Conjugate Bayesian analysis of the Gaussian distribution

Kevin P. Murphy*
murphyk@cs.ubc.ca

Last updated October 3, 2007

1 Introduction

The Gaussian or normal distribution is one of the most widely used in statistics. Estimating its parameters using Bayesian inference and conjugate priors is also widely used. The use of conjugate priors allows all the results to be derived in closed form. Unfortunately, different books use different conventions on how to parameterize the various distributions (e.g., put the prior on the precision or the variance, use an inverse gamma or inverse chi-squared, etc), which can be very confusing for the student. In this report, we summarize all of the most commonly used forms. We provide detailed derivations for some of these results; the rest can be obtained by simple reparameterization. See the appendix for the definition the distributions that are used.

2 Normal prior

Let us consider Bayesian estimation of the mean of a univariate Gaussian, whose variance is assumed to be known. (We discuss the unknown variance case later.)

2.1 Likelihood

Let $D = (x_1, \ldots, x_n)$ be the data. The likelihood is

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

Let us define the empirical mean and variance

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (2)$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad (3)$$

(Note that other authors (e.g., [GCSR04]) define $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.) We can rewrite the term in the exponent as follows

$$\sum_i (x_i - \mu)^2 = \sum_i [(x_i - \overline{x}) - (\mu - \overline{x})]^2 \quad (4)$$

$$= \sum_i (x_i - \overline{x})^2 + \sum_i (\overline{x} - \mu)^2 - 2\sum_i (x_i - \overline{x})(\mu - \overline{x}) \quad (5)$$

$$= ns^2 + n(\overline{x} - \mu)^2 \quad (6)$$

since

$$\sum_i (x_i - \overline{x})(\mu - \overline{x}) = (\mu - \overline{x})\left((\sum_i x_i) - n\overline{x}\right) = (\mu - \overline{x})(n\overline{x} - n\overline{x}) = 0 \quad (7)$$

*Thanks to Hoyt Koepke for proof reading.

Hence

$$p(D|\mu,\sigma^2) = \frac{1}{(2\pi)^{n/2}}\frac{1}{\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\left[ns^2+n(\overline{x}-\mu)^2\right]\right) \quad (8)$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{n/2}\exp\left(-\frac{n}{2\sigma^2}(\overline{x}-\mu)^2\right)\exp\left(-\frac{ns^2}{2\sigma^2}\right) \quad (9)$$

If σ^2 is a constant, we can write this as

$$p(D|\mu) \propto \exp\left(-\frac{n}{2\sigma^2}(\overline{x}-\mu)^2\right) \propto \mathcal{N}(\overline{x}|\mu,\frac{\sigma^2}{n}) \quad (10)$$

since we are free to drop constant factors in the definition of the likelihood. Thus n observations with variance σ^2 and mean $\overline{x}$ is equivalent to 1 observation $x_1 = \overline{x}$ with variance σ^2/n.

2.2 Prior

Since the likelihood has the form

$$p(D|\mu) \propto \exp\left(-\frac{n}{2\sigma^2}(\overline{x}-\mu)^2\right) \propto \mathcal{N}(\overline{x}|\mu,\frac{\sigma^2}{n}) \quad (11)$$

the **natural conjugate prior** has the form

$$p(\mu) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \propto \mathcal{N}(\mu|\mu_0,\sigma_0^2) \quad (12)$$

(Do not confuse σ_0^2, which is the variance of the prior, with σ^2, which is the variance of the observation noise.) (A natural conjugate prior is one that has the same form as the likelihood.)

2.3 Posterior

Hence the posterior is given by

$$p(\mu|D) \propto p(D|\mu,\sigma)p(\mu|\mu_0,\sigma_0^2) \quad (13)$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2\right]\times\exp\left[-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right] \quad (14)$$

$$= \exp\left[\frac{-1}{2\sigma^2}\sum_i(x_i^2+\mu^2-2x_i\mu)+\frac{-1}{2\sigma_0^2}(\mu^2+\mu_0^2-2\mu_0\mu)\right] \quad (15)$$

Since the product of two Gaussians is a Gaussian, we will rewrite this in the form

$$p(\mu|D) \propto \exp\left[-\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)+\mu\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_i x_i}{\sigma^2}\right)-\left(\frac{\mu_0^2}{2\sigma_0^2}+\frac{\sum_i x_i^2}{2\sigma^2}\right)\right] \quad (16)$$

$$\stackrel{\text{def}}{=} \exp\left[-\frac{1}{2\sigma_n^2}(\mu^2-2\mu\mu_n+\mu_n^2)\right] = \exp\left[-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2\right] \quad (17)$$

Matching coefficients of μ^2, we find σ_n^2 is given by

$$\frac{-\mu^2}{2\sigma_n^2} = \frac{-\mu^2}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right) \quad (18)$$

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2}+\frac{n}{\sigma^2} \quad (19)$$

$$\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2} = \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}} \quad (20)$$

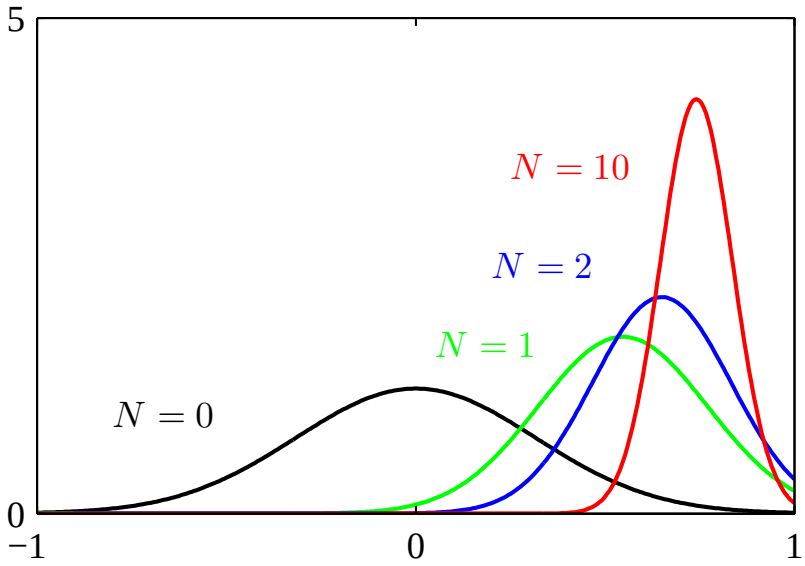


Figure 1: Sequentially updating a Gaussian mean starting with a prior centered on $\mu_0 = 0$. The true parameters are $\mu^* = 0.8$ (unknown), $(\sigma^2)^* = 0.1$ (known). Notice how the data quickly overwhelms the prior, and how the posterior becomes narrower. Source: Figure 2.12 [Bis06].

Matching coefficients of μ we get

$$\frac{-2\mu\mu_n}{-2\sigma_n^2} = \mu\left(\frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) \tag{21}$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \tag{22}$$

$$= \frac{\sigma_0^2 n\overline{x} + \sigma^2\mu_0}{\sigma^2\sigma_0^2} \tag{23}$$

Hence

$$\mu_n = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\overline{x} = \sigma_n^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2}\right) \tag{24}$$

This operation of matching first and second powers of μ is called **completing the square**.

Another way to understand these results is if we work with the **precision** of a Gaussian, which is 1/variance (high precision means low variance, low precision means high variance). Let

$$\lambda = 1/\sigma^2 \tag{25}$$

$$\lambda_0 = 1/\sigma_0^2 \tag{26}$$

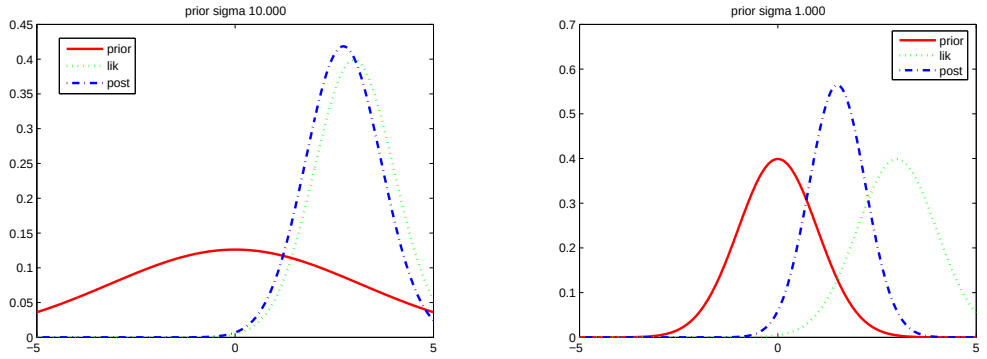
$$\lambda_n = 1/\sigma_n^2 \tag{27}$$

Then we can rewrite the posterior as

$$p(\mu|D,\lambda) = \mathcal{N}(\mu|\mu_n,\lambda_n) \tag{28}$$

$$\lambda_n = \lambda_0 + n\lambda \tag{29}$$

$$\mu_n = \frac{\overline{x}n\lambda + \mu_0\lambda_0}{\lambda_n} = w\mu_{ML} + (1-w)\mu_0 \tag{30}$$



(a) (b)

Figure 2: Bayesian estimation of the mean of a Gaussian from one sample. (a) Weak prior $\mathcal{N}(0, 10)$. (b) Strong prior $\mathcal{N}(0, 1)$. In the latter case, we see the posterior mean is "shrunk" towards the prior mean, which is 0. Figure produced by `gaussBayesDemo`.

where $n\overline{x} = \sum_{i=1}^{n} x_i$ and $w = \frac{n\lambda}{\lambda_n}$. The precision of the posterior λ_n is the precision of the prior λ_0 plus one contribution of data precision λ for each observed data point. Also, we see the mean of the posterior is a convex combination of the prior and the MLE, with weights proportional to the relative precisions.

To gain further insight into these equations, consider the effect of sequentially updating our estimate of μ (see Figure 1). After observing one data point x (so $n = 1$), we have the following posterior mean

$$\mu_1 = \frac{\sigma^2}{\sigma^2 + \sigma_0^2}\mu_0 + \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}x \tag{31}$$

$$= \mu_0 + (x - \mu_0)\frac{\sigma_0^2}{\sigma^2 + \sigma_0^2} \tag{32}$$

$$= x - (x - \mu_0)\frac{\sigma^2}{\sigma^2 + \sigma_0^2} \tag{33}$$

The first equation is a convex combination of the prior and MLE. The second equation is the prior mean ajusted towards the data x. The third equation is the data x adjusted towads the prior mean; this is called **shrinkage**. These are all equivalent ways of expressing the tradeoff between likelihood and prior. See Figure 2 for an example.

2.4 Posterior predictive

The posterior predictive is given by

$$p(x|D) = \int p(x|\mu)p(\mu|D)d\mu \tag{34}$$

$$= \int \mathcal{N}(x|\mu, \sigma^2)\mathcal{N}(\mu|\mu_n, \sigma_n^2)d\mu \tag{35}$$

$$= \mathcal{N}(x|\mu_n, \sigma_n^2 + \sigma^2) \tag{36}$$

This follows from general properties of the Gaussian distribution (see Equation 2.115 of [Bis06]). An alternative proof is to note that

$$x = (x - \mu) + \mu \tag{37}$$

$$x - \mu \sim \mathcal{N}(0, \sigma^2) \tag{38}$$

$$\mu \sim \mathcal{N}(\mu_n, \sigma_n^2) \tag{39}$$

Since $E[X_1 + X_2] = E[X_1] + E[X_2]$ and Var $[X_1 + X_2]$ = Var $[X_1]$ + Var $[X_2]$ if X_1, X_2 are independent, we have

$$X \sim \mathcal{N}(\mu_n, \sigma_n^2 + \sigma^2) \tag{40}$$

since we assume that the residual error is conditionally independent of the parameter. Thus the predictive variance is the uncertainty due to the observation noise σ^2 plus the uncertainty due to the parameters, σ_n^2.

