Conjugate Bayesian analysis of common distributions

Zijing Ou
ouzj@mail2.sysu.edu.cn

School of Computer and Engineering, Sun Yat-sen University

Contents

1 Multinomial Dirichlet Conjugacy

Data:

N The number of data items

$\boldsymbol{X}$ The data items $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, and $\boldsymbol{x}_i \triangleq [x_i^{(1)}, \ldots, x_i^{(K)}]^T$

Parameters:

$\boldsymbol{\theta}$ The event probabilities $\theta_1, \ldots, \theta_K$, $\sum\limits_{i=1}^{K} \theta_i = 1$

n Number of trials (positive integer, regard as constant here), $\sum\limits_{j=1}^{K} x_i^{(j)} = n \quad \forall \boldsymbol{x_i} \in \boldsymbol{X}$

Likelihood of Data:

$$p(\boldsymbol{X}|\boldsymbol{\theta}; n) = \prod_{i=1}^{N} \frac{\Gamma(n+1)}{\prod_{j=1}^{K} \Gamma(x_i^{(j)} + 1)} \prod_{j=1}^{K} \theta_j^{x_i^{(j)}}$$

Hyperparameter:

$\boldsymbol{\alpha}$ Concentration parameters of the Dirichlet prior $\alpha_1, \ldots, \alpha_K$

Prior:

$$\text{Dirichlet} \quad p(\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1}$$

Marginal likelihood:

$$p(\boldsymbol{X}) = \frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \left[\prod_{\Gamma(n+1)}^{K} \frac{\Gamma(n+1)}{\prod_{j=1}^{K} \Gamma(x_i^{(j)} + 1)} \right] \frac{\prod_{j=1}^{K} \Gamma(\sum_{i=1}^{N} x_i^{(j)} + \alpha_j)}{\Gamma(Nn + \sum_{j=1}^{K} \alpha_j)}$$

Posterior:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{X}) &= \frac{\Gamma(Nn + \sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\sum_{i=1}^{N} x_i^{(j)} + \alpha_j)} \prod_{j=1}^{K} \theta_j^{\sum_{i=1}^{N} x_i^{(j)} + \alpha_j - 1} \\ &= Dir(\sum_{i=1}^{N} x_i^{(1)} + \alpha_1, \ldots, \sum_{i=1}^{N} x_i^{(K)} + \alpha_K) \end{aligned}$$

Posterior Predictive:

$$p(\boldsymbol{x}|\boldsymbol{X}) = \frac{\Gamma(n+1)}{\prod_{j=1}^{K} \Gamma(x^{(j)} + 1)} \left[\frac{\Gamma(Nn + \sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\sum_{i=1}^{N} x_i^{(j)} + \alpha_j)} \right] \frac{\prod_{j=1}^{K} \Gamma(\sum_{i=1}^{N} x_i^{(j)} + \alpha_j + x^{(j)})}{\Gamma(Nn + \sum_{j=1}^{K} \alpha_j + n)}$$

2 Categorical Dirichlet Conjugacy

Data:

N The number of data items

$\boldsymbol{x}$ The data items $x_1, \ldots, x_N$, and $x_i \in \{1, \ldots, K\}$

Parameters:

$\boldsymbol{\theta}$ The event probabilities $\theta_1, \ldots, \theta_K$, $\sum_{i=1}^{K} \theta_i = 1$

Likelihood of Data:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{K} \theta_j^{\mathbb{1}(x_i=j)}$$

Hyperparameter:

$\boldsymbol{\alpha}$ Concentration parameters of the Dirichlet prior $\alpha_1, \ldots, \alpha_K$

Prior:

$$\text{Dirichlet} \quad p(\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1}$$

Marginal likelihood:

$$p(\boldsymbol{x}) = \frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \frac{\prod_{j=1}^{K} \Gamma(\sum_{i=1}^{N} \mathbb{1}(x_i = j) + \alpha_j)}{\Gamma(N + \sum_{j=1}^{K} \alpha_j)}$$

Posterior:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) &= \frac{\Gamma(N + \sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\sum_{i=1}^{N} \mathbb{1}(x_i = j) + \alpha_j)} \prod_{j=1}^{K} \theta_j^{\sum_{i=1}^{N} \mathbb{1}(x_i=j) + \alpha_j - 1} \\ &= Dir(\sum_{i=1}^{N} \mathbb{1}(x_i = 1) + \alpha_1, \ldots, \sum_{i=1}^{N} \mathbb{1}(x_i = K) + \alpha_K) \end{aligned}$$

Posterior Predictive:

$$p(x|\boldsymbol{x}) = \frac{\Gamma(N + \sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\sum_{i=1}^{N} \mathbb{1}(x_i = j) + \alpha_j)} \frac{\prod_{j=1}^{K} \Gamma(\sum_{i=1}^{N} \mathbb{1}(x_i = j) + \alpha_j + \mathbb{1}(x = j))}{\Gamma(N + \sum_{j=1}^{K} \alpha_j + 1)}$$

3 Bernoulli Beta Conjugacy

Data:

N The number of data items

$\boldsymbol{x}$ The data items $x_1, \ldots, x_N$, and $x_i \in \{0, 1\}$

Parameters:

θ Mean of data

Likelihood of Data:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \theta^{x_i} (1 - \theta)^{1 - x_i}$$

Hyperparameter:

α Parameter of Beta prior

β Parameter of Beta prior

Prior:

$$\text{Beta:} \quad p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

Marginal likelihood:

$$p(\boldsymbol{x}) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+\sum_{i=1}^N x_i)\Gamma(\beta+\sum_{i=1}^N (1-x_i))}{\Gamma(\alpha+\beta+N)}$$

Posterior:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) &= \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+\sum_{i=1}^N x_i)\Gamma(\beta+\sum_{i=1}^N (1-x_i))} \theta^{\alpha+\sum_{i=1}^N x_i - 1}(1-\theta)^{\beta+\sum_{i=1}^N (1-x_i)-1} \\ &= Beta(\theta|\alpha+\sum_{i=1}^N x_i, \beta+\sum_{i=1}^N (1-x_i)) \end{aligned}$$

Posterior Predictive:

$$p(x|\boldsymbol{x}) = (\alpha+\beta+N)\frac{\Gamma(\alpha+\sum_{i=1}^N x_i + x)\Gamma(\beta+\sum_{i=1}^N (1-x_i)+(1-x))}{\Gamma(\alpha+\sum_{i=1}^N x_i)\Gamma(\beta+\sum_{i=1}^N (1-x_i))}$$

4 Binomial Beta Conjugacy

Data:

N The number of data items

$\boldsymbol{x}$ The data items $x_1, \ldots, x_N$

Parameters:

θ Success probability for each trial

n Number of trials (positive integer, regard as constant here), $x_i \in \{0, 1, \ldots, n\}$

Likelihood of Data:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^N \frac{n!}{(n-x_i)!x_i!}\theta^{x_i}(1-\theta)^{n-x_i}$$

Hyperparameter:

α Parameter of Beta prior

β Parameter of Beta prior

Prior:

$$\text{Beta:} \quad p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

Marginal likelihood:

$$p(\boldsymbol{x}) = \prod_{i=1}^{N} \left[\frac{n!}{(n-x_i)!x_i!} \right] \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + \sum_{i=1}^{N} x_i)\Gamma(\beta + \sum_{i=1}^{N}(n - x_i))}{\Gamma(\alpha+\beta+Nn)}$$

Posterior:

$$\begin{aligned} p(\theta|\boldsymbol{x}) &= \frac{\alpha+\beta+Nn}{\Gamma(\alpha + \sum_{i=1}^{N} x_i)\Gamma(\beta + \sum_{i=1}^{N}(n - x_i))} \theta^{\alpha+\sum_{i=1}^{N} x_i - 1}(1-\theta)^{\beta+\sum_{i=1}^{N}(n-x_i)-1} \\ &= Beta(\theta|\alpha + \sum_{i=1}^{N} x_i, \beta + \sum_{i=1}^{N}(n - x_i)) \end{aligned}$$

Posterior Predictive:

$$p(x|\boldsymbol{x}) = \frac{n!}{(n-x)!x!} \frac{\Gamma(\alpha+\beta+Nn)}{\Gamma(\alpha+\beta+Nn+n)} \frac{\Gamma(\alpha + \sum_{i=1}^{N} x_i + x)\Gamma(\beta + \sum_{i=1}^{N}(n - x_i) + (n-x))}{\Gamma(\alpha + \sum_{i=1}^{N} x_i)\Gamma(\beta + \sum_{i=1}^{N}(n - x_i))}$$

5 Poisson Gamma Conjugacy

Data:

N The number of data items

$\boldsymbol{x}$ The data items $x_1, \ldots, x_N$

Parameters:

θ Mean of data

Likelihood of Data:

$$p(\boldsymbol{x}|\theta) = \prod_{i=1}^{N} \frac{\theta^{x_j} e^{-\theta}}{x_i!}$$

Hyperparameter:

α Shape parameter of Gamma prior

β Rata parameter of Gamma prior

Prior:

$$\text{Gamma:} \quad p(\theta|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$$

Marginal likelihood:

$$p(\boldsymbol{x}) = \prod_{i=1}^{N} \left[\frac{1}{x_i!} \right] \frac{\beta^{\alpha}}{\Gamma(\alpha)} \frac{\Gamma(\sum_{i=1}^{N} x_i + \alpha)}{(\beta+1)^{\sum_{i=1}^{N} x_i + \alpha}}$$

Posterior:

$$\begin{aligned} p(\theta|\boldsymbol{x}) &= \frac{(\beta+1)^{\sum_{i=1}^{N} x_i + \alpha}}{\Gamma(\sum_{i=1}^{N} x_i + \alpha)} \theta^{\sum_{i=1}^{N} x_i + \alpha - 1} e^{-(\beta+N)\theta} \\ &= Gamma(\alpha + \sum_{i=1}^{N} x_i, \beta + N) \end{aligned}$$

Posterior Predictive:

$$p(x|\boldsymbol{x}) = \frac{1}{x!} \frac{(\beta+1)^{\sum_{i=1}^{N} x_i + \alpha}}{\Gamma(\sum_{i=1}^{N} x_i + \alpha)} \frac{\Gamma(\sum_{i=1}^{N} + \alpha + 1)}{(\beta + N + 1)^{\sum_{i=1}^{N} x_i + \alpha + 1}}$$

6 Conjugacy for General Exponential Families

Data:

N The number of data items

$\boldsymbol{x}$ The data items $x_1, \ldots, x_N$

Parameters:

η The parameter of general exponential families

Likelihood of Data:

$$p(\boldsymbol{x}|\eta) = \prod_{i=1}^{N} [h(x_i)] \, exp \left\{ \eta^T \sum_{i=1}^{N} T(x_i) - NA(\eta) \right\}$$

Hyperparameter:

τ Parameters of the prior

n_0 Parameters of the prior

Prior:

$$p(\eta|\tau, n_0) = H(\tau, n_0) exp \left\{ \tau^T \eta - n_0 A(\eta) \right\}$$

Marginal likelihood:

$$p(\boldsymbol{x}) = \frac{\prod_{i=1}^{N} [h(x_i)] \, H(\tau, n_0)}{H(\tau + \sum_{i=1}^{N} T(x_i), n_0 + N)}$$

Posterior:

$$\begin{aligned} p(\eta|\boldsymbol{x}) &= H(\tau + \sum_{i=1}^{N} T(x_i), n_0 + N) exp \left\{ \eta^T (\tau + \sum_{i=1}^{N} T(x_i)) - (N + n_0) A(\eta) \right\} \\ &= p(\eta|\tau + \sum_{i=1}^{N} T(x_i), n_0 + N) \end{aligned}$$

Posterior Predictive:

$$p(x|\boldsymbol{x}) = \frac{h(x) H(\tau + \sum_{i=1}^{N} T(x_i), n_0 + N)}{H(T(x) + \tau + \sum_{i=1}^{N} T(x_i), n_o + N + 1)}$$

7 Normal Normal-Mean Conjugacy

Setting:

Univariate Gaussian with unknown mean μ and known variance σ^2.

Data:

n The number of data items

$\boldsymbol{x}$ The data items $x_1, \ldots, x_n$

$\overline{x}$ $= \frac{\sum_{i=1}^{N} x_i}{n}$

Parameters:

μ Mean of data

σ^2 Variance of data

Likelihood of Data:

$$p(\boldsymbol{x}|\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

Hyperparameter:

μ_0 Mean of μ is μ_0

σ_0^2 Variance of μ is σ_0^2

Prior:

$$\text{Normal} \quad p(\mu|\mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi}\sigma_0} exp\left\{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\}$$

Marginal likelihood:

$$p(\boldsymbol{x}) = \frac{exp\left(-\frac{1}{2\sigma^2}n\overline{x} - \frac{1}{2\sigma_0^2}\mu_0^2\right)}{(\sigma\sqrt{2\pi})^n(\sigma_0\sqrt{2\pi})} exp\left\{\frac{(\frac{1}{\sigma^2}n\overline{x} + \frac{1}{\sigma_0^2}\mu_0)^2}{2(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2})}\right\} \frac{\sqrt{2\pi}}{\sqrt{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}}$$

Posterior:

$$\begin{aligned} p(\mu|\boldsymbol{x}) &= \mathcal{N}(\mu|\frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\overline{x}, \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}) \\ &\triangleq \mathcal{N}(\mu|\mu_n, \sigma_n^2) \end{aligned}$$

Posterior Predictive:

$$p(x|\boldsymbol{x}) = \mathcal{N}(x|\mu_n, \sigma_n^2 + \sigma^2)$$

8 Normal Normal-Gamma Conjugacy

Setting:

Univariate Gaussian with unknown mean μ and unknown precision $\lambda = \sigma^{-2}$.

