Derivation of Gibbs Sampling for Finite Gaussian Mixture Model

Xiaodong Yu
University of Maryland, College Park

September 9, 2009

1 Introduction

This report illustrates many technical details of the algorithm described in "Infinite Gaussian Mixture Model" by Carl E. Rasmussen, NIPS 2000. Rasmussen's paper provides the conditional posterior distributions of the parameters in the mixture model, but lacks the details how to derive them. Furthermore, his definition of Gamma distribution introduces lots of confusions and his computation of Gamma posterior is doubtful. All these problems make it difficult for others to implement his algorithm in Matlab. The purpose of this report is to fill this gap. In the Appendix sections of this report, I also illustrate the derivations of the posterior distributions of a few standard distributions, such as Gaussian and Gamma, whose results are used in this report.

Rasmussen starts the paper with the finite Gaussian mixture model and then extend it to the case of infinite number of components. The focus of this report is the finite case, which is described in Section 2.1 of his NIPS paper.

The finite Gaussian mixture model (FGMM) with k clusters can be written as:

$$p(y|\mu_1, ..., \mu_k, s_1, ..., s_k, \pi_1, ..., \pi_k) = \sum_{j=1}^{k} \pi_j \mathcal{N}(\mu_j, s_j^{-1}). \tag{1.1}$$

where μ_j are the means and s_j the precisions, i.e., inverse variances, π_j the mixture proportions. In Rasmussen's paper and this report, only the scalar observations are considered. The graphical presentation of FGMM is in Figure 1. Besides the cluster parameters μ_j, s_j and π, there are also hyperparameters that control the priors of the cluster parameters, which are

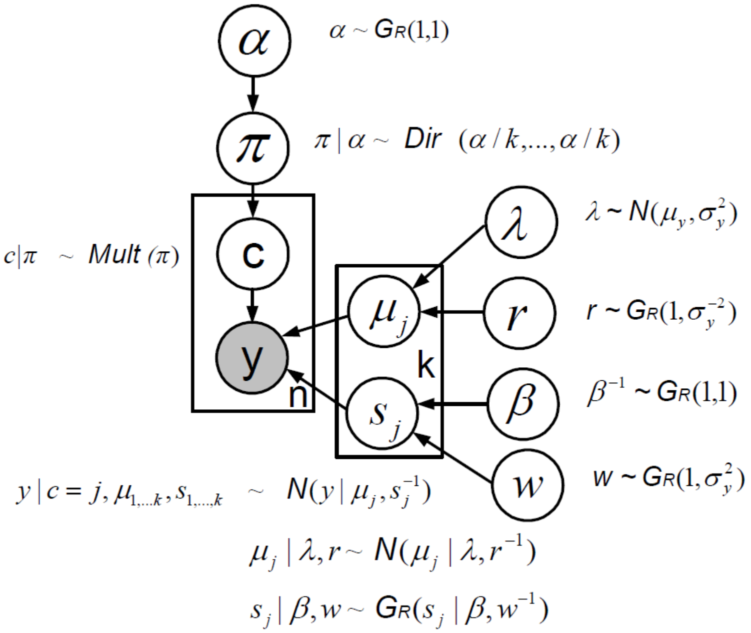
α
$\alpha \sim G_R(1,1)$
π
$\pi \mid \alpha \sim Dir\ (\alpha/k,\dots,\alpha/k)$
$c \mid \pi \sim Mult\ (\pi)$
c
y
n
μ_j
s_j
k
λ
$\lambda \sim N(\mu_y, \sigma_y^2)$
r
$r \sim G_R(1, \sigma_y^{-2})$
β
$\beta^{-1} \sim G_R(1,1)$
w
$w \sim G_R(1, \sigma_y^2)$
$y \mid c = j, \mu_{1,\dots,k}, s_{1,\dots,k} \sim N(y \mid \mu_j, s_j^{-1})$
$\mu_j \mid \lambda, r \sim N(\mu_j \mid \lambda, r^{-1})$
$s_j \mid \beta, w \sim G_R(s_j \mid \beta, w^{-1})$