2.5 Marginal likelihood

Writing $m = \mu_0$ and $\tau^2 = \sigma_0^2$ for the hyper-parameters, we can derive the marginal likelihood as follows:

$$\ell = p(\mathcal{D}|m, \sigma^2, \tau^2) = \int [\prod_{i=1}^{n} \mathcal{N}(x_i|\mu, \sigma^2)] \mathcal{N}(\mu|m, \tau^2) d\mu \tag{41}$$

$$= \frac{\sigma}{(\sqrt{2\pi}\sigma)^n \sqrt{n\tau^2 + \sigma^2}} \exp\left(-\frac{\sum_i x_i^2}{2\sigma^2} - \frac{m^2}{2\tau^2}\right) \exp\left(\frac{\frac{\tau^2 n^2 \overline{x}^2}{\sigma^2} + \frac{\sigma^2 m^2}{\tau^2} + 2n\overline{x}m}{2(n\tau^2 + \sigma^2)}\right) \tag{42}$$

The proof is below, based on the on the appendix of [DMP+06].

We have

$$\ell = p(\mathcal{D}|m, \sigma^2, \tau^2) = \int [\prod_{i=1}^{n} \mathcal{N}(x_i|\mu, \sigma^2)] \mathcal{N}(\mu|m, \tau^2) d\mu \tag{43}$$

$$= \frac{1}{(\sigma\sqrt{2\pi})^n (\tau\sqrt{2\pi})} \int \exp\left(-\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2 - \frac{1}{2\tau^2}(\mu - m)^2\right) d\mu \tag{44}$$

Let us define $S^2 = 1/\sigma^2$ and $T^2 = 1/\tau^2$. Then

$$\ell = \frac{1}{(\sqrt{2\pi}/S)^n (\sqrt{2\pi}/T)} \int \exp\left(-\frac{S^2}{2}(\sum_i x_i^2 + n\mu^2 - 2\mu \sum_i x_i) - \frac{T^2}{2}(\mu^2 + m^2 - 2\mu m)\right) d\mu \tag{45}$$

$$= c \int \exp\left(-\tfrac{1}{2}(S^2 n\mu^2 - 2S^2 \sum_i x_i + T^2\mu^2 - 2T^2\mu m)\right) d\mu \tag{46}$$

where

$$c = \frac{\exp\left(-\frac{1}{2}(S^2 \sum_i x_i^2 + T^2 m^2)\right)}{(\sqrt{2\pi}/S)^n (\sqrt{2\pi}/T)} \tag{47}$$

So

$$\ell = c \int \exp\left[-\tfrac{1}{2}(S^2 n + T^2)\left(\mu^2 - 2\mu \frac{S^2 \sum_i x_i + T^2 m}{S^2 n + T^2}\right)\right] d\mu \tag{48}$$

$$= c \exp\left(\frac{(S^2 n\overline{x} + T^2 m)^2}{2(S^2 n + T^2)}\right) \int \exp\left[-\tfrac{1}{2}(S^2 n + T^2)\left(\mu - \frac{S^2 n\overline{x} + T^2 m}{S^2 n + T^2}\right)^2\right] d\mu \tag{49}$$

$$= c \exp\left(\frac{(S^2 n\overline{x} + T^2 m)^2}{2(S^2 n + T^2)}\right) \frac{\sqrt{2\pi}}{\sqrt{S^2 n + T^2}} \tag{50}$$

$$= \frac{\exp\left(-\frac{1}{2}(S^2 \sum_i x_i^2 + T^2 m^2)\right)}{(\sqrt{2\pi}/S)^n (\sqrt{2\pi}/T)} \exp\left(\frac{(S^2 n\overline{x} + T^2 m)^2}{2(S^2 n + T^2)}\right) \frac{\sqrt{2\pi}}{\sqrt{S^2 n + T^2}} \tag{51}$$

Now

$$\frac{1}{\sqrt{(2\pi)/T}} \frac{\sqrt{2\pi}}{\sqrt{S^2 n + T^2}} = \frac{\sigma}{\sqrt{N\tau^2 + \sigma^2}} \tag{52}$$

and

$$\frac{(\frac{n\overline{x}}{\sigma^2} + \frac{m}{\tau^2})^2}{2(\frac{n}{\sigma^2} + \frac{1}{\tau^2})} = \frac{(n\overline{x}\tau^2 + m\sigma^2)^2}{2\sigma^2\tau^2(n\tau^2 + \sigma^2)} \tag{53}$$

$$= \frac{n^2\overline{x}^2\tau^2/\sigma^2 + \sigma^2 m^2/\tau^2 + 2n\overline{x}m}{2(n\tau^2 + \sigma^2)} \tag{54}$$

So

$$p(\mathcal{D}) = \frac{\sigma}{(\sqrt{2\pi}\sigma)^n\sqrt{n\tau^2+\sigma^2}} \exp\left(-\frac{\sum_i x_i^2}{2\sigma^2} - \frac{m^2}{2\tau^2}\right) \exp\left(\frac{\frac{\tau^2 n^2 \overline{x}^2}{\sigma^2} + \frac{\sigma^2 m^2}{\tau^2} + 2n\overline{x}m}{2(n\tau^2+\sigma^2)}\right) \quad (55)$$

To check this, we should ensure that we get

$$p(x|D) = \frac{p(x,D)}{p(D)} = \mathcal{N}(x|\mu_n, \sigma_n^2 + \sigma^2) \quad (56)$$

(To be completed)

2.6 Conditional prior $p(\mu|\sigma^2)$

Note that the previous prior is not, strictly speaking, conjugate, since it has the form $p(\mu)$ whereas the posterior has the form $p(\mu|D,\sigma)$, i.e., σ occurs in the posterior but not the prior. We can rewrite the prior in conditional form as follows

$$p(\mu|\sigma) = \mathcal{N}(\mu|\mu_0, \sigma^2/\kappa_0) \quad (57)$$

This means that if σ^2 is large, the variance on the prior of μ is also large. This is reasonable since σ^2 defines the measurement scale of x, so the prior belief about μ is equivalent to κ_0 observations of μ_0 on this scale. (Hence a noninformative prior is $\kappa_0 = 0$.) Then the posterior is

$$p(\mu|D) = \mathcal{N}(\mu|\mu_n, \sigma^2/\kappa_n) \quad (58)$$

where $\kappa_n = \kappa_0 + n$. In this form, it is clear that κ_0 plays a role analogous to n. Hence κ_0 is the **equivalent sample size** of the prior.

2.7 Reference analysis

To get an uninformative prior, we just set the prior variance to infinity to simulate a uniform prior on μ.

$$p(\mu) \propto 1 = \mathcal{N}(\mu|\cdot, \infty) \quad (59)$$

$$p(\mu|D) = \mathcal{N}(\mu|\overline{x}, \sigma^2/n) \quad (60)$$

3 Normal-Gamma prior

We will now suppose that both the mean m and the precision $\lambda = \sigma^{-2}$ are unknown. We will mostly follow the notation in [DeG70, p169].

3.1 Likelihood

The likelihood can be written in this form

$$p(D|\mu,\lambda) = \frac{1}{(2\pi)^{n/2}}\lambda^{n/2}\exp\left(-\frac{\lambda}{2}\sum_{i=1}^n (x_i-\mu)^2\right) \quad (61)$$

$$= \frac{1}{(2\pi)^{n/2}}\lambda^{n/2}\exp\left(-\frac{\lambda}{2}\left[n(\mu-\overline{x})^2 + \sum_{i=1}^n (x_i-\overline{x})^2\right]\right) \quad (62)$$

3.2 Prior

The conjugate prior is the **normal-Gamma**:

$$NG(\mu,\lambda|\mu_0,\kappa_0,\alpha_0,\beta_0) \stackrel{\text{def}}{=} \mathcal{N}(\mu|\mu_0,(\kappa_0\lambda)^{-1})Ga(\lambda|\alpha_0,\text{rate}=\beta_0) \quad (63)$$

$$= \frac{1}{Z_{NG}(\mu_0,\kappa_0,\alpha_0,\beta_0)}\lambda^{\frac{1}{2}}\exp(-\frac{\kappa_0\lambda}{2}(\mu-\mu_0)^2)\lambda^{\alpha_0-1}e^{-\lambda\beta_0} \quad (64)$$

$$= \frac{1}{Z_{NG}}\lambda^{\alpha_0-\frac{1}{2}}\exp\left(-\frac{\lambda}{2}\left[\kappa_0(\mu-\mu_0)^2+2\beta_0\right]\right) \quad (65)$$

$$Z_{NG}(\mu_0,\kappa_0,\alpha_0,\beta_0) = \frac{\Gamma(\alpha_0)}{\beta_0^{\alpha_0}}\left(\frac{2\pi}{\kappa_0}\right)^{\frac{1}{2}} \quad (66)$$

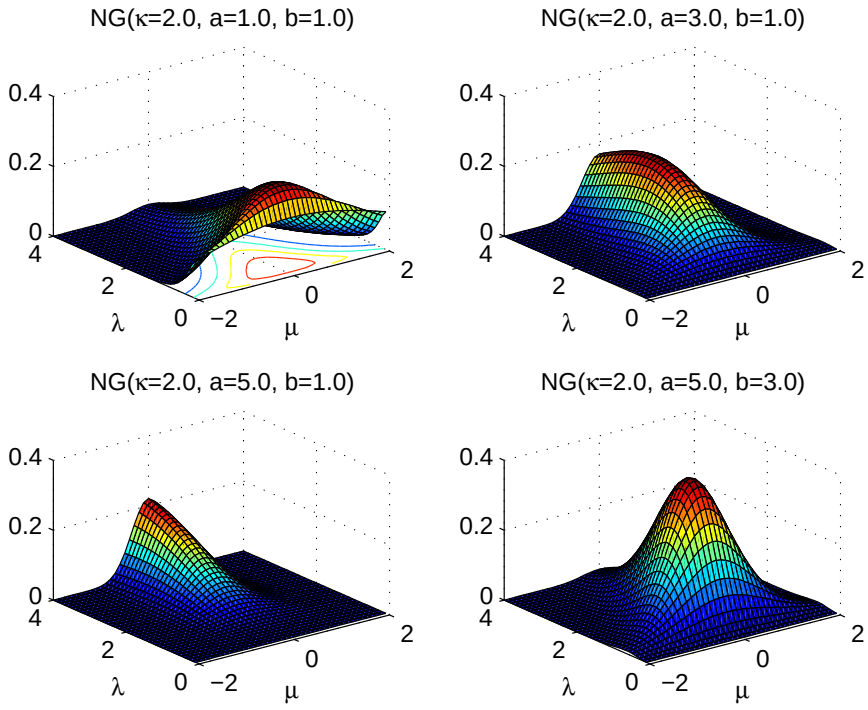


Figure 3: Some Normal-Gamma distributions. Produced by `NGplot2`.

See Figure 3 for some plots.