Data:

n The number of data items

$\boldsymbol{x}$ The data items $x_1, \ldots, x_n$

$\overline{x}$ $= \frac{\sum_{i=1}^{N} x_i}{n}$

Parameters:

μ Mean of data

$\lambda = \sigma^{-2}$ Precision (inverse variance) of data

Likelihood of Data:

$$p(\boldsymbol{x}|\mu,\lambda) = (\frac{\lambda}{2\pi})^{\frac{n}{2}} exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

Hyperparameter:

μ_0 Mean of μ is μ_0

κ_0 Parameter of precision of μ

α_0 Shape parameter of Gamma prior of λ

β_0 Rate parameter of Gamma prior of λ

Prior:

The conjugate prior is normal-Gamma distribution $NG(\mu,\lambda|\mu_0,\kappa_0,\alpha_0,\beta_0)$.

$$\begin{aligned} p(\mu,\lambda|\mu_0,\kappa_0,\alpha_0,\beta_0) &= \mathcal{N}(\mu|\mu_0,(\kappa_0\lambda)^{-1})Ga(\lambda|\alpha_0,\beta_0) \\ &= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\frac{\kappa_0}{2\pi})^{\frac{1}{2}}\lambda^{\alpha_0-\frac{1}{2}} exp\left\{-\frac{\lambda}{2}[\kappa_0(\mu-\mu_0)^2+2\beta_0]\right\} \end{aligned}$$

Marginal likelihood:

$$p(\boldsymbol{x}) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\frac{\kappa_0}{n+\kappa_0})^{\frac{1}{2}}\frac{\Gamma(\alpha_0+\frac{n}{2})}{\left[\beta_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2+\frac{n\kappa_0(\overline{x}-\mu_0)^2}{2(n+\kappa_0)}\right]^{\alpha_0+\frac{n}{2}}}(2\pi)^{-\frac{n}{2}}$$

Posterior:

$$p(\mu,\lambda|\boldsymbol{x}) = NG(\mu,\lambda|\mu_n,\kappa_n,\alpha_n,\beta_n)$$

where

$$\begin{aligned} \mu_n &= \frac{n\overline{x}+\kappa_0\mu_0}{n+\kappa_0} \\ \kappa_n &= n+\kappa_0 \\ \alpha_n &= \alpha_0+\frac{n}{2} \\ \beta_n &= \beta_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2+\frac{n\kappa_0(\overline{x}-\mu_0)^2}{2(n+\kappa_0)} \end{aligned}$$

Posterior Predictive:

Denote m new observations as $\boldsymbol{x}_m = \{x_{n+1},\ldots,x_{n+m}\}$, then

$$p(\boldsymbol{x}_m|\boldsymbol{x}) = \frac{\Gamma(\alpha_{n+m})}{\Gamma(\alpha_n)}\frac{\beta_n^{\alpha_n}}{\beta_{n+m}^{\alpha_{n+m}}}\left(\frac{\kappa_n}{\kappa_{n+m}}\right)^{\frac{1}{2}}(2\pi)^{-\frac{m}{2}}$$

9 Normal Gamma-Precision Conjugacy

Setting:

Univariate Gaussian with known mean μ and unknown precision $\lambda = \sigma^{-2}$.

Data:

n The number of data items

$\boldsymbol{x}$ The data items $x_1, \ldots, x_n$

$\overline{x}$ $= \frac{\sum_{i=1}^{N} x_i}{n}$

Parameters:

μ Mean of data

λ $= \sigma^{-2}$ Precision (inverse variance) of data

Likelihood of Data:

$$p(\boldsymbol{x}|\mu, \lambda) = (\frac{\lambda}{2\pi})^{\frac{n}{2}} exp \left\{ -\frac{\lambda}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

Hyperparameter:

α Shape parameter of Gamma prior of λ

β Rate parameter of Gamma prior of λ

Prior:

$$\text{Gamma:} \quad p(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

Marginal likelihood:

$$p(\boldsymbol{x}) = (\frac{1}{2\pi})^{n/2} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha + \frac{n}{2})}{\left[\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 + 2\beta\right]^{\alpha + \frac{n}{2}}}$$

Posterior:

$$\begin{aligned} p(\lambda|\boldsymbol{x}) &= Ga(\lambda|\alpha_n, \beta_n) \\ \alpha_n &= \alpha + \frac{n}{2} \\ \beta_n &= \beta + \frac{1}{2} \sum_{i=1}^{n} (x_i - n)^2 \end{aligned}$$

Posterior Predictive:

$$\begin{aligned} p(x|\boldsymbol{x}) &= \frac{\Gamma((2\alpha_n + 1)/2)}{\Gamma((2\alpha_n)/2)} \left(\frac{\alpha_n}{2\pi\alpha_n\beta_n} \right)^{\frac{1}{2}} \left(1 + \frac{\alpha_n (x - \mu)^2}{2\alpha_n\beta_n} \right)^{-(2\alpha_n+1)/2} \\ &= t_{2\alpha_n}(x|\mu, \sigma^2 = \frac{\beta_n}{\alpha_n}) \end{aligned}$$

10 Normal Normal-inverse-chi-square (NIX) Conjugacy

Setting:

Univariate Gaussian with unknown mean μ and unknown variance σ^2.

Data:

n The number of data items

$\boldsymbol{x}$ The data items $x_1, \ldots, x_n$

$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{n}$ empirical mean of data

Parameters:

μ Mean of data

σ^2 Variance of data

Likelihood of Data:

$$\begin{aligned} p(\boldsymbol{x}|\mu, \sigma^2) &= (\frac{1}{2\pi\sigma^2})^{\frac{n}{2}} exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} \\ &= \frac{1}{(2\pi)^{n/2}} (\sigma^2)^{-n/2} exp \left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} (x_i - \overline{x})^2 + n(\overline{x} - \mu)^2 \right] \right\} \end{aligned}$$

Hyperparameter:

μ_0 Mean of μ

κ_0 Parameter of the variance of μ

v_0 Degree of freedom of σ^2

σ_0^2 Scale parameter of σ^2

Prior:

The conjugate prior is the Normal (scale) inverse chi-square distribution $NI\chi^2(\mu, \sigma^2|\mu_0, \kappa_0, v_0, \sigma_0^2)$

$$\begin{aligned} p(\mu, \sigma^2) &= \mathcal{N}(\mu|\mu_0, \sigma^2/\kappa_0)\chi^{-2}(\sigma^2|v_0, \sigma_0^2) \\ &= \frac{\sqrt{\kappa_0}}{\sqrt{2\pi}} \frac{1}{\Gamma(v_0/2)} \left(\frac{v_0\sigma_0^2}{2} \right)^{v_0/2} \sigma^{-1} (\sigma^2)^{-(\frac{v_0}{2}+1)} exp \left\{ -\frac{1}{2\sigma^2} \left[v_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2 \right] \right\} \end{aligned}$$

Marginal likelihood:

$$p(\boldsymbol{x}) = \frac{\Gamma(v_n/2)}{\Gamma(v_0/2)} \sqrt{\frac{k_0}{k_n}} \frac{(v_0\sigma_0^2)^{v_0/2}}{(v_n\sigma_n^2)^{v_n/2}} \frac{1}{\pi^{n/2}}$$

where

$$\begin{aligned} \mu_n &= \frac{\kappa_0\mu_0 + n\overline{x}}{\kappa_n} \\ \kappa_n &= \kappa_0 + n \\ v_n &= v_0 + n \\ \sigma_n^2 &= \frac{1}{v_0 + n} \left(v_0\sigma_0^2 + \sum_{i=1}^{n} (x_i - \overline{x})^2 + \frac{n\kappa_0}{\kappa_0 + n} (\mu_0 - \overline{x})^2 \right) \end{aligned}$$

Posterior:

$$p(\mu, \sigma^2) = NI\chi^2(\mu, \sigma^2|\mu_n, \kappa_n, v_n, \sigma_n^2)$$

Posterior Predictive:

$$p(x|\boldsymbol{x}) = \frac{\Gamma((v_n+1)/2)}{\Gamma(v_n/2)} \left(\frac{\kappa_n}{(\kappa_n+1)\pi v_n \sigma_n^2}\right)^{\frac{1}{2}} \left(1 + \frac{\kappa_n(x-\mu_n)^2}{(\kappa_n+1)v_n\sigma_n^2}\right)^{-(v_n+1)/2}$$
$$= t_{v_n}(x|\mu_n, \frac{(1+\kappa_n)\sigma_n^2}{\kappa_n})$$

11 Normal Normal-inverse-Gamma Conjugacy

Setting:

Univariate Gaussian with unknown mean μ and unknown variance σ^2.

Data:

n The number of data items

$\boldsymbol{x}$ The data items $x_1, \ldots, x_n$

$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{n}$ empirical mean of data

Parameters:

μ Mean of data

σ^2 Variance of data

Likelihood of Data:

$$p(\boldsymbol{x}|\mu, \sigma^2) = (\frac{1}{2\pi\sigma^2})^{\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$
$$= \frac{1}{(2\pi)^{n/2}}(\sigma^2)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\mu)^2\right]\right\}$$

Hyperparameter:

m_0 Mean of μ

V_0 Parameter of the variance of μ

α_0 Shape parameter of σ^2

b_0 Scale parameter of σ^2

Prior:

The conjugate prior is the Normal-inverse-Gamma distribution $NIG(\mu, \sigma^2|m_0, V_0, \alpha_0, b_0)$

$$p(\mu, \sigma^2) = \mathcal{N}(\mu|\mu_0, \sigma^2 V_0) IG(\sigma^2|\alpha_0, b_0)$$
$$= \frac{1}{\sqrt{2\pi V_0}} \frac{b_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{1}{\sigma} (\sigma^2)^{-\alpha_0-1} exp\left(-\frac{1}{2\sigma^2}[V_0^{-1}(\mu-\mu_0)^2 + 2b_0]\right)$$

This is equivalent to the $NI\chi^2$ prior, where we make the following substitutions:

$$m_0 = \mu_0$$
$$V_0 = \frac{1}{\kappa_0}$$
$$\alpha_0 = \frac{v_0}{2}$$
$$b_0 = \frac{v_0\sigma_0^2}{2}$$

Marginal likelihood:

$$p(\boldsymbol{x}) = \frac{\Gamma(\alpha_n)}{\Gamma(\alpha_0)} \sqrt{\frac{V_n}{V_0}} \frac{b_0^{\alpha_0}}{b_n^{\alpha_n}} \frac{1}{(2\pi)^{n/2}}$$

where

$$\begin{aligned} m_n &= \frac{V_0^{-1} m_0 + n\overline{x}}{V_0^{-1} + n} \\ V_n^{-1} &= V_0^{-1} + n \\ \alpha_n &= \alpha_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2} \sum_{i=1}^{n} (x_i - \overline{x})^2 + \frac{V_0^{-1} n}{2(V_0^{-1} + n)} (m_0 - \overline{x})^2 \end{aligned}$$

Actually, the last term can be further expressed as

$$b_n = b_0 + \frac{1}{2} \left[m_0^2 V_0^{-1} + \sum_{i=1}^{n} x_i^2 - m_n^2 V_n^{-1} \right],$$

which is more common, but its derivation requires some tedious algebra (see this link for derivation).

Posterior:

$$p(\mu, \sigma^2 | \boldsymbol{x}) = NIG(\mu, \sigma^2 | m_n, V_n, \alpha_n, b_n)$$

Posterior Predictive:

$$\begin{aligned} p(x|\boldsymbol{x}) &= t_{2\alpha_n}(x | m_n, \frac{b_n(1+V_n)}{\alpha_n}) \\ &= \frac{\Gamma((2\alpha_n + 1)/2)}{\Gamma(2\alpha_n/2)} \left(\frac{\alpha_n}{\pi 2\alpha_n b_n (1+V_n)} \right)^{1/2} \left(1 + \frac{1}{2\alpha_n} \frac{\alpha_n (x - m_n)^2}{b_n (1 + V_n)} \right)^{-\frac{2\alpha_n + 1}{2}} \end{aligned}$$

12 Multivariate Normal Normal-Mean Conjugacy

Setting:

Multivariate Gaussian with unknown mean μ and known variance Σ.

Data:

N The number of data items

$\boldsymbol{X}$ The data items $x_1, \ldots, x_n$, $x_i \in \mathbb{R}^d$

$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ empirical mean of data

Parameters:

μ Mean of data

Σ Variance of data

Likelihood of Data:

$$p(X|\mu) = (2\pi)^{-\frac{d}{2}N} |\Sigma|^{-\frac{N}{2}} exp\left(-\frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right)$$

Hyperparameter:

μ_0 Mean of μ

Σ_0 Parameter of the variance of μ

Prior:

$$p(\mu|\mu_0, \Sigma_0) = (2\pi)^{-d/2}|\Sigma_0|^{-1/2}exp\left(-\frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)\right)$$

Marginal likelihood:

$$p(X) = (2\pi)^{-\frac{d}{2}N}\left(\frac{|\Sigma_N|}{|\Sigma_0||\Sigma|^N}\right)^{\frac{1}{2}} exp\left(-\frac{1}{2}[\mu_N^T\mu_N + \sum_{i=1}^{N} x_i^T\Sigma^{-1}x_i + \mu_0^T\Sigma_o^{-1}\mu_0]\right)$$

where

$$\mu_N = (\Sigma_0^{-1} + N\Sigma^{-1})^{-1}(N\Sigma^{-1}\overline{x} + \Sigma_0^{-1}\mu_0)$$
$$\Sigma_N = (\Sigma_0^{-1} + N\Sigma^{-1})^{-1}$$

Posterior:

$$p(\mu|X) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$$

Posterior Predictive:

$$p(x|X) = \mathcal{N}(x|\mu_N, \Sigma + \Sigma_N)$$

13 Multivariate Normal Wishart-Precision Conjugacy

Setting:

Multivariate Gaussian with known mean μ and unknown precision $\Lambda = \Sigma^{-1}$.

Data:

N The number of data items

X The data items $x_1, \ldots, x_n$, $x_i \in \mathbb{R}^d$

$\overline{x} = \frac{\sum_{i=1}^N x_i}{N}$ empirical mean of data

Parameters:

μ Mean of data

Λ Precision (inverse variance) of data

Likelihood of Data:

$$\begin{aligned}
p(X|\Lambda) &= (2\pi)^{-\frac{d}{2}N}|\Lambda|^{\frac{N}{2}} exp\left(-\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Lambda(x_i-\mu)\right) \\
&= (2\pi)^{-\frac{d}{2}N}|\Lambda|^{\frac{N}{2}} exp\left(-\frac{1}{2}tr[\Lambda \underbrace{\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T}_{S}]\right) \\
&= (2\pi)^{-\frac{d}{2}N}|\Lambda|^{\frac{N}{2}} exp\left(-\frac{1}{2}tr[\Lambda S]\right)
\end{aligned}$$

where we use the cyclic property of the trace operator and scalar $= tr$(scalar).