Figure 1: The graphical presentation of FGMM

common to all clusters. In addition, an indicator variable, c_i, is introduced, one for each observation, to represent the observation's cluster membership, taking on values between 1 to k. Taking both cluster parameters and hyperparameters into account, the parameter set of this model is

$$\boldsymbol{\theta} = \{\mu_1, ..., \mu_k, s_1, ..., s_k, \pi_1, ..., \pi_k, \lambda, r, \beta, w, \alpha, c_1, ..., c_n\}. \qquad (1.2)$$

To do Gibbs sampling, we need to derive the conditional posterior distributions for each parameters conditioned on all the other parameters, $p(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{y})$, where $\mathbf{y} = \{y_t\}_{t=1}^n$ is the set of n data points. But for a graphical model, this conditional distribution is a function only of the nodes in the Markov blanket. For the FGMM, a directed graphic model, the Markov blanket includes the parents, the children, and the co-parents, as shown in Figure 2. From this graphical representation, we can find the Markov blanket for each parameter in the FGMM model, and then figure out their

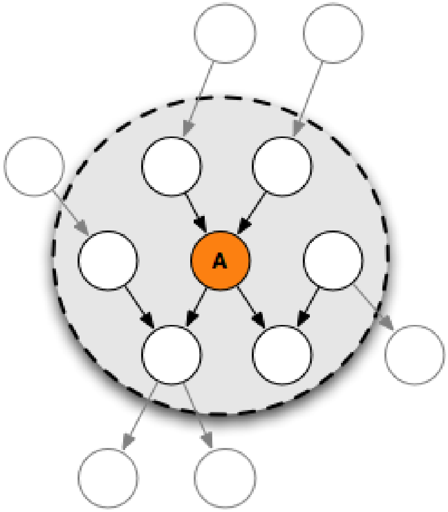

Figure 2: The Markov blanket of a directed graphic model. The picture is due to wikipedia.

conditional posterior distributions to be derived:

$$p(\mu_j|\mathbf{c}, \mathbf{y}, s_j, \lambda, r), j = 1, ..., k \tag{1.3}$$

$$p(\lambda|\mu_1, ..., \mu_k, r) \tag{1.4}$$

$$p(r|\mu_1, ..., \mu_k, \lambda) \tag{1.5}$$

$$p(s_j|\mathbf{c}, \mathbf{y}, \mu_j, \beta, w), j = 1, ..., k \tag{1.6}$$

$$p(w|s_1, ..., s_k, \beta) \tag{1.7}$$

$$p(\beta|s_1, ..., s_k, w) \tag{1.8}$$

$$p(c_i = j|\mathbf{c}_{-i}, \pi, \mathbf{y}, \mu, \mathbf{s}), i, ..., n \tag{1.9}$$

$$p(\pi|\alpha, \mathbf{c}) \tag{1.10}$$

$$p(\alpha|\pi) \tag{1.11}$$

Note that μ_j's from different components are independent to each other given the hyperparameters and the observations, and μ_j is related to only the precision of the j-th clusters, and the . That is why there is no term such as $\boldsymbol{\mu}_{-j}$ or $\mathbf{s}_{-j}$ within the conditional variables in the conditional posterior distributions of μ_j. This also applied to s_j's. But c_i's are not independent to each other. So there is a term $\mathbf{c}_{-i}$ that represents all indicator excluding c_i in the within the conditional variables in the conditional posterior distributions of c_i.

One of the most frequently used trick in the derivation is to apply the Bayesian theorem:

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \propto \text{prior} \times \text{likelihood}. \tag{1.12}$$

As we can see, we need not do the integral in the denominator. For a particular parameter, we first need to identify what are its observations and derive the likelihood from these observations. Then we assume a conjugate prior for this likelihood. Finally we multiply the likelihood and the prior to get the posterior distribution. This procedure work well for all parameters except c_i. For c_i, we need to consider the distribution conditioning on $\mathbf{c}_{-i}$. Detailed discussion will be presented in the relevant section.

