Chapter 9

The exponential family: Conjugate priors

Within the Bayesian framework the parameter θ is treated as a random quantity. This requires us to specify a *prior distribution* $p(\theta)$, from which we can obtain the *posterior distribution* $p(\theta \,|\, x)$ via Bayes theorem:

$$p(\theta \,|\, x) = \frac{p(x \,|\, \theta)p(\theta)}{p(x)}, \tag{9.1}$$

where $p(x \,|\, \theta)$ is the likelihood.

Most inferential conclusions obtained within the Bayesian framework are based in one way or another on averages computed under the posterior distribution, and thus for the Bayesian framework to be useful it is essential to be able to compute these integrals with some effective procedure. In particular, prediction of future data x_{new} is based on the predictive probability:

$$p(x_{\text{new}} \,|\, x) = \int p(x_{\text{new}} \,|\, \theta)p(\theta \,|\, x)d\theta, \tag{9.2}$$

which is an integral with respect to the posterior. (We have assumed that $X_{\text{new}} \perp\!\!\!\perp X \,|\, \theta$). Note also that forming the posterior distribution itself involves computing an integral: to normalize the posterior we must compute

$$p(x) = \int p(x \,|\, \theta)p(\theta)d\theta, \tag{9.3}$$

which is an integral with respect to the prior.

In this section we introduce the idea of a *conjugate prior*. The basic idea is as follows. Given a likelihood $p(x \,|\, \theta)$, we choose a family of prior distributions such that integrals of the form Eq. (9.3) can be obtained tractably (for every prior in the family). Moreover, we choose this family such that prior-to-posterior updating yields a posterior that is also in the family. This means that integrals of the form Eq. (9.2) can also be obtained tractably for every posterior distribution in the family. In general these two goals are in conflict. For example, the goal of invariance of prior-to-posterior updating (i.e., asking that the posterior remains in the same family of distributions of the prior) can be acheived vacuously by defining the family of all probability distributions, but this would not yield tractable integrals. On the other extreme, we could aim to obtain tractable integrals by taking the family of prior distributions to be a single distribution of a simple form (e.g., a constant), but the posterior would not generally retain this form.

In the setting of the exponential family this dilemma is readily resolved. For exponential families the likelihood is a simple standarized function of the parameter and we can define conjugate priors by mimicking the form of the likelihood. Multiplication of a likelihood and a prior that have the same exponential form yields a posterior that retains that form. Moreover, for the exponential families that are most useful in practice, these exponential forms are readily integrated. In the remainder of this section we present examples that illustrate conjugate priors for exponential family distributions.

Conjugate priors thus have appealing computational properties and for this reason they are widely used in practice. Indeed, for the complex models of the kind that are often constructed using the graphical model toolbox, computational considerations may be paramount, and there may be little choice but to use conjugate priors. On the other hand, there are also good reasons *not* to use conjugate priors and one should not be lulled into a sense of complacency when using conjugate priors. Before turning to a presentation of examples, let us briefly discuss some of the philosophical issues. we will return to this discussion in Section **??** after we have obtained a better idea of some of the options.

Recall from our earlier discussion in Section **??** the distinction between subjective Bayesian and objective Bayesian perspectives. The subjective Bayesian perspective takes the optimistic view that priors are an opportunity to express knowledge; in particular, a prior may be a posterior from a previous experiment. The objective Bayesian perspective takes the more pessimistic view that prior knowledge is often not available and that priors should be chosen to have as little impact on the analysis as possible, relative to the impact of the data. In this regard, it is important to note that conjugate priors involve making relatively strong assumptions. Indeed, in a sense to be made clear in Section **??**, conjugate priors *minimize* the impact of the data on the posterior. From the subjective perspective, this can be viewed favorably—conjugate priors provide an opportunity to express knowledge in a relatively influential way. From the objective perspective, however, conjugate priors are decidedly dangerous; objective priors aim to *maximize* the impact of the data on the posterior.

The general point to be made is that one should take care with conjugate priors. The use of conjugate priors involves relatively strong assumptions and thus it is particularly important to do sensitivity analysis to assess how strongly the posterior is influenced by the prior. If the answer is "not much," then one can proceed with some confidence. If the answer is "a lot," then one should either take great care to assess whether a domain expert is comfortable with these priors on subjective grounds or one should consider other kinds of priors (such as those discussed in Section **??**) and/or gather more data so as to diminish the effect of the prior.

9.0.1 Bernoulli distribution and beta priors

We have stated that conjugate priors can be obtained by mimicking the form of the likelihood. This is easily understood by considering examples. Let us begin with the Bernoulli distribution.

Parameterizing the Bernoullli distribution using the mean parameter θ, the likelihood takes the following form:

$$p(x \,|\, \theta) = \theta^x (1-\theta)^{1-x}. \tag{9.4}$$

Under i.i.d. sampling, this expression retains the form of a product of powers of θ and $1-\theta$, with the exponents growing. This suggests that to obtain a conjugate prior for θ, we use a distribution that is a product of powers of θ and $1-\theta$, with free parameters in the exponents:

$$p(\theta \,|\, \tau) \propto \theta^{\tau_1} (1-\theta)^{\tau_2}. \tag{9.5}$$

This expression can be normalized if $\tau_1 > -1$ and $\tau_2 > -1$. The resulting distribution is known as the *beta distribution*, another example of an exponential family distribution.

The beta distribution is traditionally parameterized using $\alpha_i - 1$ instead of τ_i in the exponents (for a reason that will become clear below), yielding the following standard form for the conjugate prior:

$$p(\theta \,|\, \alpha) = K(\alpha) \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_2 - 1}. \tag{9.6}$$

where the normalization factor $K(\alpha)$ can be obtained analytically (see Exercise **??**):

$$\begin{aligned} K(\alpha) &= \left(\int \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_2 - 1} d\theta \right)^{-1} && (9.7) \\ &= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} && (9.8) \end{aligned}$$

as a ratio of gamma functions.

