Fully Bayesian Score Calibration assuming Gaussian Distributions

Niko Brümmer
AGNITIO LABS, South Africa

May 2011

In this note, we explore a fully Bayesian recipe for calculating likelihood-ratios of univariate normal 'score' distributions.

1 Notation

We shall use bold roman or greek fonts to represent vector-valued variables or functions. We shall use either round or square brackets to assemble vectors, e.g.: $\mathbf{x} = (s, s^2) = \left[\begin{smallmatrix} s \\ s^2 \end{smallmatrix}\right]$ has 2 components and $(\mathbf{x}, \alpha) = \left[\begin{smallmatrix} \mathbf{x} \\ \alpha \end{smallmatrix}\right]$ has 3. The dot product between two vectors will be denoted as $\mathbf{x} \cdot \mathbf{y}$.

2 Bayesian integrals with exponential likelihood and conjugate prior

We shall implement our fully Bayesian recipe by using a prior that is conjugate to an exponential family likelihood, so that we can obtain closed-form solutions to the Bayesian integrals. By taking appropriate limits of prior hyperparameters, we can make the conjugate prior non-informative.

In this section, we outline the whole inference procedure symbolically and generally for an exponential family likelihood with conjugate prior. Then in the next section, we apply this recipe to the specific case of a Gaussian likelihood, with a Gaussian-gamma prior.

Let $\mathbf{s} = s_1, s_2, \ldots, s_N$ be N scores, presumed to have been generated iid from some univariate likelihood $P(s|\boldsymbol{\gamma})$, which is an exponential family distribution with parameter vector $\boldsymbol{\gamma}$. The complication is that the parameter is not given and must be inferred from training data.

The *exponential family likelihood* for the parameter vector $\boldsymbol{\gamma}$, given data $\mathbf{s} = s_1, s_2, \ldots, s_N$, is of the form:

$$P(\mathbf{s}|\boldsymbol{\gamma}) = \prod_{i=1}^{N} P(s_i|\boldsymbol{\gamma}) = g(\boldsymbol{\gamma})^N \exp\left(\mathbf{y} \cdot \mathbf{v}(\boldsymbol{\gamma})\right) \prod_{i=1}^{N} h(s_i) \tag{1}$$

where $\mathbf{y} = \sum_{i=1}^{N} \mathbf{u}(s_i)$. The function $\mathbf{u}$ maps any datum s_i to a vector-valued *sufficient statistic*, and the function $\mathbf{v}$ maps $\boldsymbol{\gamma}$ to a vector known as the *natural parameter*, which has the same size as the statistic, so that the argument of the exponent is the dot product between parameter and summed statistic. The functions h and g are positive, scalar-valued. The role of $g(\boldsymbol{\gamma})$ is to normalize the distribution so that it integrates to 1, for any valid parameter value $\boldsymbol{\gamma}$.

We shall use a *conjugate prior* of the form:

$$P(\boldsymbol{\gamma}|\mathbf{x}, m, n) = f(\mathbf{x}, m, n) q(\boldsymbol{\gamma})^m r(\boldsymbol{\gamma})^n t(\boldsymbol{\gamma}) \exp(\mathbf{x} \cdot \mathbf{v}(\boldsymbol{\gamma})) \tag{2}$$

The *prior hyperparameters* are: the vector $\mathbf{x}$ and the scalars m, n. When the prior parameters are in this form we shall refer to them as *pseudostats*, for reasons that will become apparent below. The function f now plays the important role of normalizer, so that:

$$f(\mathbf{x}, m, n) \int q(\boldsymbol{\gamma})^m r(\boldsymbol{\gamma})^n t(\boldsymbol{\gamma}) \exp(\mathbf{x} \cdot \mathbf{v}(\boldsymbol{\gamma})) \, d\boldsymbol{\gamma} = 1 \tag{3}$$

Finally, the functions q, r, t are positive, scalar-valued. The prior (2) is conjugate to the likelihood (1), by virtue of the fact that we use the same function $\mathbf{v}(\boldsymbol{\gamma})$ and by requiring that:

$$q(\boldsymbol{\gamma}) r(\boldsymbol{\gamma}) = g(\boldsymbol{\gamma}) \tag{4}$$

As an aside, notice that (2) is also an exponential family distribution, where the natural parameter is given by $(\mathbf{x}, m, n)$ and the sufficient statistic by $(\mathbf{v}(\boldsymbol{\gamma}), \log q(\boldsymbol{\gamma}), \log r(\boldsymbol{\gamma}))$.