2 Conditional Posterior of μ_j, λ and r

The component means, μ_j, has a Gaussian prior:

$$\mu_j|\lambda, r \sim \mathcal{N}(\lambda, r^{-1}). \tag{2.1}$$

we can get the μ_j's posterior by multiplying Equation (2.1) with Equation (1.1). Note that only the j-th cluster is involved, so only one Gaussian involved in the $\sum$ function in Equation (1.1). Thus the likelihood is in the same form of Equation (A.1). We can apply the results in Equation (A.9) to derive the conditional posterior of μ_j as:

$$p(\mu_j|\mathbf{c}, \mathbf{y}, s_j, \lambda, r) = \mathcal{N}\left(\frac{r\lambda + s_j \sum_{t:c_t=j} y_t}{n_j s_j + r}, \frac{1}{n_j s_j + r}\right). \tag{2.2}$$

The hyperparameter mean λ is common to all clusters. It is given a vague conjugate Gaussian prior:

$$p(\lambda) \sim \mathcal{N}(\mu_y, \sigma_y^2). \tag{2.3}$$

For λ, Equation 2.1 is its likelihood where the Gaussian precision r is known and the k of μ_j, $j = 1, ..., k$ are λ's observations. Use the result in Equation (A.9) again, we get λ's conditional posterior distribution:

$$p(\lambda|\mu_1, ..., \mu_k, r) = \mathcal{N}\left(\frac{\sigma^{-2}\mu_y + r\sum_{j=1}^{k} \mu_j}{kr + \sigma^{-2}}, \frac{1}{kr + \sigma^{-2}}\right) \tag{2.4}$$

$\mathcal{G}_R(\alpha_R,\beta_R)$	$\mathcal{G}(\alpha,\beta)$	$\mathcal{G}_M(\alpha_M,\beta_M)$
$\propto x^{\alpha_R/2-1}e^{-\frac{\alpha_R}{\beta_R}x}$	$\propto x^{\alpha-1}e^{-\beta x}$	$\propto x^{\alpha_M-1}e^{-\frac{x}{\beta_M}}$

Table 1: Comparison of three definitions of Gamma distributions.

The hyperparameter precision r is common to all clusters. It is given a vague conjugate Gamma prior:

$$p(r)=\mathcal{G}_R(1,\sigma_y^{-2})\propto r^{-1/2}exp(-r\sigma_y^2/2). \tag{2.5}$$

Notice that Rasmussen's definition of Gamma distribution $\mathcal{G}_R(\cdot)$ is different from the one we find on wikipedia (also used in this report, see Equation (**??**)) or the one defined in Matlab. Let's denote the definition of Rasmussen's Gamma distribution, ours, and Matlab's as $\mathcal{G}_R(\alpha_R,\beta_R)$, $\mathcal{G}(\alpha,\beta)$ and $\mathcal{G}_M(\alpha_M,\beta_M)$ respectively. Their definitions are compared in Table 1 From Table 1, we can conclude the the relationships among their parameters are:

$$\alpha=\alpha_R/2=\alpha_M,\beta=\alpha_R/\beta_R=1/\beta_M. \tag{2.6}$$

Thus the prior of r is equivalent to $\mathcal{G}(1/2,\sigma_y^2/2)$ or $\mathcal{G}_M(1/2,2/\sigma_y^2)$. The posterior distribution of r can be obtained by applying the results in Equation (A.23):

$$\begin{aligned}p(r|\mu_1,...,\mu_k,\lambda) &= \mathcal{G}\left(\frac{k+1}{2},\frac{\sigma_y^2+\sum_{j=1}^k(\mu_j-\lambda)^2}{2}\right) &(2.7)\\ &= \mathcal{G}_M\left(\frac{k+1}{2},\frac{2}{\sigma_y^2+\sum_{j=1}^k(\mu_j-\lambda)^2}\right). &(2.8)\end{aligned}$$

3 Conditional Posterior of s_j, β and w

The component precisions s_j are given Gamma priors:

$$p(s_j|\beta,w)=\mathcal{G}_R(\beta,w^{-1})\propto s_j^{\beta/2-1}e^{-\beta w s_j}\Rightarrow p(s_j|\beta,w)=\mathcal{G}(\beta/2,\beta w). \tag{3.1}$$

Similar to the case of μ_j, s_j has n_j observations y_t with $c_t=j$, and their mean is μ_j. Thus we can apply the result in Equation (A.23) to derive the conditional posterior of s_j:

$$p(s_j|\mathbf{c},\mathbf{y},\mu_j,\beta,w) = \mathcal{G}\left(\frac{\beta+n_j}{2}, \beta w + \sum_{t:c_t=j}(y_t-\mu_j)^2/2\right) \tag{3.2}$$

$$= \mathcal{G}_M\left(\frac{\beta+n_j}{2}, \frac{1}{\beta w + \sum_{t:c_t=j}(y_t-\mu_j)^2/2}\right). \tag{3.3}$$

w controls the rate parameter of the Gamma distribution of s_j, which is a hyperparameter common to all components. It has k observations, $s_j, j = 1, ..., k$. We can give it a conjugate Gamma prior:

$$p(w) = \mathcal{G}_R(1, \sigma_y^2) \propto w^{-1/2}e^{-w/\sigma_y^2} \Rightarrow p(w) = \mathcal{G}(1/2, 1/\sigma_y^2). \tag{3.4}$$