Hyperparameter:

v_0 Degree of freedom of the Wishart prior of μ

T_0 Scale matrix of the Wishart prior of μ

Prior:

$$\begin{aligned} p(\Lambda) &= Wi_{v_0}(\Lambda|T_0) \\ &= \frac{1}{Z_0}|\Lambda|^{(v_0-d-1)/2} exp\left\{-\frac{1}{2}tr(T_0^{-1}\Lambda)\right\} \\ Z_0 &= 2^{v_0 d/2}\Gamma_d(v_0/2)|T_0|^{v_0/2} \end{aligned}$$

Marginal likelihood:

$$\begin{aligned} p(X) &= (2\pi)^{-\frac{d}{2}N}\frac{Z_N}{Z_0} \\ Z_N &= 2^{v_N d/2}\Gamma_d(v_N/2)|T_N|^{v_N/2} \\ T_N &= (T_0^{-1}+S)^{-1} \\ v_N &= N + v_0 \\ S &= \sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T \end{aligned}$$

Posterior:

$$p(\Lambda|X) = Wi_{v_N}(\Lambda|T_N)$$

Posterior Predictive:

$$p(x|X) = t_{v_N-d+1}(x|\mu, \frac{1}{v_N-d+1}T_N^{-1})$$

14 Multivariate Normal Normal-Wishart Conjugacy

Setting:

Multivariate Gaussian with unknown mean μ and unknown precision $\Lambda = \Sigma^{-1}$.

Data:

N The number of data items

X The data items $x_1, \ldots, x_N$, $x_i \in \mathbb{R}^d$

$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{N}$ empirical mean of data

Parameters:

μ Mean of data

Λ Precision (inverse variance) of data

Likelihood of Data:

$$p(X|\Lambda) = (2\pi)^{-\frac{d}{2}N}|\Lambda|^{\frac{N}{2}} exp\left(-\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Lambda(x_i-\mu)\right)$$

Hyperparameter:

μ_0 Mean of Normal-Wishart prior

κ Scale factor of Normal-Wishart prior

v Degree of freedom of Normal-Wishart prior

T Scale matrix of Normal-Wishart prior

Prior:

$$\begin{aligned} p(\mu,\Lambda) &= NWi(\mu,\Lambda|\mu_0,\kappa,v,T) = \mathcal{N}(\mu|\mu_0,(\kappa\Lambda)^{-1})Wi_v(\Lambda|T) \\ &= \frac{1}{Z}|\Lambda|^{\frac{1}{2}}exp\left(-\frac{\kappa}{2}(\mu-\mu_0)^T\Lambda(\mu-\mu_0)\right)|\Lambda|^{(\kappa-d-1)/2}exp(-\frac{1}{2}tr(T^{-1}\Lambda)) \\ Z &= \left(\frac{2\pi}{\kappa}\right)^{\frac{d}{2}} 2^{\frac{vd}{2}}|T|^{\frac{v}{2}}\Gamma_d(\frac{v}{2}) \end{aligned}$$

Posterior:

$$\begin{aligned} p(\mu,\Lambda|X) &= NWi(\mu,\Lambda|\mu_N,\kappa_N,v_N,T_N) \\ &= \mathcal{N}(\mu|\mu_N,(\kappa_N\Lambda)^{-1})Wi_{v_N}(\Lambda|T_N) \\ \mu_N &= \frac{\kappa\mu_0+N\overline{x}}{N+\kappa} \\ \kappa_N &= \kappa+N \\ v_N &= v+N \\ T_N &= \left(T^{-1}+S+\frac{N\kappa}{N+\kappa}(\overline{x}-\mu_0)(\overline{x}-\mu_0)^T\right)^{-1} \\ S &= \sum_{i=1}^{N}(x_i-\overline{x})(x_i-\overline{x})^T \end{aligned}$$

Marginal likelihood:

$$p(X) = \frac{1}{(\pi)^{Nd/2}}\frac{\Gamma_d^{v_N/2}}{\Gamma_d^{v_0/2}}\frac{\Gamma_d^{v_N/2}}{\Gamma_d^{v_0/2}}\left(\frac{\kappa_0}{\kappa_N}\right)^{d/2}$$

Posterior Predictive:

$$p(x|X) = t_{v_N-d+1}(\mu_N,\frac{(\kappa_N+1)}{\kappa_N(v_N-d+1)}T_N^{-1})$$

15 Multivariate Normal Normal-Inverse-Wishart Conjugacy

Setting:

Multivariate Gaussian with unknown mean μ and unknown variance Σ^2.

Data:

N The number of data items

X The data items $x_1,\ldots,x_N$, $x_i \in \mathbb{R}^d$

$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{N}$ empirical mean of data

Parameters:

μ Mean of data

Σ Variance of data

Likelihood of Data:

$$p(X|\Lambda) = (2\pi)^{-\frac{d}{2}N}|\Sigma|^{-\frac{N}{2}} exp\left(-\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$$

Hyperparameter:

μ_0 Mean of Normal-inverse-Wishart prior

κ_0 Scale factor of Normal-inverse-Wishart prior

v_0 Degree of freedom of Normal-inverse-Wishart prior

Λ_0 Scale matrix of Normal-inverse-Wishart prior

Prior:

$$\begin{aligned}
p(\mu,\Lambda) &= NIW(\mu,\Lambda|\mu_0,\kappa_0,v_0,\Lambda_0) = \mathcal{N}(\mu|\mu_0,\frac{1}{\kappa_0}\Sigma)IW_{v_0}(\Sigma|\Lambda_0) \\
&= (2\pi)^{-\frac{d}{2}}|\frac{1}{\kappa_0}\Sigma|^{-\frac{1}{2}} exp\left(-\frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right) \\
&\quad \frac{|\Lambda_0|^{\frac{v_0}{2}}}{2^{v_0 d/2}\Gamma_d(\frac{v_0}{2})}|\Sigma|^{-(v_0+d+1)/2} exp\left(-\frac{1}{2}tr(\Lambda_0\Sigma^{-1})\right) \\
&= \frac{1}{Z}|\Sigma|^{-(\frac{v_0+d}{2}+1)} exp\left(-\frac{1}{2}\Big[tr(\Lambda_0\Sigma^{-1})+\kappa_0(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\Big]\right) \\
Z &= \frac{2^{v_0 d/2}\Gamma_d^{\frac{v_0}{2}}(2\pi/\kappa_0)^{d/2}}{|\Lambda_0|^{v_0/2}}
\end{aligned}$$

Posterior:

$$\begin{aligned}
p(\mu,\Sigma|X) &= NIW(\mu,\Sigma|\mu_N,\kappa_N,v_N,\Lambda_N) \\
&= \mathcal{N}(\mu|\mu_N,\frac{1}{\kappa_N}\Sigma)IW_{v_N}(\Sigma|\Lambda_N) \\
\mu_N &= \frac{\kappa_0\mu_0+N\overline{x}}{N+\kappa_0} \\
\kappa_N &= \kappa_0+N \\
v_N &= v_0+N \\
\Lambda_N &= \Lambda_0+S+\frac{\kappa_0 N}{\kappa_0+N}(\overline{x}-\mu_0)(\overline{x}-\mu_0)^T \\
S &= \sum_{i=1}^{N}(x_i-\overline{x})(x_i-\overline{x})^T
\end{aligned}$$

Marginal likelihood:

$$p(X) = \frac{1}{\pi^{Nd/2}}\frac{\Gamma_d(v_N/2)}{\Gamma_d(v_0/2)}\frac{|\Lambda_0|^{v_0/2}}{|\Lambda_N|^{v_N/2}}\left(\frac{\kappa_0}{\kappa_N}\right)^{d/2}$$

Posterior Predictive:

$$p(x|X) = t_{v_N-d+1}(\mu_N,\frac{\kappa_N+1}{v_N-d+1}\Lambda_N)$$

A Proof of Multinomial Dirichlet Conjugacy

Marginal likelihood:

$$
\begin{aligned}
p(\boldsymbol{X}) &= \int p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \int \left[\prod_{i=1}^{N} \frac{\Gamma(n+1)}{\prod_{j=1}^{K}\Gamma(x_i^{(j)}+1)} \prod_{j=1}^{K} \theta_j^{x_i^{(j)}} \right] \frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \prod_{j=1}^{K} \theta_j^{\alpha_j-1} d\boldsymbol{\theta} \\
&= \frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \left[\prod_{i=1}^{N} \frac{\Gamma(n+1)}{\prod_{j=1}^{K}\Gamma(x_i^{(j)}+1)} \right] \int \prod_{j=1}^{K} \theta_j^{\sum_{i=1}^{N} x_i^{(j)}\alpha_j-1} d\boldsymbol{\theta} \\
&= \frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \left[\prod_{i=1}^{N} \frac{\Gamma(n+1)}{\prod_{j=1}^{K}\Gamma(x_i^{(j)}+1)} \right] \frac{\prod_{j=1}^{K}\Gamma(\sum_{i=1}^{N} x_i^{(j)}+\alpha_j)}{\Gamma(Nn+\sum_{j=1}^{K}\alpha_j)},
\end{aligned}
$$

where we use the equality $\int \sum_{j=1}^{K} \theta_j^{\alpha_j-1} d\boldsymbol{\theta} = \frac{\sum_{j=1}^{K}\Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{K}\alpha_j)}$, since $\int Dir(\boldsymbol{\theta}|\alpha_1,\ldots,\theta_K)d\boldsymbol{\theta} = 1$.

Posterior:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{X}) &= \frac{p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{X})} \\
&= \frac{\left[\prod_{i=1}^{N} \frac{\Gamma(n+1)}{\prod_{j=1}^{K}\Gamma(x_i^{(j)}+1)} \prod_{j=1}^{K} \theta_j^{x_i^{(j)}} \right] \frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \prod_{j=1}^{K} \theta_j^{\alpha_j-1}}{\frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \left[\prod_{i=1}^{N} \frac{\Gamma(n+1)}{\prod_{j=1}^{K}\Gamma(x_i^{(j)}+1)} \right] \frac{\prod_{j=1}^{K}\Gamma(\sum_{i=1}^{N} x_i^{(j)}+\alpha_j)}{\Gamma(Nn+\sum_{j=1}^{K}\alpha_j)}} \\
&= \frac{\Gamma(Nn+\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\sum_{i=1}^{N} x_i^{(j)}+\alpha_j)} \prod_{j=1}^{K} \theta_j^{\sum_{i=1}^{N} x_i^{(j)}+\alpha_j-1} \\
&= Dir(\boldsymbol{\theta}|\sum_{i=1}^{N} x_i^{(1)}+\alpha_1,\ldots,\sum_{i=1}^{N} x_i^{(K)}+\alpha_K)
\end{aligned}
$$

Posterior Predictive:

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{X}) &= \int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{X})d\boldsymbol{\theta} \\
&= \int \frac{\Gamma(n+1)}{\prod_{j=1}^{K}\Gamma(x^{(j)}+1)} \prod_{j=1}^{K} \theta_j^{x^{(j)}} \frac{\Gamma(Nn+\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\sum_{i=1}^{N} x_i^{(j)}+\alpha_j)} \prod_{j=1}^{K} \theta_j^{\sum_{i=1}^{N} x_i^{(j)}+\alpha_j-1} d\boldsymbol{\theta} \\
&= \frac{\Gamma(n+1)}{\prod_{j=1}^{K}\Gamma(x^{(j)}+1)} \frac{\Gamma(Nn+\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\sum_{i=1}^{N} x_i^{(j)}+\alpha_j)} \int \prod_{j=1}^{K} \theta_j^{\sum_{i=1}^{N} x_i^{(j)}+\alpha_j+x^{(j)}-1} d\boldsymbol{\theta} \\
&= \frac{\Gamma(n+1)}{\prod_{j=1}^{K}\Gamma(x^{(j)}+1)} \frac{\Gamma(Nn+\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\sum_{i=1}^{N} x_i^{(j)}+\alpha_j)} \frac{\prod_{j=1}^{K}\Gamma(\sum_{i=1}^{N} x_i^{(j)}+\alpha_j+x^{(j)})}{\Gamma(Nn+\sum_{j=1}^{K}\alpha_j+n)}
\end{aligned}
$$

B Proof of Categorical Dirichlet Conjugacy

Marginal likelihood:

$$\begin{aligned}
p(\boldsymbol{x}) &= \int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \int \prod_{i=1}^{N}\prod_{j=1}^{K}\theta_j^{\mathbb{1}(x_i=j)} \frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \prod_{j=1}^{K}\theta_j^{\alpha_j-1}d\boldsymbol{\theta} \\
&= \frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \int \prod_{j=1}^{K}\theta_j^{\sum_{i=1}^{N}\mathbb{1}(x_i=j)+\alpha_j-1}d\boldsymbol{\theta} \\
&= \frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \frac{\prod_{j=1}^{K}\Gamma(\sum_{i=1}^{N}\mathbb{1}(x_i=j)+\alpha_j)}{\Gamma(N+\sum_{j=1}^{K}\alpha_k)}
\end{aligned}$$

Posterior:

$$\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\
&\propto \left[\prod_{i=1}^{N}\prod_{j=1}^{K}\theta_j^{\mathbb{1}(x_i=j)}\right]\prod_{j=1}^{K}\theta_j^{\alpha_j-1} \\
&= \prod_{j=1}^{K}\theta_j^{\sum_{i=1}^{N}\mathbb{1}(x_i=j)+\alpha_j-1}
\end{aligned}$$

It turns out that Category and Dirichlet distributions are conjugate. Therefore

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = Dir(\boldsymbol{\theta}|\sum_{i=1}^{N}\mathbb{1}(x_i=1)+\alpha_1,\ldots,\sum_{i=1}^{N}\mathbb{1}(x_i=K)+\alpha_K)$$