If we multiply the beta density by the Bernoulli likelihood we obtain a beta density. Consider in particular N i.i.d. Bernoulli observations, $\mathbf{x} = (x_1, \ldots, x_N)^T$:

$$\begin{aligned} p(\theta \,|\, \mathbf{x}, \alpha) &\propto \left(\prod_{n=1}^{N} \theta^{x_n} (1-\theta)^{1-x_n} \right) \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_2 - 1} && (9.9) \\ &= \theta^{\sum_{n=1}^{N} x_n + \alpha_1 - 1} (1-\theta)^{N - \sum_{n=1}^{N} x_n + \alpha_2 - 1}. && (9.10) \end{aligned}$$

This is a beta density with updated values of the parameters. In particular, it is a Beta($\sum_{n=1}^N x_n + \alpha_1, N - \sum_{n=1}^N x_n + \alpha_2$) distribution.

Note the simple nature of the prior-to-posterior updating procedure. For each observation x_n we simply add x_n to the first parameter of the beta distribution and add $1 - x_n$ to the second parameter of the beta distribution. At each step we simply retain two numbers as our representation of the posterior distribution. Note also that the form of the updating procedure provides an interpretation for the parameters α_1 and α_2. In particular, viewing the posterior as a prior from a previous experiment, we can view α_1 and α_2 as "effective counts"; α_1 can be viewed as an effective number of prior observations of $X = 1$ and α_2 can be interpreted as an effective number of prior observations of $X = 0$. (In general, however, the parameters are not restricted to integer values.)

The fact that the normalization factor of the beta distribution has an analytic form allows us to compute various averages in closed form. Consider, in particular, the mean of a beta random variable:

$$\begin{aligned}
\mathbb{E}[\theta \,|\, \alpha] &= \int \theta K(\alpha) \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1} d\theta && (9.11)\\
&= K(\alpha) \int \theta \theta^{\alpha_1}(1-\theta)^{\alpha_2 - 1} d\theta && (9.12)\\
&= \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1+1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+1+\alpha_2)} && (9.13)\\
&= \frac{\alpha_1}{\alpha_1+\alpha_2}, && (9.14)
\end{aligned}$$

using $\Gamma(a+1) = a\Gamma(a)$ in the final line. A similar calculation yields the variance:

$$\mathrm{Var}[\theta \,|\, \alpha] = \frac{\alpha_1 \alpha_2}{(\alpha_1+\alpha_2+1)(\alpha_1+\alpha_2)^2}. \quad (9.15)$$

From these results we see that the relative values of α_1 and α_2 determine the mean, whereas the magnitude $\alpha_1 + \alpha_2$ determines the variance. That is, for a fixed value of the mean, the variance goes to zero as $\alpha_1 + \alpha_2$ goes to infinity.

Applying these results to the posterior distribution in Eq. (9.10), we can compute the posterior mean:

$$\mathbb{E}[\theta \,|\, \mathbf{x}, \alpha] = \frac{\sum_{n=1}^N x_n + \alpha_1}{N + \alpha_1 + \alpha_2}. \quad (9.16)$$

and the posterior variance:

$$\mathrm{Var}[\theta \,|\, \mathbf{x}, \alpha] = \frac{(\sum_{n=1}^N x_n + \alpha_1)(N - \sum_{n=1}^N x_n + \alpha_2)}{(N+\alpha_1+\alpha_2+1)(N+\alpha_1+\alpha_2)^2}. \quad (9.17)$$

These equations yield several significant pieces of information. First, letting N tend to infinity, we see that

$$\mathbb{E}[\theta \,|\, \mathbf{x}, \alpha] \to \frac{1}{N}\sum_{n=1}^{N} x_n, \tag{9.18}$$

which is the maximum likelihood estimate of θ. Second, we have

$$\mathrm{Var}[\theta \,|\, \mathbf{x}, \alpha] \to 0, \tag{9.19}$$

showing that the posterior distribution concentrates around the maximum likelihood estimate for large N. Thus we see inklings of the reconciliation of Bayesian and frequentist statistics that is achieved in the large-sample limit. Finally, rewriting Eq. (9.16), we have:

$$\mathbb{E}[\theta \,|\, \mathbf{x}, \alpha] = \kappa \frac{\alpha_1}{\alpha_1 + \alpha_2} + (1-\kappa)\bar{x}, \tag{9.20}$$

where $\kappa = (\alpha_1 + \alpha_2)/(N + \alpha_1 + \alpha_2)$. We see that the posterior mean is a convex combination of the prior mean and the maximum likelihood estimate. As N goes to infinity, κ goes to zero and the posterior mean converges to the maximum likelihood estimate.

Finally, let us compute the predictive probability of a new data point, X_{new}, assumed conditionally independent of the observations $\mathbf{X}$. We have:

$$\begin{aligned}
p(X_{\mathrm{new}} = 1 \,|\, x, \alpha) &= \int p(X_{\mathrm{new}} = 1 \,|\, \theta) p(\theta \,|\, x, \alpha) d\theta \\
&= \int \theta \frac{\Gamma(N + \alpha_1 + \alpha_2)}{\Gamma(\sum_{n=1}^N x_n + \alpha_1)\Gamma(N - \sum_{n=1}^N x_n + \alpha_2)} \theta^{\sum_{n=1}^N x_n + \alpha_1 - 1}(1-\theta)^{N - \sum_{n=1}^N x_n + \alpha_2 - 1} d\theta \\
&= \frac{\Gamma(N + \alpha_1 + \alpha_2)}{\Gamma(\sum_{n=1}^N x_n + \alpha_1)\Gamma(N - \sum_{n=1}^N x_n + \alpha_2)} \int \theta^{\sum_{n=1}^N x_n + \alpha_1}(1-\theta)^{N - \sum_{n=1}^N x_n + \alpha_2 - 1} d\theta \\
&= \frac{\Gamma(N + \alpha_1 + \alpha_2)}{\Gamma(\sum_{n=1}^N x_n + \alpha_1)\Gamma(N - \sum_{n=1}^N x_n + \alpha_2)} \frac{\Gamma(\sum_{n=1}^N x_n + \alpha_1 + 1)\Gamma(N - \sum_{n=1}^N x_n + \alpha_2)}{\Gamma(N + \alpha_1 + \alpha_2 + 1)} \\
&= \frac{\sum_{n=1}^N x_n + \alpha_1}{N + \alpha_1 + \alpha_2}.
\end{aligned} \tag{9.21}$$

We have carried out this calculation in detail so that we could display the fact that the predictive distribution takes the form of a ratio of normalizing factors. This is a general result—one that we will demonstrate for general exponential families in Section 9.0.5. However, having done the calculation, we should note that the final result is identical to Eq. (9.16). Indeed, we could have obtained the predictive distribution in this case via a much shorter conditional expectation calculation (see Exercise **??**).

