Bayesian inference on examples

Filip Jurcicek

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

March 28, 2014

Abstract

This article briefly introduces Bayesian inference on a set of simple examples. The used techniques include analytic solutions for the simplest cases, Markov Chain Monte Carlo (MCMC) techniques, represented by Gibbs sampling, the Variational Inference and Expectation Propagation algorithms.

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 article introduces Bayesian inference on a set of simple examples. The used techniques include analytic solutions for the simplest cases, Markov Chain Monte Carlo (MCMC) techniques, represented by Gibbs sampling, the Variational Inference and Expectation Propagation algorithms.

2 The binomial distribution

TBD

3 The multinomial distribution

TBD

4 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.

4.1 Inference of the mean while the variance is known

Figure 1 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.

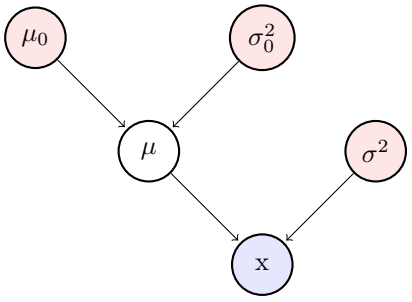

Figure 1: The graphical model for prediction of the observation x given the mean μ where the mean μ is unknown.

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 then 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:

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

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).

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

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

$$\mu_x = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2} x + \frac{\sigma^2}{\sigma_0^2 + \sigma^2} \mu_0 \tag{12}$$

$$\frac{1}{\sigma_x^2} = \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} \tag{13}$$

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 2 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.

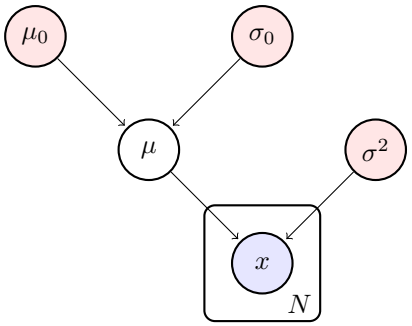

Figure 2: The graphical model representing the joint distribution for the observations $x_1, \ldots, x_N$ and the mean μ.

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

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

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 $\mu_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i$.

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

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

4.2 Inference of the variance while the mean is known

Figure 3 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

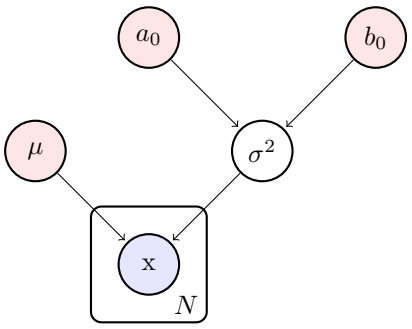

Figure 3: The graphical model for prediction of the observations $x_1, \ldots, x_N$ given the mean μ where the variance σ^2 is unknown.

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.

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 4. Substitution (16) into (9), the normal distribution using

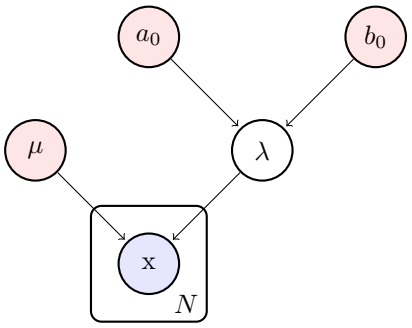

Figure 4: The graphical model for prediction of the observations $x_1, \ldots, x_N$ given the mean μ where the precision λ is unknown.

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

4.3 Inference of the mean and variance using a conjugate prior

Figure 5 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.

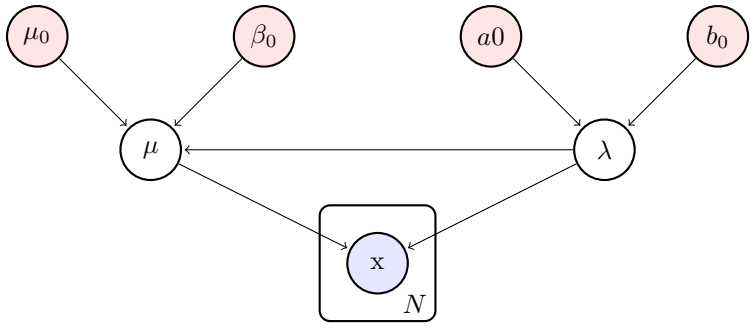

Figure 5: The graphical model for prediction of the observations $x_1, \ldots, x_N$ given the mean μ and the precision λ. Where both the mean and variance is unknown.

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:

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

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:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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:

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

4.4 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 6 depict a graphical model representing inference of the mean and precision of an univariate normal distribution using non-conjugate prior.