To derive the posterior of w when we assume β known, we need to apply the results in Equation (B.7). Note that in Equation (3.1), the rate parameter is βw. So we need to transform the prior in Equation (3.4) to be a function of βw. Use the following identity:

$$p(w) = \mathcal{G}(a,b) \propto w^{a-1}e^{-bw} \propto (\beta w)^{a-1}e^{-\frac{b}{\beta}(\beta w)}, \tag{3.5}$$

we have $p(\beta w) = \mathcal{G}(1/2, \frac{1}{\sigma_y^2\beta})$. Use the results in Equation (B.7), we have the posterior of βw

$$p(\beta w|s_1, ..., s_k, \beta) = \mathcal{G}\left(\frac{1+k\beta}{2}, \frac{1}{\beta\sigma_y^2} + \sum_{j=1}^{k} s_j\right). \tag{3.6}$$

With known β, it is trivial to obtain the sample of w after we sample βw from Equation (3.6).

β is the shape parameter of the Gamma prior distribution of s_j, which is a hyperparameter common to all components. It has k observations, $s_j, j = 1, ..., k$. But we do not have a conjugate prior for this parameter. Rasmussen gives it an inverse Gamma priors:

$$p(\beta^{-1}) = \mathcal{IG}_R(1,1) \Rightarrow p(\beta) \propto \beta^{-3/2}e^{-\frac{1}{2\beta}} \Rightarrow p(\beta) = \mathcal{IG}(5/2, 1/2). \tag{3.7}$$

To derive the posterior of β, we need to use the definition of Gamma distribution in Equation (B.1). To avoid confusion, let's temporarily use $\alpha_0 = 5/2, \beta_0 = 1/2$ as the parameter of β's prior in Equation (3.7), and

$\hat{\alpha} = \beta/2, \hat{\beta} = \beta w$ as the parameter of β's likelihood in the first step of derivation. Then, β's posterior is:

$$\begin{aligned} p(\beta|s_1, ..., s_k, w) &\propto \frac{1}{\beta^{\alpha_0+1}} e^{-\frac{\beta_0}{\beta}} \prod_{j=1}^{k} \left(\frac{\hat{\beta}^{\hat{\alpha}}}{\Gamma(\hat{\alpha})} s_j^{\hat{\alpha}-1} e^{-\hat{\beta} s_j} \right) && (3.8) \\ &= \frac{1}{\beta^{5/2+1}} e^{-\frac{1/2}{\beta}} \prod_{j=1}^{k} \left(\frac{(\beta w)^{\beta/2}}{\Gamma(\beta/2)} s_j^{\beta/2-1} e^{-\beta w s_j} \right) && (3.9) \\ &= \frac{\exp(\beta w \sum_{j=1}^{k} s_j - \frac{1}{2\beta})}{\Gamma(\beta/2)^k} \beta^{\frac{k\beta-7}{2}} (\prod_{j=1}^{k} w s_j)^{\beta/2}. && (3.10) \end{aligned}$$

This is not of standard form and can not use off-the-shelf Matlab routine to generate samples therefrom. But it can be shown that $\log(\beta)|s_1, ..., s_k, w$ is log-cancave, so we can generate samples from the distribution of $\log(\beta)$ using Adaptive Rejection Sampling techniques and then transform them to values of β.