We can compute the prior marginal on μ as follows:

$$p(\mu) \propto \int_0^\infty p(\mu, \lambda) d\lambda \tag{67}$$

$$= \int_0^\infty \lambda^{\alpha_0 + \frac{1}{2} - 1} \exp\left(-\lambda(\beta_0 + \frac{\kappa_0(\mu - \mu_0)^2}{2})\right) d\lambda \tag{68}$$

We recognize this as an unnormalized $Ga(a = \alpha_0 + \frac{1}{2}, b = \beta_0 + \frac{\kappa_0(\mu-\mu_0)^2}{2})$ distribution, so we can just write down

$$p(\mu) \propto \frac{\Gamma(a)}{b^a} \tag{69}$$

$$\propto b^{-a} \tag{70}$$

$$= (\beta_0 + \frac{\kappa_0}{2}(\mu - \mu_0)^2)^{-\alpha_0 - \frac{1}{2}} \tag{71}$$

$$= (1 + \frac{1}{2\alpha_0} \frac{\alpha_0 \kappa_0 (\mu - \mu_0)^2}{\beta_0})^{-(2\alpha_0+1)/2} \tag{72}$$

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

3.3 Posterior

The posterior can be derived as follows.

$$
\begin{aligned}
p(\mu, \lambda|D) &\propto NG(\mu, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0) p(D|\mu, \lambda) && (73)\\
&\propto \lambda^{\frac{1}{2}} e^{-(\kappa_0 \lambda(\mu-\mu_0)^2)/2} \lambda^{\alpha_0 - 1} e^{-\beta_0 \lambda} \times \lambda^{n/2} e^{-\frac{\lambda}{2}\sum_{i=1}^n (x_i-\mu)^2} && (74)\\
&\propto \lambda^{\frac{1}{2}} \lambda^{\alpha_0 + n/2 - 1} e^{-\beta_0 \lambda} e^{-(\lambda/2)[\kappa_0(\mu-\mu_0)^2 + \sum_i (x_i-\mu)^2]} && (75)
\end{aligned}
$$

From Equation 6 we have

$$
\sum_{i=1}^n (x_i - \mu)^2 = n(\mu - \overline{x})^2 + \sum_{i=1}^n (x_i - \overline{x})^2 \tag{76}
$$

Also, it can be shown that

$$
\kappa_0(\mu - \mu_0)^2 + n(\mu - \overline{x})^2 = (\kappa_0 + n)(\mu - \mu_n)^2 + \frac{\kappa_0 n (\overline{x} - \mu_0)^2}{\kappa_0 + n} \tag{77}
$$

where

$$
\mu_n = \frac{\kappa_0 \mu_0 + n\overline{x}}{\kappa_0 + n} \tag{78}
$$

Hence

$$
\begin{aligned}
\kappa_0(\mu - \mu_0)^2 + \sum_i (x_i - \mu)^2 &= \kappa_0(\mu - \mu_0)^2 + n(\mu - \overline{x})^2 + \sum_i (x_i - \overline{x})^2 && (79)\\
&= (\kappa_0 + n)(\mu - \mu_n)^2 + \frac{\kappa_0 n (\overline{x} - \mu_0)^2}{\kappa_0 + n} + \sum_i (x_i - \overline{x})^2 && (80)
\end{aligned}
$$

So

$$
\begin{aligned}
p(\mu, \lambda|D) &\propto \lambda^{\frac{1}{2}} e^{-(\lambda/2)(\kappa_0 + n)(\mu - \mu_n)^2} && (81)\\
&\quad \times \lambda^{\alpha_0 + n/2 - 1} e^{-\beta_0 \lambda} e^{-(\lambda/2)\sum_i (x_i - \overline{x})^2} e^{-(\lambda/2)\frac{\kappa_0 n (\overline{x} - \mu_0)^2}{\kappa_0 + n}} && (82)\\
&\propto \mathcal{N}(\mu|\mu_n, ((\kappa + n)\lambda)^{-1}) \times Ga(\lambda|\alpha_0 + n/2, \beta_n) && (83)
\end{aligned}
$$

where

$$
\beta_n = \beta_0 + \tfrac{1}{2}\sum_{i=1}^n (x_i - \overline{x})^2 + \frac{\kappa_0 n (\overline{x} - \mu_0)^2}{2(\kappa_0 + n)} \tag{84}
$$

In summary,

$$
\begin{aligned}
p(\mu, \lambda|D) &= NG(\mu, \lambda|\mu_n, \kappa_n, \alpha_n, \beta_n) && (85)\\
\mu_n &= \frac{\kappa_0 \mu_0 + n\overline{x}}{\kappa_0 + n} && (86)\\
\kappa_n &= \kappa_0 + n && (87)\\
\alpha_n &= \alpha_0 + n/2 && (88)\\
\beta_n &= \beta_0 + \tfrac{1}{2}\sum_{i=1}^n (x_i - \overline{x})^2 + \frac{\kappa_0 n (\overline{x} - \mu_0)^2}{2(\kappa_0 + n)} && (89)
\end{aligned}
$$

We see that the posterior sum of squares, β_n, combines the prior sum of squares, β_0, the sample sum of squares, $\sum_i (x_i - \overline{x})^2$, and a term due to the discrepancy between the prior mean and sample mean. As can be seen from Figure 3, the range of probable values for μ and σ^2 can be quite large even after for moderate n. Keep this picture in mind whenever someones claims to have "fit a Gaussian" to their data.

3.3.1 Posterior marginals

The posterior marginals are (using Equation 72)

$$p(\lambda|D) = Ga(\lambda|\alpha_n, \beta_n) \quad (90)$$

$$p(\mu|D) = T_{2\alpha_n}(\mu|\mu_n, \beta_n/(\alpha_n\kappa_n)) \quad (91)$$

3.4 Marginal likelihood

To derive the marginal likelihood, we just dererive the posterior, but this time we keep track of all the constant factors. Let $NG'(\mu, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0)$ denote an unnormalized Normal-Gamma distribution, and let $Z_0 = Z_{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0)$ be the normalization constant of the prior; similarly let Z_n be the normalization constant of the posterior. Let $N'(x_i|\mu, \lambda)$ denote an unnormalized Gaussian with normalization constant $1/\sqrt{2\pi}$. Then

$$p(\mu, \lambda|D) = \frac{1}{p(D)}\frac{1}{Z_0}NG'(\mu, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0)\left(\frac{1}{2\pi}\right)^{n/2}\prod_i N'(x_i|\mu, \lambda) \quad (92)$$

The NG' and N' terms combine to make the posterior NG':

$$p(\mu, \lambda|D) = \frac{1}{Z_n}NG'(\mu, \lambda|\mu_n, \kappa_n, \alpha_n, \beta_n) \quad (93)$$

Hence

$$p(D) = \frac{Z_n}{Z_0}(2\pi)^{-n/2} \quad (94)$$

$$= \frac{\Gamma(\alpha_n)}{\Gamma(\alpha_0)}\frac{\beta_0^{\alpha_0}}{\beta_n^{\alpha_n}}(\frac{\kappa_0}{\kappa_n})^{\frac{1}{2}}(2\pi)^{-n/2} \quad (95)$$

3.5 Posterior predictive

The posterior predictive for m new observations is given by

$$p(D_{new}|D) = \frac{p(D_{new}, D)}{p(D)} \quad (96)$$

$$= \frac{Z_{n+m}}{Z_0}(2\pi)^{-(n+m)/2}\frac{Z_0}{Z_n}(2\pi)^{n/2} \quad (97)$$

$$= \frac{Z_{n+m}}{Z_n}(2\pi)^{-m/2} \quad (98)$$

$$= \frac{\Gamma(\alpha_{n+m})}{\Gamma(\alpha_n)}\frac{\beta_n^{\alpha_n}}{\beta_{n+m}^{\alpha_{n+m}}}\left(\frac{\kappa_n}{\kappa_{n+m}}\right)^{\frac{1}{2}}(2\pi)^{-m/2} \quad (99)$$

In the special case that $m = 1$, it can be shown (see below) that this is a T-distribution

$$p(x|D) = t_{2\alpha_n}(x|\mu_n, \frac{\beta_n(\kappa_n + 1)}{\alpha_n\kappa_n}) \quad (100)$$

To derive the $m = 1$ result, we proceed as follows. (This proof is by Xiang Xuan, and is based on [GH94, p10].) When $m = 1$, the posterior parameters are

$$\alpha_{n+1} = \alpha_n + 1/2 \quad (101)$$

$$\kappa_{n+1} = \kappa_n + 1 \quad (102)$$

$$\beta_{n+1} = \beta_n + \frac{1}{2}\sum_{i=1}^{1}(x_i - \bar{x})^2 + \frac{\kappa_n(\bar{x} - \mu_n)^2}{2(\kappa_n + 1)} \quad (103)$$

Use the fact that when $m = 1$, we have $x_1 = \bar{x}$ (since there is only one observation), hence we have $\frac{1}{2}\sum_{i=1}^{1}(x_i - \bar{x})^2 = 0$. Let's use x denote D_{new}, then β_{n+1} is

$$\beta_{n+1} = \beta_n + \frac{\kappa_n(x-\mu_n)^2}{2(\kappa_n+1)} \tag{104}$$

Substituting, we have the following,

$$p(D_{new}|D) = \frac{\Gamma(\alpha_{n+1})}{\Gamma(\alpha_n)}\frac{\beta_n^{\alpha_n}}{\beta_{n+1}^{\alpha_{n+1}}}\left(\frac{\kappa_n}{\kappa_{n+1}}\right)^{\frac{1}{2}}(2\pi)^{-1/2} \tag{105}$$

$$= \frac{\Gamma(\alpha_n+1/2)}{\Gamma(\alpha_n)}\frac{\beta_n^{\alpha_n}}{(\beta_n + \frac{\kappa_n(x-\mu_n)^2}{2(\kappa_n+1)})^{\alpha_n+1/2}}\left(\frac{\kappa_n}{\kappa_n+1}\right)^{\frac{1}{2}}(2\pi)^{-1/2} \tag{106}$$

$$= \frac{\Gamma((2\alpha_n+1)/2)}{\Gamma((2\alpha_n)/2)}\left(\frac{\beta_n}{\beta_n + \frac{\kappa_n(x-\mu_n)^2}{2(\kappa_n+1)}}\right)^{\alpha_n+1/2}\frac{1}{\beta_n^{\frac{1}{2}}}\left(\frac{\kappa_n}{2(\kappa_n+1)}\right)^{\frac{1}{2}}\pi)^{-1/2} \tag{107}$$

$$= \frac{\Gamma((2\alpha_n+1)/2)}{\Gamma((2\alpha_n)/2)}\left(\frac{1}{1 + \frac{\kappa_n(x-\mu_n)^2}{2\beta_n(\kappa_n+1)}}\right)^{\alpha_n+1/2}\left(\frac{\kappa_n}{2\beta_n(\kappa_n+1)}\right)^{\frac{1}{2}}(\pi)^{-1/2} \tag{108}$$

$$= (\pi)^{-1/2}\frac{\Gamma((2\alpha_n+1)/2)}{\Gamma((2\alpha_n)/2)}\left(\frac{\alpha_n\kappa_n}{2\alpha_n\beta_n(\kappa_n+1)}\right)^{\frac{1}{2}}\left(1 + \frac{\alpha_n\kappa_n(x-\mu_n)^2}{2\alpha_n\beta_n(\kappa_n+1)}\right)^{-(2\alpha_n+1)/2} \tag{109}$$

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

$$p(D_{new}|D) = (\pi)^{-1/2}\frac{\Gamma((2\alpha_n+1)/2)}{\Gamma((2\alpha_n)/2)}\left(\frac{\Lambda}{2\alpha_n}\right)^{\frac{1}{2}}\left(1 + \frac{\Lambda(x-\mu_n)^2}{2\alpha_n}\right)^{-(2\alpha_n+1)/2} \tag{110}$$

We can see this is a T-distribution with center at μ_n, precision $\Lambda = \frac{\alpha_n\kappa_n}{\beta_n(\kappa_n+1)}$, and degree of freedom $2\alpha_n$.

3.6 Reference analysis

The reference prior for NG is

$$p(m,\lambda) \propto \lambda^{-1} = NG(m,\lambda|\mu=\cdot,\kappa=0,\alpha=-\tfrac{1}{2},\beta=0) \tag{111}$$

So the posterior is

$$p(m,\lambda|D) = NG(\mu_n=\overline{x},\kappa_n=n,\alpha_n=(n-1)/2,\beta_n=\tfrac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2) \tag{112}$$

So the posterior marginal of the mean is

$$p(m|D) = t_{n-1}(m|\overline{x},\frac{\sum_i(x_i-\overline{x})^2}{n(n-1)}) \tag{113}$$

which corresponds to the frequentist sampling distribution of the MLE $\hat{\mu}$. Thus *in this case*, the confidence interval and credible interval coincide.

4 Gamma prior

If μ is known, and only λ is unknown (e.g., when implementing Gibbs sampling), we can use the following results, which can be derived by simplifying the results for the Normal-NG model.

4.1 Likelihood

$$p(D|\lambda) \propto \lambda^{n/2} \exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \tag{114}$$

4.2 Prior

$$p(\lambda) = Ga(\lambda|\alpha,\beta) \propto \lambda^{\alpha-1}e^{-\lambda\beta} \tag{115}$$

4.3 Posterior

$$p(\lambda|D) = Ga(\lambda|\alpha_n,\beta_n) \tag{116}$$
$$\alpha_n = \alpha + n/2 \tag{117}$$
$$\beta_n = \beta + \tfrac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{118}$$

4.4 Marginal likelihood

To be completed.