Posterior Predictive:

$$\begin{aligned}
p(x|\boldsymbol{x}) &= \int p(x|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta} \\
&= \int \prod_{j=1}^{K}\theta_j^{\mathbb{1}(x=j)} \frac{\Gamma(N+\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\sum_{i=1}^{N}\mathbb{1}(x_i=j)+\alpha_j)} \prod_{j=1}^{K}\theta_j^{\sum_{i=1}^{N}\mathbb{1}(x_i=j)+\alpha_j-1}d\boldsymbol{\theta} \\
&= \frac{\Gamma(N+\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\sum_{i=1}^{N}\mathbb{1}(x_i=j)+\alpha_j)} \int \theta_j^{\mathbb{1}(x=j)+\sum_{i=1}^{N}\mathbb{1}(x_i=j)+\alpha_j-1}d\boldsymbol{\theta} \\
&= \frac{\Gamma(N+\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\sum_{i=1}^{N}\mathbb{1}(x_i=j)+\alpha_j)} \frac{\prod_{j=1}^{K}\Gamma(\mathbb{1}(x=j)+\sum_{i=1}^{N}\mathbb{1}(x_i=j)+\alpha_j)}{\Gamma(N+1+\sum_{j=1}^{N}\alpha_j)}
\end{aligned}$$

C Proof of Bernoulli Beta Conjugacy

Marginal likelihood:

$$\begin{aligned}
p(\boldsymbol{x}) &= \int p(\boldsymbol{x}|\theta)p(\theta)d\theta \\
&= \int \prod_{i=1}^{N}\theta^{x_i}(1-\theta)^{1-x_i}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int \theta^{\alpha+\sum_{i=1}^{N}x_i-1}(1-\theta)^{\beta+\sum_{i=1}^{N}(1-x_i)-1}d\theta \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+\sum_{i=1}^{N}x_i)\Gamma(\beta+\sum_{i=1}^{N}(1-x_i))}{\Gamma(\alpha+\beta+N)}
\end{aligned}$$

Posterior:

$$\begin{aligned}
p(\theta|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\theta)p(\theta) \\
&\propto \left[\prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}\right]\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{\alpha+\sum_{i=1}^{N} x_i-1}(1-\theta)^{\beta+\sum_{i=1}^{N}(1-x_i)-1}
\end{aligned}$$

Since Bernoulli and Beta are conjugate, we have

$$p(\theta|\boldsymbol{x}) = Beta(\theta|\alpha+\sum_{i=1}^{N} x_i, \beta+\sum_{i=1}^{N}(1-x_i))$$

Posterior Predictive:

$$\begin{aligned}
p(x|\boldsymbol{x}) &= \int p(x|\theta)p(\theta|\boldsymbol{x})d\theta \\
&= \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+\sum_{i=1}^{N} x_i)\Gamma(\beta+\sum_{i=1}^{N}(1-x_i))}\int \theta^{\alpha+x+\sum_{i=1}^{N} x_i-1}(1-\theta)^{\beta+(1-x_i)+\sum_{i=1}^{N}(1-x_i-1)}d\theta \\
&= \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+\sum_{i=1}^{N} x_i)\Gamma(\beta+\sum_{i=1}^{N}(1-x_i))}\frac{\Gamma(\alpha+x+\sum_{i=1}^{N} x_i)\Gamma(\beta+(1-x)+\sum_{i=1}^{N}(1-x_i))}{\Gamma(\alpha+\beta+N+1)}
\end{aligned}$$

D Proof of Binomial Beta Conjugacy

Marginal likelihood:

$$\begin{aligned}
p(\boldsymbol{x}) &= \int p(\boldsymbol{x}|\theta)p(\theta)d\theta \\
&= \prod_{i=1}^{N}\left[\frac{n!}{(n-x_i)!x_i!}\right]\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int \theta^{\alpha+\sum_{i=1}^{N} x_i-1}(1-\theta)^{\beta+\sum_{i=1}^{N}(n-x_i)-1}d\theta \\
&= \prod_{i=1}^{N}\left[\frac{n!}{(n-x_i)!x_i!}\right]\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+\sum_{i=1}^{N} x_i)\Gamma(\beta+\sum_{i=1}^{N}(n-x_i))}{\Gamma(\alpha+\beta+Nn)}
\end{aligned}$$

Posterior:

$$\begin{aligned}
p(\theta|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\theta)p(\theta) \\
&\propto \left[\prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{n-x_i}\right]\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{\alpha+\sum_{i=1}^{N} x_i-1}(1-\theta)^{\beta+\sum_{i=1}^{N}(n-x_i)-1}
\end{aligned}$$

Since the Binomial and Beta are conjugate, we have

$$p(\theta|\boldsymbol{x}) = Beta(\theta|\alpha+\sum_{i=1}^{N} x_i, \beta+\sum_{i=1}^{N}(n-x_i))$$

Posterior Predictive:

$$\begin{aligned}
p(x|\boldsymbol{x}) &= \int p(x|\theta)p(\theta|\boldsymbol{x})d\theta \\
&= \frac{n!}{(n-x)!x!}\frac{\Gamma(\alpha+\beta+Nn)}{\Gamma(\alpha+\sum_{i=1}^{N} x_i)\Gamma(\beta+\sum_{i=1}^{N}(n-x_i)}\int \theta^{x+\alpha+\sum_{i=1}^{N} x_i-1}(1-\theta)^{(n-x)+\beta+\sum_{i=1}^{N}(n-x_i)-1}d\theta \\
&= \frac{n!}{(n-x)!x!}\frac{\Gamma(\alpha+\beta+Nn)}{\Gamma(\alpha+\sum_{i=1}^{N} x_i)\Gamma(\beta+\sum_{i=1}^{N}(n-x_i)}\frac{\Gamma(x+\alpha+\sum_{i=1}^{N} x_i)\Gamma((n-x)+\beta+\sum_{i=1}^{N}(n-x_i)}{\Gamma(\alpha+\beta+Nn+n)}
\end{aligned}$$

E Proof of Poisson Gamma Conjugacy

Marginal likelihood:

$$\begin{aligned} p(\boldsymbol{x}) &= \int p(\boldsymbol{x}|\theta)p(\theta)d\theta \\ &= \int \left[\prod_{i=1}^{N} \frac{\theta^{x_i}e^{-\theta}}{x_i!}\right] \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}d\theta \\ &= \prod_{i=1}^{N}\left[\frac{1}{x_i!}\right]\frac{\beta^\alpha}{\Gamma(\alpha)}\int \theta^{\sum_{i=1}^{N}x_i+\alpha-1}e^{-(\beta+1)\theta}d\theta \\ &= \prod_{i=1}^{N}\left[\frac{1}{x_i!}\right]\frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\sum_{i=1}^{N}x_i+\alpha)}{(\beta+1)^{\alpha+\sum_{i=1}^{N}x_i}} \end{aligned}$$

Posterior:

$$\begin{aligned} p(\theta|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\theta)p(\theta) \\ &\propto \left[\prod_{i=1}^{N}\theta^{x_i}e^{-\theta}\right]\theta^{\alpha-1}e^{-\beta\theta} \\ &= \theta^{\alpha+\sum_{i=1}^{N}x_i-1}e^{-(\beta+N)\theta} \end{aligned}$$

Since Poisson and Gamma are conjugate, the posterior are also Gamma. Hence

$$p(\theta|\boldsymbol{x}) = Gamma(\theta|\alpha+\sum_{i=1}^{N}x_i, \beta+N)$$

Posterior Predictive:

$$\begin{aligned} p(x|\boldsymbol{x}) &= \int p(x|\theta)p(\theta|\boldsymbol{x})d\theta \\ &= \frac{1}{x!}\frac{(\beta+N)^{\alpha+\sum_{i=1}^{N}x_i}}{\Gamma(\alpha+\sum_{i=1}^{N}x_i)}\int \theta^{\alpha+\sum_{i=1}^{N}x_i+1-1}e^{-(\beta+N+1)\theta}d\theta \\ &= \frac{1}{x!}\frac{(\beta+N)^{\alpha+\sum_{i=1}^{N}x_i}}{\Gamma(\alpha+\sum_{i=1}^{N}x_i)}\frac{\Gamma(\alpha+\sum_{i=1}^{N}x_i+1)}{(\beta+N+1)^{\sum_{i=1}^{N}x_i+\alpha+1}} \end{aligned}$$

F Proof of Conjugacy for General Exponential Families

Marginal likelihood:

$$\begin{aligned} p(\boldsymbol{x}) &= \int p(\boldsymbol{x}|\eta)p(\eta)d\eta \\ &= \int \prod_{i=1}^{N}[h(x_i)]\, exp\left\{\eta^T\sum_{i=1}^{N}T(x_i) - NA(\eta)\right\} H(\tau,n_0)exp\left\{\tau^T\eta - n_0A(\eta)\right\}d\eta \\ &= \prod_{i=1}^{N}[h(x_i)]\,H(\tau,n_0)\int exp\left\{\eta^T(\tau+\sum_{i=1}^{N}T(x_i)) - (N+n_0)A(\eta)\right\}d\eta \\ &= \frac{\prod_{i=1}^{N}[h(x_i)]\,H(\tau,n_0)}{H(\tau+\sum_{i=1}^{N}T(x_i), n_0+N)} \end{aligned}$$

Posterior:

$$
\begin{aligned}
p(\eta|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|\eta)p(\eta)}{p(\boldsymbol{x})} \\
&= \frac{\prod_{i=1}^{N} [h(x_i)]\, exp\left\{\eta^T \sum_{i=1}^{N} T(x_i) - NA(\eta)\right\} H(\tau, n_0) exp\left\{\tau^T \eta - n_0 A(\eta)\right\}}{\frac{\prod_{i=1}^{N}[h(x_i)]H(\tau,n_0)}{H(\tau+\sum_{i=1}^{N} T(x_i), n_0+N)}} \\
&= H(\tau + \sum_{i=1}^{N} T(x_i), n_0 + N) exp\left\{\eta^T (\sum_{i=1}^{N} T(x_i) + \tau) - (N + n_0)A(\eta)\right\} \\
&= p(\eta|\tau + \sum_{i=1}^{N} T(x_i), n_0 + N)
\end{aligned}
$$

Posterior Predictive:

$$
\begin{aligned}
p(x|\boldsymbol{x}) &= \int p(\boldsymbol{x}|\eta)p(\eta|\boldsymbol{x})d\eta \\
&= h(x)H(\tau + \sum_{i=1}^{N} T(x_i), n_0 + N) \int exp\left\{\eta^T (T(x) + \sum_{i=1}^{N} T(x_i) + \tau) - (n_0 + N + 1)A(\eta)\right\} d\eta \\
&= \frac{h(x)H(\tau + \sum_{i=1}^{N} T(x_i), n_0 + N)}{H(\tau + \sum_{i=1}^{N} T(x_i) + T(x), n_0 + N + 1)}
\end{aligned}
$$

G Proof of Normal Normal-Mean Conjugacy

Marginal likelihood:

$$
\begin{aligned}
p(\boldsymbol{x}) &= \int p(\boldsymbol{x}|\mu, \sigma^2) p(\mu|\mu_0, \sigma_0^2) d\mu \\
&= \int \prod_{i=1}^{n} \mathcal{N}(x_i|\mu, \sigma^2) \mathcal{N}(\mu|\mu_0, \sigma_0^2) d\mu \\
&= \frac{1}{(\sigma\sqrt{2\pi})^n (\sigma_0\sqrt{2\pi})} \int exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\} d\mu \\
&= \frac{exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 - \frac{1}{2\sigma_0^2}\mu_0^2\right)}{(\sigma\sqrt{2\pi})^n (\sigma_0\sqrt{2\pi})} \int exp\left\{-\frac{1}{2}\left[(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2})\mu^2 - 2(\frac{n\overline{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2})\mu\right]\right\} d\mu \\
&= \frac{exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 - \frac{1}{2\sigma_0^2}\mu_0^2\right)}{(\sigma\sqrt{2\pi})^n (\sigma_0\sqrt{2\pi})} \int exp\left\{-\frac{1}{2}(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2})(\mu^2 - 2\frac{\frac{n\overline{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\mu)\right\} d\mu \\
&= \frac{exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 - \frac{1}{2\sigma_0^2}\mu_0^2\right)}{(\sigma\sqrt{2\pi})^n (\sigma_0\sqrt{2\pi})} exp\left\{\frac{(\frac{n\overline{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2})^2}{2(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2})}\right\} \int exp\left\{-\frac{1}{2}(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2})(\mu - \frac{\frac{n\overline{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}})^2\right\} d\mu \\
&= \frac{exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 - \frac{1}{2\sigma_0^2}\mu_0^2\right)}{(\sigma\sqrt{2\pi})^n (\sigma_0\sqrt{2\pi})} exp\left\{\frac{(\frac{n\overline{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2})^2}{2(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2})}\right\} \frac{\sqrt{2\pi}}{\sqrt{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}}
\end{aligned}
$$

Posterior:

$$
\begin{aligned}
p(\mu|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\mu,\sigma^2)p(\mu|\mu_0,\sigma_0^2) \\
&\propto exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\} \\
&\propto exp\left\{-\frac{1}{2}\left[(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2})\mu^2 - 2(\frac{n\overline{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})\mu\right]\right\} \\
&= exp\left\{-\frac{1}{2}(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2})(\mu^2 - 2\frac{\frac{n\overline{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\mu)\right\} \\
&\propto exp\left\{-\frac{1}{2}(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2})(\mu - \frac{\frac{n\overline{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}})^2\right\}
\end{aligned}
$$

Use the fact of conjugacy, denote $p(\mu|\boldsymbol{x})$ as $\mathcal{N}(\mu|\mu_n,\sigma_n^2)$, we have

$$
\begin{aligned}
\sigma_n^2 &= \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}} = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2} \\
\mu_n &= \frac{\frac{n\overline{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}} = \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\overline{x}
\end{aligned}
$$

Therefore

$$
p(\mu|\boldsymbol{x}) = \mathcal{N}(\mu|\mu_n,\sigma_n^2) = \mathcal{N}(\mu|\frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}, \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2})
$$