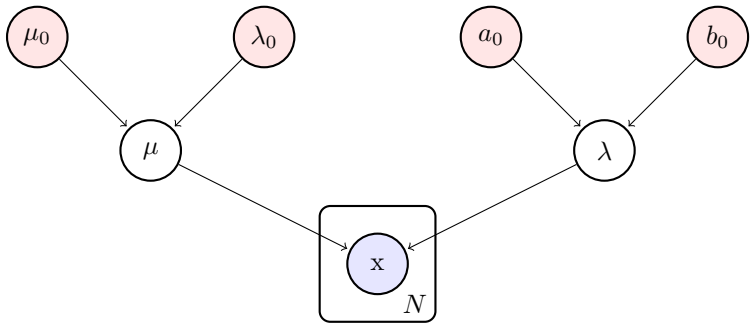

Figure 6: The graphical model for prediction of the observation x given the mean μ and the precision λ. Where both the mean and precision are unknown and the prior is not conjugate.

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:

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

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 (21). 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. In the next three sections, we will demonstrate all the mentioned techniques.

4.4.1 Gibbs sampling

Instead of direct computation of the posterior, MCMC techniques sample from the posterior distribution and then they use the samples to represent the posterior. 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.

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 $p(\mu, \lambda | \ldots)$. 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 (21).

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

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

Recall that we derived the posterior for the mean in Section 4.1 and that the posterior for the precision was derived in Section 4.2. Therefore, we already know that:

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

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 $p(\mu, \lambda | \ldots)$. Note that although the posteriors for μ and λ are independent, they exhibit a coupling since the posterior $p(\mu | \ldots)$ depends on the precision λ and the other way around.

Assume that we have obtained N samples of the mean, e.g. $\mu_1, \ldots, \mu_N$, and precision, e.g. $\lambda_1, \ldots, \lambda_N$, 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:

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

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

4.4.2 Variational Inference

In the previous section, the Gibbs sampling was used to obtain samples from the posterior of the unknown variables. Although the use of Gibbs sampling and MCMC methods is straight forward and one can get samples from the true posterior, these methods are very computationally expensive. One of the main reason is that many samples must be excluded to obtain representative samples of the posterior.

An interesting alternative to the MCMC methods, is Variational Inference (sometimes called Mean field variational inference or variational approximation inference). Instead of sampling from the posteriors, one can compute approximations of the true posteriors. In general, the task in Variational Inference is to approximate the joint distribution over all unobserved variables with a product of marginals, that is to find a

$$q(\boldsymbol{z};\theta) = \prod_i q_i(z_i;\theta_i) \tag{23}$$

such as $q(\boldsymbol{z};\theta) \approx p(\boldsymbol{z})$, where $p(\boldsymbol{z})$ is the true joint distribution. The objective in Variational Inference is to minimise Kullback–Leibler divergence $KL(q||p)$, where the KL divergence is defined as

$$\begin{aligned} KL(q||p) &= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z})} \text{ or } = \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z})} d\boldsymbol{z} \\ &= -\sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} \text{ or } = -\int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} d\boldsymbol{z}. \end{aligned}$$

One possibility is how minimise the KL divergence is to use gradient descend. However, more common approach is coordinate descend on individual $q(z_i;\theta_i)$.

In our case, the goal is to approximate the posterior (22) with the with a product of marginals in the form of the prior as defined in (21). Therefore, the approximating distribution (23) is defined as follows:

$$q(\mu, \lambda|\mu_N, \lambda_N, a_N, b_N) = N(\mu|\mu_N, \lambda_N^{-1}) Gam(\lambda|a_N, b_N), \tag{24}$$

where $z_1 = \mu$ and $z_2 = \lambda$, and the factors q_i are

$$\begin{aligned} q_1(z_1|\mu_N, \lambda_N) &\equiv q_1(\mu|\mu_N, \lambda_N) = N(\mu|\mu_N, \lambda_N^{-1}) \\ q_2(z_2|a_N, b_N) &\equiv q_2(\lambda|a_N, b_N) = Gam(\lambda|a_N, b_N) \end{aligned}$$

The true distribution is defines as follows:

$$p(\mu, \lambda|\boldsymbol{x}, \mu_0, \lambda_0, a_0, b_0) \propto N(\mu|\mu_0, \lambda_0^{-1}) Gam(\lambda|a_0, b_0) \prod_{i=1}^{N} N(z_i|\mu, \lambda^{-1}). \tag{25}$$

One can notice that we know the joint posterior distribution (25) 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 24. 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:

$$
\begin{aligned}
\frac{\partial KL(q||p)}{\partial \mu_N} = & \frac{\partial}{\partial \mu_N} \int_{\mu} N(\mu|\mu_N, \lambda_N^{-1}) \log N(\mu|\mu_N, \lambda_N^{-1}) d\mu \\
& - \frac{\partial}{\partial \mu_N} \int_{\mu} N(\mu|\mu_N, \lambda_N^{-1}) \log N(\mu|\mu_0, \lambda_0^{-1}) d\mu \\
& - \frac{\partial}{\partial \mu_N} \int_{\mu,\lambda} N(\mu|\mu_N, \lambda_N^{-1}) Gam(\lambda|a_N, b_N) \sum_{i=1}^{N} \log N(x_i|\mu, \lambda^{-1}) d\mu d\lambda
\end{aligned} \tag{26}
$$

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

$$
N(x|\mu, \lambda^{-1}) = \frac{\lambda^{1/2}}{(2\pi)^{1/2}} \exp\left\{ -\frac{\lambda}{2}(x-\mu)^2 \right\} \tag{27}
$$

$$
Gam(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp\{-b\lambda\}. \tag{28}
$$

Now substitute (27) and (28) into (26).

