N(\mu|\mu_N, \lambda_N^{-1}) &\propto \exp\left\{-\frac{\lambda_0}{2}\mu^2 + \lambda_0\mu_0\mu + E_{Gam(\lambda|a_N,b_N)}\left[-\frac{N\lambda}{2}\mu^2 + N\lambda\mu_{ML}\mu\right]\right\} \\
N(\mu|\mu_N, \lambda_N^{-1}) &\propto \exp\left\{-\frac{\lambda_0}{2}\mu^2 + \lambda_0\mu_0\mu - \frac{N \cdot E_{Gam(\lambda|a_N,b_N)}[\lambda]}{2}\mu^2 + N \cdot E_{Gam(\lambda|a_N,b_N)}[\lambda]\mu_{ML}\mu\right\} \\
N(\mu|\mu_N, \lambda_N^{-1}) &\propto \exp\left\{-\frac{\lambda_0}{2}\mu^2 - \frac{N \cdot E_{Gam(\lambda|a_N,b_N)}[\lambda]}{2}\mu^2 + \lambda_0\mu_0\mu + N \cdot E_{Gam(\lambda|a_N,b_N)}[\lambda]\mu_{ML}\mu\right\} \\
N(\mu|\mu_N, \lambda_N^{-1}) &\propto \exp\left\{-\frac{(N \cdot E_{Gam(\lambda|a_N,b_N)}[\lambda] + \lambda_0)}{2}\mu^2 + (N \cdot E_{Gam(\lambda|a_N,b_N)}[\lambda]\mu_{ML} + \lambda_0\mu_0)\mu\right\}
\end{aligned}
$$

Completing the squares, one can derive the following parameters of the approximating normal distribution $N(\mu|\mu_N, \lambda_N^{-1})$:

$$
\begin{aligned}
\mu_N &= \frac{N \cdot E_{Gam(\lambda|a_N,b_N)}[\lambda]\mu_{ML} + \lambda_0\mu_0}{N \cdot E_{Gam(\lambda|a_N,b_N)}[\lambda] + \lambda_0} \\
\lambda_N &= N \cdot E_{Gam(\lambda|a_N,b_N)}[\lambda] + \lambda_0
\end{aligned}
$$

This can be further simplified using the mean for the gamma distribution ($E_{Gam(\lambda|a,b)}[\lambda] = a/b$) as

$$
\begin{aligned}
N(\mu|\mu_N, \lambda_N^{-1}) &\equiv N(\mu|\mu_N, \lambda_N^{-1}; a_N, b_N) \\
\mu_N &= \frac{N \cdot (a_N/b_N)\mu_{ML} + \lambda_0\mu_0}{N \cdot (a_N/b_N) + \lambda_0} \\
\lambda_N &= N \cdot (a_N/b_N) + \lambda_0
\end{aligned} \tag{31}
$$

Similarly, one can derive the approximation for the posterior probability for λ.

$$
\begin{aligned}
\frac{\partial KL(q||p)}{\partial Gam(\lambda|a_N, b_N)} &= \frac{\partial}{\partial Gam(\lambda|a_N, b_N)} \int_\lambda Gam(\lambda|a_N, b_N) \log Gam(\lambda|a_N, b_N) d\lambda \\
&- \frac{\partial}{\partial Gam(\lambda|a_N, b_N)} \int_\lambda Gam(\lambda|a_N, b_N) \log Gam(\lambda|a_0, b_0) d\lambda \\
&- \frac{\partial}{\partial Gam(\lambda|a_N, b_N)} \int_{\mu,\lambda} N(\mu|\mu_N, \lambda_N^{-1}) Gam(\lambda|a_N, b_N) \sum_{i=1}^{N} \log N(x_i|\mu, \lambda^{-1}) d\mu d\lambda \\
&= (\log Gam(\lambda|a_N, b_N) + 1) - \log Gam(\lambda|a_0, b_0) - \int_\lambda N(\mu|\mu_N, \lambda_N^{-1}) \sum_{i=1}^{N} \log N(x_i|\mu, \lambda^{-1}) d\lambda
\end{aligned}
$$

Setting the derivative equal to zero, one can compute the approximation:

$$
\begin{aligned}
0 &= (\log Gam(\lambda|a_N, b_N) + 1) - \log Gam(\lambda|a_0, b_0) - \int_\lambda N(\mu|\mu_N, \lambda_N^{-1}) \sum_{i=1}^{N} \log N(x_i|\mu, \lambda^{-1}) d\lambda \\
\log Gam(\lambda|a_N, b_N) &= \log Gam(\lambda|a_0, b_0) + \int_\lambda N(\mu|\mu_N, \lambda_N^{-1}) \sum_{i=1}^{N} \log N(x_i|\mu, \lambda^{-1}) d\lambda - 1 \\
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ \log Gam(\lambda|a_0, b_0) + \int_\lambda N(\mu|\mu_N, \lambda_N^{-1}) \sum_{i=1}^{N} \log N(x_i|\mu, \lambda^{-1}) d\lambda \right\} \\
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ \int_\lambda N(\mu|\mu_N, \lambda_N^{-1}) \log \left(Gam(\lambda|a_0, b_0) \prod_{i=1}^{N} N(x_i|\mu, \lambda^{-1}) \right) d\lambda \right\}
\end{aligned}
\tag{32}
$$

Using the results (20), (32) can be written as:

$$
\begin{aligned}
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ \int_\lambda N(\mu|\mu_N, \lambda_N^{-1}) \log Gam(\lambda|a_X, b_X) d\lambda \right\} \\
a_X &= a_0 + \frac{N}{2} \\
b_X &= b_0 + \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^2 \\
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ E_{N(\mu|\mu_N, \lambda_N^{-1})} \log Gam(\lambda|a_X, b_X) \right\} \\
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ E_{N(\mu|\mu_N, \lambda_N^{-1})} \left[\left(a_0 + \frac{N}{2} - 1 \right) \log \lambda - \left(b_0 + \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^2 \right) \lambda \right] \right\} \\
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ E_{N(\mu|\mu_N, \lambda_N^{-1})} \left[\left(a_0 + \frac{N}{2} - 1 \right) \log \lambda - b_0 \lambda - \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^2 \lambda \right] \right\} \\
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ \left(a_0 + \frac{N}{2} - 1 \right) \log \lambda - b_0 \lambda - E_{N(\mu|\mu_N, \lambda_N^{-1})} \left[\frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^2 \lambda \right] \right\} \\
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ \left(a_0 + \frac{N}{2} - 1 \right) \log \lambda - b_0 \lambda - \frac{\lambda}{2} \sum_{i=1}^{N} E_{N(\mu|\mu_N, \lambda_N^{-1})} \left[(x_i - \mu)^2 \right] \right\} \\
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ \left(a_0 + \frac{N}{2} - 1 \right) \log \lambda - b_0 \lambda - \frac{\lambda}{2} \sum_{i=1}^{N} E_{N(\mu|\mu_N, \lambda_N^{-1})} \left[x_i^2 - 2 x_i \mu + \mu^2 \right] \right\} \\
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ \left(a_0 + \frac{N}{2} - 1 \right) \log \lambda - b_0 \lambda - \frac{\lambda}{2} \sum_{i=1}^{N} \left(x_i^2 - 2 x_i E_{N(\mu|\mu_N, \lambda_N^{-1})} [\mu] + E_{N(\mu|\mu_N, \lambda_N^{-1})} \left[\mu^2 \right] \right) \right\} \\
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ \left(a_0 + \frac{N}{2} - 1 \right) \log \lambda - b_0 \lambda - \frac{\lambda}{2} \sum_{i=1}^{N} \left(x_i^2 - 2 x_i \mu_N + \mu_N^2 + \lambda_N^{-1} \right) \right\} \\
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ \left(a_0 + \frac{N}{2} - 1 \right) \log \lambda - \left(b_0 + \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu_N)^2 + \frac{N \lambda_N^{-1}}{2} \right) \lambda \right\} \\
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ \log \lambda^{\left(a_0 + \frac{N}{2} - 1 \right)} - \left(b_0 + \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu_N)^2 + \frac{N \lambda_N^{-1}}{2} \right) \lambda \right\} \\
Gam(\lambda|a_N, b_N) &\propto \lambda^{a_0 + \frac{N}{2} - 1} \exp\left\{ - \left(b_0 + \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu_N)^2 + \frac{N \lambda_N^{-1}}{2} \right) \lambda \right\}
\end{aligned}
$$

Now, one can derive parameters for the approximating gamma distribution $Gam(\lambda|a_N, b_N)$:

$$
\begin{aligned}
Gam(\lambda|a_N, b_N) &\equiv Gam(\lambda|a_N, b_N; \mu_N, \lambda_N^{-1}) \\
a_N &= a_0 + \frac{N}{2} \\
b_N &= b_0 + \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu_N)^2 + \frac{N \lambda_N^{-1}}{2}
\end{aligned}
\tag{33}
$$

6 Expectation Propagation

Expectation Propagation (EP) can be used to approximate an distribution by a simpler *parametric distribution*, in a similar way as Variational Inference (VI). It is based on the minimisation of the KL-divergence, but in its direct way $KL(p||q)$ instead of $KL(q||p)$ (the one used by VI). Ideally, we would like to minimise $KL(p||q)$ directly:

$$KL(p||q) = \int p(\boldsymbol{w}|D) \log \left\{ \frac{p(\boldsymbol{w}|D)}{q(\boldsymbol{w})} \right\} d\boldsymbol{w}$$

This involves computing averages with respect to the exact posterior which is intractable. However, we compute the approximation because we assume that we cannot handle the true posterior distribution in the first place. Therefore, EP minimises the KL divergence between $\hat{p}(\boldsymbol{w})$ and $q(\boldsymbol{w})$, where $\hat{p}(\boldsymbol{w})$ is an approximation of $p(\boldsymbol{w}|D)$.

$$KL(\hat{p}||q) = \int \hat{p}(\boldsymbol{w}) \log \left\{ \frac{\hat{p}(\boldsymbol{w})}{q(\boldsymbol{w})} \right\} d\boldsymbol{w}$$

The facts and assumptions:

- EP is a generalisation of LBP to graphical models which may contain continuous variables
- The distribution q is restricted to belong to a family of probability distributions that is *closed under the product operation* - the exponential distribution family.

6.1 The exponential distribution family

Most of the simplest parametric distributions belong to the exponential family. This includes the normal, exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, categorical, Poisson, Wishart, Inverse Wishart and others.