Posterior Predictive:

$$
\begin{aligned}
p(x|\boldsymbol{x}) &= \int \mathcal{N}(x|\mu,\sigma^2)\mathcal{N}(\mu|\mu_n,\sigma_n^2)d\mu \\
&= \frac{1}{(\sqrt{2\pi}\sigma)(\sqrt{2\pi}\sigma_n)}\int exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} exp\left\{-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2\right\} d\mu \\
&= \frac{exp(-\frac{x^2}{2\sigma^2}-\frac{\mu_n^2}{2\sigma_n^2})}{(\sqrt{2\pi}\sigma)(\sqrt{2\pi}\sigma_n)}\int exp\left\{-\frac{1}{2}(\frac{1}{\sigma^2}+\frac{1}{\sigma_n^2})(\mu^2 - 2\frac{\frac{x}{\sigma^2}+\frac{\mu_n}{\sigma_n^2}}{\frac{1}{\sigma^2}+\frac{1}{\sigma_n^2}}\mu)\right\} d\mu \\
&= \frac{exp(-\frac{x^2}{2\sigma^2}-\frac{\mu_n^2}{2\sigma_n^2})}{(\sqrt{2\pi}\sigma)(\sqrt{2\pi}\sigma_n)} exp\left\{\frac{(\frac{x}{\sigma^2}+\frac{\mu_n}{\sigma_n^2})^2}{2(\frac{1}{\sigma^2}+\frac{1}{\sigma_n^2})}\right\}\int exp\left\{-\frac{1}{2}(\frac{1}{\sigma^2}+\frac{1}{\sigma_n^2})(\mu - \frac{\frac{x}{\sigma^2}+\frac{\mu_n}{\sigma_n^2}}{\frac{1}{\sigma^2}+\frac{1}{\sigma_n^2}})^2\right\} d\mu \\
&= \frac{exp(-\frac{x^2}{2\sigma^2}-\frac{\mu_n^2}{2\sigma_n^2})}{(\sqrt{2\pi}\sigma)(\sqrt{2\pi}\sigma_n)} exp\left\{\frac{(\frac{x}{\sigma^2}+\frac{\mu_n}{\sigma_n^2})^2}{2(\frac{1}{\sigma^2}+\frac{1}{\sigma_n^2})}\right\} \frac{\sqrt{2\pi}}{\sqrt{\frac{1}{\sigma^2}+\frac{1}{\sigma_n^2}}} \\
&= \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2+\sigma_n^2}} exp\left\{-\frac{1}{2(\sigma^2+\sigma_n^2)}(x-\mu_n)^2\right\} \\
&= \mathcal{N}(x|\mu_n,\sigma^2+\sigma_n^2)
\end{aligned}
$$

H Proof of Normal Normal-Gamma Conjugacy

Likelihood of Data:

For the purpose of simplifying derivation, we rewrite the likelihood of data as

$$\begin{aligned}
p(\boldsymbol{x}|\mu,\lambda) &= (\frac{\lambda}{2\pi})^{\frac{n}{2}} exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\
&= (\frac{\lambda}{2\pi})^{\frac{n}{2}} exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{n}[(x_i-\overline{x})-(\mu-\overline{x})]^2\right\} \\
&= (\frac{\lambda}{2\pi})^{\frac{n}{2}} exp\left\{-\frac{\lambda}{2}\left[\sum_{i=1}^{n}(x_i-\overline{x})^2+\sum_{i=1}^{n}(\mu-\overline{x})(\mu+\overline{x}-2x_i)\right]\right\} \\
&= (\frac{\lambda}{2\pi})^{\frac{n}{2}} exp\left\{-\frac{\lambda}{2}\left[n(\mu-\overline{x})^2+\sum_{i=1}^{n}(x_i-\overline{x})^2\right]\right\}
\end{aligned}$$

Marginal likelihood:

$$\begin{aligned}
p(\boldsymbol{x}) &= \int p(\boldsymbol{x}|\mu,\lambda)NG(\mu,\lambda|\mu_0,\kappa_0,\alpha_0,\beta_0)d\mu d\lambda \\
&= \int (\frac{\lambda}{2\pi})^{\frac{n}{2}} exp\left\{-\frac{\lambda}{2}\left[n(\mu-\overline{x})^2+\sum_{i=1}^{n}(x_i-\overline{x})^2\right]\right\}\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\frac{\kappa_0}{2\pi})^{\frac{1}{2}}\lambda^{\alpha_0-\frac{1}{2}} exp\left\{-\frac{\lambda}{2}[\kappa_0(\mu-\mu_0)^2+2\beta_0]\right\}d\mu d\lambda \\
&= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\frac{\kappa_0}{2\pi})^{\frac{1}{2}}(2\pi)^{-\frac{n}{2}}\int \lambda^{\alpha_0+\frac{n}{2}-\frac{1}{2}} exp\left\{-\frac{\lambda}{2}\left[n(\mu-\overline{x})^2+\kappa_0(\mu-\mu_0)^2+2\beta_0+\sum_{i=1}^{n}(x_i-\overline{x})^2\right]\right\}d\mu d\lambda \\
&= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\frac{\kappa_0}{2\pi})^{\frac{1}{2}}(2\pi)^{-\frac{n}{2}}\int \lambda^{\alpha_0+\frac{n}{2}-\frac{1}{2}} exp\left\{-\frac{\lambda}{2}\left[(n+\kappa_0)\mu^2-2(n\overline{x}+\kappa_0\mu_0)\mu+n\overline{x}^2+\kappa_0\mu_0^2+2\beta_0+\sum_{i=1}^{n}(x_i-\overline{x})^2\right]\right\}d\mu d\lambda \\
&= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\frac{\kappa_0}{2\pi})^{\frac{1}{2}}(2\pi)^{-\frac{n}{2}}\int \lambda^{\alpha_0+\frac{n}{2}-\frac{1}{2}} exp\left\{-\frac{\lambda}{2}\left[(n+\kappa_0)(\mu-\frac{n\overline{x}+\kappa_0\mu_0}{n+\kappa_0})^2+2(\beta_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2+\frac{n\kappa_0(\overline{x}-\mu_0)^2}{2(n+\kappa_0)})\right]\right\} \\
&= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\frac{\kappa_0}{n+\kappa_0})^{\frac{1}{2}}\frac{\Gamma(\alpha_0+\frac{n}{2})}{\left[\beta_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2+\frac{n\kappa_0(\overline{x}-\mu_0)^2}{2(n+\kappa_0)}\right]^{\alpha_0+\frac{n}{2}}}(2\pi)^{-\frac{n}{2}}
\end{aligned}$$

Posterior:

$$\begin{aligned}
p(\mu,\lambda) &\propto p(\boldsymbol{x}|\mu,\lambda)p(\mu,\lambda) \\
&\propto \lambda^{\frac{n}{2}} exp\left\{-\frac{\lambda}{2}\left[n(\mu-\overline{x})^2+\sum_{i=1}^{n}(x_i-\overline{x})^2\right]\right\}\lambda^{\alpha_0-\frac{1}{2}} exp\left\{-\frac{\lambda}{2}[\kappa_0(\mu-\mu_0)^2+2\beta_0]\right\} \\
&= \lambda^{\alpha_0+\frac{n}{2}-\frac{1}{2}} exp\left\{-\frac{\lambda}{2}\left[(n+\kappa_0)(\mu-\frac{n\overline{x}+\kappa_0\mu_0}{n+\kappa_0})^2+2(\beta_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2+\frac{n\kappa_0(\overline{x}-\mu_0)^2}{2(n+\kappa_0)})\right]\right\}
\end{aligned}$$

Due to the fact of conjugacy, we have

$$p(\mu,\lambda|\boldsymbol{x}) = NG(\mu,\lambda|\mu_n,\kappa_n,\alpha_n,\beta_n)$$

where

$$\begin{aligned}
\mu_n &= \frac{n\overline{x}+\kappa_0\mu_0}{n+\kappa_0} \\
\kappa_n &= n+\kappa_0 \\
\alpha_n &= \alpha_0+\frac{n}{2} \\
\beta_n &= \beta_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2+\frac{n\kappa_0(\overline{x}-\mu_0)^2}{2(n+\kappa_0)}
\end{aligned}$$

Posterior Predictive:

We have know the expression of marginal likelihood:

$$p(\boldsymbol{x}) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\frac{\kappa_0}{\kappa_n})^{\frac{1}{2}}\frac{\Gamma(\alpha_n)}{\beta_n^{\alpha_n}}(2\pi)^{-\frac{n}{2}}$$

Therefore

$$\begin{aligned} p(\boldsymbol{x}_m|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}_m, \boldsymbol{x})}{p(\boldsymbol{x})} \\ &= \frac{\Gamma(\alpha_{n+m})}{\Gamma(\alpha_0)}\frac{\beta_n^{\alpha_n}}{\beta_{n+m}^{\alpha_{a+m}}}(\frac{\kappa_n}{\kappa_{n+m}})^{\frac{1}{2}}(2\pi)^{-\frac{m}{2}} \end{aligned}$$

In the special case of $m = 1$, we show that the posterior predictive can be expressed as the non-standardized Student's t-distribution with mean μ_n, scale parameter $\frac{\beta_n(\kappa_n+1)}{\alpha_n\kappa_n}$ and freedom degree $2\alpha_n$, *i.e.*,

$$p(x|\boldsymbol{x}) = t_{2\alpha_n}p(x|\mu_n, \frac{\beta_n(\kappa_n+1)}{\alpha_n\kappa_n})$$

To do this, we use the following equations (see this link for the details of proof):

$$\begin{aligned} \alpha_{n+1} &= \alpha_n + 1/2 \\ \kappa_{n+1} &= \kappa_n + 1 \\ \beta_{n+1} &= \beta_n + \frac{\kappa_n(x-\mu_n)^2}{2(\kappa_n+1)} \end{aligned}$$

Substituting, we have

$$\begin{aligned} p(x|\boldsymbol{x}) &= \frac{\Gamma(\alpha_{n+1})}{\Gamma(\alpha_n)}\frac{\beta_n^{\alpha_n}}{\beta_{n+1}^{\alpha_{n+1}}}\left(\frac{\kappa_n}{\kappa_{n+1}}\right)^{\frac{1}{2}}(2\pi)^{-1/2} \\ &= \frac{\Gamma(\alpha_n+1/2)}{\Gamma(\alpha_n)}\frac{\beta_n^{\alpha_n}}{(\beta_n+\frac{\kappa_n(x-\mu_n)^2}{2(\kappa_n+1)})^{\alpha_n+1/2}}\left(\frac{\kappa_n}{\kappa_{n+1}}\right)^{\frac{1}{2}}(2\pi)^{-1/2} \\ &= \frac{\Gamma((2\alpha_n+1)/2)}{\Gamma((2\alpha_n)/2)}\left(\frac{\beta_n}{\beta_n+\frac{\kappa_n(x-\mu_n)^2}{2(\kappa_n+1)}}\right)^{\alpha_n+1/2}\frac{1}{\beta_n^{1/2}}\left(\frac{\kappa_n}{2(\kappa_n+1)}\right)^{\frac{1}{2}}(\pi)^{-1/2} \\ &= \frac{\Gamma((2\alpha_n+1)/2)}{\Gamma((2\alpha_n)/2)}\left(\frac{1}{1+\frac{\kappa_n(x-\mu_n)^2}{2\beta_n(\kappa_n+1)}}\right)^{\alpha_n+1/2}\left(\frac{\kappa_n}{2\beta_n(\kappa_n+1)}\right)^{\frac{1}{2}}(\pi)^{-1/2} \\ &= (\pi)^{-1/2}\frac{\Gamma((2\alpha_n+1)/2)}{\Gamma((2\alpha_n)/2)}\left(\frac{\alpha_n\kappa_n}{2\alpha_n\beta_n(\kappa_n+1)}\right)^{\frac{1}{2}}\left(1+\frac{\alpha_n\kappa_n(x-\mu_n)^2}{2\alpha_n\beta_n(\kappa_n+1)}\right)^{-(2\alpha_n+1)/2} \end{aligned}$$

Let $\Lambda \triangleq \frac{\alpha_n\kappa_n}{\beta_n(\kappa_n+1)}$, we have

$$p(x|\boldsymbol{x}) = (\pi)^{-1/2}\frac{\Gamma((2\alpha_n+1)/2)}{\Gamma((2\alpha_n)/2)}\left(\frac{\Lambda}{2\alpha_n}\right)^{\frac{1}{2}}\left(1+\frac{\Lambda(x-\mu_n)^2}{2\alpha_n}\right)^{-(2\alpha_n+1)/2}$$

We can see that this is a T-distribution with center at μ_n, precision Λ and degree of freedom $2\alpha_n$.