There are many ways to formulate conjugate priors for a given likelihood function. A simpler conjugate prior (with fewer hyperparameters) would result if we did not factorize g as in (4). A more complex conjugate prior (with more hyperparameters) would result if we factored g into more than two factors. One could even use mixtures of distributions as a conjugate prior. We chose the binary factorization of $g(\boldsymbol{\gamma})$ to correspond to the complexity of the Gaussian-gamma prior that we will use below.

The *parameter posterior* is of the form:

$$\begin{aligned} P(\boldsymbol{\gamma}|\mathbf{s},\mathbf{x},m,n) &= P(\boldsymbol{\gamma}|\mathbf{y},N,\mathbf{x},m,n) \\ &\propto P(\boldsymbol{\gamma}|\mathbf{x},m,n)P(\mathbf{s}|\boldsymbol{\gamma}) \\ &\propto q(\boldsymbol{\gamma})^{m+N}r(\boldsymbol{\gamma})^{n+N}t(\boldsymbol{\gamma})\exp\left((\mathbf{x}+\mathbf{y})\cdot\mathbf{v}(\boldsymbol{\gamma})\right) \end{aligned} \tag{5}$$

where on right-hand sides of the $\propto$ symbols, we have conveniently omitted factors that are not dependent on $\boldsymbol{\gamma}$. This posterior is of the same form as the prior (2) and can therefore be normalized as:

$$\begin{aligned} &P(\boldsymbol{\gamma}|\mathbf{y},N,\mathbf{x},m,n) = P(\boldsymbol{\gamma}|\mathbf{x}+\mathbf{y},m+N,n+N) \\ &= f(\mathbf{x}+\mathbf{y},m+N,n+N)q(\boldsymbol{\gamma})^{m+N}r(\boldsymbol{\gamma})^{n+N}t(\boldsymbol{\gamma})\exp((\mathbf{x}+\mathbf{y})\cdot\mathbf{v}(\boldsymbol{\gamma})) \end{aligned} \tag{6}$$

The *predictive distribution* for a new datum s is:

$$\begin{aligned} &P(s|\mathbf{y},N,\mathbf{x},m,n) = \int P(\boldsymbol{\gamma}|\mathbf{y},N,\mathbf{x},m,n)P(s|\boldsymbol{\gamma})\,d\boldsymbol{\gamma} \\ &= h(s)f(\mathbf{x}+\mathbf{y},m+N,n+N) \\ &\quad\times\int q(\boldsymbol{\gamma})^{m+N+1}r(\boldsymbol{\gamma})^{n+N+1}t(\boldsymbol{\gamma})\exp\left((\mathbf{x}+\mathbf{y}+\mathbf{u}(s))\cdot\mathbf{v}(\boldsymbol{\gamma})\right)\,d\boldsymbol{\gamma} \\ &= h(s)\frac{f(\mathbf{x}+\mathbf{y},m+N,n+N)}{f(\mathbf{x}+\mathbf{y}+\mathbf{u}(s),m+N+1,n+N+1)} \end{aligned} \tag{7}$$

where the integral was solved by noting that the integrand is of the same form as the integrand of (3).

For our purpose, namely to form likelihood-ratios of predictive distributions, $h(s)$ will cancel and is therefore unimportant. (In fact, for the Gaussian distribution, h is constant.) The work involved in finding a concrete solution is therefore to define the function f and its arguments, the sufficient statistics and prior hyperparameters.

3 Gaussian likelihood with Gaussian-gamma prior

3.1 Gaussian Likelihood

We parametrize the Gaussian likelihood in terms of the mean, μ and the precision (inverse variance), λ. The likelihood, given data $\mathbf{s} = s_1, s_2, \ldots, s_N$,

is:

$$\begin{aligned} P(\mathbf{s}|\mu,\lambda) &= \prod_{n=1}^{N} \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(s_n-\mu)^2\right) \\ &= \left[\left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left(\lambda\mu F - \frac{\lambda}{2}S\right) \end{aligned} \tag{8}$$

where we have defined the *first and second order statistics*:

$$F = \sum_{n=1}^{N} s_n, \qquad\qquad S = \sum_{n=1}^{N} s_n^2 \tag{9}$$

The count, N is also referred to as the zero-order statistic. For later convenience, we also define the *centred second order statistic*:

$$\bar{S} = \sum_{n=1}^{N} \left(s_n - \frac{F}{N}\right)^2 = S - N\left(\frac{F}{N}\right)^2 \tag{10}$$