4 Conditional Posterior of c and α

Generally, the conjugate prior of a multinomial distribution is a Dirichlet distribution:

$$p(\pi_1, ..., \pi_k|\alpha) = \mathrm{Dir}(\alpha_1, ..., \alpha_k) = \frac{\Gamma(\sum_{j=1}^{k} \alpha_j)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} \pi_j^{\alpha_j} d\pi_j. \qquad (4.1)$$

For the symmetric Dirichlet distribution in this model,

$$p(\pi_1, ..., \pi_k|\alpha) = \mathrm{Dir}(\alpha/k, ..., \alpha/k) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \prod_{j=1}^{k} \pi_j^{\alpha/k-1}. \qquad (4.2)$$

Since $\int p(\pi_1, ..., \pi_k|\alpha) d\pi_1 ... d\pi_k = 1$, we have

$$\int \prod_{j=1}^{k} \pi_j^{\alpha_j} d\pi_j = \frac{\prod_{j=1}^{k} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{k} \alpha_j)}. \qquad (4.3)$$

$c_1, ..., c_n$ follow the multinomial distribution:

$$p(c_1, ..., c_n|\pi_1, ..., \pi_k) = \prod_{j=1}^{n} \pi_j^{n_j}, \qquad (4.4)$$

where n_j is the number of points in the j Integrate out $\pi_1, ..., \pi_k$ using the result in Equation (4.2) and Equation (4.4), we get

$$
\begin{aligned}
p(c_1, ..., c_n, |\alpha) &= \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \int \prod_{j=1}^{k} \pi_j^{n_j+\alpha/k-1} d\pi_j & (4.5)\\
&= \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} / \frac{\Gamma(\alpha+n)}{\prod_{j=1}^{k} \Gamma(n_j+\alpha/k)^k} & (4.6)\\
&= \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \prod_{j=1}^{k} \frac{\Gamma(n_j+\alpha/k)}{\Gamma(\alpha/k)} & (4.7)
\end{aligned}
$$

where n_j denote the number of points in the j-th cluster The conditional distribution of $c_1, ..., c_n$ is

$$
\begin{aligned}
p(c_i = j|\mathbf{c}_{-i}, \alpha) &= \frac{p(\mathbf{c}|\alpha)}{p(\mathbf{c}_{-i}|\alpha)} & (4.8)\\
&= \frac{\frac{1}{\Gamma(n+\alpha)}}{\frac{1}{\Gamma(n+\alpha-1)}} \times \frac{\Gamma(n_j+\alpha/k)}{\Gamma(n_{j,-i}+\alpha/k)} & (4.9)\\
&= \frac{1}{n+\alpha-1} \times \frac{n_{j,-i}+\alpha/k}{1} & (4.10)\\
&= \frac{n_{j,-i}+\alpha/k}{n+\alpha-1} & (4.11)
\end{aligned}
$$

where $n_{j,-i}$ denote the number of points in the j-th cluster excluding the i-th point. Since we are considering the case $c_i = j$, the terms in the product of Equation(4.7) are all the same except the the j-th term, this leads to the equality in Equation(4.9). Use the identity $\Gamma(x+1) = x\Gamma(x)$, we get the equality in Equation(4.10).

The results through Equation (4.11) to (4.7) are essential to handle the case of infinite number of clusters. If we can not integrate out π, we will need to sample it. When we have infinite number of clusters, π will be of infinite dimensions, making it difficult to sample. With the results in Equation (4.11), we can work directly with the finite number of samples.

Equation (4.11) is the conditional prior for c_i. Multiply it with the likelihood in Equation (1.1), we can get its conditional posterior distribution:

$$
\begin{aligned}
p(c_i = j|\mathbf{c}_{-i}, \pi, \mathbf{y}, \mu, \mathbf{s}) &= p(c_i = j|\mathbf{c}_{-i}, \alpha, y_i, \mu_j, s_j) & (4.12)\\
&= p(c_i = j|\mathbf{c}_{-i}, \alpha) p(y_i|\mu_j, s_j) & (4.13)\\
&\propto \frac{n_{j,-i}+\alpha/k}{n+\alpha-1} s_j^{1/2} \exp(-s_j(y_i-\mu_j)^2/2) & (4.14)
\end{aligned}
$$

A Derivation of Posterior Distribution of Gaussian Distribution

For a univariate Gaussian distribution, $\mathcal{N}(\mu, \sigma^2)$, given n observation $\mathbf{x} = \{x_i, i = 1, ..., n\}$, the likelihood is

$$L(\mathbf{x}|\mu, \sigma^2) \propto \frac{1}{(\sigma^2)^n} \exp\left(-\frac{\sum_{i=1}^{n/2}(x_i - \mu)^2}{2\sigma^2}\right) \tag{A.1}$$