4.5 Posterior predictive

$$p(x|D) = t_{2\alpha_n}(x|\mu,\sigma^2=\beta_n/\alpha_n) \tag{119}$$

4.6 Reference analysis

$$p(\lambda) \propto \lambda^{-1} = Ga(\lambda|0,0) \tag{120}$$
$$p(\lambda|D) = Ga(\lambda|n/2,\tfrac{1}{2}\sum_{i=1}^{m}(x_i-\mu)^2) \tag{121}$$

5 Normal-inverse-chi-squared (NIX) prior

We will see that the natural conjugate prior for σ^2 is the inverse-chi-squared distribution.

5.1 Likelihood

The likelihood can be written in this form

$$p(D|\mu,\sigma^2) = \frac{1}{(2\pi)^{n/2}}(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left[n\sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\mu)^2\right]\right) \tag{122}$$

5.2 Prior

The normal-inverse-chi-squared prior is

$$p(\mu,\sigma^2) = NI\chi^2(\mu_0,\kappa_0,\nu_0,\sigma_0^2) \tag{123}$$
$$= \mathcal{N}(\mu|\mu_0,\sigma^2/\kappa_0)\times\chi^{-2}(\sigma^2|\nu_0,\sigma_0^2) \tag{124}$$
$$= \frac{1}{Z_p(\mu_0,\kappa_0,\nu_0,\sigma_0^2)}\sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)}\exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2+\kappa_0(\mu_0-\mu)^2]\right) \tag{125}$$
$$Z_p(\mu_0,\kappa_0,\nu_0,\sigma_0^2) = \frac{\sqrt{(}2\pi)}{\sqrt{\kappa}_0}\Gamma(\nu_0/2)\left(\frac{2}{\nu_0\sigma_0^2}\right)^{\nu_0/2} \tag{126}$$

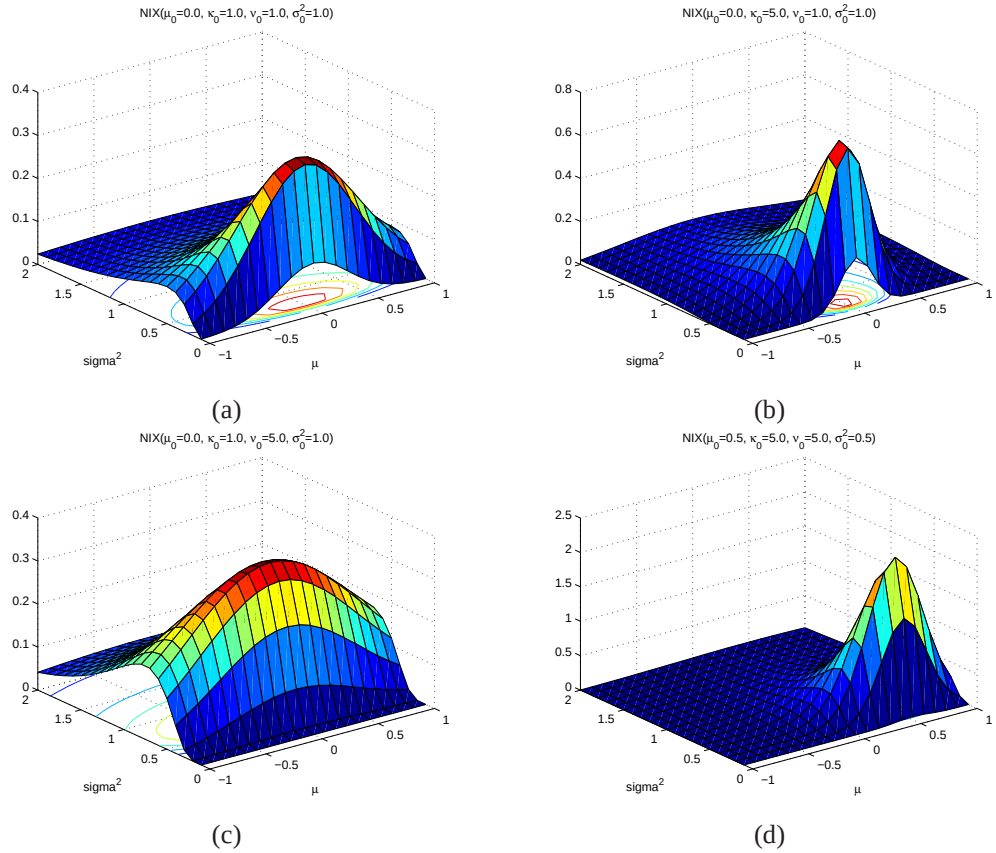


Figure 4: The $NI\chi^2(\mu_0, \kappa_0, \nu_0, \sigma_0^2)$ distribution. μ_0 is the prior mean and κ_0 is how strongly we believe this; σ_0^2 is the prior variance and ν_0 is how strongly we believe this. (a) $\mu_0 = 0, \kappa_0 = 1, \nu_0 = 1, \sigma_0^2 = 1$. Notice that the contour plot (underneath the surface) is shaped like a "squashed egg". (b) We increase the strenght of our belief in the mean, so it gets narrower: $\mu_0 = 0, \kappa_0 = 5, \nu_0 = 1, \sigma_0^2 = 1$. (c) We increase the strenght of our belief in the variance, so it gets narrower: $\mu_0 = 0, \kappa_0 = 1, \nu_0 = 5, \sigma_0^2 = 1$. (d) We strongly believe the mean and variance are 0.5: $\mu_0 = 0.5, \kappa_0 = 5, \nu_0 = 5, \sigma_0^2 = 0.5$. These plots were produced with `NIXdemo2`.

See Figure 4 for some plots. The hyperparameters μ_0 and σ^2/κ_0 can be interpreted as the location and scale of μ, and the hyperparameters $\mathbf{u}_0$ and σ_0^2 as the degrees of freedom and scale of σ^2.

For future reference, it is useful to note that the quadratic term in the prior can be written as

$$Q_0(\mu) = S_0 + \kappa_0(\mu - \mu_0)^2 \quad (127)$$

$$= \kappa_0\mu^2 - 2(\kappa_0\mu_0)\mu + (\kappa_0\mu_0^2 + S_0) \quad (128)$$

where $S_0 = \nu_0\sigma_0^2$ is the prior sum of squares.

5.3 Posterior

(The following derivation is based on [Lee04, p67].) The posterior is

$$\begin{aligned} p(\mu,\sigma^2|D) &\propto \mathcal{N}(\mu|\mu_0,\sigma^2/\kappa_0)\chi^{-2}(\sigma^2|\nu_0,\sigma_0^2)p(D|\mu,\sigma^2) && (129)\\ &\propto \left[\sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)}\exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2+\kappa_0(\mu_0-\mu)^2]\right)\right] && (130)\\ &\quad\times\left[(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left[ns^2+n(\overline{x}-\mu)^2\right]\right)\right] && (131)\\ &\propto \sigma^{-3}(\sigma^2)^{-(\nu_n/2)}\exp\left(-\frac{1}{2\sigma^2}[\nu_n\sigma_n^2+\kappa_n(\mu_n-\mu)^2]\right)=NI\chi^2(\mu_n,\kappa_n,\nu_n,\sigma_n^2) && (132)\end{aligned}$$

Matching powers of σ^2, we find

$$\nu_n = \nu_0+n \quad (133)$$

To derive the other terms, we will complete the square. Let $S_0=\nu_0\sigma_0^2$ and $S_n=\nu_n\sigma_n^2$ for brevity. Grouping the terms inside the exponential, we have

$$S_0+\kappa_0(\mu_0-\mu)^2+ns^2+n(\overline{x}-\mu)^2 = (S_0+\kappa_0\mu_0^2+ns^2+n\overline{x}^2)+\mu^2(\kappa_0+n)-2(\kappa_0\mu_0+n\overline{x})\mu \quad (134)$$

Comparing to Equation 128, we have

$$\begin{aligned}\kappa_n &= \kappa_0+n && (135)\\ \kappa_n\mu_n &= \kappa_0\mu_0+n\overline{x} && (136)\\ S_n+\kappa_n\mu_n^2 &= (S_0+\kappa_0\mu_0^2+ns^2+n\overline{x}^2) && (137)\\ S_n &= S_0+ns^2+\kappa_0\mu_0^2+n\overline{x}^2-\kappa_n\mu_n^2 && (138)\end{aligned}$$

One can rearrange this to get

$$\begin{aligned}S_n &= S_0+ns^2+(\kappa_0^{-1}+n^{-1})^{-1}(\mu_0-\overline{x})^2 && (139)\\ &= S_0+ns^2+\frac{n\kappa_0}{\kappa_0+n}(\mu_0-\overline{x})^2 && (140)\end{aligned}$$

We see that the posterior sum of squares, $S_n=\nu_n\sigma_n^2$, combines the prior sum of squares, $S_0=\nu_0\sigma_0^2$, the sample sum of squares, ns^2, and a term due to the uncertainty in the mean.

In summary,

$$\begin{aligned}\mu_n &= \frac{\kappa_0\mu_0+n\overline{x}}{\kappa_n} && (141)\\ \kappa_n &= \kappa_0+n && (142)\\ \nu_n &= \nu_0+n && (143)\\ \sigma_n^2 &= \frac{1}{\nu_n}(\nu_0\sigma_0^2+\sum_i(x_i-\overline{x})^2+\frac{n\kappa_0}{\kappa_0+n}(\mu_0-\overline{x})^2) && (144)\end{aligned}$$

The posterior mean is given by

$$\begin{aligned}E[\mu|D] &= \mu_n && (145)\\ E[\sigma^2|D] &= \frac{\nu_n}{\nu_n-2}\sigma_n^2 && (146)\end{aligned}$$

The posterior mode is given by (Equation 14 of [BL01]):

$$\begin{aligned}\text{mode}[\mu|D] &= \mu_n && (147)\\ \text{mode}[\sigma^2|D] &= \frac{\nu_n\sigma_n^2}{\nu_n-1} && (148)\end{aligned}$$

The modes of the marginal posterior are

$$\text{mode}[\mu|D] = \mu_n \tag{149}$$

$$\text{mode}[\sigma^2|D] = \frac{\nu_n \sigma_n^2}{\nu_n + 2} \tag{150}$$

5.3.1 Marginal posterior of σ^2

First we integrate out μ, which is just a Gaussian integral.

$$p(\sigma^2|D) = \int p(\sigma^2, \mu|D) d\mu \tag{151}$$

$$\propto \sigma^{-1}(\sigma^2)^{-(\nu_n/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_n\sigma_n^2]\right) \int \exp\left(-\frac{\kappa_n}{2\sigma^2}(\mu_n - \mu)^2]\right) d\mu \tag{152}$$

$$\propto \sigma^{-1}(\sigma^2)^{-(\nu_n/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_n\sigma_n^2]\right) \frac{\sigma\sqrt{(}2\pi)}{\sqrt{\kappa}_n} \tag{153}$$

$$\propto (\sigma^2)^{-(\nu_n/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_n\sigma_n^2]\right) \tag{154}$$

$$= \chi^{-2}(\sigma^2|\nu_n, \sigma_n^2) \tag{155}$$

5.3.2 Marginal posterior of μ

Let us rewrite the posterior as

$$p(\mu, \sigma^2|D) = C\phi^{-\alpha}\phi^{-1} \exp\left(-\frac{1}{2\phi}[\nu_n\sigma_n^2 + \kappa_n(\mu_n - \mu)^2]\right) \tag{156}$$

where $\phi = \sigma^2$ and $\alpha = (\nu_n + 1)/2$. This follows since

$$\sigma^{-1}(\sigma^2)^{-(\nu_n/2+1)} = \sigma^{-1}\sigma^{-\nu_n}\sigma^{-2} = \phi^{-\frac{\nu_n+1}{2}}\phi^{-1} = \phi^{-\alpha-1} \tag{157}$$