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

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:

$$\begin{aligned}0 &= (\log N(\mu|\mu_N,\lambda_N^{-1})+1)-\log N(\mu|\mu_0,\lambda_0^{-1})-\int_\lambda Gam(\lambda|a_N,b_N)\sum_{i=1}^N\log N(x_i|\mu,\lambda^{-1})d\lambda\\ \log N(\mu|\mu_N,\lambda_N^{-1}) &= \log N(\mu|\mu_0,\lambda_0^{-1})+\int_\lambda Gam(\lambda|a_N,b_N)\sum_{i=1}^N\log N(x_i|\mu,\lambda^{-1})d\lambda-1\\ N(\mu|\mu_N,\lambda_N^{-1}) &\propto \exp\left\{\log N(\mu|\mu_0,\lambda_0^{-1})+\int_\lambda Gam(\lambda|a_N,b_N)\sum_{i=1}^N\log N(x_i|\mu,\lambda^{-1})d\lambda\right\}\\ N(\mu|\mu_N,\lambda_N^{-1}) &\propto \exp\left\{\int_\lambda Gam(\lambda|a_N,b_N)\log\left(N(\mu|\mu_0,\lambda_0^{-1})\prod_{i=1}^N N(x_i|\mu,\lambda^{-1})\right)d\lambda\right\}\end{aligned}\tag{29}$$

Using the results (17), (29) 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

$$
\begin{aligned}
N(\mu|\mu_N,\lambda_N^{-1}) &\equiv N(\mu|\mu_N,\lambda_N^{-1};a_N,b_N) \qquad (30)\\
\mu_N &= \frac{N\cdot(a_N/b_N)\mu_{ML}+\lambda_0\mu_0}{N\cdot(a_N/b_N)+\lambda_0}\\
\lambda_N &= N\cdot(a_N/b_N)+\lambda_0
\end{aligned}
$$

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:

$$
\begin{aligned}
0 &= (\log Gam(\lambda|a_N, b_N) + 1) - \log Gam(\lambda|a_0, b_0) - \int_\lambda N(\mu|\mu_N, \lambda_N^{-1}) \sum_{i=1}^{N} \log N(x_i|\mu, \lambda^{-1}) d\lambda \\
\log Gam(\lambda|a_N, b_N) &= \log Gam(\lambda|a_0, b_0) + \int_\lambda N(\mu|\mu_N, \lambda_N^{-1}) \sum_{i=1}^{N} \log N(x_i|\mu, \lambda^{-1}) d\lambda - 1 \\
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ \log Gam(\lambda|a_0, b_0) + \int_\lambda N(\mu|\mu_N, \lambda_N^{-1}) \sum_{i=1}^{N} \log N(x_i|\mu, \lambda^{-1}) d\lambda \right\} \\
Gam(\lambda|a_N, b_N) &\propto \exp\left\{ \int_\lambda N(\mu|\mu_N, \lambda_N^{-1}) \log \left(Gam(\lambda|a_0, b_0) \prod_{i=1}^{N} N(x_i|\mu, \lambda^{-1}) \right) d\lambda \right\}
\end{aligned}
\tag{31}
$$

Using the results (20), (31) 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})} \left[\mu \right] + 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 $Gam(\lambda|a_N, b_N)$:

$$Gam(\lambda|a_N, b_N) \equiv Gam(\lambda|a_N, b_N; \mu_N, \lambda_N^{-1}) \tag{32}$$
$$a_N = a_0 + \frac{N}{2}$$
$$b_N = b_0 + \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu_N)^2 + \frac{N\lambda_N^{-1}}{2}$$

One can see that the formulae (30) and (32) depend on each other, therefore the process of Variational Inference is based on iterative computation of the posterior approximations (30) and (32), where the initial values are guessed.

In the text above, the Variational Inference algorithm was derived for estimating posteriors of parameters of a normal distribution with non-conjugate prior. However, this approach can be rather tedious if it has to be done the same way for every new model. Therefore, it is convenient to derive some general results. The aim of Variational Inference is to minimise the KL divergence between the approximating distribution $q(\boldsymbol{z};\theta) = \prod_i q_i(z_i;\theta_i)$ and the true distribution $p(\boldsymbol{z})$. This can be done by manipulating the $KL(q|p)$ divergence and throwing away all terms that do not depend on $q_i(z_i;\theta_i)$:

$$\begin{aligned}
KL(q||p) &= \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z})} d\boldsymbol{z} \\
&= \int_{\boldsymbol{z}} \prod_j q_j(z_j;\theta_j) \log \frac{\prod_k q_k(z_k;\theta_k)}{p(\boldsymbol{z})} d\boldsymbol{z} \\
&= \int_{\boldsymbol{z}} \left(\prod_j q_j(z_j;\theta_j)\right) \left(\sum_k \log q_k(z_k;\theta_k)\right) d\boldsymbol{z} - \int_{\boldsymbol{z}} \prod_j q_j(z_j;\theta_j) \log p(\boldsymbol{z}) d\boldsymbol{z} \\
&= \sum_k \int_{\boldsymbol{z}} \prod_j q_j(z_j;\theta_j) \log q_k(z_k;\theta_k) d\boldsymbol{z} - \int_{\boldsymbol{z}} \prod_j q_j(z_j;\theta_j) \log p(\boldsymbol{z}) d\boldsymbol{z} \\
&= \int_{\boldsymbol{z}} \prod_j q_j(z_j;\theta_j) \log q_i(z_i;\theta_i) d\boldsymbol{z} - \int_{\boldsymbol{z}} \prod_j q_j(z_j;\theta_j) \log p(\boldsymbol{z}) d\boldsymbol{z} + C_1 \\
&= \int_{\boldsymbol{z}} q_i(z_j;\theta_j) \log q_i(z_i;\theta_i) \prod_{j\neq i} q_i(z_j;\theta_j) d\boldsymbol{z} - \int_{\boldsymbol{z}} \prod_j q_j(z_j;\theta_j) \log p(\boldsymbol{z}) d\boldsymbol{z} + C_1 \\
&= \int_{z_i} q_i(z_j;\theta_j) \log q_i(z_i;\theta_i) \int_{\boldsymbol{z}\setminus z_i} \prod_{j\neq i} q_i(z_j;\theta_j) d\boldsymbol{z} - \int_{\boldsymbol{z}} \prod_j q_j(z_j;\theta_j) \log p(\boldsymbol{z}) d\boldsymbol{z} + C_1 \\
&= \int_{z_i} q_i(z_i;\theta_i) \log q_i(z_i;\theta_i) dz_i - \int_{\boldsymbol{z}} \prod_j q_j(z_j;\theta_j) \log p(\boldsymbol{z}) d\boldsymbol{z} + C_1 \\
&= \int_{z_i} q_i(z_i;\theta_i) \log q_i(z_i;\theta_i) dz_i - \int_{z_i} q_i(z_i;\theta_i) \int_{\boldsymbol{z}\setminus z_i} \prod_{j\neq i} q_j(z_j;\theta_j) \log p(\boldsymbol{z}) d\boldsymbol{z} + C_1 \\
&= \int_{z_i} q_i(z_i;\theta_i) \log q_i(z_i;\theta_i) dz_i - \int_{z_i} q_i(z_i;\theta_i) \log \left(\exp\left\{\int_{\boldsymbol{z}\setminus z_i} \prod_{j\neq i} q_j(z_j;\theta_j) \log p(\boldsymbol{z}) d\boldsymbol{z}\setminus z_i\right\}\right) dz_i + C_1 \\
&= \int_{z_i} q_i(z_i;\theta_i) \log \frac{q_i(z_i;\theta_i)}{\exp\left\{\int_{\boldsymbol{z}\setminus z_i} \prod_{j\neq i} q_j(z_j;\theta_j) \log p(\boldsymbol{z}) d\boldsymbol{z}\setminus z_i\right\}} dz_i + C_1 \\
&= KL\left(q_i || \exp\left\{\int_{\boldsymbol{z}\setminus z_i} \prod_{j\neq i} q_j(z_j;\theta_j) \log p(\boldsymbol{z}) d\boldsymbol{z}\setminus z_i\right\}\right) + C_1 \\
&= KL(q_i || \exp E_{q\setminus q_i}[\log p(\boldsymbol{z})]) + C_1
\end{aligned}$$

where $\boldsymbol{z}\setminus z_i = [z_0, \ldots, z_{i-1}, z_{i+1,\ldots}]$, $q\setminus q_i = q(\boldsymbol{z})/q_i(z_i;\theta_i) = \prod_{j\neq i} q_j(z_j;\theta_j)$ and $C_1 = \sum_{k\neq i} \int_{\boldsymbol{z}} \prod_j q_j(z_j;\theta_j) \log q_k(z_k;\theta_k) d\boldsymbol{z}$.

Since the KL divergence is minimised when the two arguments are the equal, the optimal approximation for $q_i(z_i;\theta_i)$ is:

$$q_i(z_i;\theta_i) \propto \exp E_{q\setminus q_i}[\log p(\boldsymbol{z})] \tag{33}$$

Now going back to our example, (29) and (31) can be computed from (33) when all necessary terms are substituted and simplified.

Theoretical properties of the Variational inference are very favourable. Although it is an approximation, it is guaranteed to converge to a local optimum. Let assume that $\boldsymbol{x}$ are observed data and $\boldsymbol{z}$ are unknown variables/parameters. Our probabilistic model specifies joint distribution $p(\boldsymbol{x}, \boldsymbol{z})$ and our goal is to find an approximation to $p(\boldsymbol{z}|\boldsymbol{x})$. Then, the log marginal probability of the data can be decomposed as follows:

$$\log p(\boldsymbol{x}) = L(q) + KL(q||p),$$

where we have defined

$$L(q) = \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} d\boldsymbol{z}$$
$$KL(q||p) = -\int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z})} d\boldsymbol{z}$$

Because 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 $\log p(\boldsymbol{x})$. The goal of Variational inference is the variational lower bound $L(q)$ with respect to the approximate $q(\boldsymbol{z})$ distribution, or to minimise the $KL(q||p)$ divergence.