9.0.2 Multinomial distribution and Dirichlet priors

A very similar development can be carried out for the multinomial distribution. In this case the likelihood takes the form

$$p(x \mid \theta) = \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_K^{x_K}. \tag{9.22}$$

This yields the following conjugate prior:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_K^{\alpha_K - 1}. \tag{9.23}$$

For $\alpha_i > 0$ this expression can be normalized (integrating over the simplex $\sum_{k=1}^{K} \theta_k = 1$). The result is the *Dirichlet distribution*:

$$p(\theta \mid \alpha) = K(\alpha) \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_K^{\alpha_K - 1}, \tag{9.24}$$

where

$$K(\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \tag{9.25}$$

and where $\alpha_{\cdot}^{K} \alpha_k$. The Dirichlet distribution plays an important role in several chapters of the book. We thus provide a detailed treatment of the Dirichlet distribution in Section **??**, including a derivation of the normalization constant $K(\alpha)$. In this section we restrict ourselves to presenting a few key results that make clear the parallel between Bernoulli/beta conjugacy and multinomial/Dirichlet conjugacy.

Note in this regard that the beta distribution is the special case of the Dirichlet distribution for $K = 2$, and, as can be checked, the results of this section reduce to those of Section 9.0.1 for $K = 2$.

Consider N i.i.d. multinomial observations, $\mathbf{x} = (x_1, \ldots, x_N)^T$. Multiplying the multinomial likelihood by the Dirichlet prior yields

$$p(\theta \mid \mathbf{x}, \alpha) \quad \propto \quad \theta_1^{\sum_{n=1}^{N} x_{n1} + \alpha_1 - 1} \theta_2^{\sum_{n=1}^{N} x_{n2} + \alpha_2 - 1} \cdots \theta_K^{\sum_{n=1}^{N} x_{nK} + \alpha_K - 1}. \tag{9.26}$$

We once again see the simple form of prior-to-posterior updating obtained using a conjugate prior. In particular, each observation x_n has one component that is equal to one (all the other components are zero), and to incorporate that observation into the posterior we simply add one to the corresponding component of α_k.

We can use Eq. (9.25) to compute the mean of the Dirichlet distribution (see Exercise **??**):

$$\mathbb{E}[\theta_i \mid \alpha] = \frac{\alpha_i}{\alpha_{\cdot}}, \tag{9.27}$$

as well as the variance of the ith component (see Exercise **??**):

$$\mathrm{Var}[\theta_i \mid \alpha] = \frac{\alpha_i(\alpha_{\cdot} - \alpha_i)}{\alpha_{\cdot}^2(\alpha_{\cdot} + 1)}. \tag{9.28}$$

and the covariances (see Exercise **??**). Given that the posterior distribution is also Dirichlet, we can use these results to obtain the posterior mean:

$$\begin{aligned} \mathbb{E}[\theta_i \,|\, \mathbf{x}, \alpha] &= \frac{\sum_{n=1}^N x_{ni} + \alpha_i}{N + \alpha.} && (9.29) \\ &= \kappa \frac{\alpha_i}{\alpha.} + (1-\kappa)\bar{x}_i, && (9.30) \end{aligned}$$

where $\bar{x}_i = \sum_{n=1}^N x_{ni}/N$ is the maximum likelihood estimate of θ_i and $\kappa = (\alpha.)/(N + \alpha.)$. Once again we see that the posterior mean is a convex combination of the prior mean and the maximum likelihood estimate. Also, as N goes to infinity it can be checked that the posterior concentrates around the maximum likelihood estimate.

9.0.3 Poisson distribution and gamma priors

As another example of a discrete exponential family distribution let us consider the Poisson distribution:

$$p(x \,|\, \theta) = \frac{\theta^x e^{-\theta}}{x!}. \qquad (9.31)$$

The corresponding conjugate prior retains the shape of the Poisson likelihood:

$$p(\theta \,|\, \alpha) \propto \theta^{\alpha_1 - 1} e^{-\alpha_2 \theta}, \qquad (9.32)$$

which we recognize as the *gamma distribution*:

$$p(\theta \,|\, \alpha) = K(\alpha) \theta^{\alpha_1 - 1} e^{-\alpha_2 \theta}, \qquad (9.33)$$

where

$$K(\alpha) = \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)}. \qquad (9.34)$$

See Exercise **??** for the computation of this normalization factor.

Given N i.i.d. Poisson observations, $\mathbf{x} = (x_1, \ldots, x_N)^T$, we see that the prior-to-posterior update involves adding the sufficient statistic $\sum_{n=1}^N x_n$ to the parameter α_1 and adding the number of observations N to the parameter α_2.

The mean of the gamma distribution is readily computed based on Eq. (9.34)

$$\mathbb{E}[\theta \,|\, \alpha] = \frac{\alpha_1}{\alpha_2} \qquad (9.35)$$

as is the variance:

$$\mathrm{Var}[\theta_i \,|\, \alpha] = \frac{\alpha_1}{\alpha_2^2}. \qquad (9.36)$$

Plugging in the parameters from the posterior distribution we find that the posterior mean has a familiar linear form:

$$\mathbb{E}[\theta \,|\, \mathbf{x}, \alpha] = \kappa \frac{\alpha_1}{\alpha_2} + (1 - \kappa)\bar{x}, \tag{9.37}$$

where $\kappa = \alpha_2/(N + \alpha_2)$. It can also be checked that the posterior variance goes to zero (at rate $1/N$).