Now make the substitutions

$$A = \nu_n\sigma_n^2 + \kappa_n(\mu_n - \mu)^2 \tag{158}$$

$$x = \frac{A}{2\phi} \tag{159}$$

$$\frac{d\phi}{dx} = -\frac{A}{2}x^{-2} \tag{160}$$

so

$$p(\mu|D) = \int C\phi^{-(\alpha+1)}e^{-A/2\phi}d\phi \tag{161}$$

$$= -(A/2)\int C(\frac{A}{2x})^{-(\alpha+1)}e^{-x}x^{-2}dx \tag{162}$$

$$\propto A^{-\alpha}\int x^{\alpha-1}e^{-x}dx \tag{163}$$

$$\propto A^{-\alpha} \tag{164}$$

$$= (\nu_n\sigma_n^2 + \kappa_n(\mu_n - \mu)^2)^{-(\nu_n+1)/2} \tag{165}$$

$$\propto \left[1 + \frac{\kappa_n}{\nu_n\sigma_n^2}(\mu - \mu_n)^2\right]^{-(\nu_n+1)/2} \tag{166}$$

$$\propto t_{\nu_n}(\mu|\mu_n, \sigma_n^2/\kappa_n) \tag{167}$$

5.4 Marginal likelihood

Repeating the derivation of the posterior, but keeping track of the normalization constants, gives the following.

$$
\begin{aligned}
p(D) &= \int\int P(D|\mu,\sigma^2)P(\mu,\sigma^2)d\mu d\sigma^2 && (168)\\
&= \frac{Z_p(\mu_n,\kappa_n,\nu_n,\sigma_n^2)}{Z_p(\mu_0,\kappa_0,\nu_0,\sigma_0^2)}\frac{1}{Z_l^N} && (169)\\
&= \frac{\sqrt{\kappa_0}}{\sqrt{\kappa_n}}\frac{\Gamma(\nu_n/2)}{\Gamma(\nu_0/2)}\left(\frac{\nu_0\sigma_0^2}{2}\right)^{\nu_0/2}\left(\frac{2}{\nu_n\sigma_n^2}\right)^{\nu_n/2}\frac{1}{(2\pi)^{(n/2)}} && (170)\\
&= \frac{\Gamma(\nu_n/2)}{\Gamma(\nu_0/2)}\sqrt{\frac{\kappa_0}{\kappa_n}}\frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_n\sigma_n^2)^{\nu_n/2}}\frac{1}{\pi^{n/2}} && (171)
\end{aligned}
$$

5.5 Posterior predictive

$$
\begin{aligned}
p(x|D) &= \int\int p(x|\mu,\sigma^2)p(\mu,\sigma^2|D)d\mu d\sigma^2 && (172)\\
&= \frac{p(x,D)}{p(D)} && (173)\\
&= \frac{\Gamma((\nu_n+1)/2)}{\Gamma(\nu_n/2)}\sqrt{\frac{\kappa_n}{\kappa_n+1}}\frac{(\nu_n\sigma_n^2)^{\nu_n/2}}{(\nu_n\sigma_n^2+\frac{\kappa_n}{\kappa_n+1}(x-\mu_n)^2))^{(\nu_n+1)/2}}\frac{1}{\pi^{1/2}} && (174)\\
&= \frac{\Gamma((\nu_n+1)/2)}{\Gamma(\nu_n/2)}\left(\frac{\kappa_n}{(\kappa_n+1)\pi\nu_n\sigma_n^2}\right)^{\frac{1}{2}}\left(1+\frac{\kappa_n(x-\mu_n)^2}{(\kappa_n+1)\nu_n\sigma_n^2}\right)^{-(\nu_n+1)/2} && (175)\\
&= t_{\nu_n}(\mu_n,\frac{(1+\kappa_n)\sigma_n^2}{\kappa_n}) && (176)
\end{aligned}
$$

5.6 Reference analysis

The reference prior is $p(\mu,\sigma^2)\propto(\sigma^2)^{-1}$ which can be modeled by $\kappa_0=0$, $\nu_0=-1$, $\sigma_0=0$, since then we get

$$p(\mu,\sigma^2)\propto\sigma^{-1}(\sigma^2)^{-(-\frac{1}{2}+1)}e^0=\sigma^{-1}(\sigma^2)^{-1/2}=\sigma^{-2} \quad (177)$$

(See also [DeG70, p197] and [GCSR04, p88].)

With the reference prior, the posterior is

$$
\begin{aligned}
\mu_n &= \overline{x} && (178)\\
\nu_n &= n-1 && (179)\\
\kappa_n &= n && (180)\\
\sigma_n^2 &= \frac{\sum_i(x_i-\overline{x})^2}{n-1} && (181)\\
p(\mu,\sigma^2|D) &\propto \sigma^{-n-2}\exp\left(-\frac{1}{2\sigma^2}[\sum_i(x_i-\overline{x})^2+n(\overline{x}-\mu)^2]\right) && (182)
\end{aligned}
$$

The posterior marginals are

$$
\begin{aligned}
p(\sigma^2|D) &= \chi^{-2}(\sigma^2|n-1,\frac{\sum_i(x_i-\overline{x})^2}{n-1}) && (183)\\
p(\mu|D) &= t_{n-1}(\mu|\overline{x},\frac{\sum_i(x_i-\overline{x})^2}{n(n-1)}) && (184)
\end{aligned}
$$

which are very closely related to the sampling distribution of the MLE. The posterior predictive is

$$p(x|D) = t_{n-1}\left(\overline{x}, \frac{(1+n)\sum_i (x_i - \overline{x})^2}{n(n-1)}\right) \tag{185}$$

Note that [Min00] argues that Jeffrey's principle says the uninformative prior should be of the form

$$\lim_{k \longrightarrow 0} \mathcal{N}(\mu|\mu_0, \sigma^2/k)\chi^{-2}(\sigma^2|k, \sigma_0^2) \propto (2\pi\sigma^2)^{-\frac{1}{2}}(\sigma^2)^{-1} \propto \sigma^{-3} \tag{186}$$

This can be achieved by setting $\nu_0 = 0$ instead of $\nu_0 = -1$.

6 Normal-inverse-Gamma (NIG) prior

Another popular parameterization is the following:

$$\begin{aligned} p(\mu, \sigma^2) &= \text{NIG}(m, V, a, b) && (187)\\ &= \mathcal{N}(\mu|m, \sigma^2 V) IG(\sigma^2|a, b) && (188) \end{aligned}$$

6.1 Likelihood

The likelihood can be written in this form

$$p(D|\mu, \sigma^2) = \frac{1}{(2\pi)^{n/2}}(\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left[ns^2 + n(\overline{x} - \mu)^2\right]\right) \tag{189}$$

6.2 Prior

$$\begin{aligned} p(\mu, \sigma^2) &= \text{NIG}(m_0, V_0, a_0, b_0) && (190)\\ &= \mathcal{N}(\mu|m_0, \sigma^2 V_0) IG(\sigma^2|a_0, b_0) && (191) \end{aligned}$$

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

$$\begin{aligned} m_0 &= \mu_0 && (192)\\ V_0 &= \frac{1}{\kappa_0} && (193)\\ a_0 &= \frac{\nu_0}{2} && (194)\\ b_0 &= \frac{\nu_0 \sigma_0^2}{2} && (195) \end{aligned}$$

6.3 Posterior

We can show that the posterior is also NIG:

$$\begin{aligned} p(\mu, \sigma^2|D) &= NIG(m_n, V_n, a_n, b_n) && (196)\\ V_n^{-1} &= V_0^{-1} + n && (197)\\ \frac{m_n}{V_n} &= V_0^{-1} m_0 + n\overline{x} && (198)\\ a_n &= a_0 + n/2 && (199)\\ b_n &= b_0 + \frac{1}{2}[m_0^2 V_0^{-1} + \sum_i x_i^2 - m_n^2 V_n^{-1}] && (200) \end{aligned}$$

The NIG posterior follows directly from the $NI\chi^2$ results using the specified substitutions. (The b_n term requires some tedious algebra...)

6.3.1 Posterior marginals

To be derived.

6.4 Marginal likelihood

For the marginal likelihood, substituting into Equation 171 we have

$$p(D) = \frac{\Gamma(a_n)}{\Gamma(a_0)}\sqrt{\frac{V_n}{V_0}}\frac{(2b_0)^{a_0}}{(2b_n)^{a_n}}\frac{1}{\pi^{n/2}} \quad (201)$$

$$= \frac{|V_n|^{\frac{1}{2}}}{|V_0|^{\frac{1}{2}}}\frac{b_0^{a_0}}{b_n^{a_n}}\frac{\Gamma(a_n)}{\Gamma(a_0)}\frac{1}{\pi^{n/2}}2^{a_0-a_n} \quad (202)$$

$$= \frac{|V_n|^{\frac{1}{2}}}{|V_0|^{\frac{1}{2}}}\frac{b_0^{a_0}}{b_n^{a_n}}\frac{\Gamma(a_n)}{\Gamma(a_0)}\frac{1}{\pi^{n/2}2^n} \quad (203)$$

6.5 Posterior predictive

For the predictive density, substituting into Equation 176 we have

$$\frac{\kappa_n}{(1+\kappa_n)\sigma_n^2} = \frac{1}{(\frac{1}{\kappa_n}+1)\sigma_n^2} \quad (204)$$

$$= \frac{2a_n}{2b_n(1+V_n)} \quad (205)$$

So

$$p(y|D) = t_{2a_n}(m_n, \frac{b_n(1+V_n)}{a_n}) \quad (206)$$

These results follow from [DHMS02, p240] by setting $x = 1$, $\beta = \mu$, $B^TB = n$, $B^TX = n\overline{x}$, $X^TX = \sum_i x_i^2$. Note that we use a difference parameterization of the student-t. Also, our equations for $p(D)$ differ by a 2^{-n} term.

7 Multivariate Normal prior

If we assume Σ is known, then a conjugate analysis of the mean is very simple, since the conjugate prior for the mean is Gaussian, the likelihood is Gaussian, and hence the posterior is Gaussian. The results are analogous to the scalar case. In particular, we use the general result from [Bis06, p92] with the following substitutions:

$$x = \mu, y = \overline{x}, \Lambda^{-1} = \Sigma_0, A = I, b = 0, L^{-1} = \Sigma/N \quad (207)$$

7.1 Prior

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0) \quad (208)$$

7.2 Likelihood

$$p(D|\mu, \Sigma) \propto \mathcal{N}(\overline{x}|\mu, \frac{1}{N}\Sigma) \quad (209)$$

7.3 Posterior

$$p(\mu|D, \Sigma) = \mathcal{N}(\mu|\mu_N, \Sigma_N) \quad (210)$$

$$\Sigma_N = \left(\Sigma_0^{-1} + N\Sigma^{-1}\right)^{-1} \quad (211)$$

$$\mu_N = \Sigma_N\left(N\Sigma^{-1}\overline{x} + \Sigma_0^{-1}\mu_0\right) \quad (212)$$

7.4 Posterior predictive

$$p(x|D) = \mathcal{N}(x|\mu_N, \Sigma + \Sigma_N) \tag{213}$$

7.5 Reference analysis

$$p(\mu) \propto 1 = \mathcal{N}(\mu|\cdot, \infty I) \tag{214}$$
$$p(\mu|D) = \mathcal{N}(\overline{x}, \Sigma/n) \tag{215}$$

8 Normal-Wishart prior

The multivariate analog of the normal-gamma prior is the normal-Wishart prior. Here we just state the results without proof; see [DeG70, p178] for details. We assume X is a d-dimensional.