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

$$\begin{aligned}\log p(\boldsymbol{x}) &= \log \int_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} = \log \int_{\boldsymbol{z}} q(\boldsymbol{z}) \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} d\boldsymbol{z} \\ &\geq \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} d\boldsymbol{z} \\ &\geq L(q)\end{aligned}$$

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.

4.4.3 Expectation Propagation

TBD

5 Hierarchical Bayesian model for the real observations

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 7. 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 are communication 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 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.

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 8.

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

1. $\mu_0 \sim N(\cdot|\mu_{-1}, \sigma_{-1})$
2. $\eta_0 \sim N(\cdot|\eta_{-1}, \gamma_{-1})$
3. $\mu_s \sim N(\cdot|\mu_0, \sigma_0)$

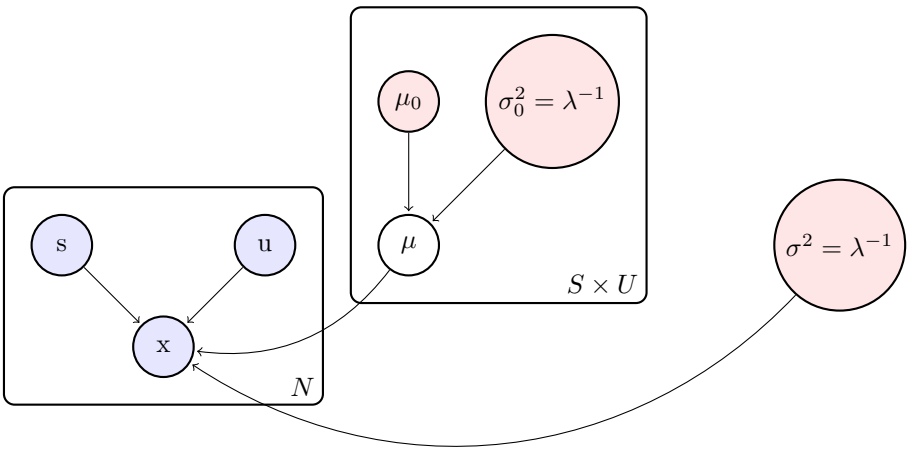

Figure 7: The probabilistic model for the prediction of the observation x given s and u.

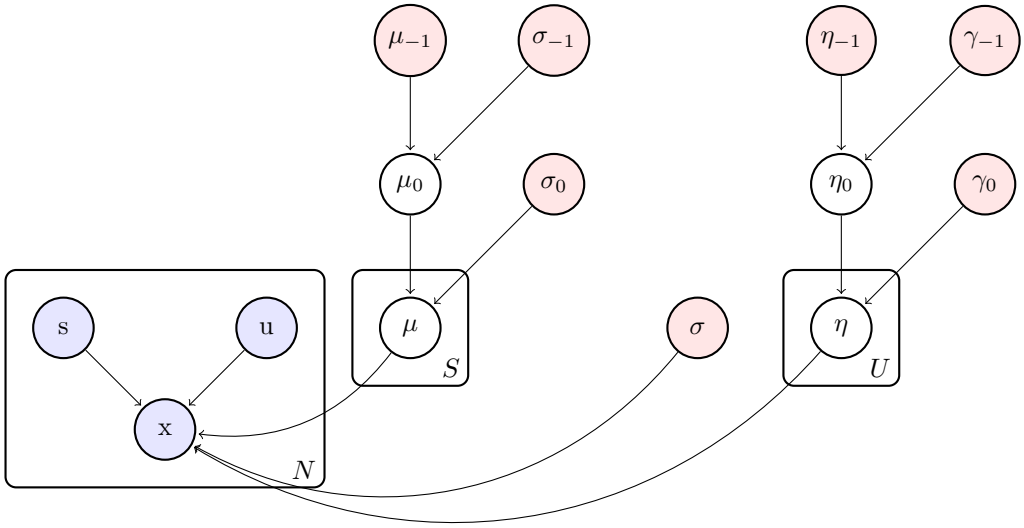

Figure 8: Graphical model factoring the system and user parameters represented by a hierarchical Bayesian model. The model is modelling only the mean and the variance is assumed to be known.

4. $\eta_u \sim N(\cdot|\eta_0, \gamma_0)$
5. $x \sim N(\cdot|\mu_s + \eta_u, \sigma)$

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:

$$\begin{aligned} &p(\boldsymbol{x}, \boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0 | \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0) = \\ &= p(\boldsymbol{\mu}|\mu_0, \sigma_0) p(\mu_0|\mu_{-1}, \sigma_{-1}) p(\boldsymbol{\eta}|\eta_0, \gamma_0) p(\eta_0|\eta_{-1}, \gamma_{-1}) \prod_{i=1}^{N} p(x_i|s_i, \boldsymbol{\mu}, u_i, \boldsymbol{\eta}, \sigma) \end{aligned} \tag{34}$$