9.0.4 Univariate Gaussian distribution and normal-inverse-gamma priors

As a final example we consider a somewhat more elaborate case: the univariate Gaussian distribution. Recall that this distribution is a two-parameter exponential family of the following form:

$$p(x \,|\, \mu, \sigma^2) \propto (\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}. \tag{9.38}$$

In general we wish to treat both parameters as random and thus we need to obtain a bivariate conjugate prior for μ and σ^2. To build up this prior, however, we consider two simpler and somewhat artificial cases, the first in which the variance is assumed known and the second in which the mean is assumed known.

Conjugacy for the mean

Inspecting the functional form in Eq. (9.38) from the point of view of its dependence on μ, we see that we have an unnormalized Gaussian density—the exponential of the negative of a quadratic form in μ. Given that the product of two such factors is also an unnormalized Gaussian density, we are led to trying a Gaussian as a conjugate prior for μ. Thus we assume

$$p(\mu \,|\, \mu_0, \sigma_0^2) \propto (\sigma_0^2)^{-1/2} e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}, \tag{9.39}$$

where μ_0 and σ_0^2 are the mean and variance of μ.

Let us begin by assuming that we have only a single data point x. To compute the posterior, $p(\mu \,|\, x)$, we need to take the product of Eq. (9.38) and Eq. (9.39) and rearrange so as to obtain a density for μ. This can be approached as an exercise in completing the square (Exercise **??**). It can also be approached in a somewhat more revealing way by treating (X, μ) as a bivariate Gaussian and using general rules for computing conditional probabilities under a multivariate Gaussian distribution. We thus reach forward to Chapter **??**, where the following result is established (using facts about the Schur complement of matrices). Letting (Z_1, Z_2) be jointly Gaussian, the conditional of Z_1 given Z_2 is also Gaussian, with conditional mean:

$$\mathbb{E}[Z_1 \,|\, Z_2] = \mathbb{E}[Z_1] + \frac{\mathrm{Cov}[Z_1, Z_2]}{\mathrm{Var}[Z_2]}(Z_2 - \mathbb{E}[Z_2]), \tag{9.40}$$

and conditional variance:

$$\mathrm{Var}[Z_1 \,|\, Z_2] = \mathrm{Var}[Z_1] - \frac{\mathrm{Cov}^2[Z_1, Z_2]}{\mathrm{Var}[Z_2]}. \tag{9.41}$$

We use these results in the following way. Write

$$\begin{aligned} X &= \mu + \sigma\epsilon && (9.42) \\ \mu &= \mu_0 + \sigma_0 \delta, && (9.43) \end{aligned}$$

where $\epsilon \sim N(0,1)$ and $\delta \sim N(0,1)$ are independent Gaussian random variables. These equations imply that μ is Gaussian with mean μ_0 and variance σ_0^2 and that X is conditionally Gaussian with mean μ and variance σ^2; i.e., this is simply another way to write Eq. (9.38) and Eq. (9.39). Moreover, we can now easily calculate:

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[\mu] + \sigma\mathbb{E}[\epsilon] = \mu_0 && (9.44) \\ \mathrm{Var}[X] &= \mathbb{E}[X - \mu_0]^2 = \mathbb{E}[\mu - \mu_0 + \sigma\epsilon]^2 = \sigma^2 + \sigma_0^2 && (9.45) \\ \mathrm{Cov}[X, \mu] &= \mathbb{E}[(X - \mu_0)(\mu - \mu_0)] = \mathbb{E}[(\mu - \mu_0 + \sigma\epsilon)(\mu - \mu_0)] = \sigma_0^2. && (9.46) \end{aligned}$$

Treating X as Z_2 and μ as Z_1 in Eq. (9.40) and Eq. (9.41), we obtain:

$$\begin{aligned} \mu_{post} &:= \mathbb{E}[\mu \,|\, X = x] = \mu_0 + \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}(x - \mu_0) = \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2} x + \frac{\sigma^2}{\sigma^2 + \sigma_0^2}\mu_0 && (9.47) \\ \sigma_{post}^2 &:= \mathrm{Var}[\mu \,|\, X = x] = \sigma_0^2 - \frac{\sigma_0^4}{\sigma^2 + \sigma_0^2} = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + \sigma_0^2} && (9.48) \end{aligned}$$

for the posterior mean and posterior variance respectively.

The posterior mean in Eq. (9.47) is a convex combination of the observation x and the prior mean μ_0. If σ_0^2 is large relative to σ^2, meaning that we are less sure of our prior than our data, then the posterior mean is closer to x than to μ_0. On the other hand, if σ_0^2 is small relative to σ^2, then the posterior mean is closer to the prior mean than to the data.

We can also express these results in terms of the inverse of the variance—the *precision.* In particular, plugging $\tau = 1/\sigma^2$ and $\tau_0 = 1/\sigma_0^2$ into Eq. (9.47) yields:

$$\mathbb{E}[\mu \,|\, X = x] = \frac{\tau}{\tau + \tau_0} x + \frac{\tau_0}{\tau + \tau_0}\mu_0. \tag{9.49}$$

This has the same interpretation as before, but is slightly more direct: the precision of the data multiplies the data x and the precision of the prior multiplies the prior mean μ_0. Moreover, denoting the inverse of the posterior variance by τ_{post}, we can rewrite Eq. (9.48) as:

$$\tau_{post} = \tau + \tau_0, \tag{9.50}$$

which has a direct interpretation: precisions add.

Let us now compute the posterior distribution in the case in which multiple data points are observed. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector of conditionally independent random variables (conditioning on μ), with $X_i \sim N(\mu, \sigma^2)$. We now have:

$$p(\mathbf{x} \,|\, \mu, \tau) \propto \tau^{1/2} e^{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2}. \tag{9.51}$$

We rewrite the exponent using a standard trick:

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu)^2 &= \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2 \\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2 \end{aligned} \tag{9.52}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean. The first term yields a constant factor when exponentiated, and we see that the problem reduces to an equivalent problem involving the univariate random variable $\bar{X}$. (That the problem reduces to $\bar{X}$ will not be a surprise after you have read Section **??**, where you will learn that $\bar{X}$ is a *sufficient statistic* for μ). Eq. (9.52) also shows that $\bar{X}$ is Gaussian with mean μ and variance σ^2/n. We plug these values into Eq. (9.47) and Eq. (9.48), which yields:

$$\mu_{post} = \frac{\sigma_0^2}{\sigma^2/n+\sigma_0^2}\bar{x} + \frac{\sigma^2/n}{\sigma^2/n+\sigma_0^2}\mu_0 \tag{9.53}$$

and

$$\sigma^2_{post} = \frac{\sigma^2\sigma_0^2/n}{\sigma^2/n+\sigma_0^2}. \tag{9.54}$$

These results can be expressed slightly more neatly using precisions:

$$\mu_{post} = \frac{n\tau}{n\tau+\tau_0}\bar{x} + \frac{\tau_0}{n\tau+\tau_0}\mu_0 \tag{9.55}$$

and

$$\tau_{post} = n\tau + \tau_0, \tag{9.56}$$

where again we see the additive property of precisions and note that each additional data point increases the overall precision by τ.