8.1 Likelihood

$$p(D|\boldsymbol{\mu}, \Lambda) = (2\pi)^{-nd/2}|\Lambda|^{n/2}\exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}(x_i - \boldsymbol{\mu})^T\Lambda(x_i - \boldsymbol{\mu})\right) \tag{216}$$

8.2 Prior

$$p(\boldsymbol{\mu}, \Lambda) = NWi(\boldsymbol{\mu}, \Lambda|\mu_0, \kappa, \nu, T) = \mathcal{N}(\boldsymbol{\mu}|\mu_0, (\kappa\Lambda)^{-1})Wi_\nu(\Lambda|T) \tag{217}$$
$$= \frac{1}{Z}|\Lambda|^{\frac{1}{2}}\exp\left(-\frac{\kappa}{2}(\boldsymbol{\mu} - \mu_0)^T\Lambda(\boldsymbol{\mu} - \mu_0)\right)|\Lambda|^{(\kappa-d-1)/2}\exp\left(-\tfrac{1}{2}\text{tr}(T^{-1}\Lambda)\right) \tag{218}$$
$$Z = \left(\frac{\kappa}{2\pi}\right)^{d/2}|T|^{\kappa/2}2^{d\kappa/2}\Gamma_d(\kappa/2) \tag{219}$$

Here T is the prior covariance. To see the connection to the scalar case, make the substitutions

$$\alpha_0 = \frac{\nu_0}{2}, \beta_0 = \frac{T_0}{2} \tag{220}$$

8.3 Posterior

$$p(\boldsymbol{\mu}, \Lambda|D) = \mathcal{N}(\boldsymbol{\mu}|\mu_n, (\kappa_n\Lambda)^{-1})Wi_{\nu_n}(\Lambda|T_n) \tag{221}$$
$$\mu_n = \frac{\kappa\mu_0 + n\overline{x}}{\kappa + n} \tag{222}$$
$$T_n = T + S + \frac{\kappa n}{\kappa + n}(\mu_0 - \overline{x})(\mu_0 - \overline{x})^T \tag{223}$$
$$S = \sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T \tag{224}$$
$$\nu_n = \nu + n \tag{225}$$
$$\kappa_n = \kappa + n \tag{226}$$

Posterior marginals

$$p(\Lambda|D) = Wi_{\nu_n}(T_n) \tag{227}$$
$$p(\boldsymbol{\mu}|D) = t_{\nu_n - d + 1}(\boldsymbol{\mu}|\mu_n, \frac{T_n}{\kappa_n(\nu_n - d + 1)}) \tag{228}$$

The MAP estimates are given by

$$(\hat{\mu}, \hat{\Lambda}) = \arg\max_{\mu,\Lambda} p(D|\mu, \Lambda) NWi(\mu, \Lambda) \tag{229}$$

$$\hat{\mu} = \sum_{i=1}^{n} x_i + \kappa_0 \mu_0 N + \kappa_0 \tag{230}$$

$$\hat{\Sigma} = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^T + \kappa_0(\hat{\mu} - \mu_0)(\hat{\mu} - \mu_0)^T + T_0^{-1}}{N + \nu_0 - d} \tag{231}$$

This reduces to the MLE if $\kappa_0 = 0$, $\nu_0 = d$ and $|T_0| = 0$.

8.4 Posterior predictive

$$p(x|D) = t_{\nu_n - d + 1}(\mu_n, \frac{T_n(\kappa_n + 1)}{\kappa_n(\nu_n - d + 1)}) \tag{232}$$

If $d = 1$, this reduces to Equation 100.

8.5 Marginal likelihood

This can be computed as a ratio of normalization constants.

$$p(D) = \frac{Z_n}{Z_0} \frac{1}{(2\pi)^{nd/2}} \tag{233}$$

$$= \frac{1}{\pi^{nd/2}} \frac{\Gamma_d(\nu_n/2)}{\Gamma_d(\nu_0/2)} \frac{|T_0|^{\nu_0/2}}{|T_n|^{\nu_n/2}} \left(\frac{\kappa_0}{\kappa_n}\right)^{d/2} \tag{234}$$

This reduces to Equation 95 if $d = 1$.

8.6 Reference analysis

We set

$$\mu_0 = 0, \;\; \kappa_0 = 0, \;\; \nu_0 = -1, \;\; |T_0| = 0 \tag{235}$$

to give

$$p(\boldsymbol{\mu}, \Lambda) \propto |\Lambda|^{-(d+1)/2} \tag{236}$$

Then the posterior parameters become

$$\mu_n = \overline{x}, \;\; T_n = S, \;\; \kappa_n = n, \;\; \nu_n = n - 1 \tag{237}$$

the posterior marginals become

$$p(\boldsymbol{\mu}|D) = t_{n-d}(\boldsymbol{\mu}|\overline{x}, \frac{S}{n(n-d)}) \tag{238}$$

$$p(\Lambda|D) = Wi_{n-d}(\Lambda|S) \tag{239}$$

and the posterior predictive becomes

$$p(x|D) = t_{n-d}(\overline{x}, \frac{S(n+1)}{n(n-d)}) \tag{240}$$

9 Normal-Inverse-Wishart prior

The multivariate analog of the normal inverse chi-squared (NIX) distribution is the normal inverse Wishart (NIW) (see also [GCSR04, p85]).

9.1 Likelihood

The likelihood is

$$p(D|\mu,\Sigma) \propto |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right) \tag{241}$$

$$= |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}tr(\Sigma^{-1}S)\right) \tag{242}$$

$$\tag{243}$$

where S is the matrix of sum of squares (scatter matrix)

$$S = \sum_{i=1}^{N}(x_i-\overline{x})(x_i-\overline{x})^T \tag{244}$$

9.2 Prior

The natural conjugate prior is normal-inverse-wishart

$$\Sigma \sim IW_{\nu_0}(\Lambda_0^{-1}) \tag{245}$$

$$\mu|\Sigma \sim N(\mu_0, \Sigma/\kappa_0) \tag{246}$$

$$p(\mu,\Sigma) \stackrel{\text{def}}{=} NIW(\mu_0,\kappa_0,\Lambda_0,\nu_0) \tag{247}$$

$$= \frac{1}{Z}|\Sigma|^{-((\nu_0+d)/2+1)} \exp\left(-\frac{1}{2}tr(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right) \tag{248}$$

$$Z = \frac{2^{v_0 d/2}\Gamma_d(\nu_0/2)(2\pi/\kappa_0)^{d/2}}{|\Lambda_0|^{\nu_0/2}} \tag{249}$$

9.3 Posterior

The posterior is

$$p(\mu,\Sigma|D,\mu_0,\kappa_0,\Lambda_0,\nu_0) = NIW(\mu,\Sigma|\mu_n,\kappa_n,\Lambda_n,\nu_n) \tag{250}$$

$$\mu_n = \frac{\kappa_0\mu + 0 + n\overline{y}}{\kappa_n} = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\overline{y} \tag{251}$$

$$\kappa_n = \kappa_0 + n \tag{252}$$

$$\nu_n = \nu_0 + n \tag{253}$$

$$\Lambda_n = \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0+n}(\overline{x}-\mu_0)(\overline{x}-\mu_0)^T \tag{254}$$

The marginals are

$$\Sigma|D \sim IW(\Lambda_n^{-1},\nu_n) \tag{255}$$

$$\mu|D = t_{\nu_n-d+1}(\mu_n, \frac{\Lambda_n}{\kappa_n(\nu_n-d+1)}) \tag{256}$$

To see the connection with the scalar case, note that Λ_n plays the role of $\nu_n\sigma_n^2$ (posterior sum of squares), so

$$\frac{\Lambda_n}{\kappa_n(\nu_n-d+1)} = \frac{\Lambda_n}{\kappa_n\nu_n} = \frac{\sigma^2}{\kappa_n} \tag{257}$$

9.4 Posterior predictive

$$p(x|D) = t_{\nu_n - d + 1}(\mu_n, \frac{\Lambda_n(\kappa_n + 1)}{\kappa_n(\nu_n - d + 1)}) \tag{258}$$

To see the connection with the scalar case, note that

$$\frac{\Lambda_n(\kappa_n + 1)}{\kappa_n(\nu_n - d + 1)} = \frac{\Lambda_n(\kappa_n + 1)}{\kappa_n \nu_n} = \frac{\sigma^2(\kappa_n + 1)}{\kappa_n} \tag{259}$$

9.5 Marginal likelihood

The posterior is given by

$$\begin{aligned} p(\mu, \Sigma|D) &= \frac{1}{p(D)} \frac{1}{Z_0} NIW'(\mu, \Sigma|\alpha_0) \frac{1}{(2\pi)^{nd/2}} N'(D|\mu, \Sigma) && (260) \\ &= \frac{1}{Z_n} NIW'(\mu, \Sigma|\alpha_n) && (261) \end{aligned}$$

where

$$\begin{aligned} NIW'(\mu, \Sigma|\alpha_0) &= |\Sigma|^{-((\nu_0+d)/2+1)} \exp\left(-\frac{1}{2} tr(\Lambda_0 \Sigma^{-1}) - \frac{\kappa_0}{2}(\mu - \mu_0)^T \Sigma^{-1}(\mu - \mu_0)\right) && (262) \\ N'(D|\mu, \Sigma) &= |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2} tr(\Sigma^{-1} S)\right) && (263) \end{aligned}$$

is the unnormalized prior and likelihood. Hence

$$\begin{aligned} p(D) &= \frac{Z_n}{Z_0} \frac{1}{(2\pi)^{nd/2}} = \frac{2^{\nu_n d/2} \Gamma_d(\nu_n/2)(2\pi/\kappa_n)^{d/2}}{|\Lambda_n|^{\nu_n/2}} \frac{|\Lambda_0|^{\nu_0/2}}{2^{\nu_0 d/2} \Gamma_d(\nu_0/2)(2\pi/\kappa_0)^{d/2}} \frac{1}{(2\pi)^{nd/2}} && (264) \\ &= \frac{1}{(2\pi)^{nd/2}} \frac{2^{\nu_n d/2}}{2^{\nu_0 d/2}} \frac{(2\pi/\kappa_n)^{d/2}}{(2\pi/\kappa_0)^{d/2}} \frac{\Gamma_d(\nu_n/2)}{\Gamma_d(\nu_0/2)} && (265) \\ &= \frac{1}{\pi^{nd/2}} \frac{\Gamma_d(\nu_n/2)}{\Gamma_d(\nu_0/2)} \frac{|\Lambda_0|^{\nu_0/2}}{|\Lambda_n|^{\nu_n/2}} \left(\frac{\kappa_0}{\kappa_n}\right)^{d/2} && (266) \end{aligned}$$

This reduces to Equation 171 if $d = 1$.

9.6 Reference analysis

A noninformative (Jeffrey's) prior is $p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2}$ which is the limit of $\kappa_0 \to 0$, $\nu_0 \to -1$, $|\Lambda_0| \to 0$ [GCSR04, p88]. Then the posterior becomes

$$\begin{aligned} \mu_n &= \overline{x} && (267) \\ \kappa_n &= n && (268) \\ \nu_n &= n - 1 && (269) \\ \Lambda_n &= S = \sum_i (x_i - \overline{x})(x_i - \overline{x})^T && (270) \\ p(\Sigma|D) &= IW_{n-1}(\Sigma|S) && (271) \\ p(\mu|D) &= t_{n-d}(\mu|\overline{x}, \frac{S}{n(n-d)}) && (272) \\ p(x|D) &= t_{n-d}(x|\overline{x}, \frac{S(n+1)}{n(n-d)}) && (273) \end{aligned}$$

Note that [Min00] argues that Jeffrey's principle says the uninformative prior should be of the form

$$\lim_{k \to 0} \mathcal{N}(\mu|\mu_0, \Sigma/k) IW_k(\Sigma|k\Sigma) \propto |2\pi\Sigma|^{-\frac{1}{2}} |\Sigma|^{-(d+1)/2} \propto |\Sigma|^{-(\frac{d}{2}+1)} \tag{274}$$

This can be achieved by setting $\nu_0 = 0$ instead of $\nu_0 = -1$.

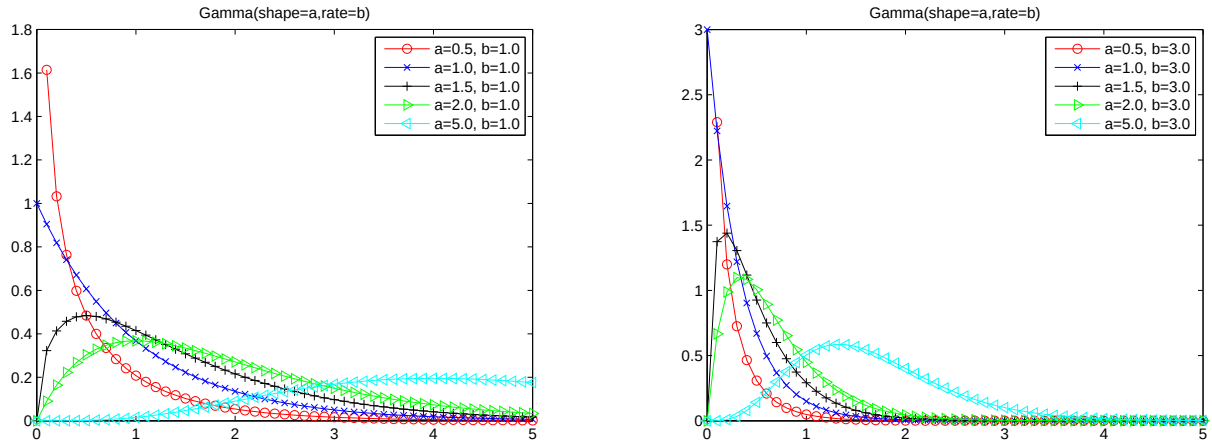


Figure 5: Some $Ga(a, b)$ distributions. If $a < 1$, the peak is at 0. As we increase b, we squeeze everything leftwards and upwards. Figures generated by `gammaDistPlot2`.