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 $\boldsymbol{\eta}$:

$$p(\boldsymbol{x}, \boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0 | \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0) =$$
$$= p(\mu_0|\mu_{-1}, \sigma_{-1}) \prod_{j=1}^{S} p(\mu_j|\mu_0, \sigma_0) \cdot p(\eta_0|\eta_{-1}, \gamma_{-1}) \prod_{k=1}^{U} p(\eta_k|\eta_0, \gamma_0) \cdot \prod_{i=1}^{N} \prod_{j=1}^{S} \prod_{k=1}^{U} p(x_i|\mu_j, \eta_k, \sigma)^{s_{ij} u_{ik}} \tag{35}$$

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:

$$p(\mu_0|\mu_{-1}, \sigma_{-1}) = N(\mu_0|\mu_{-1}, \sigma_{-1}) \tag{36}$$
$$p(\mu_j|\mu_0, \sigma_0) = N(\mu_j|\mu_0, \sigma_0) \tag{37}$$
$$p(\eta_0|\eta_{-1}, \gamma_{-1}) = N(\eta_0|\eta_{-1}, \gamma_{-1}) \tag{38}$$
$$p(\eta_k|\eta_0, \gamma_0) = N(\eta_k|\eta_0, \gamma_0) \tag{39}$$
$$p(x_i|\mu_j, \eta_k, \sigma) = N(x_i|\mu_j + \eta_k, \sigma) \tag{40}$$

5.1 Gibbs sampling

In this section, we will describe inference using Gibbs sampling in the model described above. The Gibbs sampling was already detailed in Section 4.4.1. 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:

$$p(\mu_0|\boldsymbol{\mu}, \mu_{-1}, \sigma_{-1}, \sigma_0) \tag{41}$$
$$p(\mu_j|\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\mu}_{-j}, \mu_0, \boldsymbol{u}, \boldsymbol{\eta}, \sigma, \sigma_0) \quad \forall j \in \{1, \ldots, S\} \tag{42}$$
$$p(\eta_0|\boldsymbol{\eta}, \eta_{-1}, \gamma_{-1}, \gamma_0) \tag{43}$$
$$p(\eta_k|\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\mu}, \boldsymbol{u}, \boldsymbol{\eta}_{-k}, \eta_0, \sigma, \gamma_0) \quad \forall k \in \{1, \ldots, U\} \tag{44}$$

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).

5.1.1 Posterior of the hyper-parameters

The easiest way to start is to compute posterior of μ_0. Using the joint distribution (34) and the Bayes rule:

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

Therefore the posterior is computed as:

$$\begin{aligned} &p(\mu_0|\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\eta}, \eta_0, \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0) = \\ &= \frac{p(\boldsymbol{x}, \boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0 | \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0)}{p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\eta}, \eta_0 | \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0)} \\ &= \frac{p(\boldsymbol{x}, \boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0 | \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0)}{\int p(\boldsymbol{x}, \boldsymbol{\mu}, \mu_0, \boldsymbol{\eta}, \eta_0 | \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0) d\mu_0} \end{aligned} \tag{45}$$

When (34) is substituted into (45), then it results in:

$$\begin{aligned} &p(\mu_0|\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\eta}, \eta_0, \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0) = \\ &= \frac{p(\boldsymbol{x}|s, \boldsymbol{\mu}, u, \boldsymbol{\eta}, \sigma) p(\boldsymbol{\mu}|\mu_0, \sigma_0) p(\mu_0|\mu_{-1}, \sigma_{-1}) p(\boldsymbol{\eta}|\eta_0, \gamma_0) p(\eta_0|\eta_{-1}, \gamma_{-1})}{\int p(\boldsymbol{x}|s, \boldsymbol{\mu}, u, \boldsymbol{\eta}, \sigma) p(\boldsymbol{\mu}|\mu_0, \sigma_0) p(\mu_0|\mu_{-1}, \sigma_{-1}) p(\boldsymbol{\eta}|\eta_0, \gamma_0) p(\eta_0|\eta_{-1}, \gamma_{-1}) d\mu_0} \\ &= \frac{p(\boldsymbol{\mu}|\mu_0, \sigma_0) p(\mu_0|\mu_{-1}, \sigma_{-1})}{\int p(\boldsymbol{\mu}|\mu_0, \sigma_0) p(\mu_0|\mu_{-1}, \sigma_{-1}) d\mu_0} \\ &= p(\mu_0|\boldsymbol{\mu}, \mu_{-1}, \sigma_{-1}, \sigma_0) \end{aligned} \tag{46}$$

One can see that the result is exactly what we need for Gibbs sampling as described in (41). Now using (35) and substituting (37) and (36) into (46), results in:

$$\begin{aligned} p(\mu_0|\boldsymbol{\mu}, \mu_{-1}, \sigma_{-1}, \sigma_0) &\propto p(\boldsymbol{\mu}|\mu_0, \sigma_0) p(\mu_0|\mu_{-1}, \sigma_{-1}) \\ &\propto N(\mu_0|\mu_{-1}, \sigma_{-1}) \prod_{j=1}^{S} N(\mu_j|\mu_0, \sigma_0) \end{aligned} \tag{47}$$

Recall that we already derived the posterior for the mean in Section 4.1. 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.