If $q(\boldsymbol{w})$ is from the exponential family then

$$q(\boldsymbol{w}) = h(\boldsymbol{w}) g(\boldsymbol{\eta}) \exp \left\{ \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{w}) \right\} \tag{34}$$

where

- $\boldsymbol{\eta}$ is a vector of natural parameters of q,
- $\boldsymbol{u}(\boldsymbol{w})$ are the sufficient statistics,
- $g(\boldsymbol{\eta})$ is a log normaliser which satisfies:

$$g(\boldsymbol{\eta}) \int h(\boldsymbol{w}) \exp \left\{ \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{w}) \right\} d\boldsymbol{w} = 1 \tag{35}$$

Consider minimising KL-divergence between $p(\boldsymbol{w})$ and $q(\boldsymbol{w})$, where $p(\boldsymbol{w})$ is a fixed distribution and $q(\boldsymbol{w})$ is a member of the exponential family:

$$KL(p||q) = -\log g(\boldsymbol{\eta}) - \boldsymbol{\eta}^T E_{p(\boldsymbol{w})}[\boldsymbol{u}(\boldsymbol{w})] + \text{const}$$

We can minimise $KL(p||q)$ with respect to the natural parameters $\boldsymbol{\eta}$

$$\begin{aligned}
\frac{\partial KL(p||q)}{\partial \boldsymbol{\eta}} &= 0 \\
\frac{\partial}{\partial \boldsymbol{\eta}} - \log g(\boldsymbol{\eta}) &= E_{p(\boldsymbol{w})}[\boldsymbol{u}(\boldsymbol{w})] \\
-\nabla \log g(\boldsymbol{\eta}) &= E_{p(\boldsymbol{w})}[\boldsymbol{u}(\boldsymbol{w})].
\end{aligned} \tag{36}$$

Differentiating both sides of (35) with respect to $\boldsymbol{\eta}$ we get:

$$\begin{aligned}
\nabla \left[g(\boldsymbol{\eta}) \int h(\boldsymbol{w}) \exp \left\{ \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{w}) \right] d\boldsymbol{w} \right\} &= \nabla 1 \\
\nabla g(\boldsymbol{\eta}) \int h(\boldsymbol{w}) \exp \left\{ \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{w}) \right\} d\boldsymbol{w} + g(\boldsymbol{\eta}) \int h(\boldsymbol{w}) \exp \left\{ \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{w}) \right\} \boldsymbol{u}(\boldsymbol{w}) d\boldsymbol{w} &= 0 \\
-\nabla g(\boldsymbol{\eta}) \frac{1}{g(\boldsymbol{\eta})} g(\boldsymbol{\eta}) \int h(\boldsymbol{w}) \exp \left\{ \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{w}) \right\} d\boldsymbol{w} &= g(\boldsymbol{\eta}) \int h(\boldsymbol{w}) \exp \left\{ \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{w}) \right\} \boldsymbol{u}(\boldsymbol{w}) d\boldsymbol{w} \\
-\nabla g(\boldsymbol{\eta}) \frac{1}{g(\boldsymbol{\eta})} \cdot 1 &= g(\boldsymbol{\eta}) \int h(\boldsymbol{w}) \exp \left\{ \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{w}) \right\} \boldsymbol{u}(\boldsymbol{w}) d\boldsymbol{w} \\
-\frac{1}{g(\boldsymbol{\eta})} \nabla g(\boldsymbol{\eta}) &= g(\boldsymbol{\eta}) \int h(\boldsymbol{w}) \exp \left\{ \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{w}) \right\} \boldsymbol{u}(\boldsymbol{w}) d\boldsymbol{w} \\
-\nabla \log g(\boldsymbol{\eta}) &= E_{q(\boldsymbol{w})}[\boldsymbol{u}(\boldsymbol{w})]
\end{aligned}$$

Comparing this result with (36), we get:

$$E_{p(\boldsymbol{w})}[\boldsymbol{u}(\boldsymbol{w})] = E_{q(\boldsymbol{w})}[\boldsymbol{u}(\boldsymbol{w})]$$

Minimising $KL(p||q)$ is equivalent to matching expected sufficient statistics. This result is systematically exploited in EP to carry out approximate inference. This method is called moment matching.

Example: If $q(\boldsymbol{w})$ is a Normal distribution $N(\boldsymbol{w}|\boldsymbol{\mu}, \Sigma)$, we minimise the KL divergence by setting the mean $\boldsymbol{\mu}$ and covariance Σ of $q(\boldsymbol{w})$ to the mean and covariance of $p(\boldsymbol{w})$. That is:

- $E_q[\boldsymbol{w}] \equiv E_p[\boldsymbol{w}]$
- $E_q[\boldsymbol{w}\boldsymbol{w}^T] \equiv E_p[\boldsymbol{w}\boldsymbol{w}^T]$.

6.2 The Expectation Propagation algorithm

6.2.1 Factorisation of the joint distribution

In EP, we assume that the joint distribution $p(\boldsymbol{w}, D)$ of the latent variables $\boldsymbol{w}$ and the observed variables D factors as

$$p(\boldsymbol{w}, D) = \prod_i f_i(\boldsymbol{w}),$$

where each factor f_i depends on $\boldsymbol{w}$ or a subset of these variables and D. Factors f_i can be arbitrary; however, they are very often represented by likelihood or prior for $\boldsymbol{w}$.

Please recall that $p(\boldsymbol{w}, D)$ can be written very often as:

$$p(\boldsymbol{w}, D) = p(\boldsymbol{w}) \prod_i^N p(x_i|\boldsymbol{w})$$

Given $p(\boldsymbol{w}, D)$, the posterior for $\boldsymbol{w}$ is obtained after normalising by $p(D)$:

$$p(\boldsymbol{w}|D) = \frac{1}{p(D)} \prod_i f_i(\boldsymbol{w})$$

$$p(D) = \int \prod_i f_i(\boldsymbol{w}) d\boldsymbol{w}$$

Example: Show factor graph and factors for Unknown Mean and Variance of a normal distribution.

Figure 11: Example of factor graph and factors for Unknown Mean and Variance of a normal distribution.

6.2.2 Approximation to the posterior distribution

As in VI, EP assumes that q factorises with respect to a partition of w into M disjoint groups w_j, with $j = \{1, \ldots, M\}$

$$q(\boldsymbol{w}) = \prod_j^M q_j(w_j)$$

In addition, it also considers that EP approximates $p(\boldsymbol{w}|D)$ using a product of simpler factors: which can be also written as

$$q(\boldsymbol{w}) = \frac{1}{Z} \prod_i^N \tilde{f}_i(\boldsymbol{w})$$

where each approximate factor $\tilde{f}_i$ approximates the corresponding exact factor f_i. The $\tilde{f}_i$ are in an exponential family but need not be normalised. For example, the $\tilde{f}_i$ can be un-normalised Normal distribution. Because the exponential family is closed under the product operation, the product of the $\tilde{f}_i(\boldsymbol{w})$ has a simple form and can be easily normalised.

We need to use the $\tilde{f}_i$ approximations as the original factors f_i may be too complex.

Example: Consider the clutter problem:

- $p(w) = N(w; 0, 100)$
- $p(x|w) = 0.5 \cdot N(x; w, 1) + 0.5 \cdot N(x; 0, 10)$
- $p(w|\boldsymbol{x}) \propto p(w) \prod_i^N p(x_i|w)$

Note that the $p(w|\boldsymbol{x})$ is a mixture of 2^N terms.

6.2.3 Minimising the KL divergence

We already noted that we cannot minimise the $KL(p||q)$ directly.

$$\begin{aligned} KL(p||q) &= \int p(\boldsymbol{w}|D) \log \left\{ \frac{p(\boldsymbol{w}|D)}{q(\boldsymbol{w})} \right\} d\boldsymbol{w} \\ &= KL\left(\frac{1}{p(D)} \prod_i f_i(\boldsymbol{w}) || \frac{1}{Z} \prod_i \tilde{f}_i(\boldsymbol{w}) \right) \end{aligned}$$

Therefore, EP minimises the KL divergence between f_i and $\tilde{f}_i$ in the context of all the other approximate factors $\tilde{f}_j$, $j \neq i$.

1. Assume that we want to update the factor $\tilde{f}_j$.

2. We remove $\tilde{f}_i$ from q to obtain:

$$q^{\setminus i}(\boldsymbol{w}) = \frac{q(\boldsymbol{w})}{\tilde{f}_i(\boldsymbol{w})} = \frac{1}{Z} \prod_{j \neq i} \tilde{f}_j(\boldsymbol{w}) \propto \prod_{j \neq i} \tilde{f}_j(\boldsymbol{w}).$$

where $q^{\setminus i}(\boldsymbol{w})$ is called a cavity distribution. Please note that

- $q^{\setminus i}(\boldsymbol{w})$ is un-normalised distribution because we removed potentially un-normalised $\tilde{f}_i$,
- division of q by $\tilde{f}_i$ may be often faster then multiplying N factors $\tilde{f}_j(\boldsymbol{w})$.

3. We want to update the factor $\tilde{f}_i$ so that

$$q_{new}(\boldsymbol{w}) \propto \tilde{f}_i(\boldsymbol{w}) q^{\setminus i}(\boldsymbol{w}) \propto \tilde{f}_i(\boldsymbol{w}) \prod_{j \neq i} \tilde{f}_j(\boldsymbol{w})$$

is as close as possible in terms of the KL divergence to

$$\hat{p}(\boldsymbol{w}) \propto f_i(\boldsymbol{w}) q^{\setminus i}(\boldsymbol{w}) \propto f_i(\boldsymbol{w}) \prod_{j \neq i} \tilde{f}_j(\boldsymbol{w}),$$

where $q^{\setminus i}$ is kept fixed (all factors $j \neq i$ are fixed). Considering f_i in the context of all the other approximate factors $\tilde{f}_j$, $j \neq i$ ensures that $\tilde{f}_i$ is accurate where $q^{\setminus i} = \prod_{j \neq i} \tilde{f}_j$ takes large values.

4. To obtain a normalised distribution $\hat{p}(\boldsymbol{w})$, we compute the normalisation constant Z_i

$$Z_i = \int f_i(\boldsymbol{w}) q^{\setminus i}(w) d\boldsymbol{w}$$

and set

$$\hat{p}(\boldsymbol{w}) = \frac{1}{Z_i} f_i(\boldsymbol{w}) q^{\setminus i}(\boldsymbol{w}).$$

5. We minimise $KL(\hat{p}(\boldsymbol{w})||q_{new})$ divergence by with respect to q_{new} for factor $\tilde{f}_i$:

$$KL\left(\frac{1}{Z_i} f_i(\boldsymbol{w}) q^{\setminus i}(\boldsymbol{w}) || q_{new}(\boldsymbol{w}) \right).$$

This can be done by matching expected sufficient statistics between q_{new} and $(1/Z_i) f_i q^{\setminus i}$. For this, expectations with respect to $(1/Z_i) f_i q^{\setminus i}$ must be tractable. In other words, we will construct a new "distribution" with all the moments of the q_{new} equal to moments $(1/Z_i) f_i q^{\setminus i}$.