10 Appendix: some standard distributions

10.1 Gamma distribution

The gamma distribution is a flexible distribution for positive real valued rv's, $x > 0$. It is defined in terms of two parameters. There are two common parameterizations. This is the one used by Bishop [Bis06] (and many other authors):

$$Ga(x|\text{shape} = a, \text{rate} = b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}, \quad x, a, b > 0 \tag{275}$$

The second parameterization (and the one used by Matlab's `gampdf`) is

$$Ga(x|\text{shape} = \alpha, \text{scale} = \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} = Ga^{rate}(x|\alpha, 1/\beta) \tag{276}$$

Note that the shape parameter controls the shape; the scale parameter merely defines the measurement scale (the horizontal axis). The rate parameter is just the inverse of the scale. See Figure 5 for some examples. This distribution has the following properties (using the rate parameterization):

$$\text{mean} = \frac{a}{b} \tag{277}$$

$$\text{mode} = \frac{a-1}{b} \text{ for } a \geq 1 \tag{278}$$

$$\text{var} = \frac{a}{b^2} \tag{279}$$

10.2 Inverse Gamma distribution

Let $X \sim Ga(\text{shape} = a, \text{rate} = b)$ and $Y = 1/X$. Then it is easy to show that $Y \sim IG(\text{shape} = a, \text{scale} = b)$, where the inverse Gamma distribution is given by

$$IG(x|\text{shape} = a, \text{scale} = b) = \frac{b^a}{\Gamma(a)} x^{-(a+1)} e^{-b/x}, \quad x, a, b > 0 \tag{280}$$

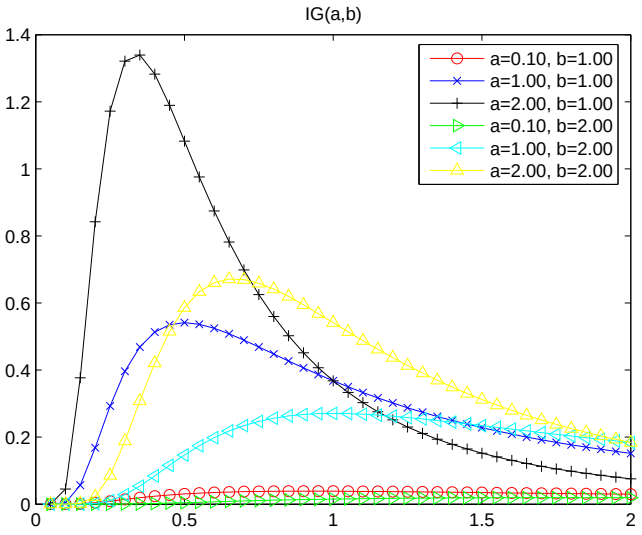


Figure 6: Some inverse gamma distributions (a=shape, b=rate). These plots were produced by `invchi2plot`.

The distribution has these properties

$$\text{mean} = \frac{b}{a-1}, \; a > 1 \tag{281}$$

$$\text{mode} = \frac{b}{a+1} \tag{282}$$

$$\text{var} = \frac{b^2}{(a-1)^2(a-2)}, \; a > 2 \tag{283}$$

See Figure 6 for some plots. We see that increasing b just stretches the horizontal axis, but increasing a moves the peak up and closer to the left.

There is also another parameterization, using the rate (inverse scale):

$$IG(x|\text{shape} = \alpha, \text{rate} = \beta) = \frac{1}{\beta^a}\Gamma(a)x^{-(\alpha+1)}e^{-1/(\beta x)}, \quad x, \alpha, \beta > 0 \tag{284}$$

10.3 Scaled Inverse-Chi-squared

The scaled inverse-chi-squared distribution is a reparameterization of the inverse Gamma [GCSR04, p575].

$$\chi^{-2}(x|\nu, \sigma^2) = \frac{1}{\Gamma(\nu/2)}\left(\frac{\nu\sigma^2}{2}\right)^{\nu/2} x^{-\frac{\nu}{2}-1}\exp[-\frac{\nu\sigma^2}{2x}], \; x > 0 \tag{285}$$

$$= IG(x|\text{shape}=\frac{\nu}{2}, \text{scale}=\frac{\nu\sigma^2}{2}) \tag{286}$$

where the parameter $\nu > 0$ is called the degrees of freedom, and $\sigma^2 > 0$ is the scale. See Figure 7 for some plots. We see that increasing ν lifts the curve up and moves it slightly to the right. Later, when we consider Bayesian parameter estimation, we will use this distribution as a conjugate prior for a scale parameter (such as the variance of a Gaussian); increasing ν corresponds to increasing the effective strength of the prior.

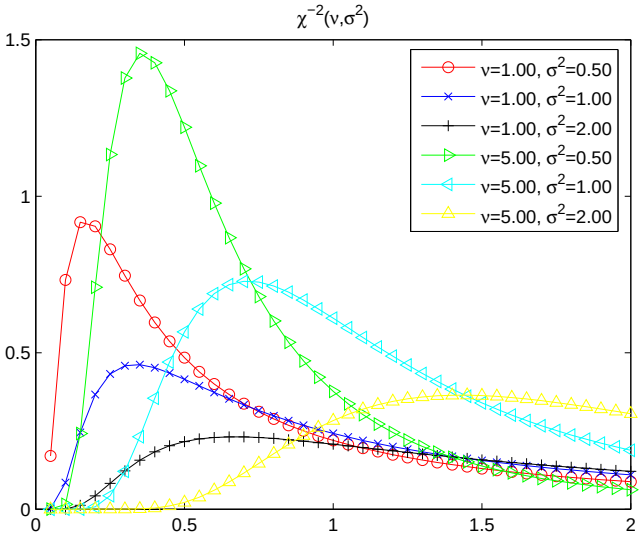


Figure 7: Some inverse scaled χ^2 distributions. These plots were produced by `invchi2plot`.

The distribution has these properties

$$\text{mean} = \frac{\nu\sigma^2}{\nu-2} \text{ for } \nu > 2 \tag{287}$$

$$\text{mode} = \frac{\nu\sigma^2}{\nu+2} \tag{288}$$

$$\text{var} = \frac{2\nu^2\sigma^4}{(\nu-2)^2(\nu-4)} \text{for } \nu > 4 \tag{289}$$

The **inverse chi-squared** distribution, written $\chi_\nu^{-2}(x)$, is the special case where $\nu\sigma^2 = 1$ (i.e., $\sigma^2 = 1/\nu$). This corresponds to $IG(a = \nu/2, b = \text{scale} = 1/2)$.

10.4 Wishart distribution

Let $\mathbf{X}$ be a p dimensional symmetric positive definite matrix. The Wishart is the multidimensional generalization of the Gamma. Since it is a distribution over matrices, it is hard to plot as a density function. However, we can easily sample from it, and then use the eigenvectors of the resulting matrix to define an ellipse. See Figure 8.

There are several possible parameterizations. Some authors (e.g., [Bis06, p693], [DeG70, p.57],[GCSR04, p574], wikipedia) as well as WinBUGS and Matlab (`wishrnd`), define the Wishart in terms of degrees of freedom $\nu \geq p$ and the scale matrix S as follows:

$$Wi_\nu(\mathbf{X}|\mathbf{S}) = \frac{1}{Z}|\mathbf{X}|^{(\nu-p-1)/2} \exp[-\tfrac{1}{2}\text{tr}(\mathbf{S}^{-1}\mathbf{X})] \tag{290}$$

$$Z = 2^{\nu p/2}\Gamma_p(\nu/2)|S|^{\nu/2} \tag{291}$$

where $\Gamma_p(a)$ is the generalized gamma function

$$\Gamma_p(\alpha) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left(\frac{2\alpha+1-i}{2}\right) \tag{292}$$

(So $\Gamma_1(\alpha) = \Gamma(\alpha)$.) The mean and mode are given by (see also [Pre05])

$$\text{mean} = \nu S \tag{293}$$

$$\text{mode} = (\nu-p-1)\mathbf{S}, \ \nu > p+1 \tag{294}$$

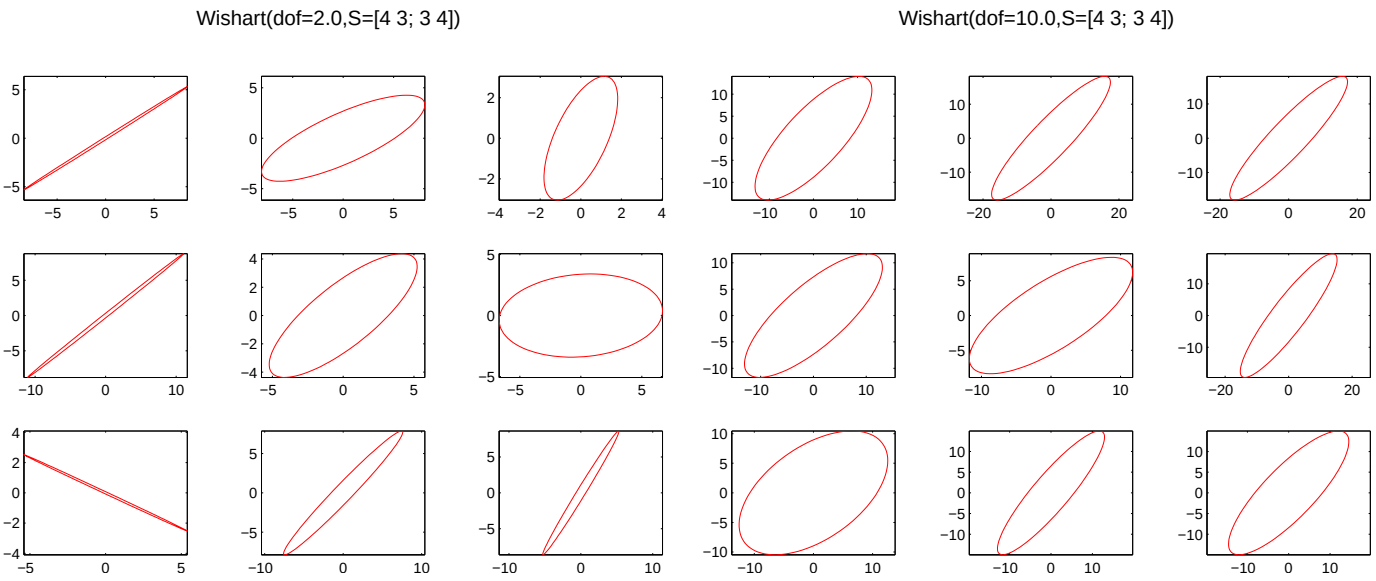


Figure 8: Some samples from the Wishart distribution. Left: $\nu = 2$, right: $\nu = 10$. We see that if if $\nu = 2$ (the smallest valid value in 2 dimensions), we often sample nearly singular matrices. As ν increases, we put more mass on the S matrix. If $S = I_2$, the samples would look (on average) like circles. Generated by `wishplot`.

In 1D, this becomes $Ga(\text{shape} = \nu/2, \text{rate} = S/2)$.

Note that if $X \sim Wi_{\mathbf{u}}(S)$, and $Y = X^{-1}$, then $Y \sim IW_\nu(S^{-1})$ and $E[Y] = \frac{S}{\nu-d-1}$.