Its posterior distributions can be categorized into the following cases:

A.1 Known variance σ^2 or known precision τ, unknown mean μ

In this case, we assume a conjugate prior for the parameter μ, which is a Gaussian distribution, $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Since σ^2 and σ_0^2 are known, they are considered as constant and factored out in the terms in front of the $\exp(\cdot)$ function. The posterior can then be written as:

$$p(\mu|\mathbf{x}, \sigma^2, \mu_0, \sigma_0^2) \propto \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right) \times \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right). \tag{A.2}$$

The only unknown parameter is μ, so the terms not involved in μ can be factored out, and the posterior becomes:

$$p(\mu|\mathbf{x}, \sigma^2, \mu_0, \sigma_0^2) \propto \exp\left(-\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) + \mu\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma^2}\right)\right). \tag{A.3}$$

Let

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}, && \text{(A.4)} \\ \tilde{\mu} &= \tilde{\sigma}^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma^2}\right). && \text{(A.5)} \end{aligned}$$

The posterior becomes:

$$\begin{aligned} p(\mu|\mathbf{x}, \sigma^2, \mu_0, \sigma_0^2) &= \exp\left(\frac{\mu^2}{\tilde{\sigma}^2} + \frac{2\mu\tilde{\mu}}{2\tilde{\sigma}^2}\right) && \text{(A.6)} \\ &\propto \exp\left(-\frac{(\mu - \tilde{\mu})^2}{2\tilde{\sigma}^2}\right). && \text{(A.7)} \end{aligned}$$

Thus,

$$p(\mu|\mathbf{x}, \sigma^2, \mu_0, \sigma_0^2) = \mathcal{N}\left(\frac{\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n x_i}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}\right) \tag{A.8}$$

We can also use the precisions $\tau = \frac{1}{\sigma^2}$, $\tau_0 = \frac{1}{\sigma_0^2}$, to replace the variances in the above equations. The problem can be stated as: for a univariate Gaussian with know precision τ, $\mathcal{N}(\mu, \tau)$, whose conjugate prior for the mean μ is a Gaussian $\mathcal{N}(\mu_0, \tau_0)$, its posterior distribution given n observations $\mathbf{x} = \{x_i, i = 1, ..., n\}$ is also a Gaussian:

$$p(\mu|\mathbf{x}, \tau, \mu_0, \tau_0) = \mathcal{N}\left(\frac{\tau_0\mu_0 + \tau\sum_{i=1}^n x_i}{n\tau + \tau_0}, \frac{1}{n\tau + \tau_0}\right). \tag{A.9}$$

A.2 Known mean μ, unknown variance σ^2 or unknown precision τ

In this case, the likelihood can be expressed as:

$$L(\mathbf{x}|\mu, \sigma^2) \propto \frac{1}{(\sigma^2)^{n/2}} \exp\left(-\frac{nS}{2\sigma^2}\right) \tag{A.10}$$

where $S = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2$ is a constant.

we can assume a conjugate prior for σ^2, which is an inverse Gamma distribution

$$p(x|\alpha, \beta) = \mathcal{IG}(\alpha, \beta) \propto x^{-(\alpha-1)} e^{-\beta/x}. \tag{A.11}$$

The posterior becomes:

$$\begin{aligned} p(\sigma^2|\mu, \mathbf{x}, \alpha, \beta) &\propto \frac{1}{(\sigma^2)^{\alpha-1}} \exp(-\frac{\beta}{\sigma^2}) \times \frac{1}{(\sigma^2)^{n/2}} \exp\left(-\frac{nS}{2\sigma^2}\right) && \text{(A.12)} \\ &= \frac{1}{(\sigma^2)^{\alpha-1+n/2}} \exp(-\frac{\beta + nS/2}{\sigma^2}). && \text{(A.13)} \end{aligned}$$

Thus,

$$p(\sigma^2|\mu, \mathbf{x}, \alpha, \beta) = \mathcal{IG}(\alpha - 1 + n/2, \beta + nS/2). \tag{A.14}$$

We can also assume a conjugate scaled inverse-χ^2 prior for σ^2, which is defined as

$$p(x|\nu, \sigma_0^2) = \chi^2_{\mathcal{SI}}(\nu, \sigma_0^2) \propto x^{-(1+\nu/2)} \exp\left(-\frac{\nu\sigma_0^2}{2x}\right). \tag{A.15}$$