Finally, let us compute the predictive probability of a new data point, X_{new}, assumed conditionally independent of the observations $\mathbf{X}$. We have:

$$\begin{aligned} p(x_{new} \,|\, \mathbf{x}, \tau) &= \int p(x_{new} \,|\, \mathbf{x}, \mu, \tau)p(\mu \,|\, \mathbf{x}, \tau)d\mu \qquad (9.57) \\ &= \int p(x_{new} \,|\, \mu, \tau)p(\mu \,|\, \mathbf{x}, \tau)d\mu. \qquad (9.58) \end{aligned}$$

This can again be expressed in terms of random variables:

$$X_{new} = \mu + \sigma\epsilon \tag{9.59}$$

$$\mu = \mu_{post} + \sigma_{post}\delta, \tag{9.60}$$

where now the distribution for μ is its posterior distribution. We see that X_{new} has a Gaussian distribution (it is the sum of Gaussian random variables), with predictive mean:

$$\mathbb{E}[X_{new}] = \mathbb{E}[\mu] + \sigma\mathbb{E}[\epsilon] = \mu_{post} \tag{9.61}$$

and predictive variance:

$$\text{Var}[X_{new}] = \mathbb{E}[X_{new} - \mu_{post}]^2 = \mathbb{E}[\mu - \mu_{post} + \sigma\delta]^2 = \sigma^2_{post} + \sigma^2. \tag{9.62}$$

Here we see that variances add—precisions are most natural for the posterior distribution, but variances are most natural for the predictive distribution.

Conjugacy for the variance

Let us now consider the Gaussian distribution from the point of view of the parameter σ^2:

$$p(\mathbf{x} \,|\, \mu, \sigma^2) \propto (\sigma^2)^a e^{-b/\sigma^2}. \tag{9.63}$$

where a is equal to $-1/2$ and b is equal to $\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2$. This has the flavor of the gamma distribution, but the random variable σ^2 is in the denominator rather than the numerator in the exponential. This is an *inverse gamma* distribution (see Appendix **??**).

We thus assume that the prior distribution for the variance is an inverse gamma distribution:

$$p(\sigma^2 \,|\, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}e^{-\beta/\sigma^2}, \tag{9.64}$$

where α and β are hyperparameters. To obtain the posterior distribution, we multiply by the likelihood and normalize. Dropping constants, we have:

$$p(\sigma^2 \,|\, \mathbf{x}, \mu, \alpha, \beta) \propto p(\mathbf{x} \,|\, \mu, \sigma^2)\; p(\sigma^2 \,|\, \alpha, \beta) \tag{9.65}$$

$$\propto (\sigma^2)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2/\sigma^2} (\sigma^2)^{-\alpha-1} e^{-\beta/\sigma^2} \tag{9.66}$$

$$= (\sigma^2)^{-(\alpha+\frac{n}{2})-1} e^{-(\beta + \frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2)/\sigma^2}. \tag{9.67}$$

Comparing to Eq. (9.64), we see that the posterior is an inverse gamma distribution with parameters $\alpha + \frac{n}{2}$ and $\beta + \frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2$.

This result can also be expressed using the scaled inverse chi-square distribution, an alternative parameterization of the inverse gamma distribution; see Exercise **??**.

We will also find it convenient to derive the posterior update in terms of the precision instead of the variance. Clearly, if the variance has an inverse gamma distribution, the precision has a gamma distribution. This can be obtained by a change of variables; alternatively we can simply multiple a $\text{Ga}(\alpha, \beta)$ prior by the likelihood for the precision:

$$
\begin{aligned}
p(\tau \,|\, \mathbf{x}, \mu, \alpha, \beta) &\propto p(\mathbf{x} \,|\, \mu, \tau)\; p(\tau \,|\, \alpha, \beta) && (9.68)\\
&\propto \tau^{\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2)\tau} \tau^{\alpha-1} e^{-\beta\tau} && (9.69)\\
&= \tau^{\alpha+\frac{n}{2}-1} e^{-(\beta+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2)\tau}, && (9.70)
\end{aligned}
$$

which is a gamma distribution with parameters $\alpha + \frac{n}{2}$ and $\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$.

Let us now compute the predictive probability; i.e., the probability of a new data point X_{new}, conditioning on the observed data:

$$
\begin{aligned}
p(x_{new} \,|\, \mathbf{x}, \mu, \alpha, \beta) &= \int p(x_{new} \,|\, \mathbf{x}, \mu, \tau) p(\tau \,|\, \mathbf{x}, \mu, \alpha, \beta) d\tau && (9.71)\\
&= \int p(x_{new} \,|\, \mu, \tau) p(\tau \,|\, \mathbf{x}, \mu, \alpha, \beta) d\tau. && (9.72)
\end{aligned}
$$

This is a scale mixture of a Gaussian with respect to a gamma density. As we show in Appendix **??**, this scale mixture is a Student t distribution. Specifically, we have:

$$
X_{new} \sim \text{St}\left(\mu, \frac{\alpha + n/2}{\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2}, 2\alpha + n\right) \tag{9.73}
$$

using the standard parameterization for the Student t.

Conjugacy for the mean and variance

We turn to the more realistic case in which both the mean and variance are unknown. We wish to place a prior on these quantities that is jointly conjugate. That is, the joint posterior distribution for μ and σ^2 should be in the same family as the joint prior distribution for these quantities.