6. Then based on the the new approximation of the q_{new}, we update $\tilde{f}_i$ using

$$\tilde{f}_i(\boldsymbol{w}) = K\frac{q_{new}(\boldsymbol{w})}{q^{\setminus i}(\boldsymbol{w})}$$

Please recall that $q_{new} \propto \tilde{f}_i(\boldsymbol{w})q^{\setminus i}(\boldsymbol{w})$. The normalisation constant K can be obtained by multiplying both sides by $q^{\setminus i}(\boldsymbol{w})$ and integration.

$$\begin{aligned} K\frac{q_{new}(\boldsymbol{w})}{q^{\setminus i}(\boldsymbol{w})} &= \tilde{f}_i(\boldsymbol{w}) \\ Kq_{new}(\boldsymbol{w}) &= \tilde{f}_i(\boldsymbol{w})q^{\setminus i}(\boldsymbol{w}) \\ K\int q_{new}(\boldsymbol{w})d\boldsymbol{w} &= \int \tilde{f}_i(\boldsymbol{w})q^{\setminus i}(\boldsymbol{w})d\boldsymbol{w} \\ K &= \int \tilde{f}_i(\boldsymbol{w})q^{\setminus i}(\boldsymbol{w})d\boldsymbol{w} \end{aligned}$$

Please note that $q_{new}(\boldsymbol{w})$ is a normalised distribution. Comparing zero moments one can found that:

$$\int \tilde{f}_i(\boldsymbol{w})q^{\setminus i}(\boldsymbol{w}) = \int f_i(\boldsymbol{w})q^{\setminus i}(\boldsymbol{w})$$

Since the right side is Z_i, we know that $K = Z_i$. This is especially important as for computation of K we need to know $\tilde{f}_i(\boldsymbol{w})$. However without K, we do know it.

Several passes are made trough the factors until they converge.

6.2.4 Summary

EP computes $q(\boldsymbol{w})$ – approximation to $p(\boldsymbol{w}|D)$.

1. Initialise q and each $\tilde{f}_i$ to be uniform.
2. Repeat until convergence of all the $\tilde{f}_i$:
   (a) Choose a factor $\tilde{f}_i$ to refine.
   (b) Remove $\tilde{f}_i$ from q by division $q^{\setminus i} = q/\tilde{f}_i$.
   (c) Compute Z_i and find q_{new} by minimising $KL(\hat{p}||q_{new})$.
   (d) Compute and store the new factor $\tilde{f}_i = Z_i\frac{q_{new}}{q^{\setminus i}}$.
3. Evaluate the approximation to the model evidence:

$$p(D) \approx Z = \int \prod_i \tilde{f}_i(\boldsymbol{w})d\boldsymbol{w}.$$

Considerations:

- The minimisation of the KL is done by moment matching (not necessarily).
- EP may not converge and the $\tilde{f}_i$ may oscillate forever (same as in LBP).
- Convergence can be improved by damping the EP updates.
- No need to replace all the factors in the joint distribution with approximations. For example, if one factor is already in the exponential family, the approximate factor is always the same and exact.
- EP considers global aspects of p by approximately minimising $KL(p||q)$.

6.3 Supporting math: Gaussian Identities

The product and ratio of Gaussians is again Gaussian.

$$N(\mu_1, \Sigma_1) \cdot N(\mu_2, \Sigma_2) = CN(\mu, \Sigma),$$

$$\begin{aligned}
\Sigma &= (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}, \\
\mu &= \Sigma \left(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2\right), \\
C &= \sqrt{\frac{|\Sigma|}{(2\pi)^d |\Sigma_1||\Sigma_2|}} \exp\left\{-\frac{1}{2}\left(\mu_1^T\Sigma_1^{-1}\mu_1 + \mu_2^T\Sigma_2^{-1}\mu_2 - \mu^T\Sigma^{-1}\mu\right)\right\}.
\end{aligned}$$

$$N(\mu_1, \Sigma_1)/N(\mu_2, \Sigma_2) = CN(\mu, \Sigma),$$

$$\begin{aligned}
\Sigma &= (\Sigma_1^{-1} - \Sigma_2^{-1})^{-1}, \\
\mu &= \Sigma \left(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2\right), \\
C &= \sqrt{\frac{|\Sigma||\Sigma_2|}{(2\pi)^d |\Sigma_1|}} \exp\left\{-\frac{1}{2}\left(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_2^T\Sigma_2^{-1}\mu_2 - \mu^T\Sigma^{-1}\mu\right)\right\}.
\end{aligned}$$

For more information see `http://research.microsoft.com/en-us/um/people/minka/papers/ep/minka-ep-quickref.pdf`

6.4 Supporting math: Gaussian Moments

Let $f(x)$ be an arbitrary factor of x and let

$$\hat{p}(x) = \frac{1}{Z} t(x) N(x|\mu, \Sigma),$$

where

$$Z = \int t(x) N(x|\mu, \Sigma).$$

Then, we have that

$$E_{\hat{p}}[x] = \mu + \Sigma \frac{\partial \log Z}{\partial \mu},$$

$$Var_{\hat{p}}[x] = E_{\hat{p}}[xx^T] - E_{\hat{p}}[x]E_{\hat{p}}[x]^T = \Sigma - \Sigma\left(\frac{\partial \log Z}{\partial \mu}\left(\frac{\partial \log Z}{\partial \mu}\right)^T - 2\frac{\partial \log Z}{\partial \Sigma}\right)\Sigma.$$

These expressions are very useful to find the parameters of q_{new} in EP.

6.5 Example: EP - Unknown Mean of a normal distribution

In this section, we will try to compute the posterior distribution for the mean parameter of the normal distribution. Figure 12 depict a graphical model representing inference of the mean of an univariate normal distribution with fixed precision.

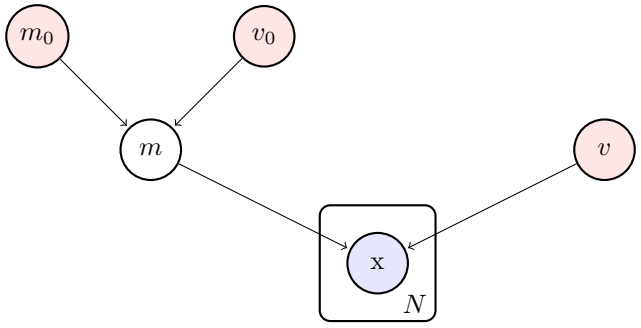


Figure 12: The graphical model for prediction of the observation x given the mean m and the variance v.

Prior As we set the priors for m to be:

$$p(m) = N(m|m_0, v_0),$$

therefore we get:

$$\begin{aligned} p(\boldsymbol{w}) &= p(m|m_0, v_0), \\ &= N(m|m_0, v_0). \end{aligned}$$

where $\boldsymbol{w} = [m,]$.

Likelihood The likelihood is defined as follows:

$$p(D|\boldsymbol{w}) = \prod_i p(x_i|w) = \prod_i N(x_i|m, v),$$

where $D = \{x_1, x_2, \ldots, x_N\}$.

Joint distribution The join distribution is defined as follows:

$$\begin{aligned} p(\boldsymbol{w}, D) &= p(m, D) \\ &= p(m|m_0, v_0) \prod_i p(x_i|w) \\ &= N(m|m_0, v_0) \prod_i N(x_i|m, v), \\ &= f_0(\boldsymbol{w}) \prod_i f_i(\boldsymbol{w}). \end{aligned}$$

where

$$\begin{aligned} f_0(\boldsymbol{w}) &= N(m|m_0, v_0), \\ f_i(\boldsymbol{w}) &= N(x_i|m, v). \end{aligned}$$

Posterior The posterior distribution is defined as follows:

$$\begin{aligned} p(\boldsymbol{w}|D) &= \frac{1}{p(D)} p(\boldsymbol{w}, D) \\ &= \frac{1}{p(D)} \prod_i f_i(w). \end{aligned}$$

However, it is in-tractable. Therefore, we choose the posterior to have for of the prior:

$$\begin{aligned} q(\boldsymbol{w}) &= q_m(m), \\ &= q_m(m|m_N, v_N), \\ &= N(m|m_N, v_N). \end{aligned}$$

We also approximate $q(\boldsymbol{w})$ as a product of simple factors:

$$q(\boldsymbol{w}) = \frac{1}{Z} \prod_i \tilde{f}_i(\boldsymbol{w}),$$

where

$$\begin{aligned} \tilde{f}_0(\boldsymbol{w}) &= N(m|\tilde{m}_0, \tilde{v}_0), \\ \tilde{f}_i(\boldsymbol{w}) &= \tilde{s}_i N(m|\tilde{m}_i, \tilde{v}_i) \quad i = 1, \ldots, N. \end{aligned}$$

Note that the $\tilde{s}_i$ is de-normalisation constant to make sure that our approximate $\tilde{f}_i$ fits well the original f_i. This is useful as the original factor f_i is normalised with respect to x_i, while the approximation is $\tilde{f}_i$ with respect to m.

Initialisation We initialise our approximation of $q(\boldsymbol{w})$ by setting

- $\tilde{f}_0$ to the prior $N(m|m_0, v_0)$,
- $\tilde{f}_i$ to have zero mean and high low precision.

Computation of the cavity distribution Computing the cavity distribution $q^{\setminus j}(\boldsymbol{w})$ we get:

$$q^{\setminus i}(\boldsymbol{w}) = \frac{q(\boldsymbol{w})}{\tilde{f}_i(\boldsymbol{w})}.$$

Since both $q(\boldsymbol{w})$ and $\tilde{f}_i(\boldsymbol{w})$ take form of Normal distributions, we can use the formulas for Normal identities to obtain un-normalised Normal shaped $q^{\setminus i}(\boldsymbol{w})$:

$$q^{\setminus i}(\boldsymbol{w}) \propto N(m|m_N^{\setminus i}, v_N^{\setminus i})$$

where:

$$\begin{aligned} v_N^{\setminus i} &= (v_N^{-1} - \tilde{v}_i^{-1})^{-1}, \\ m_N^{\setminus i} &= v_n^{\setminus i}(v_N^{-1} m_N - \tilde{v}_i^{-1}\tilde{m}_i). \end{aligned}$$

Computation of the new posterior The first step is to compute the Z_i:

$$\begin{aligned} Z_i &= \int f_i(m) q^{\setminus i}(m) dm \\ &= \int N(x_i|m, v) N(m|m_N^{\setminus i}, v_N^{\setminus i}) dm \\ &= N(x_i|m_N^{\setminus i}, v + v_N^{\setminus i}) \end{aligned}$$

which is obtained from the convolution of two Gaussians.