5.1.2 Posterior of the parameters

Now, we will compute the posterior for $\boldsymbol{\mu}$. Using the joint distribution (34) 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:

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

When (35) is substituted into (48), then it results in:

$$\begin{aligned} &p(\mu_j|\boldsymbol{x}, \boldsymbol{\mu}_{-j}, \mu_0, \boldsymbol{\eta}, \eta_0, \boldsymbol{s}, \boldsymbol{u}, \mu_{-1}, \sigma_{-1}, \sigma_0, \sigma, \eta_{-1}, \gamma_{-1}, \gamma_0) = \\ &= \frac{p(\mu_0|\mu_{-1}, \sigma_{-1}) \prod_{l=1}^{S} p(\mu_l|\mu_0, \sigma_0) \cdot p(\eta_0|\eta_{-1}, \gamma_{-1}) \prod_{k=1}^{U} p(\eta_k|\eta_0, \gamma_0) \cdot \prod_{i=1}^{N} \prod_{l=1}^{S} \prod_{k=1}^{U} p(x_i|\mu_l, \eta_k, \sigma)^{s_{il}u_{ik}}}{\int p(\mu_0|\mu_{-1}, \sigma_{-1}) \prod_{l=1}^{S} p(\mu_l|\mu_0, \sigma_0) \cdot p(\eta_0|\eta_{-1}, \gamma_{-1}) \prod_{k=1}^{U} p(\eta_k|\eta_0, \gamma_0) \cdot \prod_{i=1}^{N} \prod_{l=1}^{S} \prod_{k=1}^{U} p(x_i|\mu_l, \eta_k, \sigma)^{s_{il}u_{ik}} d\mu_j} \\ &= \frac{\prod_{l=1}^{S} p(\mu_l|\mu_0, \sigma_0) \cdot \prod_{i=1}^{N} \prod_{l=1}^{S} \prod_{k=1}^{U} p(x_i|\mu_l, \eta_k, \sigma)^{s_{il}u_{ik}}}{\int \prod_{l=1}^{S} p(\mu_l|\mu_0, \sigma_0) \cdot \prod_{i=1}^{N} \prod_{l=1}^{S} \prod_{k=1}^{U} p(x_i|\mu_l, \eta_k, \sigma)^{s_{il}u_{ik}} d\mu_j} \\ &= \frac{p(\mu_j|\mu_0, \sigma_0) \prod_{l=1, l\neq j}^{S} p(\mu_l|\mu_0, \sigma_0) \cdot \prod_{i=1}^{N} \prod_{k=1}^{U} p(x_i|\mu_j, \eta_k, \sigma)^{s_{ij}u_{ik}} \prod_{l=1, l\neq j}^{S} p(x_i|\mu_l, \eta_k, \sigma)^{s_{il}u_{ik}}}{\int p(\mu_j|\mu_0, \sigma_0) \prod_{l=1, l\neq j}^{S} p(\mu_l|\mu_0, \sigma_0) \cdot \prod_{i=1}^{N} \prod_{k=1}^{U} p(x_i|\mu_j, \eta_k, \sigma)^{s_{ij}u_{ik}} \prod_{l=1, l\neq j}^{S} p(x_i|\mu_l, \eta_k, \sigma)^{s_{il}u_{ik}} d\mu_j} \\ &= \frac{p(\mu_j|\mu_0, \sigma_0) \cdot \prod_{i=1}^{N} \prod_{k=1}^{U} p(x_i|\mu_j, \eta_k, \sigma)^{s_{ij}u_{ik}}}{\int p(\mu_j|\mu_0, \sigma_0) \cdot \prod_{i=1}^{N} \prod_{k=1}^{U} p(x_i|\mu_j, \eta_k, \sigma)^{s_{ij}u_{ik}} d\mu_j} \\ &= p(\mu_j|\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\mu}_{-j}, \mu_0, \boldsymbol{u}, \boldsymbol{\eta}, \sigma, \sigma_0) \end{aligned} \tag{49}$$

Substituting (37) and (40) into (49) 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.

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

$$\propto \exp\left\{ -\frac{1}{2}\left(\frac{1}{\sigma_0^2}+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}s_{ij}u_{ik}\right)\left(\mu_j^2-2\mu_j\left[\frac{1}{\sigma_0^2}+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}s_{ij}u_{ik}\right]^{-1}\left[\frac{1}{\sigma_0^2}\mu_0+\sum_{i=1}^{N}\sum_{k=1}^{U}\frac{1}{\sigma^2}(x_i-\eta_k)s_{ij}u_{ik}\right]\right)\right\} \tag{50}$$

Completing the squares of the exponent in (50) 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.

5.1.3 Inference

As described in Section 4.4.1, 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 5.1.1 and Section 5.1.2, 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:

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

6 Mixture model

TBD

7 Hidden Markov Model (HMM)

TBD

8 The Laplace Approximation

TBD

9 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

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

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 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.

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

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

9.1 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.

9.2 Variational Mean Field

An alternative is to assume that q factorises 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.