Property of Normal-Gamma prior:

We have seen that the normal-Gamma prior is

$$NG(\mu, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0) = \mathcal{N}(\mu|\mu_0, (\kappa_0\lambda)^{-1})Ga(\lambda|\alpha_0, \beta_0).$$

It is easy to verify that

$$\begin{aligned} p(\lambda) &= Ga(\lambda|\alpha_0, \beta_0) \\ p(\mu|\lambda) &= \mathcal{N}(\mu_0, (\kappa_0\lambda)^{-1}). \end{aligned}$$

But the marginal distribution of μ is a non-standardized Student's t-distribution. To see that we have

$$\begin{aligned}
p(\mu) &= \int_0^\infty p(\mu, \lambda) d\lambda \\
&\propto \int_0^\infty \lambda^{\alpha_0 + \frac{1}{2} - 1} exp\left(\lambda(\beta_0 + \frac{\kappa_0(\mu - \mu_0)^2}{2}) \right) d\lambda
\end{aligned}$$

We recognize this is an unnormalized $Ga(\lambda|\alpha_0 + \frac{1}{2}, \beta_0 + \frac{\kappa_0(\mu-\mu_0)^2}{2})$, so we can write down

$$\begin{aligned}
p(\mu) &\propto (\beta_0 + \frac{\kappa_0(\mu - \mu_0)^2}{2})^{-\alpha_0 - \frac{1}{2}} \\
&\propto (1 + \frac{1}{2\alpha_0} \frac{\alpha_0 \kappa_0 (\mu - \mu_0)^2}{\beta_0})^{-(2\alpha_0+1)/2}
\end{aligned}$$

which we recognize as a $t_{2\alpha_0}(\mu|\mu_0, \frac{\beta_0}{\alpha_0\kappa_0})$ distribution

$$p(\mu) = \frac{\Gamma(\frac{2\alpha_0+1}{2})}{\Gamma(\frac{2\alpha_0}{2})} \left(\frac{\alpha_0\kappa_0}{2\alpha_0 \pi \beta_0} \right)^{\frac{1}{2}} (1 + \frac{1}{2\alpha_0} \frac{\alpha_0\kappa_0(\mu - \mu_0)^2}{\beta_0})^{-(2\alpha_0+1)/2}$$

I Proof of Normal Gamma-Precision Conjugacy

Marginal likelihood:

$$\begin{aligned}
p(\boldsymbol{x}) &= \int p(\boldsymbol{x}|\lambda) p(\lambda) d\lambda \\
&= (\frac{1}{2\pi})^{\frac{1}{2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \int \lambda^{\alpha + \frac{n}{2} - 1} exp\left\{ -\frac{\lambda}{2} \left[\sum_{i=1}^n (x_i - \mu)^2 + 2\beta \right] \right\} d\lambda \\
&= (\frac{1}{2\pi})^{\frac{1}{2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha + \frac{n}{2})}{\left[\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 + \beta \right]^{\alpha + \frac{n}{2}}}
\end{aligned}$$

Posterior:

$$\begin{aligned}
p(\lambda|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\lambda) p(\lambda) \\
&\propto \lambda^{\frac{n}{2}} exp\left\{ -\frac{\lambda}{2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \lambda^{\alpha - 1} e^{-\beta\lambda} \\
&= \lambda^{\alpha + \frac{n}{2} - 1} exp\left\{ -\frac{\lambda}{2} \left[\sum_{i=1}^n (x_i - \mu)^2 + 2\beta \right] \right\}
\end{aligned}$$

We recognize it as an unnormalized Gamma distribution, therefore

$$\begin{aligned}
p(\lambda|\boldsymbol{x}) &= Ga(\lambda|\alpha + \frac{n}{2}, \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 + \beta) \\
&\triangleq Ga(\lambda|\alpha_n, \beta_n)
\end{aligned}$$

Posterior Predictive:

$$\begin{aligned}
p(x|\boldsymbol{x}) &= \int p(x|\lambda) p(\lambda|\boldsymbol{x}) d\lambda \\
&= (\frac{1}{2\pi})^{\frac{1}{2}} \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \int \lambda^{\alpha_n + \frac{1}{2} - 1} exp\left\{ -\frac{\lambda}{2} [(x - \mu)^2 + 2\beta_n] \right\} d\lambda \\
&= (\frac{1}{2\pi})^{\frac{1}{2}} \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \frac{\Gamma(\alpha_n + \frac{1}{2})}{\left[\frac{1}{2}(x - \mu)^2 + \beta_n \right]^{\alpha_n + \frac{1}{2}}} \\
&= \frac{\Gamma((2\alpha_n + 1)/2)}{\Gamma((2\alpha_n)/2)} \left(\frac{\alpha_n}{2\pi\alpha_n\beta_n} \right)^{\frac{1}{2}} \left(1 + \frac{\alpha_n(x - \mu)^2}{2\alpha_n\beta_n} \right)^{-(2\alpha_n+1)/2} \\
&= t_{2\alpha_n}(x|\mu, \sigma^2 = \frac{\beta_n}{\alpha_n})
\end{aligned}$$

J Proof of Normal Normal-inverse-chi-square (NIX) Conjugacy

Marginal likelihood:

$$
\begin{aligned}
p(\boldsymbol{x}) &= \int p(\boldsymbol{x}|\mu,\sigma^2) NI\chi^2(\mu,\sigma^2|\mu_0,\kappa_0,v_0,\sigma_0^2) d\mu d\sigma^2 \\
&= \frac{1}{(2\pi)^{n/2}} \frac{\sqrt{\kappa_0}}{\sqrt{2\pi}} \frac{1}{\Gamma(v_0/2)} \left(\frac{v_0\sigma_0^2}{2}\right)^{v_0/2} \int \sigma^{-1}(\sigma^2)^{-(v_0+n)/2+1} exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\mu)^2 + v_0\sigma_0^2 + \kappa_0(\mu_0-\mu)^2\right]\right\} \\
&= \frac{1}{(2\pi)^{n/2}} \frac{\sqrt{\kappa_0}}{\sqrt{2\pi}} \frac{1}{\Gamma(v_0/2)} \left(\frac{v_0\sigma_0^2}{2}\right)^{v_0/2} \int \sigma^{-1}(\sigma^2)^{-\frac{v_0+n}{2}+1} \\
&\qquad exp\Big\{-\frac{1}{2\sigma^2}\Big[(v_0+n)\Big(\frac{1}{v_0+n}\big[v_0\sigma_0^2 + \sum_{i=1}^{n}(x_i-\overline{x})^2 + \frac{n\kappa_0}{\kappa_0+n}(\mu_0-\overline{x})^2\big]\Big) + (n+\kappa_0)(\mu - \frac{n\overline{x}+\kappa_0\mu_0}{n+\kappa_0})^2\Big]\Big\} d\mu d\sigma^2 \\
&= \frac{1}{(2\pi)^{n/2}} \frac{\sqrt{\kappa_0}}{\sqrt{2\pi}} \frac{1}{\Gamma(v_0/2)} \left(\frac{v_0\sigma_0^2}{2}\right)^{v_0/2} \frac{\sqrt{2\pi}}{\sqrt{\kappa_n}} \Gamma(v_n/2) \left(\frac{2}{v_n\sigma_n^2}\right)^{v_n/2} \\
&= \frac{\Gamma(v_n/2)}{\Gamma(v_0/2)} \sqrt{\frac{\kappa_0}{\kappa_n}} \frac{(v_0\sigma_0^2)^{v_n/2}}{(v_n\sigma_n^2)^{v_n/2}} \frac{1}{\pi^{n/2}},
\end{aligned}
$$

where

$$
\begin{aligned}
\mu_n &= \frac{\kappa_0\mu_0 + n\overline{x}}{\kappa_n} \\
\kappa_n &= \kappa_0 + n \\
v_n &= v_0 + n \\
\sigma_n^2 &= \frac{1}{v_0+n}\left(v_0\sigma_0^2 + \sum_{i=1}^{n}(x_i-\overline{x})^2 + \frac{n\kappa_0}{\kappa_0+n}(\mu_0-\overline{x})^2\right)
\end{aligned}
$$

Posterior:

$$
\begin{aligned}
p(\mu,\sigma^2|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\mu,\sigma^2)p(\mu,\sigma^2) \\
&\propto \left[(\sigma^2)^{-n/2} exp\left(-\frac{1}{2\sigma^2}\big[\sum_{i=1}^{b}(x_i-\overline{x})^2 + n(\overline{x}-\mu)^2\big]\right)\right] \times \left[\sigma^{-1}(\sigma^2)^{-\frac{v_0}{2}+1} exp\left(-\frac{1}{2\sigma^2}\big[v_0\sigma_0^2 + \kappa_0(\mu_0-\mu)^2\big]\right)\right] \\
&= \sigma^{-1}(\sigma^2)^{-(v_0+n)/2+1} exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\mu)^2 + v_0\sigma_0^2 + \kappa_0(\mu_0-\mu)^2\right]\right\} \\
&= \sigma^{-1}(\sigma^2)^{-\frac{v_0+n}{2}+1} exp\Big\{-\frac{1}{2\sigma^2}\Big[(v_0+n)\Big(\frac{1}{v_0+n}\big[v_0\sigma_0^2 + \sum_{i=1}^{n}(x_i-\overline{x})^2 + \frac{n\kappa_0}{\kappa_0+n}(\mu_0-\overline{x})^2\big]\Big) + (n+\kappa_0)(\mu - \frac{n\overline{x}+\kappa_0\mu_0}{n+\kappa_0})^2\Big]\Big\}
\end{aligned}
$$

We recognize this is an unnormalized normal-inverse-chi-square distribution, therefore

$$
p(\mu,\sigma^2|\boldsymbol{x}) = NI\chi^2(\mu,\sigma^2|\mu_n,\kappa_n,v_n,\sigma_n^2)
$$

Posterior Predictive:

Using the following equations

$$
\begin{aligned}
\kappa_{n+1} &= \kappa_n + 1 \\
v_{n+1} &= v_n + 1 \\
\sigma_{n+1}^2 &= \frac{1}{v_n+1}\left(v_n\sigma_n^2 + \frac{k_n}{k_n+1}(\mu_n - x)^2\right),
\end{aligned}
$$

where x is the new observation. Then, we have

$$\begin{aligned}
p(x|\boldsymbol{x}) &= \frac{p(x,\boldsymbol{x})}{p(\boldsymbol{x})} \\
&= \frac{\Gamma((v_n+1)/2)}{\Gamma(v_n/2)}\sqrt{\frac{\kappa_n}{\kappa_n+1}}\frac{(v_n\sigma_n^2)^{v_n/2}}{(v_n\sigma_n^2+\frac{k_n}{k_n+1}(\mu_n-x)^2)^{(v_n+1)/2}}\frac{1}{\pi^{1/2}} \\
&= \frac{\Gamma((v_n+1)/2)}{\Gamma(v_n/2)}\sqrt{\frac{\kappa_n}{(\kappa_n+1)\pi v_n\sigma_n^2}}\left(1+\frac{\kappa_n(x-\mu_n)^2}{(\kappa_n+1)v_n\sigma_n^2}\right)^{-(v_n+1)/2} \\
&= t_{v_n}(x|\mu_n,\frac{(1+\kappa_n)\sigma_n^2}{\kappa_n})
\end{aligned}$$

Property of NIX prior:

We have defined the Normal-inverse-chi-squared prior as

$$p(\mu,\sigma^2|\mu_0,\kappa_0,v_0,\sigma_0^2) = \mathcal{N}(\mu|\mu_0,\sigma^2/\kappa_0)\chi^{-2}(\sigma^2|v_0,\sigma_0^2)$$

It is easy to verify that

$$\begin{aligned}
p(\sigma^2) &= \chi^{-2}(\sigma^2|v_0,\sigma_0^2) \\
p(\mu|\sigma^2) &= \mathcal{N}(\mu|\mu_0,\sigma^2/\kappa_0)
\end{aligned}$$

But the marginal distribution of μ is a non-standardized Student's t-distribution. To see that, we have

$$\begin{aligned}
p(\mu) &= \int \mathcal{N}(\mu|\mu_0,\sigma^2/\kappa_0)\chi^{-2}(\sigma^2|v_0,\sigma_0^2)d\sigma^2 \\
&\propto \int (\sigma^2)^{-(\frac{v_0+1}{2}+1)}exp\left(-\frac{1}{2\sigma^2}[v_0\sigma_0^2+\kappa_0(\mu_0-\mu)^2]\right)d\sigma^2
\end{aligned}$$

Denote $\phi=\sigma^2$, $\alpha=\frac{v_0+1}{2}$ and $A=v_0\sigma_0^2+\kappa_0(\mu_0-\mu)^2$, we have

$$p(\mu) \propto \int \phi^{-\alpha-1}e^{-\frac{A}{2\phi}}d\phi$$

Denote $x=\frac{A}{2\phi}$, then

$$\frac{d\phi}{dx} = -\frac{A}{2}x^{-2}$$

Note that only A is correlated with μ, hence

$$\begin{aligned}
p(\mu) &\propto \int \left(\frac{A}{2x}\right)^{-\alpha-1}e^{-x}(-\frac{A}{2})x^{-2}dx \\
&\propto A^{-\alpha}\int x^{\alpha-1}e^{-x}dx
\end{aligned}$$

We recognize that the integral term is an unnormalized Gamma distribution, so

$$\begin{aligned}
p(\mu) &\propto A^{\alpha} \\
&= (v_0\sigma_0^2+\kappa_0(\mu_0-\mu)^2)^{-\frac{v_0+1}{2}} \\
&\propto \left[1+\frac{\kappa_0}{v_0\sigma_0^2}(\mu-\mu_0)^2\right]^{-\frac{v_0+1}{2}}
\end{aligned}$$

We recognize that it is an unnormalized student's t-distribution, *i.e.*,

$$\begin{aligned}
p(\mu) &= t_{v_0}(\mu|\mu_0,\sigma_0^2/\kappa_0) \\
&= \frac{\Gamma(\frac{v_0+1}{2})}{\Gamma(\frac{v_0}{2})}\left(\frac{\kappa_0}{\pi v_0\sigma_0^2}\right)^{1/2}\left[1+\frac{\kappa_0}{v_0\sigma_0^2}(\mu-\mu_0)^2\right]^{-\frac{v_0+1}{2}}
\end{aligned}$$

K Proof of Normal Normal-inverse-Gamma Conjugacy

Marginal likelihood:

$$
\begin{aligned}
p(\boldsymbol{x}) &= \int p(\boldsymbol{x}|\mu,\sigma^2) NIG(\mu,\sigma^2|m_0,V_0,\alpha_0,b_0) d\mu d\sigma^2 \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\sqrt{2\pi V_0}} \frac{b_0^{\alpha_0}}{\Gamma(\alpha_0)} \int \sigma^{-1}(\sigma^2)^{-(\alpha_0+\frac{n}{2})-1} exp\left(-\frac{1}{2\sigma^2}\left[V_0^{-1}(\mu-m_0)^2+2b_0+\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-\mu)^2\right]\right) d\mu d\sigma^2 \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\sqrt{2\pi V_0}} \frac{b_0^{\alpha_0}}{\Gamma(\alpha_0)} \int \sigma^{-1}(\sigma^2)^{-(\alpha_0+\frac{n}{2})-1} \\
&\quad exp\Big\{-\frac{1}{2\sigma^2}\Big[(V_0^{-1}+n)(\mu-\frac{V_0^{-1}m_0+n\overline{x}}{V_0^{-1}+n})^2+\big(b_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2+\frac{V_0^{-1}n}{2(V_0^{-1}+n)}(m_0-\overline{x})^2\big)\Big]\Big\} d\mu d\sigma^2 \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\sqrt{2\pi V_0}} \frac{b_0^{\alpha_0}}{\Gamma(\alpha_0)} \sqrt{2\pi V_n} \frac{\Gamma(\alpha_n)}{b_n^{\alpha_n}} \\
&= \frac{\Gamma(\alpha_n)}{\Gamma(\alpha_0)} \sqrt{\frac{V_n}{V_0}} \frac{b_0^{\alpha_0}}{b_n^{\alpha_n}} \frac{1}{(2\pi)^{n/2}}
\end{aligned}
$$

where

$$
\begin{aligned}
m_n &= \frac{V_0^{-1}m_0+n\overline{x}}{V_0^{-1}+n} \\
V_n^{-1} &= V_0^{-1}+n \\
\alpha_n &= \alpha_0+\frac{n}{2} \\
b_n &= b_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2+\frac{V_0^{-1}n}{2(V_0^{-1}+n)}(m_0-\overline{x})^2
\end{aligned}
$$