Next, we compute q_{new} by finding the mean and the variance of $\hat{p}(m) \propto f_i(m) q^{\setminus i}(m)$. Using the Gaussian moments, we obtain the mean and the variance of the new approximate posterior $q_{new}(m)$:

$$\begin{aligned} \hat{p}(m) &= \frac{1}{Z_i} f_i(m) q^{\setminus i}(m) \\ &= \frac{1}{Z_i} N(x_i|m, v) N(m|m_N^{\setminus i}, v_N^{\setminus i}) \\ &= N(m|m_{new}, v_{new}) \end{aligned}$$

$$\begin{aligned} m_{new} &= m_N^{\setminus i} + v_N^{\setminus i} \frac{\partial \log Z_i}{\partial m_N^{\setminus i}} \\ v_{new} &= v_N^{\setminus i} - \left(v_N^{\setminus i}\right)^2 \left(\left(\frac{\partial \log Z_i}{\partial m_N^{\setminus i}} \right)^2 - 2 \frac{\partial \log Z_i}{\partial v_N^{\setminus i}} \right) \end{aligned}$$

Computation of the derivatives of $\log Z_i$:

$$\frac{\partial \log Z_i}{\partial m_N^{\setminus i}} =$$

$$\frac{\partial \log Z_i}{\partial v_N^{\setminus i}} =$$

Computing the q_{new} distribution.
TBD

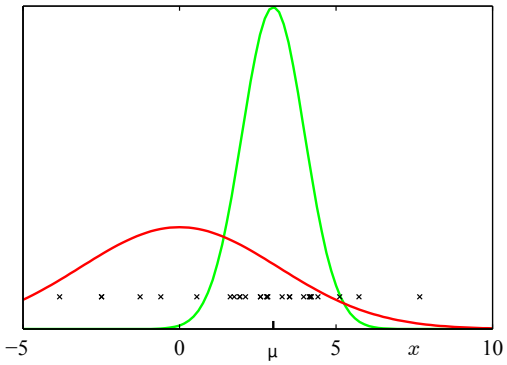
Update the approximate factor Updating the $\tilde{f}_j$ factor.
TBD

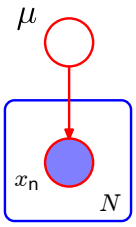
6.6 Example: EP - The clutter problem

We consider the problem of inferring the mean $\boldsymbol{\mu}$ of a multivariate Gaussian when the Gaussian observations are embedded in background Gaussian clutter.
In this problem $\boldsymbol{w} = \boldsymbol{\mu}$ and D are the observations $\boldsymbol{x}$, which are generated from:

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = (1 - w) N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{I}) + w N(\boldsymbol{x}|\boldsymbol{0}, \boldsymbol{I}a),$$

where $w = 0.5$ is the proportion of clutter and $a = 10$.





Prior The prior for $\boldsymbol{\mu}$ is:

$$p(\boldsymbol{\mu}) = N(\boldsymbol{\mu}|0, \boldsymbol{I}b),$$

with $b = 100$ (little informative).

Likelihood The likelihood is defined as follows:

$$p(D|\boldsymbol{w}) = \prod_i p(x_i|w) = \prod_i \left[(1-w)N(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{I}) + wN(\boldsymbol{x}_i|\boldsymbol{0}, \boldsymbol{I}a)\right],$$

where $D = \{x_1, x_2, \ldots, x_N\}$.

Joint distribution The joint distribution of $\boldsymbol{\mu}$ and the evidence $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ is

$$\begin{aligned} p(\boldsymbol{w}, D) &= p(\boldsymbol{\mu}, D) \\ &= p(\boldsymbol{\mu}) \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\mu}) \\ &= N(\boldsymbol{\mu}|0, \boldsymbol{I}b) \prod_i \left[(1-w)N(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{I}) + wN(\boldsymbol{x}_i|\boldsymbol{0}, \boldsymbol{I}a)\right] \\ &= f_0(\boldsymbol{\mu}) \prod_{i=1}^{N} f_i(\boldsymbol{\mu}), \end{aligned}$$

a mixture of 2^N terms. Computing $p(\boldsymbol{\mu}|D)$ is intractable for large N.

Posterior The posterior distribution is defined as follows:

$$\begin{aligned} p(\boldsymbol{w}|D) &= \frac{1}{p(D)} p(\boldsymbol{w}, D) \\ &= \frac{1}{p(D)} \prod_i f_i(w). \end{aligned}$$

However, this is intractable.

We choose a parametric form for q that belongs to the exponential family:

$$q(\boldsymbol{\mu}) = N(\boldsymbol{\mu}|\boldsymbol{m}, v\boldsymbol{I}),$$

with parameters $\boldsymbol{m}$ and v.
We also approximate $q(\boldsymbol{w})$ as a product of simple factors:

$$q(\boldsymbol{w}) = \frac{1}{Z} \prod_i \tilde{f}_i(\boldsymbol{w}),$$

where

$$\tilde{f}_0(\boldsymbol{\mu}) = N(\boldsymbol{\mu}|\tilde{\boldsymbol{m}}_0, \tilde{v}_0\boldsymbol{I}),$$
$$\tilde{f}_i(\boldsymbol{\mu}) = \tilde{s}_i N(\boldsymbol{\mu}|\tilde{\boldsymbol{m}}_i, \tilde{v}_i\boldsymbol{I}),$$

with parameters $\{\tilde{\boldsymbol{m}}_i\}_{i=0}^N$, $\{\tilde{s}_i\}_{i=0}^N$, and $\{\tilde{v}_i\}_{i=0}^N$.

Note that the $\tilde{f}_i$ are not densities and negative values for $\tilde{v}_i$ are valid.

Initialisation f_0 can be approximated exactly and the optimal choice for $\tilde{f}_0$ is $\tilde{f}_0 = f_0$. Once initialised, this term needs not be updated by EP anymore.

The $\tilde{f}_i$ are initialised to be non-informative, q is also non-informative:

$$\begin{aligned}
\tilde{s}_i &= (2\pi\tilde{v}_i)^{\frac{D}{2}} \\
\boldsymbol{m} &= \boldsymbol{0}\,, & v &= b, \\
\tilde{\boldsymbol{m}}_i &= \boldsymbol{0}\,, & \tilde{v}_i &\to \infty & &\text{for } i = 1, \ldots, N.,
\end{aligned}$$

where we have used the Gaussian identities. After refining $\tilde{f}_0$, q is equal to the prior $p(\boldsymbol{\mu})$.

Computation of the cavity distribution To compute $q^{\setminus i}$ one can derive that:

$$\begin{aligned}
q^{\setminus i}(\boldsymbol{\mu}) &\propto \frac{q(\boldsymbol{\mu})}{\tilde{f}_i(\boldsymbol{\mu})} \\
&\propto N(\boldsymbol{\mu}|\boldsymbol{m}^{\setminus i}, \boldsymbol{I}v^{\setminus i}),
\end{aligned}$$

where we use the Gaussian identities again to get

$$\begin{aligned}
v^{\setminus i} &= (v^{-1} - \tilde{v}_i^{-1})^{-1}, \\
\boldsymbol{m}^{\setminus i} &= v^{\setminus i}(v^{-1}\boldsymbol{m} - \tilde{v}_i^{-1}\tilde{\boldsymbol{m}}_i).
\end{aligned}$$

We used the fact that both $q(\boldsymbol{\mu})$ and $\tilde{f}_i(\boldsymbol{\mu})$ take the form of Normal distributions.

Computation of the new posterior The first step is to compute Z_i:

$$\begin{aligned}
Z_i &= \int f_i(\boldsymbol{\mu}) q^{\setminus i}(\boldsymbol{\mu}) d\boldsymbol{\mu} \\
&= \int \left[(1-w)N(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{I}) + wN(\boldsymbol{x}_i|\boldsymbol{0}, \boldsymbol{I}a)\right] N(\boldsymbol{\mu}|\boldsymbol{m}^{\setminus i}, \boldsymbol{I}v^{\setminus i}) d\boldsymbol{\mu} \\
&= \int (1-w)N(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{I})N(\boldsymbol{\mu}|\boldsymbol{m}^{\setminus i}, \boldsymbol{I}v^{\setminus i}) d\boldsymbol{\mu} + \int wN(\boldsymbol{x}_i|\boldsymbol{0}, \boldsymbol{I}a)N(\boldsymbol{\mu}|\boldsymbol{m}^{\setminus i}, \boldsymbol{I}v^{\setminus i}) d\boldsymbol{\mu} \\
&= (1-w)\int N(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{I})N(\boldsymbol{\mu}|\boldsymbol{m}^{\setminus i}, \boldsymbol{I}v^{\setminus i}) d\boldsymbol{\mu} + wN(\boldsymbol{x}_i|\boldsymbol{0}, \boldsymbol{I}a)\int N(\boldsymbol{\mu}|\boldsymbol{m}^{\setminus i}, \boldsymbol{I}v^{\setminus i}) d\boldsymbol{\mu} \\
&= (1-w)N(\boldsymbol{x}_i|\boldsymbol{m}^{\setminus i}, (v^{\setminus i}+1)\boldsymbol{I}) + wN(\boldsymbol{x}_i|\boldsymbol{0}, a\boldsymbol{I}).
\end{aligned}$$

which is obtained from the convolution of two Gaussians.

Next, we compute q_{new} by finding the mean and the variance of $\hat{p}(\mu) \propto f_i(\mu)q^{\setminus i}(\mu)$:

$$\begin{aligned}
\boldsymbol{m}_{new} &= \boldsymbol{m}^{\setminus i} + \rho_i \frac{v^{\setminus i}}{v^{\setminus i}+1}(\boldsymbol{x}_i - \boldsymbol{m})\,, \\
v_{new} &= v^{\setminus i} - \rho_i \frac{(v^{\setminus i})^2}{v^{\setminus i}+1} + \rho_i(1-\rho_i)\frac{(v^{\setminus i})^2||\boldsymbol{x}_i - \boldsymbol{m}^{\setminus i}||^2}{D(v^{\setminus i}+1)^2}\,,
\end{aligned}$$

where we have used the Gaussian moments and identities and

$$\rho_i = 1 - \frac{w}{Z_i}N(\boldsymbol{x}_i|\boldsymbol{0}, a\boldsymbol{I})$$

can be interpreted as the probability of $\boldsymbol{x}_i$ not being clutter.

Note that derivation of the above solution is doable; however, very tedious. Therefore, see `https://github.com/bayesian-inference/notes/blob/master/expectation_propagation__the_clutter_problem/clutter.pdf` for details.

Update of the approximate factor $\tilde{f}_i$ is updated to be equal to $Z_i q_{new}/q^{\setminus i}$:

$$\begin{aligned}
\tilde{v}_i &= \left((v_{new})^{-1} - (v^{\setminus i})^{-1}\right)^{-1}, \\
\tilde{\boldsymbol{m}}_i &= \tilde{v}_i \left(v_{new}^{-1}\boldsymbol{m}_{new} - (v^{\setminus i})^{-1}\boldsymbol{m}^{\setminus i}\right), \\
\tilde{s}_i &= \frac{Z_i}{N(\tilde{\boldsymbol{m}}_i|\boldsymbol{m}^{\setminus i}, (\tilde{v}_i + v^{\setminus i})\boldsymbol{I})},
\end{aligned}$$

where we used the Gaussian identities.

At convergence we evaluate the approximation of the marginal likelihood:

$$p(D) \approx \int \prod_{i=0}^{N} \tilde{f}_i(\boldsymbol{\mu}) d\boldsymbol{\mu} = (2\pi v_{new})^{D/2} \exp(B/2) \prod_{i=0}^{N} \left[\tilde{s}_i (2\pi\tilde{v}_i)^{-D/2}\right],$$

where $B = \boldsymbol{m}_{new}^{\mathrm{T}} v_{new}^{-1} \boldsymbol{m}_{new} - \sum_{i=0}^{N} \tilde{\boldsymbol{m}}^{\mathrm{T}} (\tilde{v}_i)^{-1} \tilde{\boldsymbol{m}}$ and we have used the Gaussian identities.