Our results thus far indicate that we should consider a Gaussian distribution for the mean and an inverse gamma distribution for the variance (or a gamma distribution for the precision). As a first try in defining a jointly conjugate prior, we might consider a product of these distributions; i.e., we might assume prior independence. Unfortunately, even if μ and σ^2 are independent in the prior, it turns out that they are dependent in the posterior (Exercise **??**). This is not entirely surprising, given that μ, σ^2 and $\mathbf{x}$ form a v-structure in which $\mathbf{x}$ is observed.

Thus, to obtain a conjugate prior we need to consider some form of dependence between the mean and variance. One natural form of dependence involves linking the scales associated with the random variables $\mathbf{x}$ and μ. Thus, given the variance of $\mathbf{x}$, we may wish to define the variance of μ as a multiple of that variance. As we will see, this is particularly natural if we wish to view the prior assessment of the variability of μ in terms of "virtual observations" of data.

To simplify the algebra we will work with precisions instead of variances. We make the following specifications:

$$\begin{align} X_i &\sim N(\mu, \tau) \qquad i = 1, \ldots, n \tag{9.74} \\ \mu &\sim N(\mu_0, n_0\tau) \tag{9.75} \\ \tau &\sim \text{Ga}(\alpha, \beta), \tag{9.76} \end{align}$$

where α, β, μ_0 and n_0 are hyperparameters and where the X_i are assumed independent given μ and τ. We refer to this prior as a *normal-gamma* distribution.

To compute the posterior of μ and τ given data $\mathbf{x} = (x_1, \ldots, x_n)$, we first compute the conditional posterior of μ given τ and then find the marginal posterior of τ.

The conditional posterior of μ given τ is easy to compute. When τ is fixed, we are back in the setting of Section 9.0.4 and we can simply copy the results, plugging in $n_0\tau$ in place of τ_0:

$$\begin{align} \mathbb{E}[\mu \,|\, \mathbf{X} = \mathbf{x}] &= \frac{n\tau}{n\tau + n_0\tau}\bar{x} + \frac{n_0\tau}{n\tau + n_0\tau}\mu_0 \tag{9.77} \\ &= \frac{n}{n + n_0}\bar{x} + \frac{n_0}{n + n_0}\mu_0 \tag{9.78} \end{align}$$

and

$$\tau_{post} = n\tau + n_0\tau = (n + n_0)\tau. \tag{9.79}$$

We see that n_0 can be viewed as a "virtual sample size" on which the prior assessment of variability in μ is based; it is as if n_0 prior observations of data points with precision τ have been made in determining the prior precision of μ.

We now proceed to working out the marginal posterior of τ. We first write down all factors in the likelihood and prior that involve either τ or μ:

$$\begin{align} p(\tau, \mu \,|\, \mathbf{x}, \mu_0, n_0, \alpha, \beta) &\propto p(\tau \,|\, \alpha, \beta) p(\mu \,|\, \tau, \mu_0, n_0) p(\mathbf{x} \,|\, \mu, \tau) \nonumber \\ &\propto \left(\tau^{\alpha-1} e^{-\beta\tau}\right) \left(\tau^{1/2} e^{-\frac{n_0\tau}{2}(\mu-\mu_0)^2}\right) \left(\tau^{n/2} e^{-\frac{\tau}{2}\sum_{i=1}^n (x_i-\mu)^2}\right) \nonumber \\ &\propto \tau^{\alpha+n/2-1} e^{-(\beta+\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x})^2)\tau} \tau^{1/2} e^{-\frac{\tau}{2}[n_0(\mu-\mu_0)^2 + n(\bar{x}-\mu)^2]}, \tag{9.80} \end{align}$$

where we have used Eq. (9.52).

We now want to integrate out μ. As in Section 9.0.4, this is again a complete-the-squares problem, which we solve indirectly by working with random variables instead of densities. In particular, the last factor in Eq. (9.80) can be viewed as arising from the following specification:

$$\begin{aligned}\bar{X} &\sim \mu + \frac{1}{\sqrt{n\tau}}\epsilon && (9.81)\\ \mu &\sim \mu_0 + \frac{1}{\sqrt{n_0\tau}}\delta, && (9.82)\end{aligned}$$

where $\epsilon \sim N(0,1)$ and $\delta \sim N(0,1)$ are independent Gaussian random variables. Integrating out μ leaves us with the marginal of $\bar{X}$. But this is a Gaussian random variable with mean

$$\mathbb{E}[\bar{X}] = \mu_0 \qquad (9.83)$$

and variance

$$\begin{aligned}\mathrm{Var}[\bar{X}] &= \mathbb{E}[\bar{X} - \mu_0]^2 && (9.84)\\ &= \mathbb{E}[(\bar{X} - \mu) + (\mu - \mu_0)]^2 && (9.85)\\ &= \frac{1}{n\tau} + \frac{1}{n_0\tau} && (9.86)\\ &= \frac{n + n_0}{nn_0\tau}. && (9.87)\end{aligned}$$

Thus integration over μ leaves us with a τ-dependent factor $\exp\{-\frac{nn_0\tau}{2(n+n_0)}(\bar{x} - \mu_0)^2\}$. It also creates a normalization factor $\tau^{-1/2}$, but this cancels the $\tau^{1/2}$ factor present in Eq. (9.80).

Returning to Eq. (9.80), we have:

$$p(\tau \mid \mathbf{x}, \mu_0, n_0, \alpha, \beta) \propto \tau^{\alpha+n/2-1} e^{-(\beta + \frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2 + \frac{nn_0\tau}{2(n+n_0)}(\bar{x}-\mu_0))^2}. \qquad (9.88)$$

This is a gamma distribution, with parameters $\alpha + n/2$ and $\beta + \frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2 + \frac{nn_0}{2(n+n_0)}(\bar{x} - \mu_0)^2$.

In summary, by starting with a normal-gamma prior, we obtain a normal-gamma posterior; i.e., we have found a conjugate prior for the mean and precision of the Gaussian.

9.0.5 The general case

We now consider the general case of an exponential family distribution. Writing the exponential family density in the canonical form, we have:

$$p(x \mid \eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\}. \qquad (9.89)$$

Given a random sample, $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, we obtain:

$$p(\mathbf{x} \mid \theta) = \left(\prod_{n=1}^N h(x_n)\right) \exp\left\{\eta^T \left(\sum_{n=1}^N T(x_n)\right) - NA(\eta)\right\} \qquad (9.90)$$

as the likelihood function.