In [BS94, p.138], and the `wishpdf` in Tom Minka's lightspeed toolbox, they use the following parameterization

$$Wi(\mathbf{X}|a, \mathbf{B}) = \frac{|\mathbf{B}|^a}{\Gamma_p(a)}|\mathbf{X}|^{a-(p+1)/2} \exp[-\text{tr}(\mathbf{BX})] \tag{295}$$

We require that $\mathbf{B}$ is a $p \times p$ symmetric positive definite matrix, and $2a > p-1$. If $p = 1$, so $\mathbf{B}$ is a scalar, this reduces to the $Ga(\text{shape} = a, \text{rate} = b)$ density.

To get some intuition for this distribution, recall that $\text{tr}(AB)$ is a vector which contains the inner product of the rows of A and the columns of B. In Matlab notation we have

```
trace(A B) = [a(1,:)*b(:,1), ..., a(n,:)*b(:,n)]
```

If $X \sim Wi_\nu(S)$, then we are performing a kind of template matching between the columns of X and S^{-1} (recall that both X and S are symmetric). This is a natural way to define the distance between two matrices.

10.5 Inverse Wishart

This is the multidimensional generalization of the inverse Gamma. Consider a $d \times d$ positive definite (covariance) matrix $\mathbf{X}$ and a dof parameter $\nu > d-1$ and psd matrix $\mathbf{S}$. Some authors (eg [GCSR04, p574]) use this parameterization:

$$IW_\nu(\mathbf{X}|\mathbf{S}^{-1}) = \frac{1}{Z}|\mathbf{X}|^{-(\nu+d+1)/2} \exp\left(-\frac{1}{2}Tr(\mathbf{SX}^{-1})\right) \tag{296}$$

$$Z = \frac{|\mathbf{S}|^{\nu/2}}{2^{\nu d/2}\Gamma_d(\nu/2)} \tag{297}$$

where

$$\Gamma_d(\nu/2) = \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma(\frac{\nu+1-i}{2}) \tag{298}$$

The distribution has mean

$$E\ X = \frac{\mathbf{S}}{\nu - d - 1} \tag{299}$$

In Matlab, use `iwishrnd`. In the 1d case, we have

$$\chi^{-2}(\Sigma|\nu_0, \sigma_0^2) = IW_{\nu_0}(\Sigma|(\nu_0\sigma_0^2)^{-1}) \tag{300}$$

Other authors (e.g., [Pre05, p117]) use a slightly different formulation (with $2d < \nu$)

$$IW_\nu(\mathbf{X}|\mathbf{Q}) = \left(2^{(\nu-d-1)d/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma((\nu-d-j)/2)\right)^{-1} \tag{301}$$

$$\times|\mathbf{Q}|^{(\nu-d-1)/2}|\mathbf{X}|^{-\nu/2}\exp\left(-\frac{1}{2}Tr(\mathbf{X}^{-1}\mathbf{Q})\right) \tag{302}$$

which has mean

$$E\ \mathbf{X} = \frac{\mathbf{Q}}{\nu - 2d - 2} \tag{303}$$

10.6 Student t distribution

The generalized t-distribution is given as

$$t_\nu(x|\mu, \sigma^2) = c\left[1 + \frac{1}{\nu}(\frac{x-\mu}{\sigma})^2\right]^{-(\frac{\nu+1}{2})} \tag{304}$$

$$c = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)}\frac{1}{\sqrt{\nu\pi}\sigma} \tag{305}$$

where c is the normalization consant. μ is the mean, $\nu > 0$ is the **degrees of freedom**, and $\sigma^2 > 0$ is the scale. (Note that the ν parameter is often written as a subscript.) In Matlab, use `tpdf`.

The distribution has the following properties:

$$\text{mean} = \mu,\ \nu > 1 \tag{306}$$

$$\text{mode} = \mu \tag{307}$$

$$\text{var} = \frac{\nu\sigma^2}{(\nu-2)},\ \ \nu > 2 \tag{308}$$

Note: if $x \sim t_\nu(\mu, \sigma^2)$, then

$$\frac{x-\mu}{\sigma} \sim t_\nu \tag{309}$$

which corresponds to a standard t-distribution with $\mu = 0, \sigma^2 = 1$:

$$t_\nu(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}(1 + x^2/\nu)^{-(\nu+1)/2} \tag{310}$$

In Figure 9, we plot the density for different parameter values. As $\nu \to \infty$, the T approaches a Gaussian. T-distributions are like Gaussian distributions with **heavy tails**. Hence they are more robust to outliers (see Figure 10).

If $\nu = 1$, this is called a **Cauchy distribution**. This is an interesting distribution since if $X \sim Cauchy$, then $E[X]$ does not exist, since the corresponding integral diverges. Essentially this is because the tails are so heavy that samples from the distribution can get very far from the center μ.

It can be shown that the t-distribution is like an infinite sum of Gaussians, where each Gaussian has a different precision:

$$p(x|\mu, a, b) = \int \mathcal{N}(x|\mu, \tau^{-1})Ga(\tau|a, \text{rate} = b)d\tau \tag{311}$$

$$= t_{2a}(x|\mu, b/a) \tag{312}$$

(See exercise 2.46 of [Bis06].)

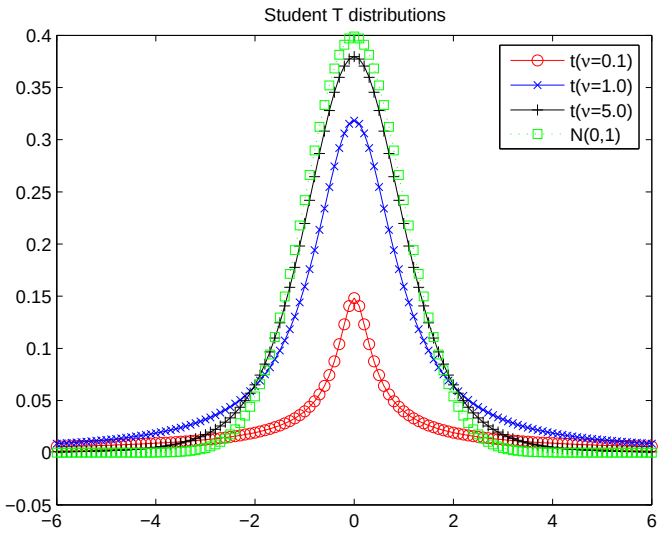


Figure 9: Student t-distributions $T(\mu, \sigma^2, \nu)$ for $\mu = 0$. The effect of σ is just to scale the horizontal axis. As $\nu \to \infty$, the distribution approaches a Gaussian. See `studentTplot`.

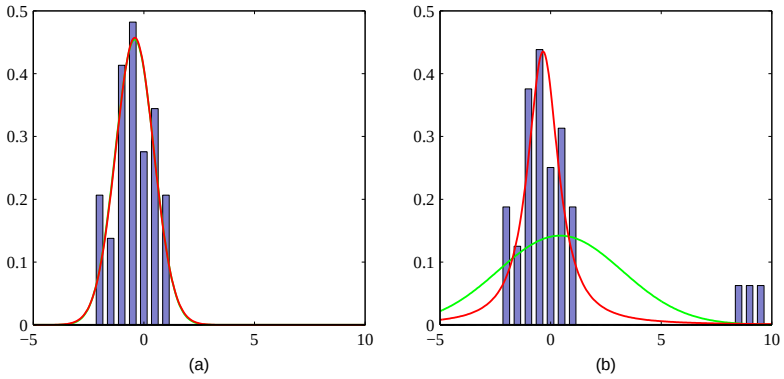


Figure 10: Fitting a Gaussian and a Student distribution to some data (left) and to some data with outliers (right). The Student distribution (red) is much less affected by outliers than the Gaussian (green). Source: [Bis06] Figure 2.16.

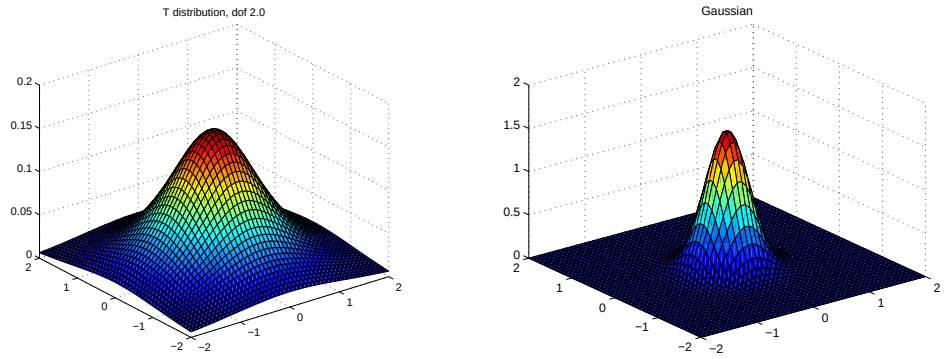


Figure 11: Left: T distribution in 2d with dof=2 and $\Sigma = 0.1I_2$. Right: Gaussian density with $\Sigma = 0.1I_2$ and $\mu = (0, 0)$; we see it goes to zero faster. Produced by `multivarTplot`.

10.7 Multivariate t distributions

The multivariate T distribution in d dimensions is given by

$$t_\nu(x|\mu, \Sigma) = \frac{\Gamma(\nu/2 + d/2)}{\Gamma(\nu/2)} \frac{|\Sigma|^{-1/2}}{v^{d/2}\pi^{d/2}} \times \left[1 + \frac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]^{-(\frac{\nu+d}{2})} \tag{313}$$

$$\tag{314}$$

where Σ is called the scale matrix (since it is not exactly the covariance matrix). This has fatter tails than a Gaussian: see Figure 11. In Matlab, use `mvtpdf`.

The distribution has the following properties

$$E\ x = \mu \text{ if } \nu > 1 \tag{315}$$

$$\text{mode}\ x = \mu \tag{316}$$

$$\text{Cov}\ \ x = \frac{\nu}{\nu - 2}\Sigma \text{ for } \nu > 2 \tag{317}$$

(The following results are from [Koo03, p328].) Suppose $Y \sim T(\mu, \Sigma, \nu)$ and we partition the variables into 2 blocks. Then the marginals are

$$Y_i \sim T(\mu_i, \Sigma_{ii}, \nu) \tag{318}$$

and the conditionals are

$$Y_1|y_2 \sim T(\mu_{1|2}, \Sigma_{1|2}, \nu + d_1) \tag{319}$$

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2) \tag{320}$$

$$\Sigma_{1|2} = h_{1|2}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T) \tag{321}$$

$$h_{1|2} = \frac{1}{\nu + d_2}\left[\nu + (y_2 - \mu_2)^T\Sigma_{22}^{-1}(y_2 - \mu_2)\right] \tag{322}$$

We can also show linear combinations of Ts are Ts:

$$Y \sim T(\mu, \Sigma, \nu) \Rightarrow AY \sim T(A\mu, A\Sigma A', \nu) \tag{323}$$

We can sample from a $y \sim T(\mu, \Sigma, \nu)$ by sampling $x \sim T(0, 1, \nu)$ and then transforming $y = \mu + R^T x$, where $R = \text{chol}(\Sigma)$, so $R^T R = \Sigma$.

References

[Bis06] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[BL01] P. Baldi and A. Long. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, 2001.

[BS94] J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley, 1994.

[DeG70] M. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, 1970.

[DHMS02] D. Denison, C. Holmes, B. Mallick, and A. Smith. *Bayesian methods for nonlinear classification and regression*. Wiley, 2002.

[DMP+06] F. Demichelis, P. Magni, P. Piergiorgi, M. Rubin, and R. Bellazzi. A hierarchical Naive Bayes model for handling sample heterogeneity in classification problems: an application to tissue microarrays. *BMC Bioinformatics*, 7:514, 2006.

[GCSR04] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman and Hall, 2004. 2nd edition.

[GH94] D. Geiger and D. Heckerman. Learning Gaussian networks. Technical Report MSR-TR-94-10, Microsoft Research, 1994.

[Koo03] Gary Koop. *Bayesian econometrics*. Wiley, 2003.

[Lee04] Peter Lee. *Bayesian statistics: an introduction*. Arnold Publishing, 2004. Third edition.

[Min00] T. Minka. Inferring a Gaussian distribution. Technical report, MIT, 2000.

[Pre05] S. J. Press. *Applied multivariate analysis, using Bayesian and frequentist methods of inference*. Dover, 2005. Second edition.