6.7 Example: EP - The probit regression model

Suppose we have independent data points $\boldsymbol{x}_{i=1}^{n}$, each consisting of d features, thus building a matrix $X \in R^{n\times d}$. Each data point $\boldsymbol{x}_i$ has a label $y_i \in \{-1, 1\}, i = 1 \dots n$, which gives a vector of labels $\boldsymbol{v}$. Then, the data D is a set $\{(x_0, y_0), (x_1, y_1), \dots\}$.

We want to model this data using a *probit model*:

$$p(y_i|\boldsymbol{x}_i, \boldsymbol{w}) = \Phi(y_i \boldsymbol{w}^T \boldsymbol{x}_i),$$

where Φ denotes a standard Gaussian cumulative distribution function:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt.$$

Prior We choose the prior for $\boldsymbol{w}$ to be:

$$\begin{aligned}
p(\boldsymbol{w}) &= N(\boldsymbol{w}|\boldsymbol{0}, \boldsymbol{I}v_0) \\
&= \prod_{j=1}^{d} N(w_j|0, v_0)
\end{aligned}$$

Likelihood The likelihood is defined as follows:

$$\begin{aligned}
p(D|\boldsymbol{w}) &= p(\boldsymbol{y}|X, \boldsymbol{w}), \\
&= \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i, \boldsymbol{w}) = \prod_{i=1}^{n} \Phi(y_i \boldsymbol{w}^T \boldsymbol{x}_i).
\end{aligned}$$

Joint distribution The joint distribution is defined as follows:

$$\begin{aligned}
p(\boldsymbol{w}, D) &= p(\boldsymbol{y}, X, \boldsymbol{w}), \\
&= p(\boldsymbol{w}) \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i, \boldsymbol{w}), \\
&= N(\boldsymbol{w}|\boldsymbol{0}, \boldsymbol{I}v_0) \prod_{i=1}^{n} \Phi(y_i \boldsymbol{w}^T \boldsymbol{x}_i), \\
&= f_0(\boldsymbol{w}) \prod_{i=1}^{n} f_i(\boldsymbol{w}),
\end{aligned}$$

where

$$\begin{aligned}
f_0(\boldsymbol{w}) &= N(\boldsymbol{w}|\boldsymbol{0}, \boldsymbol{I}v_0), \\
f_i(\boldsymbol{w}) &= \Phi(y_i \boldsymbol{w}^T \boldsymbol{x}_i).
\end{aligned}$$

Posterior The posterior distribution is defined as follows:

$$\begin{aligned} p(\boldsymbol{w}|D) &= \frac{1}{p(D)} p(\boldsymbol{w}, D) \\ &= \frac{1}{Z} p(w) \prod_i p(y_i|\boldsymbol{x}_i, \boldsymbol{w}), \\ &= \frac{1}{p(D)} f_0(\boldsymbol{w}) \prod_i f_i(\boldsymbol{w}). \end{aligned}$$

However, this posterior is in-tractable. Therefore, we choose the posterior to have form of the prior:

$$q(\boldsymbol{w}) = N(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{I}\boldsymbol{v}) = \prod_j N(w_j|m_j, v_j).$$

Note that $\boldsymbol{m} = \{m_j\}_{j=1}^d$ and $\boldsymbol{v} = \{v_j\}_{j=1}^d$ denote means and variances in the individual dimensions.
We also approximate $q(\boldsymbol{w})$ as a product of simple factors:

$$q(\boldsymbol{w}) = \frac{1}{Z} \tilde{f}_0(\boldsymbol{w}) \prod_{i=1}^N \tilde{f}_i(\boldsymbol{w}),$$

where

$$\begin{aligned} \tilde{f}_0(\boldsymbol{w}) &= f_0(\boldsymbol{w}) = N(\boldsymbol{w}|\boldsymbol{0}, \boldsymbol{I}v_0), \\ \tilde{f}_i(\boldsymbol{w}) &= \boldsymbol{s}_i N(\boldsymbol{w}|\boldsymbol{m}_i, \boldsymbol{I}\boldsymbol{v}_i) = \prod_{j=1}^d s_{ij} N(w_j|m_{ij}, v_{ij}) \quad i = 1, \ldots, N. \end{aligned}$$

Note that $\boldsymbol{m}_i = \{m_{ij}\}_{j=1}^d$ for $i = 1, \ldots, N$ and $\boldsymbol{v}_i = \{v_{ij}\}_{j=1}^d$ for $i = 1, \ldots, N$ denote means and variances in the individual dimensions of the approximated factors.

Also, note that the s_{ij} is a de-normalisation constant to make sure that our approximate $\tilde{f}_i$ fits well the original f_i. This is useful as the original factor f_i is normalised with respect to (y_i, x_i), while the approximation is $\tilde{f}_i$ with respect to $\boldsymbol{w}$.

Initialisation We initialise our approximation $q(\boldsymbol{w})$ by setting $\tilde{f}_0(\boldsymbol{w})$ to the prior and $\tilde{f}_i(\boldsymbol{w})$ to uniform distributions.[1] The parameters of the approximate factors then may look as follows:

$$\begin{aligned} m_{0j} &= 0, & v_{0j} &= v_0 & j &= 1, \ldots, d \\ m_{ij} &= 0, & v_{ij} &= \infty & j &= 1, \ldots, d, \quad i = 1, \ldots, N \end{aligned}$$

Now our posterior approximation is in fact equal to our prior (if we view it as prior $\cdot \prod_{i=1}^n$ uniform).

Computation of the cavity distribution Computing the cavity distribution $q^{\setminus i}(\boldsymbol{w})$ we get:

$$q^{\setminus i}(\boldsymbol{w}) = \frac{q(\boldsymbol{w})}{\tilde{f}_i(\boldsymbol{w})}.$$

Since both $q(\boldsymbol{w})$ and $\tilde{f}_i(\boldsymbol{w})$ take form of Normal distributions, we can use the formulas for Normal identities to obtain un-normalised Normal shaped $q^{\setminus i}(\boldsymbol{w})$:

$$q^{\setminus i}(\boldsymbol{w}) \propto N(\boldsymbol{w}|\boldsymbol{m}^{\setminus i}, \boldsymbol{v}^{\setminus i})$$

where:

$$\begin{aligned} v_j^{\setminus i} &= (v_j^{-1} - v_{ij}^{-1})^{-1}, \\ m_j^{\setminus i} &= v_j^{\setminus i}(v_j^{-1} m_j - v_{ij}^{-1} m_{ij}). \end{aligned}$$

Note that m_{ij}, v_{ij} refer to the current approximation of $\tilde{f}_i(\boldsymbol{w})$ and m_j, v_j refer to the current approximation of $q(\boldsymbol{w})$.

[1] Or as close to uniform distributions as we can get in practice since $\tilde{f}_i(\boldsymbol{w})$ are assumed to be Gaussian.

Computation of the new posterior The first step is to compute Z_i:

$$\begin{aligned} Z_i &= \int f_i(\boldsymbol{w}) q^{\backslash i}(\boldsymbol{w}) d\boldsymbol{w} \\ &= \int p(y_i|\boldsymbol{x}_i, \boldsymbol{w}) q^{\backslash i}(\boldsymbol{w}) d\boldsymbol{w} \\ &= \int \Phi(y_i \boldsymbol{w}^T \boldsymbol{x}_i) \prod_{i=1}^{d} N(w_j | m_j^{\backslash i}, v_j^{\backslash i}) d\boldsymbol{w} \\ &= \Phi\left(\frac{y_i \sum_{j=1}^{d} m_j^{\backslash i} x_{ij}}{\sqrt{\sum_{j=1}^{d} v_j^{\backslash i} x_{ij}^2 + 1}} \right) \end{aligned}$$

Note that the derivation above is not trivial.

Now using the Gaussian moments, we can compute the $q_{new}(\boldsymbol{w})$. We obtain the mean and the variance of the new approximate posterior $q_{new}(\boldsymbol{w})$:

$$\begin{aligned} m_j^{new} &= m_j^{\backslash i} + v_j^{\backslash i} \frac{\partial \log Z_i}{\partial m_j^{\backslash i}} \\ v_j^{new} &= v_j^{\backslash i} - \left(v_j^{\backslash i}\right)^2 \left(\left(\frac{\partial \log Z_i}{\partial m_j^{\backslash i}} \right)^2 - 2 \frac{\partial \log Z_i}{\partial v_j^{\backslash i}} \right) \end{aligned}$$

TBD: Compute the derivatives of $\log Z_i$.

Update of the approximate factor We now have the new approximate posterior $q^{new}(\boldsymbol{w})$ and need to obtain our new approximate factor $\tilde{f}_i^{new}(\boldsymbol{w})$ for later use.

$$\begin{aligned} q_{new}(\boldsymbol{w}) &= \frac{1}{Z_i} \tilde{f}_i(\boldsymbol{w}) q^{\backslash i}(\boldsymbol{w}) \\ \tilde{f}_i(\boldsymbol{w}) &= Z_i \frac{q^{new}(\boldsymbol{w})}{q^{\backslash i}(\boldsymbol{w})} \end{aligned}$$

The parameters $m_{ij}^{new}, v_{ij}^{new}, s_{ij}^{new}$ are obtained from the Gaussian identities formulas:

$$\begin{aligned} v_{ij}^{new} &= \left(v_j^{-1} - \left(v_j^{\backslash i}\right)^{-1} \right)^{-1}, \\ m_{ij}^{new} &= v_{ij}^{new} \cdot \left(m_j v_j^{-1} - m_j^{\backslash i} \left(v_j^{\backslash i}\right)^{-1} \right), \\ s_{ij}^{new} &= Z_i \cdot C_j, \end{aligned}$$

where

$$C_j = \sqrt{\frac{v_{ij}^{new} v_j^{\backslash i}}{(2\pi)^d v_j}} \exp\left(-\frac{1}{2} \left(m_j^2 v_j^{-1} - \left(m_j^{\backslash i}\right)^2 \left(v_j^{\backslash i}\right)^{-1} - \left(m_{ij}^{new}\right)^2 \left(v_{ij}^{new}\right)^{-1} \right) \right)$$

We can now use $\tilde{f}_i^{new}(\boldsymbol{w})$ and $q^{new}(\boldsymbol{w})$ in the next iterations.

7 Markov Chain Monte Carlo

8 MCMC: Gibbs sampling

In this section, we will introduce a specific case of a MCMC technique called Gibbs sampling. Since direct sampling from the posterior is usually intractable, Gibbs sampling draws from a posterior for each hidden variable given the rest of the observed and hidden variables.