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

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 $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_i]$. This approach is known in the literature as variational mean field or global variation 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\} dw_j \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:

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

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.

$$
\log q_j(w_j; \theta_j) = E_{q_{i\neq j}} \left[\log p(\boldsymbol{w}, D) \right] + C_3
$$

9.3 Example: Unknown Mean and Variance of a normal distribution, with improper priors

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

The likelihood of μ and τ is

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

where $\overline{x}$ is empirical mean.
Set the priors for μ and τ to be improper priors:

$$
\begin{aligned}
p(\mu) &= 1/\sigma_\mu \\
p(\tau) &= 1/\tau
\end{aligned}
$$

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 factorises

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

and solve for the optimal factors

$$
\begin{aligned}
\log q_\mu(\mu) &= E_{q_\tau}\left[\log p(D,\mu,\tau)\right] \\
\log q_\tau(\tau) &= E_{q_\mu}\left[\log p(D,\mu,\tau)\right]
\end{aligned}
$$

9.3.1 $\log q_\mu(\mu)$

Derivation:

$$\begin{aligned}
\log q_\mu(\mu) &= E_{q_\tau}\left[\log p(D,\mu,\tau)\right] \\
&= E_{q_\tau}\left[\log p(D|\mu,\tau)p(\mu)p(\tau)\right] \\
&= E_{q_\tau}\left[\log p(D|\mu,\tau) + \log p(\mu) + \log p(\tau)\right] \\
&= \int q_\tau(\tau)\left[\log p(D|\mu,\tau) + \log p(\mu) + \log p(\tau)\right] d\tau \\
&= \int q_\tau(\tau)\log p(D|\mu,\tau)d\tau + \int q_\tau(\tau)\log p(\mu)d\tau + \int q_\tau(\tau)\log p(\tau)d\tau \\
&= \int q_\tau(\tau)\log p(D|\mu,\tau)d\tau + \log p(\mu)\int q_\tau(\tau)d\tau + C_1 \\
&= \int q_\tau(\tau)\log p(D|\mu,\tau)d\tau + \log(1/\sigma_\mu)\cdot 1 + C_1 \\
&= \int q_\tau(\tau)\log p(D|\mu,\tau)d\tau + C_2 \\
&= \int q_\tau(\tau)\left(\frac{N}{2}\log\tau - \frac{\tau}{2}\left[N(\mu-\overline{x})^2 + \sum_{i=1}^{N}(x_i-\overline{x})^2\right]\right)d\tau + C_2 \\
&= \int q_\tau(\tau)\left(-\frac{N\tau}{2}(\mu-\overline{x})^2\right)d\tau + C_3 \\
&= \left(-\frac{N}{2}(\mu-\overline{x})^2\right)\int q_\tau(\tau)\tau d\tau + C_3 \\
&= -\frac{N}{2}(\mu-\overline{x})^2 E_\tau[\tau] + C_3
\end{aligned}$$

By introspection, one can observe that

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

where $\lambda = NE_\tau[\tau]$.

9.3.2 $\log q_\tau(\tau)$

Derivation:

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

By introspection, one can observe that

$$\begin{aligned} q_\tau(\tau) &\propto \tau^{\frac{N}{2}-1} \exp\left\{ -\frac{N}{2}\left(\lambda^{-1} + \sum_{i=1}^{N}(x_i - \overline{x})^2\right)\tau \right\} \\ &= Gam(\tau; a, b) \end{aligned}$$

where

$$\begin{aligned} Gam(\tau; a, b) &= b^a \frac{1}{\Gamma(a)} \tau^{a-1} \exp(-b\tau) \\ a &= \frac{N}{2} \\ b &= \frac{N}{2}\left(\lambda^{-1} + \sum_{i=1}^{N}(x_i - \overline{x})^2\right) \end{aligned}$$

9.3.3 Summary

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

$$\begin{aligned} q_\mu(\mu) &= N(\mu|\overline{x}, \lambda^{-1}) \\ q_\tau(\tau) &= Gam(\tau|a, b) = b^a \frac{1}{\Gamma(a)} \tau^{a-1} \exp\{-b\tau\} \end{aligned}$$

where

$$\begin{aligned} \lambda &= N E_{q_\tau}[\tau] = N\frac{a}{b} \\ a &= \frac{N}{2} \\ b &= \frac{N}{2}\left(E_{q_\mu}[\mu^2] - 2E_{q_\mu}[\mu]\overline{x} + \overline{x}^2 + \sum_{i=1}^{N}(x_i - \overline{x})^2\right) \\ &= \frac{N}{2}\left(\lambda^{-1} + \sum_{i=1}^{N}(x_i - \overline{x})^2\right) \end{aligned}$$

We iteratively optimise q_μ and q_τ until convergence.

9.4 Example: Unknown Mean and Variance of a normal distribution, with conjugate priors

Set the priors for μ and τ to be improper priors:

$$\begin{aligned} p(\mu) &= N(\mu|\mu_0, \lambda^{-1}) \\ p(\tau) &= Gam(\tau|a, b) \end{aligned}$$

We enforce that the posterior approximation factorises

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

and solve for the optimal factors

$$\begin{aligned} \log q_\mu(\mu) &= E_{q_\tau}\left[\log p(D, \mu, \tau)\right] \\ \log q_\tau(\tau) &= E_{q_\mu}\left[\log p(D, \mu, \tau)\right] \end{aligned}$$