Note that $\bar{S}$ is determined by all of N, F, S. The following update formula will be useful[1]. If the data is partitioned into two subsets, with respective stats N_1, F_1, S_1 and N_2, F_2, S_2, so that $N = N_1 + N_2$, $F = F_1 + F_2$ and $S = S_1 + S_2$, then:

$$\bar{S} = \bar{S}_1 + \bar{S}_2 + \frac{N_1 N_2}{N_1 + N_2}\left(\frac{F_1}{N_1} - \frac{F_2}{N_2}\right)^2 \tag{11}$$

As an aside: equating the partial derivatives of the log-likelihood to zero, it is straight-forward to see that (8) is maximized at $\mu = \frac{F}{N}$ and $\frac{1}{\lambda} = \frac{\bar{S}}{N}$.

To satisfy the generic notation of the previous section, with $\boldsymbol{\gamma} = (\mu, \lambda)$, we could[2] choose:

$$\mathbf{v}(\mu,\lambda) = \begin{bmatrix} \lambda\mu \\ -\frac{\lambda}{2} \end{bmatrix}, \qquad \mathbf{u}(s) = \begin{bmatrix} s \\ s^2 \end{bmatrix}, \qquad g(\mu,\lambda) = \lambda^{\frac{1}{2}} \exp\left(-\frac{\lambda\mu^2}{2}\right) \tag{12}$$

while $h(s) = \sqrt{\frac{1}{2\pi}}$ is constant. The dimensionality of the sufficient statistic is 2, so that the total number of scalar prior hyperparameters in a prior of the form $P(\boldsymbol{\gamma}|\mathbf{x}, m, n)$ is 4.

[1]For the last form, google for Tom Minka's 2001 online report: 'Inferring a Gaussian distribution'.

[2]This is one of many ways to do it.

3.2 Gaussian-gamma prior

The *Gaussian-gamma distribution*, denoted $\mathcal{GG}$, forms a 4-parameter conjugate prior for the unknown mean and precision of the Gaussian likelihood:

$$\mathcal{GG}(\mu, \lambda|\mu_0, \beta, a, b) = \mathcal{N}\big(\mu|\mu_0, (\beta\lambda)^{-1}\big)\mathrm{Gam}(\lambda|a, b) \tag{13}$$

where the gamma distribution, Gam, is defined in appendix A. Note that the four scalar hyperparameters, μ_0 and $\beta, a, b > 0$, are *not* in pseudostats form. To find the pseudostats, we express (13) in the generic form (2) by expanding and re-organizing:

$$\begin{aligned}
&\mathcal{GG}(\mu, \lambda|\mu_0, \beta, a, b) \\
&= \left(\frac{\beta\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\beta\lambda}{2}(\mu-\mu_0)^2\right) \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda) \\
&= \frac{\beta^{\frac{1}{2}}b^a}{\Gamma(a)} \times \left[\lambda^{\frac{1}{2}}\right]^{2a} \times \left[\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\beta} \times \frac{1}{\sqrt{2\pi\lambda}} \\
&\quad \times \exp\left(\lambda\mu\,(\beta\mu_0) - \frac{\lambda}{2}\big(2b + \beta\mu_0^2\big)\right) \\
&= f(x_1, x_2, m, n) q(\lambda)^m r(\mu, \lambda)^n t(\lambda) \exp\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \cdot \mathbf{v}(\mu, \lambda)\right)
\end{aligned} \tag{14}$$

By comparing the last two lines, we can identify the *pseudostats*:

$$x_1 = \beta\mu_0, \qquad x_2 = 2b + \beta\mu_0^2, \qquad m = 2a, \qquad n = \beta \tag{15}$$

and the functions:

$$q(\lambda) = \lambda^{\frac{1}{2}}, \qquad r(\mu, \lambda) = \exp\left(-\frac{\lambda\mu^2}{2}\right), \qquad t(\lambda) = \frac{1}{\sqrt{2\pi\lambda}} \tag{16}$$

These functions satisfy $q(\lambda)r(\mu, \lambda) = g(\mu, \lambda)$, with g as defined in (12). Finally, to express the function f, we need to invert (15):

$$\beta = n, \qquad a = \frac{m}{2}, \qquad \mu_0 = \frac{x_1}{n}, \qquad b = \frac{\bar{x}_2}{2} \tag{17}$$

where for convenience, we have defined the centred second order pseudostat:

$$\bar{x}_2 = x_2 - n\left(\frac{x_1}{n}\right)^2 \tag{18}$$

We can now express f:

$$f(x1, x2, m, n) = \frac{\beta^{\frac{1}{2}}b^a}{\Gamma(a)} = \frac{\sqrt{n}}{\Gamma\left(\frac{m}{2}\right)}\left(\frac{\bar{x}_2}{2}\right)^{\frac{m}{2}} \tag{19}$$