8.1 Example: Inference of the mean and variance using a non-conjugate prior

Here, we will try to solve the the task described in Section 2.3.4. Figure 8 depict a graphical model representing inference of the mean and precision of an univariate normal distribution using non-conjugate prior. Instead of direct analytic computation of the posterior, we will use Gibbs sampling from the posterior distributions and use the samples to represent the posterior.

In our case, we want to draw samples from the posterior distribution $p(\mu, \lambda|\boldsymbol{x}, \mu_0, \lambda_0, a_0, b_0)$. Therefore, we will iteratively sample from the posterior $p(\mu|\lambda, \boldsymbol{x}, \mu_0, \lambda_0, a_0, b_0)$ and $p(\lambda|\mu, \boldsymbol{x}, \mu_0, \lambda_0, a_0, b_0)$, where the initial values for μ and λ are set manually. After some burn-in period, e.g. M samples, the samples will be distributed according to the posterior $p(\mu, \lambda|\dots)$. These generated samples can be then used on its own, e.g. for visualisation or the estimation of the parameters $\tilde{\mu}_N, \tilde{\lambda}_N, \tilde{a}_N, \tilde{b}_N$ of the approximate posterior defined in the form of the prior (24).

Note that given the conditional independence defined by the graphical model, the posteriors have the following form:

$$p(\mu|\lambda, \boldsymbol{x}, \mu_0, \lambda_0, a_0, b_0) = p(\mu|\boldsymbol{x}, \lambda, \mu_0, \lambda_0)$$
$$p(\lambda|\mu, \boldsymbol{x}, \mu_0, \lambda_0, a_0, b_0) = p(\lambda|\boldsymbol{x}, \mu, a_0, b_0)$$

Recall that we derived the posterior for the mean in Section 2.3.1 and that the posterior for the precision was derived in Section 2.3.2. Therefore, we already know that:

$$p(\mu|\boldsymbol{x}, \lambda, \mu_0, \lambda_0) = N(\mu|\mu_N, \lambda_N^{-1})$$
$$\mu_N = \frac{N\lambda\mu_{ML} + \lambda_0\mu_0}{N\lambda + \lambda_0}$$
$$\lambda_N = N\lambda + \lambda_0$$

where $\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\sigma^2 = 1/\lambda$.

Now we have to iteratively sample $p(\mu|\boldsymbol{x}, \lambda, \mu_0, \lambda_0)$ and $p(\lambda|\boldsymbol{x}, \mu, a_0, b_0)$ to obtain the samples from the posterior $p(\mu, \lambda|\dots)$. Note that although the posteriors for μ and λ are independent, they exhibit a coupling since the posterior $p(\mu|\dots)$ depends on the precision λ and the other way around.

Assume that we have obtained N samples of the mean, e.g. $\mu_1, \dots, \mu_N$, and precision, e.g. $\lambda_1, \dots, \lambda_N$, using Gibbs sampling described above. Then, the parameters $\tilde{\mu}_N, \tilde{\lambda}_N, \tilde{a}_N, \tilde{b}_N$ of the approximate posterior

$$p(\mu, \lambda|\tilde{\mu}_N, \tilde{\lambda}_N, \tilde{a}_N, \tilde{b}_N) = N(\mu|\tilde{\mu}_N, \tilde{\lambda}_N^{-1})Gam(\lambda|\tilde{a}_N, \tilde{b}_N)$$

can be computed as follows:

$$\tilde{\mu}_N = \frac{1}{N}\sum_{i=i}^{N}\mu_i$$
$$\tilde{\lambda}_N^{-1} = \frac{1}{N}\sum_{i=i}^{N}(\mu_i - \tilde{\mu}_N)^2$$
$$\tilde{a}_N = \text{no closed form solution}$$
$$\tilde{b}_N = \frac{a_N}{\frac{1}{N}\sum_{i=i}^{N}\mu_i}$$

In the case of $\tilde{a}_N$, a numerical maximisation of likelihood has to be performed since there is no closed form solution.

8.2 Example: Hierarchical Bayesian model for the real observations

This section does not use correct notation for σ^2 for normal distribution.

In this section, we will model the observations using hierarchical Bayesian model. The observations are still real; however, we observe additional information about the observations. The situation can by described by the model depicted on Figure 13. In this case, we model real valued observations x which depend on s and u. This can be a model of a fundamental frequency (an inverse of a pitch period) of speech of the user u when communicating with the system s. It is known that the fundamental frequency is defined by the physical

properties of the user's vocal tract. However, users tends to adapt the frequency based on the partner they communicate with.

The model is equivalent to predicting x given s and u using the probability distribution $p(x|s,u)$. We can assume that the observations x are generated from a normal distribution where the mean of the distribution depends on both the system and user. If we had enough data, then we could estimate a specific mean for each combination of the system and the user. We would need $S \times U$ parameters, where S represents the number of systems and U represents the number of users, to specify the distribution $p(x|s,u) = N(x|\mu_{s,u}, \sigma)$. However, we aim to develop a more compact probabilistic model.

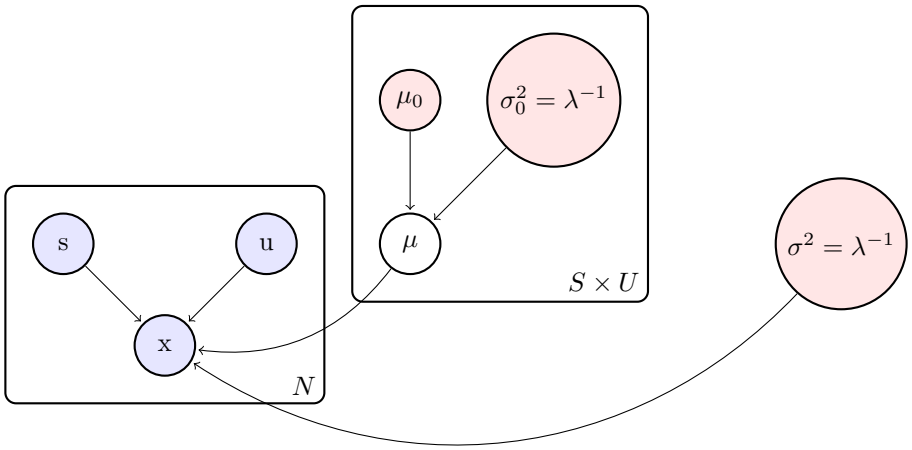


Figure 13: The probabilistic model for the prediction of the observation x given s and u.

Instead, we will try to make use of the knowledge that there is similarity between the observations for the same systems as well as that there is similarity between the observations for the same users. More precisely, we will assume that the probability distribution of the observations can be described by the distribution $N(x|\mu_s + \eta_u, \sigma)$. In this case, we will need only $S + U$ parameters. In addition, we will add unknown priors for μ_s and η_s which will be inferred from the data. These priors will enable sharing information about the means among the systems and the means among the users, e.g. the prior for one user will be affected by observations from other users. Such model is depicted on Figure 14.

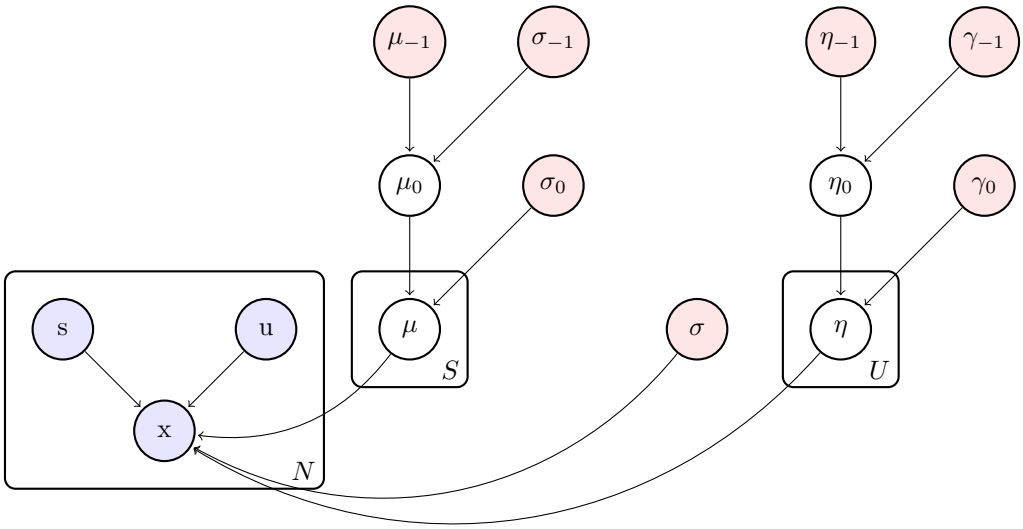


Figure 14: Graphical model factoring the system and user parameters represented by a hierarchical Bayesian model. The model is modelling only the mean and the variance is assumed to be known.

The model depicted on Figure 14 assumes the following generative process:

1. $\mu_0 \sim N(\cdot|\mu_{-1}, \sigma_{-1}^2)$
2. $\eta_0 \sim N(\cdot|\eta_{-1}, \gamma_{-1})$

3. $\mu_s \sim N(\cdot|\mu_0, \sigma_0^2)$
4. $\eta_u \sim N(\cdot|\eta_0, \gamma_0^2)$
5. $x \sim N(\cdot|\mu_s + \eta_u, \sigma^2)$

where the parameters μ_{-1}, σ_{-1}, σ_0, σ, η_{-1}, γ_{-1}, and γ_0 are priors set manually and s, u, and x are the observations.

Given the parameters μ_{-1}, σ_{-1}, σ_0, σ, η_{-1}, γ_{-1}, γ_0 and the observations $\boldsymbol{s}$ and $\boldsymbol{u}$, the joint distribution of the observations $\boldsymbol{x}$, the system mean values $\boldsymbol{\mu}$, the prior of the system mean values μ_0, the user mean values $\boldsymbol{\eta}$, the prior of the user mean values η_0 is given by:

$$\begin{aligned} &p(\boldsymbol{x}, \boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0 | \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0) = \\ &= p(\boldsymbol{\mu}|\mu_0, \sigma_0) p(\mu_0|\mu_{-1}, \sigma_{-1}) p(\boldsymbol{\eta}|\eta_0, \gamma_0) p(\eta_0|\eta_{-1}, \gamma_{-1}) \prod_{i=1}^{N} p(x_i|s_i, \boldsymbol{\mu}, u_i, \boldsymbol{\eta}, \sigma) \end{aligned} \tag{37}$$

Note that the system s (more precisely s_i) and the user u (u_i) are represented by a unit-basis vectors that have a single component equal to one and all other components equal to zero. For example, the jth system in ith sample is represented by S-vector $\boldsymbol{s}$ such that $s_{ij} = 1$ and $s_{ik} = 0$ for $k \neq j$ for all i.