Posterior:

$$
\begin{aligned}
p(\mu,\sigma^2|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\mu,\sigma^2)p(\mu,\sigma^2) \\
&\propto \left[(\sigma^2)^{-n/2} exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{b}(x_i-\overline{x})^2+n(\overline{x}-\mu)^2\right]\right)\right] \times \left[\sigma^{-1}(\sigma^2)^{-\alpha_0-1} exp\left(-\frac{1}{2\sigma^2}\left[V_0^{-1}(\mu-\mu_0)^2+2b_0\right]\right)\right] \\
&= \sigma^{-1}(\sigma^2)^{-(\alpha_0+\frac{n}{2})-1} exp\left(-\frac{1}{2\sigma^2}\left[V_0^{-1}(\mu-m_0)^2+2b_0+\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-\mu)^2\right]\right) \\
&= \sigma^{-1}(\sigma^2)^{-(\alpha_0+\frac{n}{2})-1} exp\Big\{-\frac{1}{2\sigma^2}\Big[(V_0^{-1}+n)(\mu-\frac{V_0^{-1}m_0+n\overline{x}}{V_0^{-1}+n})^2+\big(b_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2+\frac{V_0^{-1}n}{2(V_0^{-1}+n)}(m_0-\overline{x})^2\big)\Big]\Big\}
\end{aligned}
$$

We recognize this is an unnormalized Normal-inverse-Gamma distribution, therefore

$$
p(\mu,\sigma^2|\boldsymbol{x}) = NIG(\mu,\sigma^2|m_n,V_n,\alpha_n,b_n)
$$

Posterior Predictive:

To derivate the posterior predictive, we use the following quations

$$
\begin{aligned}
m_{n+1} &= \frac{V_n^{-1}m_n+x}{V_n^{-1}+1} \\
V_{n+1}^{-1} &= V_n^{-1}+1 \\
\alpha_{n+1} &= \alpha_n+\frac{1}{2} \\
b_{n+1} &= b_n+\frac{V_n^{-1}}{2(V_n^{-1}+1)}(m_n-x)^2
\end{aligned}
$$

where x is the new observation. Then we have

$$
\begin{aligned}
p(x|\boldsymbol{x}) &= \frac{p(x,\boldsymbol{x})}{p(\boldsymbol{x})} \\
&= \frac{\Gamma(\alpha_{n+1})}{\Gamma(\alpha_n)}\sqrt{\frac{V_{n+1}}{V_n}}\frac{b_n^{\alpha_n}}{b_{n+1}^{\alpha_{n+1}}}\frac{1}{\sqrt{2\pi}} \\
&= \frac{\Gamma((2\alpha_n+1)/2)}{\Gamma(2\alpha_n/2)}\left(\frac{\alpha_n V_n^{-1}}{2\alpha_n \pi b_n(V_n^{-1}+1)}\right)^{\frac{1}{2}}\left[1+\frac{1}{2\alpha_n}\frac{\alpha_n V_n^{-1}}{b_n(V_n^{-1}+1)}(x-m_n)^2\right]^{-\frac{2\alpha_n+1}{2}} \\
&= t_{2\alpha_n}(x|m_n, \frac{b_n(V_n^{-1}+1)}{\alpha_n V_n^{-1}}) \\
&= t_{2\alpha_n}(x|m_n, \frac{b_n(V_n^{+}1)}{\alpha_n})
\end{aligned}
$$

Property of NIG prior:

We have defined the Normal-inverse-Gamma prior as

$$
\begin{aligned}
p(\mu,\sigma^2) &= \mathcal{N}(\mu|\mu_0,\sigma^2 V_0)IG(\sigma^2|\alpha_0,b_0) \\
&= \frac{1}{\sqrt{2\pi V_0}}\frac{b_0^{\alpha_0}}{\Gamma(\alpha_0)}\frac{1}{\sigma}(\sigma^2)^{-\alpha_0-1}exp\left(-\frac{1}{2\sigma^2}[V_0^{-1}(\mu-\mu_0)^2+2b_0]\right)
\end{aligned}
$$

It is easy to verify that

$$
\begin{aligned}
p(\sigma^2) &= IG(\sigma^2|\alpha_0,b_0) \\
p(\mu|\sigma^2) &= \mathcal{N}(\mu|\mu_0,\sigma^2/V_0),
\end{aligned}
$$

and the marginal distribution of μ is a non-standardized Student's t-distribution.

$$
\begin{aligned}
p(\mu) &= \int \mathcal{N}(\mu|\mu_0,\sigma^2 V_0)IG(\sigma^2|\alpha_0,b_0)d\sigma^2 \\
&\propto \int (\sigma^2)^{-(\frac{2\alpha_0+1}{2}+1)}exp\left(-\frac{1}{2\sigma^2}[V_0^{-1}(\mu-\mu_0)^2+2b_0]\right)d\sigma^2
\end{aligned}
$$

Denoting $\phi=\sigma^2$, $\alpha=\frac{2\alpha_0+1}{2}$ and $A=V_0^{-1}(\mu-\mu_0)^2+2b_0$, we have

$$
\begin{aligned}
p(\mu) &\propto \int \phi^{-\alpha-1}e^{-\frac{A}{2\phi}}d\phi \\
&= \int \left(\frac{A}{2x}\right)^{-\alpha-1}e^{-x}(-\frac{A}{2})x^{-2}dx \\
&\propto A^{-\alpha}\int x^{\alpha-1}e^{-x}dx \\
&\propto A^{-\alpha} \\
&= (V_0^{-1}(\mu-\mu_0)^2+2b_0)^{-\frac{2\alpha_0+1}{2}} \\
&\propto \left[1+\frac{\alpha_0(\mu-\mu_0)^2}{2\alpha_0 b_0 V_0}\right]^{-\frac{2\alpha_0+1}{2}}
\end{aligned}
$$

where we set $x=\frac{A}{2\phi}$ (note that only A is relevant to μ). We recognize it is an unnormalized student's t-distribution, *i.e.*,

$$
\begin{aligned}
p(\mu) &= t_{2\alpha_0}(\mu|\mu_0,\frac{b_0V_0}{\alpha_0}) \\
&= \frac{\Gamma((2\alpha_0+1)/2)}{\Gamma(2\alpha_0/2)}\left(\frac{\alpha_0}{\pi 2\alpha_0 b_0 V_0}\right)^{\frac{1}{2}}\left[1+\frac{1}{2\alpha_0}\frac{\alpha_0(\mu-\mu_0)^2}{b_0V_0}\right]^{-\frac{2\alpha_0+1}{2}}
\end{aligned}
$$

L Proof of Multivariate Normal Normal-Mean Conjugacy

Marginal likelihood:

$$
\begin{aligned}
p(X) &= \int p(X|\mu,\Sigma)p(\mu|\mu_0,\Sigma_0) \\
&= (2\pi)^{-\frac{d}{2}(N+1)}|\Sigma_0|^{-\frac{1}{2}}|\Sigma|^{-\frac{N}{2}} \int exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)+(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)\right]\right\} d\mu
\end{aligned}
$$

Denote $\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)+(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)$ as A, then

$$
\begin{aligned}
A &= \sum_{i=1}^{N}(x_i^T\Sigma^{-1}x_i+\mu^T\Sigma^{-1}\mu-2x_i^T\Sigma^{-1}\mu)+\mu^T\Sigma_0^{-1}\mu+\mu_0^T\Sigma_0^{-1}\mu_0-2\mu^T\Sigma_0^{-1}\mu_0 \\
&= \mu^T N\Sigma^{-1}\mu-2\mu^T\Sigma^{-1}N\overline{x}+\sum_{i=1}^{N}x_i^T\Sigma^{-1}x_i+\mu^T\Sigma_0^{-1}\mu-2\mu^T\Sigma_0^{-1}\mu_0+\mu_0^T\Sigma_0^{-1}\mu_0 \\
&= \mu^T(N\Sigma^{-1}+\Sigma_0^{-1})\mu-2\mu^T(\Sigma^{-1}N\overline{x}+\Sigma_0^{-1}\mu_0)+\sum_{i=1}^{N}x_i^T\Sigma^{-1}x_i+\mu_0^T\Sigma_0^{-1}\mu_0 \\
&= \Bigg[\mu-\underbrace{(N\Sigma^{-1}+\Sigma_0^{-1})^{-1}(\Sigma^{-1}N\overline{x}+\Sigma_0^{-1}\mu_0)}_{\mu_N}\Bigg]^T\underbrace{(N\Sigma^{-1}+\Sigma_0^{-1})}_{\Sigma_N^{-1}}\Bigg[\mu-\underbrace{(N\Sigma^{-1}+\Sigma_0^{-1})^{-1}(\Sigma^{-1}N\overline{x}+\Sigma_0^{-1}\mu_0)}_{\mu_N}\Bigg] \\
&\quad -||(N\Sigma^{-1}+\Sigma_0^{-1})^{-1}(\Sigma^{-1}N\overline{x}+\Sigma_0^{-1}\mu_0)||^2+\sum_{i=1}^{N}x_i^T\Sigma^{-1}x_i+\mu_0^T\Sigma_0^{-1}\mu_0 \\
&= (\mu-\mu_N)^T\Sigma_N^{-1}(\mu-\mu_N)-\mu_N^T\mu_N+\sum_{i=1}^{N}x_i^T\Sigma^{-1}x_i+\mu_0^T\Sigma_0^{-1}\mu_0.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
p(X) &= (2\pi)^{-\frac{d}{2}(N+1)}|\Sigma_0|^{-\frac{1}{2}}|\Sigma|^{-\frac{N}{2}} exp\left(-\frac{1}{2}[-\mu_N^T\mu_N+\sum_{i=1}^{N}x_i^T\Sigma^{-1}x_i+\mu_0^T\Sigma_0^{-1}\mu_0]\right) \\
&\quad \int exp\left\{-\frac{1}{2}(\mu-\mu_N)^T\Sigma_N^{-1}(\mu-\mu_N)\right\} d\mu \\
&= (2\pi)^{-\frac{d}{2}(N+1)}|\Sigma_0|^{-\frac{1}{2}}|\Sigma|^{-\frac{N}{2}} exp\left(-\frac{1}{2}[-\mu_N^T\mu_N+\sum_{i=1}^{N}x_i^T\Sigma^{-1}x_i+\mu_0^T\Sigma_0^{-1}\mu_0]\right)(2\pi)^{\frac{d}{2}}|\Sigma_N|^{\frac{1}{2}} \\
&= (2\pi)^{-\frac{d}{2}N}\left(\frac{|\Sigma_N|}{|\Sigma_0||\Sigma|^N}\right)^{\frac{1}{2}} exp\left(-\frac{1}{2}[\mu_N^T\mu_N+\sum_{i=1}^{N}x_i^T\Sigma^{-1}x_i+\mu_0^T\Sigma_o^{-1}\mu_0]\right)
\end{aligned}
$$

where

$$
\begin{aligned}
\mu_N &= (\Sigma_0^{-1}+N\Sigma^{-1})^{-1}(N\Sigma^{-1}\overline{x}+\Sigma_0^{-1}\mu_0) \\
\Sigma_N &= (\Sigma_0^{-1}+N\Sigma^{-1})^{-1}
\end{aligned}
$$

Posterior:

$$
\begin{aligned}
p(\mu|X) &\propto p(X|\mu,\Sigma)p(\mu|\mu_0,\Sigma_0) \\
&\propto exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)+(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)\right]\right\} \\
&\propto exp\left\{-\frac{1}{2}(\mu-\mu_N)^T\Sigma_N^{-1}(\mu-\mu_N)\right\}
\end{aligned}
$$

We recognize this is an unnormalized Gaussian distribution. Therefore,

$$
p(\mu|X) = \mathcal{N}(\mu|\mu_N,\Sigma_N)
$$

M Proof of Multivariate Normal Wishart-Precision Conjugacy

Marginal likelihood:

$$\begin{aligned}
p(X) &= \int p(X|\mu,\Lambda)p(\Lambda)d\Lambda \\
&= \int (2\pi)^{-\frac{d}{2}N}|\Lambda|^{\frac{N}{2}} exp\left(-\frac{1}{2}tr(\Lambda S)\right)\frac{1}{Z_0}|\Lambda|^{(v_0-d-1)/2} exp\left(-\frac{1}{2}tr(T_0^{-1}\Lambda)\right)d\Lambda \\
&= (2\pi)^{-\frac{d}{2}N}\frac{1}{Z_0}\int |\Lambda|^{(v_0+N-d-1)/2} exp\left(-\frac{1}{2}tr((S+T_0^{-1})\Lambda)\right)d\Lambda \\
&= (2\pi)^{-\frac{d}{2}N}\frac{Z_N}{Z_0}
\end{aligned}$$

where

$$\begin{aligned}
Z_0 &= 2^{v_0 d/2}\Gamma_d(v_0/2)|T_0|^{v_0/2} \\
Z_N &= 2^{v_N d/2}\Gamma_d(v_N/2)|T_N|^{v_N/2} \\
v_N &= v_0 + N \\
T_N &= (S+T_0^{-1})^{-1}
\end{aligned}$$

Posterior:

$$\begin{aligned}
p(\Lambda|X) &\propto p(X|\Lambda)p(\Lambda) \\
&\propto |\Lambda|^{(v_0+N-d-1)/2} exp\left(-\frac{1}{2}tr((S+T_0^{-1})\Lambda)\right)
\end{aligned}$$