To obtain a probability density for θ we mimic the likelihood:

$$p(\eta \,|\, \tau, n_0) = H(\tau, n_0) \exp\{\tau^T \eta - n_0 A(\eta)\}, \tag{9.91}$$

where the underlying measure is Lebesgue measure. Here τ and n_0 are parameters and $H(\tau, n_0)$ denotes the normalizing factor. It is possible to show that this distribution is normalizable if $n_0 > 0$ and if τ/n_0 lies in the interior of the convex hull of the support of the measure η (Diaconis and Ylvisaker, 1979).[1]

To see that Eq. (9.91) is a conjugate prior we compute the posterior density:

$$p(\eta \,|\, x, \tau, n_0) \propto \exp\left\{ \left(\tau + \sum_{n=1}^{N} T(x_n) \right)^T \eta - (n_0 + N) A(\eta) \right\}, \tag{9.92}$$

which retains the form of Eq. (9.89), thus the prior is conjugate. We can summarize the prior-to-posterior conversion with the following update rules:

$$\begin{aligned} \tau &\rightarrow \tau + \sum_{n=1}^{N} T(x_n) && (9.93) \\ n_0 &\rightarrow n_0 + N && (9.94) \end{aligned}$$

for the parameters of the posterior.

We can also obtain conjugate priors for non-canonical representations. In particular, if we replace η by $\phi(\theta)$ in Eq. (9.89), then we obtain a conjugate prior for θ by simply replacing η in Eq. (9.91) with $\phi(\theta)$ (and changing the definition of $H(\tau, n_0)$ so that it is obtained by integrating over θ). Note that this is *not* the same prior as would be obtained by applying the change-of-variables procedure to Eq. (9.92). Such a procedure generally yields a well-defined prior, but that prior is not generally a conjugate prior for θ. (We discuss the relationship between these two priors further in Section 9.0.6 below).

Finally, let us write a general expression for the predictive distribution of a new data point X_{new} when the posterior is based on a conjugate prior. Writing $\tau_{\text{post}} = \tau + \sum_{n=1}^{N} T(x_n)$, we

[1] The latter condition has a natural interpretation. We will show in Section 9.0.6 that τ/n_0 is the expectation of the mean, and the mean certainly lies the convex hull of the support of the measure.

have

$$\begin{aligned} p(x_{\text{new}} \,|\, x) &= \int p(x_{\text{new}} \,|\, \eta) p(\eta \,|\, x, \tau, n_0) d\eta \\ &= \int h(x_{\text{new}}) \exp\{\eta^T T(x_{\text{new}}) - A(\eta)\} H(\tau_{\text{post}}, n_0 + N) \exp\{\tau_{\text{post}}^T \eta - (n_0 + N) A(\eta)\} d\eta \\ &= H(\tau_{\text{post}}, n_0 + N) \int h(x_{\text{new}}) \exp\{(\tau_{\text{post}} + T(x_{\text{new}}))^T \eta - (n_0 + N + 1) A(\eta)\} d\eta \\ &= \frac{H(\tau_{\text{post}}, n_0 + N)}{H(\tau_{\text{post}} + T(x_{\text{new}}), n_0 + N + 1)}. \end{aligned} \tag{9.95}$$

We should that the general form of predictive distribution is a ratio of normalizing factors. It should be kept in mind, however, that this is based on the canonical parameterization. For other parameterizations it may be easiest to compute the predictive distribution directly, as we saw in the case of the Gaussian distribution.

9.0.6 Linearity of the posterior expectation of the mean

In several of the examples we presented in Section **??** we saw that the posterior expectation of the mean is linear in the sufficient statistic; more precisely, it is a convex combination of the prior expectation and the maximum likelihood estimate. How general is this appealing result?

Let us first consider the general case of a canonical exponential family and the corresponding conjugate prior. We wish to compute the expectation of the mean with respect to the posterior distribution of η. Given conjugacy, the posterior distribution is in the same family as the prior, so it suffices to compute this expectation under the prior. We then substitute the appropriate parameters for the posterior in place of those for the prior. Thus we wish to compute:

$$\mathbb{E}[\mu \,|\, \tau, n_0] = \mathbb{E}[\nabla A(\eta) \,|\, \tau, n_0]. \tag{9.96}$$

To carry out this computation, we first compute the gradient of the prior density:

$$\nabla p(\eta \,|\, \tau, n_0) = p(\eta \,|\, \tau, n_0)(\tau - n_0 \nabla A(\eta)). \tag{9.97}$$

Note that if we integrate this expression with respect to η, we obtain $\tau - n_0 \mathbb{E}[\nabla A(\eta) \,|\, \tau, n_0]$ on the right-hand side. So our problem is solved if we can integrate $\nabla p(\eta \,|\, \tau, n_0)$.

It turns out that the integral of $\nabla p(\eta \,|\, \tau, n_0)$ is equal to zero. This result is a consequence of Green's theorem (a general form of the fundamental theorem of calculus), and it relies on our standing assumption that the exponential family under consideration is regular. A rigorous proof of this result was first presented in Diaconis and Ylvisaker (1979); for a simplified proof see Brown (1986). Assuming this result to be true, we have:

$$\mathbb{E}[\mu \,|\, \tau, n_0] = \tau / n_0 \tag{9.98}$$

for the expectation of the mean under the prior.

To convert this to our desired result, we recall that under the posterior τ is replaced by $\tau + \sum_{n=1}^{N} T(x_n)$ and n_0 is replaced by $n_0 + N$. We therefore obtain:

$$\begin{aligned} \mathbb{E}[\mu \,|\, \mathbf{x}, \tau, n_0] &= \frac{\tau + \sum_{n=1}^{N} T(x_n)}{n_0 + N} & (9.99) \\ &= \kappa \frac{\tau}{n_0} + (1-\kappa)\hat{\mu}_{ML}, & (9.100) \end{aligned}$$

where $\kappa = n_0/(n_0 + N)$. We thus see that the posterior expectation of the mean is a convex combination of the prior expectation and the maximum likelihood estimate in general for conjugate priors.