This can be further factored according to the components of the vectors $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$:

$$\begin{aligned} &p(\boldsymbol{x}, \boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0 | \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0) = \\ &= p(\mu_0|\mu_{-1}, \sigma_{-1}) \prod_{j=1}^{S} p(\mu_j|\mu_0, \sigma_0) \cdot p(\eta_0|\eta_{-1}, \gamma_{-1}) \prod_{k=1}^{U} p(\eta_k|\eta_0, \gamma_0) \cdot \prod_{i=1}^{N} \prod_{j=1}^{S} \prod_{k=1}^{U} p(x_i|\mu_j, \eta_k, \sigma)^{s_{ij} u_{ik}} \end{aligned} \tag{38}$$

Note that now the system and user are represented by indexes j and k respectively. Given the generative model and the assumptions on the normal distribution of the observations, the probability distributions are represented as follows:

$$p(\mu_0|\mu_{-1}, \sigma_{-1}) = N(\mu_0|\mu_{-1}, \sigma_{-1}) \tag{39}$$
$$p(\mu_j|\mu_0, \sigma_0) = N(\mu_j|\mu_0, \sigma_0) \tag{40}$$
$$p(\eta_0|\eta_{-1}, \gamma_{-1}) = N(\eta_0|\eta_{-1}, \gamma_{-1}) \tag{41}$$
$$p(\eta_k|\eta_0, \gamma_0) = N(\eta_k|\eta_0, \gamma_0) \tag{42}$$
$$p(x_i|\mu_j, \eta_k, \sigma) = N(x_i|\mu_j + \eta_k, \sigma) \tag{43}$$

Now, we will describe inference using Gibbs sampling in the model described above. In summary, the samples from the joint posterior for all hidden variables $p(\boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0 | \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0)$ can be obtained by iterative sampling from posteriors for individual hidden variables. This turns out to be very often simpler than sampling from the full joint distribution.

To apply Gibbs sampling method, posterior distributions for each hidden variable $\boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0$ must be derived. More precisely, we must derive the following posteriors:

$$p(\mu_0|\boldsymbol{\mu}, \mu_{-1}, \sigma_{-1}, \sigma_0) \tag{44}$$
$$p(\mu_j|\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\mu}_{-j}, \mu_0, \boldsymbol{u}, \boldsymbol{\eta}, \sigma, \sigma_0) \quad \forall j \in \{1, \ldots, S\} \tag{45}$$
$$p(\eta_0|\boldsymbol{\eta}, \eta_{-1}, \gamma_{-1}, \gamma_0) \tag{46}$$
$$p(\eta_k|\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\mu}, \boldsymbol{u}, \boldsymbol{\eta}_{-k}, \eta_0, \sigma, \gamma_0) \quad \forall k \in \{1, \ldots, U\} \tag{47}$$

where $\boldsymbol{\mu}_{-i}$ is the vector $\boldsymbol{\mu}$ without μ_i and $\boldsymbol{\eta}_{-i}$ is the vector $\boldsymbol{\eta}$ without η_i.

Note that in situation where all latent variables are know except for the latent variable for which we want to compute the posterior, the posterior depend only on the parents, children and parents of the children (aka Markov blanket).

8.2.1 Posterior of the hyper-parameters

The easiest way to start is to compute posterior of μ_0. Using the joint distribution (37) and the Bayes rule:

$$\begin{aligned} &p(\boldsymbol{x}, \boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0 | \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0) = \\ &= p(\mu_0|\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\eta}, \eta_0, \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0) p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\eta}, \eta_0 | \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0) \end{aligned}$$

Therefore the posterior is computed as:

$$\begin{aligned} &p(\mu_0|\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\eta}, \eta_0, \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0) = \\ &= \frac{p(\boldsymbol{x}, \boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0 | \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0)}{p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\eta}, \eta_0 | \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0)} \\ &= \frac{p(\boldsymbol{x}, \boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0 | \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0)}{\int p(\boldsymbol{x}, \boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0 | \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0) d\mu_0} \end{aligned} \tag{48}$$

When (37) is substituted into (48), then it results in:

$$\begin{aligned}
&p(\mu_0|\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\eta},\eta_0,\boldsymbol{s},\boldsymbol{u},\mu_{-1},\sigma_{-1},\sigma_0,\sigma,\eta_{-1},\gamma_{-1},\gamma_0) = \\
&= \frac{p(\boldsymbol{x}|s,\boldsymbol{\mu},u,\boldsymbol{\eta},\sigma)p(\boldsymbol{\mu}|\mu_0,\sigma_0)p(\mu_0|\mu_{-1},\sigma_{-1})p(\boldsymbol{\eta}|\eta_0,\gamma_0)p(\eta_0|\eta_{-1},\gamma_{-1})}{\int p(\boldsymbol{x}|s,\boldsymbol{\mu},u,\boldsymbol{\eta},\sigma)p(\boldsymbol{\mu}|\mu_0,\sigma_0)p(\mu_0|\mu_{-1},\sigma_{-1})p(\boldsymbol{\eta}|\eta_0,\gamma_0)p(\eta_0|\eta_{-1},\gamma_{-1})d\mu_0} \\
&= \frac{p(\boldsymbol{\mu}|\mu_0,\sigma_0)p(\mu_0|\mu_{-1},\sigma_{-1})}{\int p(\boldsymbol{\mu}|\mu_0,\sigma_0)p(\mu_0|\mu_{-1},\sigma_{-1})d\mu_0} \\
&= p(\mu_0|\boldsymbol{\mu},\mu_{-1},\sigma_{-1},\sigma_0)
\end{aligned} \tag{49}$$

One can see that the result is exactly what we need for Gibbs sampling as described in (44). Now using (38) and substituting (40) and (39) into (49), results in:

$$\begin{aligned}
p(\mu_0|\boldsymbol{\mu},\mu_{-1},\sigma_{-1},\sigma_0) &\propto p(\boldsymbol{\mu}|\mu_0,\sigma_0)p(\mu_0|\mu_{-1},\sigma_{-1}) \\
&\propto N(\mu_0|\mu_{-1},\sigma_{-1})\prod_{j=1}^{S} N(\mu_j|\mu_0,\sigma_0)
\end{aligned} \tag{50}$$

Recall that we already derived the posterior for the mean in Section 2.3.1. Therefore, we already know that:

$$\begin{aligned}
p(\mu_0|\boldsymbol{\mu},\mu_{-1},\sigma_{-1},\sigma_0) &= N(\mu_0|\mu_S,\sigma_S^2) \\
\mu_S &= \frac{S\sigma_{-1}^2}{S\sigma_{-1}^2+\sigma_0^2}\bar{\mu} + \frac{\sigma_0^2}{\sigma_0^2+S\sigma_{-1}^2}\mu_{-1} \\
\frac{1}{\sigma_S^2} &= \frac{S}{\sigma_0^2} + \frac{1}{\sigma_{-1}^2}
\end{aligned}$$

where $\bar{\mu} = \frac{1}{S}\sum_{j=1}^{S}\mu_j$ and S is the number of the modelled systems. Similar results can be obtained for η_0:

$$\begin{aligned}
p(\eta_0|\boldsymbol{\eta},\eta_{-1},\gamma_{-1},\gamma_0) &= N(\eta_0|\eta_U,\gamma_U^2) \\
\eta_U &= \frac{S\gamma_{-1}^2}{S\gamma_{-1}^2+\gamma_0^2}\bar{\eta} + \frac{\gamma_0^2}{\gamma_0^2+S\gamma_{-1}^2}\eta_{-1} \\
\frac{1}{\gamma_U^2} &= \frac{S}{\gamma_0^2} + \frac{1}{\gamma_{-1}^2}
\end{aligned}$$

where $\bar{\eta} = \frac{1}{U}\sum_{k=1}^{U}\eta_k$ and U is the number of the modelled users.

8.2.2 Posterior of the parameters

Now, we will compute the posterior for $\boldsymbol{\mu}$. Using the joint distribution (37) and the Bayes rule:

$$\begin{aligned}
&p(\boldsymbol{x},\boldsymbol{\mu},\mu_0,\boldsymbol{\eta},\eta_0|\boldsymbol{s},\boldsymbol{u},\mu_{-1},\sigma_{-1},\sigma_0,\sigma,\eta_{-1},\gamma_{-1},\gamma_0) = \\
&= p(\mu_j|\boldsymbol{x},\boldsymbol{\mu}_{-j},\mu_0,\boldsymbol{\eta},\eta_0,\boldsymbol{s},\boldsymbol{u},\mu_{-1},\sigma_{-1},\sigma_0,\sigma,\eta_{-1},\gamma_{-1},\gamma_0)p(\boldsymbol{x},\boldsymbol{\mu}_{-j},\boldsymbol{\eta},\eta_0|\boldsymbol{s},\boldsymbol{u},\mu_{-1},\sigma_{-1},\sigma_0,\sigma,\eta_{-1},\gamma_{-1},\gamma_0)
\end{aligned}$$

Therefore the posterior is computed as:

$$\begin{aligned}
&p(\mu_j|\boldsymbol{x},\boldsymbol{\mu}_{-j},\mu_0,\boldsymbol{\eta},\eta_0,\boldsymbol{s},\boldsymbol{u},\mu_{-1},\sigma_{-1},\sigma_0,\sigma,\eta_{-1},\gamma_{-1},\gamma_0) = \\
&= \frac{p(\boldsymbol{x},\boldsymbol{\mu},\mu_0,\boldsymbol{\eta},\eta_0|\boldsymbol{s},\boldsymbol{u},\mu_{-1},\sigma_{-1},\sigma_0,\sigma,\eta_{-1},\gamma_{-1},\gamma_0)}{p(\boldsymbol{x},\boldsymbol{\mu}_{-j},\boldsymbol{\eta},\eta_0|\boldsymbol{s},\boldsymbol{u},\mu_{-1},\sigma_{-1},\sigma_0,\sigma,\eta_{-1},\gamma_{-1},\gamma_0)} \\
&= \frac{p(\boldsymbol{x},\boldsymbol{\mu},\mu_0,\boldsymbol{\eta},\eta_0|\boldsymbol{s},\boldsymbol{u},\mu_{-1},\sigma_{-1},\sigma_0,\sigma,\eta_{-1},\gamma_{-1},\gamma_0)}{\int p(\boldsymbol{x},\boldsymbol{\mu}_{-j},\mu_j,\mu_0,\boldsymbol{\eta},\eta_0|\boldsymbol{s},\boldsymbol{u},\mu_{-1},\sigma_{-1},\sigma_0,\sigma,\eta_{-1},\gamma_{-1},\gamma_0)d\mu_j}
\end{aligned} \tag{51}$$