We recognize this is an unnormalized Wishart distribution, hence

$$p(\Lambda|X) = Wi_{v_N}(\Lambda|T_N)$$

where

$$\begin{aligned}
v_N &= v_0 + N \\
T_N &= (S+T_0^{-1})^{-1}
\end{aligned}$$

N Proof of Multivariate Normal Normal-Wishart Conjugacy

Posterior:

$$\begin{aligned}
p(\mu,\Lambda|X) &\propto p(X|\mu,\Lambda)p(\mu,\Lambda) \\
&\propto |\Lambda|^{\frac{N}{2}} exp\left(-\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Lambda(x_i-\mu)\right) \\
&\quad |\Lambda|^{\frac{1}{2}} exp\left(-\frac{\kappa}{2}(\mu-\mu_0)^T\Lambda(\mu-\mu_0)\right)|\Lambda|^{(v-d-1)/2} exp\left(-\frac{1}{2}tr(T^{-1}\Lambda)\right) \\
&\propto |\Lambda|^{\frac{1}{2}}|\Lambda|^{(v-d-1)/2} exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{N}(x_i^T\Lambda x_i - 2x_i^T\Lambda\mu + \mu^T\Lambda\mu) + \kappa(\mu^T\Lambda\mu - 2\mu^T\Lambda\mu_0 + \mu_0^T\Lambda\mu_0) + tr(T^{-1}\Lambda)\right]\right\}
\end{aligned}$$

Denote $\sum_{i=1}^{N}(x_i^T\Lambda x_i - 2x_i^T\Lambda\mu + \mu^T\Lambda\mu) + \kappa(\mu^T\Lambda\mu - 2\mu^T\Lambda\mu_0 + \mu_0^T\Lambda\mu_0) + tr(T^{-1}\Lambda)$ as A, we have

$$
\begin{aligned}
A &= (N+\kappa)\mu^T\Lambda\mu - 2\mu^T\Lambda(\kappa\mu_0 + N\overline{x}) + \kappa\mu_0^T\Lambda\mu_0 + \sum_{i=1}^{N} x_i^T\Lambda x_i + tr(T^{-1}\Lambda) \\
&= (N+\kappa)(\mu - \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa})^T\Lambda(\mu - \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa}) - \frac{1}{N+\kappa}(\kappa\mu_0 + N\overline{x})^T\Lambda(\kappa\mu_0 + N\overline{x}) \\
&\quad + \kappa\mu_0^T\Lambda\mu_0 + \sum_{i=1}^{N} x_i^T\Lambda x_i + tr(T^{-1}\Lambda) \\
&= (N+\kappa)(\mu - \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa})^T\Lambda(\mu - \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa}) - \frac{1}{N+\kappa}(\kappa\mu_0 + N\overline{x})^T\Lambda(\kappa\mu_0 + N\overline{x}) \\
&\quad \sum_{i=1}^{N}(x_i^T\Lambda x_i - x_i^T\Lambda\overline{x} - \overline{x}^T\Lambda x_i + \overline{x}^T\Lambda\overline{x}) + N\overline{x}^T\Lambda\overline{x} + \kappa\mu_0^T\Lambda\mu_0 + tr(T^{-1}\Lambda) \\
&= (N+\kappa)(\mu - \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa})^T\Lambda(\mu - \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa}) + \sum_{i=1}^{N}(x_i - \overline{x})^T\Lambda(x_i - \overline{x}) \\
&\quad - \frac{1}{N+\kappa}(\kappa\mu_0 + N\overline{x})^T\Lambda(\kappa\mu_0 + N\overline{x}) + N\overline{x}^T\Lambda\overline{x} + \kappa\mu_0^T\Lambda\mu_0 + tr(T^{-1}\Lambda) \\
&= (N+\kappa)(\mu - \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa})^T\Lambda(\mu - \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa}) + \sum_{i=1}^{N}(x_i - \overline{x})^T\Lambda(x_i - \overline{x}) \\
&\quad \frac{N\kappa}{N+\kappa}(\overline{x}^T\Lambda\overline{x} - \overline{x}^T\Lambda\mu_0 - \mu_0^T\Lambda\overline{x} + \mu_0^T\Lambda\mu_0) + tr(T^{-1}\Lambda) \\
&= (N+\kappa)(\mu - \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa})^T\Lambda(\mu - \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa}) + \sum_{i=1}^{N}(x_i - \overline{x})^T\Lambda(x_i - \overline{x}) \\
&\quad + \frac{N\kappa}{N+\kappa}(\overline{x} - \mu_0)^T\Lambda(\overline{x} - \mu_0) + tr(T^{-1}\Lambda) \\
&= (N+\kappa)(\mu - \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa})^T\Lambda(\mu - \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa}) \\
&\quad + tr\left\{\left[\underbrace{\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})^T}_{S} + \frac{N\kappa}{N+\kappa}(\overline{x} - \mu_0)(\overline{x} - \mu_0)^T + T^{-1}\right]\Lambda\right\}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
p(\mu, \Lambda | X) \propto |\Lambda|^{\frac{1}{2}} exp\left(-\frac{N+\kappa}{2}(\mu - \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa})^T\Lambda(\mu - \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa})\right) \\
|\Lambda|^{\frac{v+N-d-1}{2}} exp\left\{-\frac{1}{2}tr\left[\left(S + \frac{N\kappa}{N+\kappa}(\overline{x} - \mu_0)(\overline{x} - \mu_0)^T + T^{-1}\right)\Lambda\right]\right\}
\end{aligned}
$$

We recognize this is an unnormalized Normal-Wishart distribution, hence

$$
\begin{aligned}
p(\mu, \Lambda | X) &= NWi(\mu, \Lambda | \mu_N, \kappa_N, v_N, T_N) \\
&= \mathcal{N}(\mu | \mu_N, (\kappa_N\Lambda)^{-1}) Wi_{v_N}(\Lambda | T_N) \\
\mu_N &= \frac{\kappa\mu_0 + N\overline{x}}{N+\kappa} \\
\kappa_N &= \kappa + N \\
v_N &= v + N \\
T_N &= \left(T^{-1} + S + \frac{N\kappa}{N+\kappa}(\overline{x} - \mu_0)(\overline{x} - \mu_0)^T\right)^{-1} \\
S &= \sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})^T
\end{aligned}
$$

Marginal likelihood:

$$
\begin{aligned}
p(X) &= \frac{p(X|\mu,\Sigma)p(\mu,\Sigma)}{p(\mu,\Sigma|X)} \\
&= \frac{\mathcal{N}(X|\mu,\Sigma)NWi(\mu,\Sigma|\mu_0,\kappa_0,v_0,\Lambda_0)}{NWi(\mu,\Sigma|\mu_N,\kappa_N,v_N,\Lambda_N)} \\
&= \frac{Z_N}{Z_0}\frac{1}{(2\pi)^{Nd/2}} \\
&= \frac{(2\pi/\kappa_N)^{d/2}2^{v_N d/2}|T_N|^{v_N/2}\Gamma_d^{v_N/2}}{(2\pi/\kappa_0)^{d/2}2^{v_0 d/2}|T_0|^{v_0/2}\Gamma_d^{v_0/2}}\frac{1}{(2\pi)^{Nd/2}} \\
&= \frac{1}{(\pi)^{Nd/2}}\frac{\Gamma_d^{v_N/2}}{\Gamma_d^{v_0/2}}\frac{\Gamma_d^{v_N/2}}{\Gamma_d^{v_0/2}}\left(\frac{\kappa_0}{\kappa_N}\right)^{d/2}
\end{aligned}
$$

Property of Normal-Wishart prior:

We have defined the Normal-Wishart prior as

$$
\begin{aligned}
p(\mu,\Lambda) &= NWi(\mu,\Lambda|\mu_0,\kappa,v,T) = \mathcal{N}(\mu|\mu_0,(\kappa\Lambda)^{-1})Wi_v(\Lambda|T) \\
&= \frac{1}{Z}|\Lambda|^{\frac{1}{2}}exp\left(-\frac{\kappa}{2}(\mu-\mu_0)^T\Lambda(\mu-\mu_0)\right)|\Lambda|^{(\kappa-d-1)/2}exp(-\frac{1}{2}tr(T^{-1}\Lambda)) \\
Z &= \left(\frac{2\pi}{\kappa}\right)^{\frac{d}{2}}2^{\frac{vd}{2}}|T|^{\frac{v}{2}}\Gamma_d(\frac{v}{2})
\end{aligned}
$$

Then its margin distribution is

$$
\begin{aligned}
p(\Lambda) &= Wi_v(\Lambda|T) \\
p(\mu|\Lambda) &= \mathcal{N}(\mu|\mu_0,(\kappa\Lambda)^{-1}) \\
p(\mu) &= t_{v-d+1}(\mu_0,\frac{T^{-1}}{\kappa(v-d+1)})
\end{aligned}
$$

O Proof of Multivariate Normal Normal-Inverse-Wishart Conjugacy

Posterior:

$$
\begin{aligned}
p(\mu,\Sigma|X) &\propto p(X|\mu,\Sigma)p(\mu,\Sigma) \\
&\propto |\Sigma|^{-\frac{N}{2}}exp\left(-\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right) \\
&\quad |\Sigma|^{-(\frac{v_0+d}{2}+1)}exp\left(-\frac{1}{2}\Big[tr(\Lambda_0\Sigma^{-1})+\kappa_0(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\Big]\right) \\
&= |\Sigma|^{-(\frac{v_0+N+d}{2}+1)}exp\left(-\frac{1}{2}\Big[\underbrace{tr(\Lambda_0\Sigma^{-1})+\kappa_0(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)+\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}_{A}\Big]\right)
\end{aligned}
$$

The derivation of A is the same as that in Normal-Wishart prior, hence

$$
\begin{aligned}
A &= (N+\kappa_0)(\mu-\frac{\kappa_0\mu_0+N\overline{x}}{N+\kappa_0})^T\Sigma^{-1}(\mu-\frac{\kappa_0\mu_0+N\overline{x}}{N+\kappa_0}) \\
&\quad + tr\Big(\big[\underbrace{\sum_{i=1}^{N}(x_i-\overline{x})(x_i-\overline{x})^T}_{S}+\frac{\kappa_0 N}{N+\kappa_0}(\overline{x}-\mu_0)(\overline{x}-\mu_0)^T+\Lambda_0\big]\Sigma^{-1}\Big)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
p(\mu,\Sigma|X) \propto \Sigma|^{-(\frac{v_0+N+d}{2}+1)} exp\Big(-\frac{1}{2}\Big[(N+\kappa_0)(\mu - \frac{\kappa_0\mu_0+N\overline{x}}{N+\kappa_0})^T\Sigma^{-1}(\mu - \frac{\kappa_0\mu_0+N\overline{x}}{N+\kappa_0}) \\
tr\big([S+\frac{\kappa_0 N}{N+\kappa_0}(\overline{x}-\mu_0)(\overline{x}-\mu_0)^T+\Lambda_0]\Sigma^{-1}\big)\Big]\Big)
\end{aligned}
$$

We recognize this is an unnormalized Normal-inver-Wishart distribution, therefore

$$
\begin{aligned}
p(\mu,\Sigma|X) &= NIW(\mu,\Sigma|\mu_N,\kappa_N,v_N,\Lambda_N) \\
&= \mathcal{N}(\mu|\mu_N,\frac{1}{\kappa_N}\Sigma)IW_{v_N}(\Sigma|\Lambda_N) \\
\mu_N &= \frac{\kappa_0\mu_0+N\overline{x}}{N+\kappa_0} \\
\kappa_N &= \kappa_0 + N \\
v_N &= v_0 + N \\
\Lambda_N &= \Lambda_0 + S + \frac{\kappa_0 N}{\kappa_0+N}(\overline{x}-\mu_0)(\overline{x}-\mu_0)^T \\
S &= \sum_{i=1}^{N}(x_i-\overline{x})(x_i-\overline{x})^T
\end{aligned}
$$

Marginal likelihood:

$$
\begin{aligned}
p(X) &= \frac{p(X|\mu,\Sigma)p(\mu,\Sigma)}{p(\mu,\Sigma|X)} \\
&= \frac{\mathcal{N}(X|\mu,\Sigma)NIW(\mu,\Sigma|\mu_0,\kappa_0,v_0,\Lambda_0)}{NIW(\mu,\Sigma|\mu_N,\kappa_N,v_N,\Lambda_N)} \\
&= \frac{Z_N}{Z_0}\frac{1}{(2\pi)^{Nd/2}} \\
&= \frac{2^{v_N d/2}\Gamma_d(v_N/2)(2\pi/\kappa_N)^{d/2}}{|\Lambda_N|^{v_N/2}}\frac{|\Lambda_0|^{v_0/2}}{2^{v_0 d/2}\Gamma_d(v_0/2)(2\pi/\kappa_0)^{d/2}}\frac{1}{(2\pi)^{Nd/2}} \\
&= \frac{1}{\pi^{Nd/2}}\frac{\Gamma_d(v_N/2)}{\Gamma_d(v_0/2)}\frac{|\Lambda_0|^{v_0/2}}{|\Lambda_N|^{v_N/2}}\left(\frac{\kappa_0}{\kappa_N}\right)^{d/2}
\end{aligned}
$$

Property of Normal-inverse-Wishart prior:

We have defined the Normal-inverse-Wishart prior as

$$
\begin{aligned}
p(\mu,\Lambda) &= NIW(\mu,\Lambda|\mu_0,\kappa_0,v_0,\Lambda_0) = \mathcal{N}(\mu|\mu_0,\frac{1}{\kappa_0}\Sigma)IW_{v_0}(\Sigma|\Lambda_0) \\
&= \frac{1}{Z}|\Sigma|^{-(\frac{v_0+d}{2}+1)}exp\left(-\frac{1}{2}\Big[tr(\Lambda_0\Sigma^{-1})+\kappa_0(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\Big]\right) \\
Z &= \frac{2^{v_0 d/2}\Gamma_d^{\frac{v_0}{2}}(2\pi/\kappa_0)^{d/2}}{|\Lambda_0|^{v_0/2}}
\end{aligned}
$$

Then its margin distribution is

$$
\begin{aligned}
p(\Lambda) &= IW_{v_0}(\Sigma|\Lambda_0) \\
p(\mu|\Lambda) &= \mathcal{N}(\mu|\mu_0,\frac{1}{\kappa_0}\Sigma)) \\
p(\mu) &= t_{v_0-d+1}(\mu_0,\frac{\Lambda_0}{\kappa_0(v_0-d+1)})
\end{aligned}
$$