Diaconis and Ylvisaker (1979) also established a converse to this result. In particular, under certain regularity conditions (which hold for example for all continuous distributions in the exponential family), if the posterior expectation of the mean is linear in the sufficient statistic, then the prior distribution for the canonical parameter must be a standard conjugate prior.

While these results provide insight into the nature of conjugacy, they can't be the entire story. Recall, in particular, that in Section 9.0.1 we placed a conjugate prior on the mean parameter of the Bernoulli distribution. Although the mean parameter is not the same as the canonical parameter, we nonetheless obtained a linear posterior expectation for the mean.

Consider a change of parameterization, $\theta = \phi(\eta)$ for a one-to-one function ϕ. In choosing a prior for θ we have two choices. First, we can retain the standard conjugate prior and use the change-of-variables formula to obtain the prior for θ. Given that we have simply changed variables, we clearly still obtain a linear posterior expectation in this case. Alternatively, we can place a standard conjugate prior on θ directly. Having done so, we can then ask what happens if we transform back to the canonical parameter, using the change-of-variables formula. By the Diaconis and Ylvisaker (1979) results, we know that we obtain a linear posterior expectation if and only if the resulting prior is the standard conjugate prior. In sum, we are asking for the transformed prior,

$$p(\theta(\eta) \,|\, \tau, n_0) \left| \frac{\partial \theta}{\partial \eta} \right|, \tag{9.101}$$

to be a standard conjugate prior for η. This happens when both factors on the right-hand side mimic the likelihood. But the first factor already mimics the likelihood (because of our choice of a conjugate prior for θ). Thus we obtain the desired result if the Jacobian $|\partial\theta/\partial\eta|$ mimics the likelihood.

? and **?** have provided general conditions under which the Jacobian mimics the likelihood. In the univariate case the result is easily stated: the Jacobian mimics the likelihood if and only if the variance of the sufficient statistic is a quadratic function of the mean. In the multivariate case things are somewhat more complicated, but again the variance function plays a fundamental role. We pursue this issue in the exercises (Exercise **??** and Exercise **??**) and in the Bibliography section.

9.1 Appendix

9.1.1 Inverse gamma distribution

Let X have the gamma distribution $\text{Ga}(\alpha, \beta)$. What is the distribution of $Z = 1/X$? The Jacobian $|dx/dz|$ of the transformation $z = 1/x$ is $1/z^2$. Multiplying the gamma density

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{9.102}$$

by this Jacobian we obtain the density of Z:

$$p(z) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} e^{-\beta/z}. \tag{9.103}$$

We refer to the distribution having this density as the *inverse gamma distribution*, denoted $\text{IG}(\alpha, \beta)$.

9.1.2 Student t distribution

There are two classical ways to derive the Student t distribution: as a scale mixture of Gaussians and as a ratio of a Gaussian and the square root of a gamma variable. The first derivation, which is the one that we present here, is often associated with Bayesian statistics, where it arises when integrating over a conjugate prior for the variance of a Gaussian. The latter derivation plays an important role in frequentist statistics, where it arises as the sampling distribution of the ratio of the sample mean and the square root of the sample variance (see, e.g., Casella and Berger, 2001).

Let X have a $N(\mu, \sigma^2)$ distribution, where the location parameter μ is assume to be known and fixed. We place a conjugate prior on σ^2 and integrate. In particular, working with the precision $\tau = 1/\sigma^2$, we place a gamma prior on τ:

$$p(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau} \tag{9.104}$$

and compute the marginal probability $p(x \mid \mu, \alpha, \beta)$:

$$\begin{aligned}
p(x \mid \mu, \alpha, \beta) &= \int \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau} \left(\frac{\tau}{2\pi}\right)^{1/2} e^{-\frac{\tau}{2}(x-\mu)^2} d\tau && (9.105)\\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(2\pi)^{1/2}} \int \tau^{\alpha-\frac{1}{2}} e^{-\beta\tau} e^{-\frac{\tau}{2}(x-\mu)^2} d\tau && (9.106)\\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(2\pi)^{1/2}} \int \tau^{\alpha-\frac{1}{2}} e^{-(\beta+\frac{1}{2}(x-\mu)^2)\tau} d\tau && (9.107)\\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(2\pi)^{1/2}} \frac{\Gamma(\alpha+\frac{1}{2})}{(\beta+\frac{1}{2}(x-\mu)^2)^{\alpha+\frac{1}{2}}} && (9.108)\\
&= \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)} \frac{1}{(2\pi\beta)^{1/2}} \frac{1}{(1+\frac{1}{2\beta}(x-\mu)^2)^{\alpha+\frac{1}{2}}}, && (9.109)
\end{aligned}$$

where in passing from Eq. (9.107) to Eq. (9.108) we have recognized an unnormalized gamma integral.

We refer to the distribution having the density in Eq. (9.109) as the *Student t distribution*. It is also common to reparameterize this density by defining

$$p := \frac{\alpha}{2} \qquad \lambda := \frac{\alpha}{\beta}, \tag{9.110}$$

which yields:

$$p(x \mid \mu, \lambda, p) = \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})} \left(\frac{\lambda}{\pi p}\right)^{1/2} \frac{1}{\left(1+\frac{\lambda}{p}(x-\mu)^2\right)^{\frac{p+1}{2}}}. \tag{9.111}$$

as the density of the Student t distribution. We denote this distribution by $\mathrm{St}(\mu, \lambda, p)$.

The *Cauchy distribution* is a special case of the Student t distribution, obtained by setting $p = 1$. It is also interesting to note that if we let $p \to \infty$ in Eq. (9.111), the final factor approaches the exponential of $-\frac{\lambda}{2}(x-\mu)^2)$, and indeed it can be shown that the Gaussian distribution is a limiting case of the Student t distribution, obtained by letting $p \to \infty$ (see Exercise **??**).

Bibliography

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families.* Institute of Mathematical Statistics.

Casella, G. and Berger, R. L. (2001). *Statistical Inference.* Duxbury Press, North Scituate, MA.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Annals of Statistics*, 7:269–281.