When (38) is substituted into (51), then it results in:

$$\begin{aligned}
&p(\mu_j|\boldsymbol{x},\boldsymbol{\mu}_{-j},\mu_0,\boldsymbol{\eta},\eta_0,\boldsymbol{s},\boldsymbol{u},\mu_{-1},\sigma_{-1},\sigma_0,\sigma,\eta_{-1},\gamma_{-1},\gamma_0) = \\
&= \frac{p(\mu_0|\mu_{-1},\sigma_{-1})\prod_{l=1}^{S}p(\mu_l|\mu_0,\sigma_0)\cdot p(\eta_0|\eta_{-1},\gamma_{-1})\prod_{k=1}^{U}p(\eta_k|\eta_0,\gamma_0)\cdot\prod_{i=1}^{N}\prod_{l=1}^{S}\prod_{k=1}^{U}p(x_i|\mu_l,\eta_k,\sigma)^{s_{il}u_{ik}}}{\int p(\mu_0|\mu_{-1},\sigma_{-1})\prod_{l=1}^{S}p(\mu_l|\mu_0,\sigma_0)\cdot p(\eta_0|\eta_{-1},\gamma_{-1})\prod_{k=1}^{U}p(\eta_k|\eta_0,\gamma_0)\cdot\prod_{i=1}^{N}\prod_{l=1}^{S}\prod_{k=1}^{U}p(x_i|\mu_l,\eta_k,\sigma)^{s_{il}u_{ik}}d\mu_j} \\
&= \frac{\prod_{l=1}^{S}p(\mu_l|\mu_0,\sigma_0)\cdot\prod_{i=1}^{N}\prod_{l=1}^{S}\prod_{k=1}^{U}p(x_i|\mu_l,\eta_k,\sigma)^{s_{il}u_{ik}}}{\int\prod_{l=1}^{S}p(\mu_l|\mu_0,\sigma_0)\cdot\prod_{i=1}^{N}\prod_{l=1}^{S}\prod_{k=1}^{U}p(x_i|\mu_l,\eta_k,\sigma)^{s_{il}u_{ik}}d\mu_j} \\
&= \frac{p(\mu_j|\mu_0,\sigma_0)\prod_{l=1,l\neq j}^{S}p(\mu_l|\mu_0,\sigma_0)\cdot\prod_{i=1}^{N}\prod_{k=1}^{U}p(x_i|\mu_j,\eta_k,\sigma)^{s_{ij}u_{ik}}\prod_{l=1,l\neq j}^{S}p(x_i|\mu_l,\eta_k,\sigma)^{s_{il}u_{ik}}}{\int p(\mu_j|\mu_0,\sigma_0)\prod_{l=1,l\neq j}^{S}p(\mu_l|\mu_0,\sigma_0)\cdot\prod_{i=1}^{N}\prod_{k=1}^{U}p(x_i|\mu_j,\eta_k,\sigma)^{s_{ij}u_{ik}}\prod_{l=1,l\neq j}^{S}p(x_i|\mu_l,\eta_k,\sigma)^{s_{il}u_{ik}}d\mu_j} \\
&= \frac{p(\mu_j|\mu_0,\sigma_0)\cdot\prod_{i=1}^{N}\prod_{k=1}^{U}p(x_i|\mu_j,\eta_k,\sigma)^{s_{ij}u_{ik}}}{\int p(\mu_j|\mu_0,\sigma_0)\cdot\prod_{i=1}^{N}\prod_{k=1}^{U}p(x_i|\mu_j,\eta_k,\sigma)^{s_{ij}u_{ik}}d\mu_j} \\
&= p(\mu_j|\boldsymbol{x},\boldsymbol{s},\boldsymbol{\mu}_{-j},\mu_0,\boldsymbol{u},\boldsymbol{\eta},\sigma,\sigma_0)
\end{aligned} \tag{52}$$

Substituting (40) and (43) into (52) results into:

$$
\begin{aligned}
&p(\mu_j|\boldsymbol{x},\boldsymbol{s},\boldsymbol{\mu}_{-j},\mu_0,\boldsymbol{u},\boldsymbol{\eta},\sigma,\sigma_0) \qquad =\\
&\propto p(\mu_j|\mu_0,\sigma_0)\cdot\prod_{i=1}^{N}\prod_{k=1}^{U}p(x_i|\mu_j,\eta_k,\sigma)^{s_{ij}u_{ik}}\\
&\propto N(\mu_j|\mu_0,\sigma_0)\cdot\prod_{i=1}^{N}\prod_{k=1}^{U}N(x_i|\mu_j+\eta_k,\sigma)^{s_{ij}u_{ik}}\\
&\propto \frac{1}{(2\pi\sigma_0^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma_0^2}(\mu_j-\mu_0)^2\right\}\cdot\prod_{i=1}^{N}\prod_{k=1}^{U}\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu_j-\eta_k)^2\right\}^{s_{ij}u_{ik}}\\
&\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu_j-\mu_0)^2\right\}\cdot\prod_{i=1}^{N}\prod_{k=1}^{U}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu_j-\eta_k)^2\right\}^{s_{ij}u_{ik}}\\
&\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu_j-\mu_0)^2\right\}\cdot\prod_{i=1}^{N}\prod_{k=1}^{U}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu_j-\eta_k)^2 s_{ij}u_{ik}\right\}\\
&\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu_j-\mu_0)^2\right\}\exp\left\{-\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{2\sigma^2}(x_i-\mu_j-\eta_k)^2 s_{ij}u_{ik}\right\}\\
&\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu_j-\mu_0)^2-\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{2\sigma^2}(x_i-\mu_j-\eta_k)^2 s_{ij}u_{ik}\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2}(\mu_j-\mu_0)^2+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}(x_i-\mu_j-\eta_k)^2 s_{ij}u_{ik}\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2}(\mu_j^2-2\mu_j\mu_0+\mu_0^2)+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}(x_i^2+\eta_k^2+\mu_j^2+2\mu_j\eta_k-2\mu_j x_i-2\eta_k x_i)s_{ij}u_{ik}\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2}\mu_j^2-2\frac{1}{\sigma_0^2}\mu_j\mu_0+\frac{1}{\sigma_0^2}\mu_0^2\right.\right.\\
&\qquad\left.\left.+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}\mu_j^2 s_{ij}u_{ik}+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}2\mu_j(\eta_k-x_i)s_{ij}u_{ik}+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}(x_i^2+\eta_k^2-2\eta_k x_i)s_{ij}u_{ik}\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2}\mu_j^2-2\frac{1}{\sigma_0^2}\mu_j\mu_0+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}\mu_j^2 s_{ij}u_{ik}+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}2\mu_j(\eta_k-x_i)s_{ij}u_{ik}\right)\right\}
\end{aligned}
$$

Note that $\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}(x_i^2+\eta_k^2-2\eta_k x_i)s_{ij}u_{ik}$ and $\frac{1}{\sigma_0^2}\mu_0^2$ are independent of μ_j and therefore a multiplying constant. Next, we just reorder the expression.

$$
\begin{aligned}
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2}\mu_j^2+\mu_j^2\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}s_{ij}u_{ik}-2\frac{1}{\sigma_0^2}\mu_j\mu_0+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}2\mu_j(\eta_k-x_i)s_{ij}u_{ik}\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(\mu_j^2\left[\frac{1}{\sigma_0^2}+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}s_{ij}u_{ik}\right]-2\mu_j\left[\frac{1}{\sigma_0^2}\mu_0+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}(x_i-\eta_k)s_{ij}u_{ik}\right]\right)\right\}
\end{aligned}
$$

$$\propto \exp\left\{ -\frac{1}{2}\left(\frac{1}{\sigma_0^2}+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}s_{ij}u_{ik}\right)\left(\mu_j^2-2\mu_j\left[\frac{1}{\sigma_0^2}+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}s_{ij}u_{ik}\right]^{-1}\left[\frac{1}{\sigma_0^2}\mu_0+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}(x_i-\eta_k)s_{ij}u_{ik}\right]\right)\right\} \tag{53}$$

Completing the squares of the exponent in (53) gives:

$$\begin{aligned}
p(\mu_j|\boldsymbol{x},\boldsymbol{s},\boldsymbol{\mu}_{-j},\mu_0,\boldsymbol{u},\boldsymbol{\eta},\sigma,\sigma_0) &= N(\mu_j|\mu_{jN},\sigma_{jN}^2)\\
\mu_{jN} &= \left[\frac{1}{\sigma_0^2}+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}s_{ij}u_{ik}\right]^{-1}\left[\frac{1}{\sigma_0^2}\mu_0+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}(x_i-\eta_k)s_{ij}u_{ik}\right]\\
\frac{1}{\sigma_{jN}^2} &= \frac{1}{\sigma_0^2}+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}s_{ij}u_{ik}
\end{aligned}$$

Note that similar results can be obtained for η_k.

8.2.3 Inference

As described, the Gibbs sampling algorithm proceeds by iterative sampling from posteriors of individual hidden variables. Since we already derived posteriors for $\boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0$ in Section 8.2.1 and Section 8.2.2, we can use these posteriors to obtain samples from the true posterior

$$p(\boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0|\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0).$$

After some burn-in period, e.g. M samples, we collect N samples of $\boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0$ and use them to estimate the parameters the approximating posterior

$$\begin{aligned}
&p(\boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0|\tilde{\mu}_{0N}, \tilde{\sigma}_{0N}^2, \tilde{\boldsymbol{\mu}}_N, \tilde{\boldsymbol{\sigma}}_N^2, \tilde{\eta}_{0N}, \tilde{\gamma}_{0N}^2, \tilde{\boldsymbol{\eta}}_N, \tilde{\boldsymbol{\gamma}}_N^2) =\\
&= N(\mu_0|\tilde{\mu}_{0N}, \tilde{\sigma}_{0N}^2)\prod_{j=1}^{S} N(\mu_j|\tilde{\mu}_{jN}, \tilde{\sigma}_{jN}^2)\cdot N(\eta_0|\tilde{\eta}_{0N}, \tilde{\gamma}_{0N}^2)\prod_{k=1}^{S} N(\eta_k|\tilde{\eta}_{kN}, \tilde{\gamma}_{kN}^2)
\end{aligned}$$

where the parameters can be computed as follows:

$$\begin{aligned}
\tilde{\mu}_{0N} &= \frac{1}{N}\sum_{i=1}^{N}\mu_{0i}\\
\tilde{\sigma}_{0N}^2 &= \frac{1}{N}\sum_{i=1}^{N}(\mu_{0i}-\tilde{\mu}_{0N})^2\\
\tilde{\mu}_{jN} &= \frac{1}{N}\sum_{i=1}^{N}\mu_{ji} && \forall j \in \{1,\dots,S\}\\
\tilde{\sigma}_{jN}^2 &= \frac{1}{N}\sum_{i=1}^{N}(\mu_{ji}-\tilde{\mu}_{jN})^2 && \forall j \in \{1,\dots,S\}\\
\tilde{\eta}_{0N} &= \frac{1}{N}\sum_{i=1}^{N}\eta_{0i}\\
\tilde{\gamma}_{0N}^2 &= \frac{1}{N}\sum_{i=1}^{N}(\eta_{0i}-\tilde{\eta}_{0N})^2\\
\tilde{\eta}_{kN} &= \frac{1}{N}\sum_{i=1}^{N}\eta_{ki} && \forall k \in \{1,\dots,U\}\\
\tilde{\gamma}_{kN}^2 &= \frac{1}{N}\sum_{i=1}^{N}(\eta_{ki}-\tilde{\eta}_{kN})^2 && \forall k \in \{1,\dots,U\}
\end{aligned}$$