The posterior becomes

$$\begin{aligned} p(\sigma^2|\mu,\mathbf{x},\nu,\sigma_0^2) &\propto \frac{1}{(\sigma^2)^{(1+\nu/2)}}\exp\left(-\frac{\nu\sigma_0^2}{2\sigma^2}\right)\frac{1}{(\sigma^2)^{n/2}}\exp\left(-\frac{nS}{2\sigma^2}\right) && \text{(A.16)}\\ &= \frac{1}{(\sigma^2)^{(1+(\nu+n)/2)}}\exp\left(-\frac{\nu\sigma_0^2+nS}{2\sigma^2}\right) && \text{(A.17)}\\ &= \frac{1}{(\sigma^2)^{(1+(\nu+n)/2)}}\exp\left(-\frac{(\nu+n)\frac{\nu\sigma_0^2+nS}{\nu+n}}{2\sigma^2}\right). && \text{(A.18)}\end{aligned}$$

Thus,

$$p(\sigma^2|\mu,\mathbf{x},\nu,\sigma_0^2) = \chi^2_{\mathcal{SI}}(\nu+n,\frac{\nu\sigma_0^2+nS}{\nu+n}). \tag{A.19}$$

If we use the precision τ to replace the variance σ^2 as the parameter in the Gaussian distribution, we can assume a conjugate prior for τ, which is a Gamma distribution

$$p(x|\alpha,\beta) = \mathcal{G}(\alpha,\beta) \propto x^{\alpha-1}e^{-\beta x}. \tag{A.20}$$

The posterior becomes:

$$\begin{aligned} p(\tau|\mu,\mathbf{x},\alpha,\beta) &\propto \tau^{\alpha-1}\exp(-\beta\tau)\times\tau^{n/2}\exp(-nS\tau/2) && \text{(A.21)}\\ &= \tau^{\alpha-1+n/2}\exp(-\tau(\beta+nS/2)). && \text{(A.22)}\end{aligned}$$

Thus,

$$p(\tau|\mu,\mathbf{x},\alpha,\beta) = \mathcal{G}(\alpha+n/2,\beta+nS/2). \tag{A.23}$$

B Derivation of Posterior Distribution of Gamma Distribution

In this report, I use the definition of Gamma distribution as follows:

$$p(x|\alpha,\beta) = \mathcal{G}(\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, \tag{B.1}$$

where α is called as a shape parameter and β a rate parameter.

Given n observations $\mathbf{x} = \{x_i, i = 1, ..., n\}$, the likelihood is:

$$p(\mathbf{x}|\alpha,\beta) = \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n}(\prod_{i=1}^n x_i)^{\alpha-1}e^{-\beta\sum_{i=1}^n x_i}. \tag{B.2}$$

Its posterior distributions can be categorized into the following cases:

B.1 Known shape parameter α, unknown rate parameter β

In this case, the likelihood can be simplified as:

$$p(\mathbf{x}|\alpha,\beta) = \beta^{n\alpha} e^{-\beta \sum_{i=1}^{n} x_i}. \tag{B.3}$$

We can assume a conjugate prior for *beta*, which is a Gamma distribution:

$$p(\beta|\alpha_0,\beta_0) = \mathcal{G}(\alpha_0,\beta_0) \propto \beta^{\alpha_0-1} e^{-\beta_0\beta}, \tag{B.4}$$

The posterior distribution of *beta* can be obtained by multiplying Equation (B.3) with Equation (B.4):

$$\begin{aligned} p(\beta|\mathbf{x},\alpha,\alpha_0,\beta_0) &\propto \beta^{n\alpha} e^{-\beta \sum_{i=1}^{n} x_i} \times \beta^{\alpha_0-1} e^{-\beta_0\beta} && \text{(B.5)} \\ &= \beta^{\alpha_0+n\alpha-1} e^{-\beta(\beta_0+\sum_{i=1}^{n} x_i)}. && \text{(B.6)} \end{aligned}$$

Thus,

$$p(\beta|\mathbf{x},\alpha,\alpha_0,\beta_0) = \mathcal{G}(\alpha_0+n\alpha, \beta_0+\sum_{i=1}^{n} x_i). \tag{B.7}$$