When $\mathcal{GG}$ is parametrized with pseudostats, we shall denote this as $\mathcal{GG}_{ps}$. To get a feel for the behaviour of $\mathcal{GG}(\mu, \lambda|\mu_0, \beta, a, b) = \mathcal{GG}_{ps}(\mu, \lambda|m, n, x_1, x_2)$, we express some of its expectations in terms of both parametrizations:

$$\langle\lambda\rangle = \frac{a}{b} = \frac{m}{\bar{x}_2} \tag{20}$$

$$\langle(\lambda - \langle\lambda\rangle)^2\rangle = \frac{a}{b^2} = \frac{2m}{\bar{x}_2^2} \tag{21}$$

$$\langle\lambda^{-1}\rangle = \frac{b}{a-1} = \frac{\bar{x}_2}{m-2} \tag{22}$$

$$\langle\mu\rangle = \mu_0 = \frac{x_1}{n} \tag{23}$$

$$\langle(\mu - \langle\mu\rangle)^2\rangle = \frac{b}{\beta(a-1)} = \frac{\langle\lambda^{-1}\rangle}{n} = \frac{\bar{x}_2}{n(m-2)} \tag{24}$$

The marginal for the precision, λ, is (by construction) a gamma distribution. The marginal for the variance, λ^{-1}, is inverse gamma. The marginal for μ is a three-parameter student's t-distribution, where m is the *degrees of freedom* parameter. The student's t becomes more Gaussian as m increases, or more heavy-tailed as m decreases. Note that some of the expectations above become undefined when $m \leq 2$. See appendices A and B for derivations.

3.3 Parameter posterior

Before we proceed to the predictive distribution, we examine the parameter posterior. To do this, we sum the data statistics, N, F, S, and prior pseudostats, m, n, x_1, x_2, to form the *posterior pseudostats*:

$$n^* = n + N, \qquad m^* = m + N \tag{25}$$

$$x_1^* = x_1 + F, \qquad x_2^* = x_2 + S \tag{26}$$

Again we define the posterior centred 2nd-order pseudostat:

$$\bar{x}_2^* = x_2^* - n^*\left(\frac{x_1^*}{n^*}\right)^2 = \bar{x}_2 + \bar{S} + \frac{nN}{n+N}\left(\frac{x_1}{n} - \frac{F}{N}\right)^2 \tag{27}$$

where we used (11). We can now compare the prior expectations listed above to the *posterior expectations*, w.r.t. $\mathcal{GG}_{ps}(\mu, \lambda | m^*, n^*, x_1^*, x_2^*)$:

$$\langle \lambda \rangle = \frac{m+N}{\bar{x}_2^*} \tag{28}$$

$$\langle \lambda^{-1} \rangle = \frac{\bar{x}_2^*}{m+N-2} \tag{29}$$

$$\langle \mu \rangle = \frac{x_1+F}{n+N} \tag{30}$$

$$\langle (\mu - \langle \mu \rangle)^2 \rangle = \frac{\bar{x}_2^*}{n(m+N-2)} \tag{31}$$

$$\langle (\lambda - \langle \lambda \rangle)^2 \rangle = \frac{2(m+N)}{(\bar{x}_2^*)^2} \tag{32}$$

3.4 Predictive distribution

To formulate the predictive distribution we need to do the following:

1. Assign prior hyperparameters. This could be done by assigning values to β, μ_0, a, b and then using (15) to find the *pseudostats*: $\mathbf{x} = (x_1, x_2)$ and m, n. Alternatively one could directly assign values to the pseudostats.

2. Extract the sufficient statistics from N 'training' data points, to get $\mathbf{y} = (y_1, y_2) = (F, S)$, as defined in (9).

3. Plug everything into (7), where $h(s) = (2\pi)^{-\frac{1}{2}}$ and f is given by (19) and $\mathbf{u}(s) = (s, s^2)$.

For now, we encapsulate steps 1 and 2 by simply using the above-defined posterior pseudostats. We now further define the *augmented pseudostats*, where s is the new (to be predicted) score:

$$n' = n^* + 1, \qquad m' = m^* + 1 \tag{33}$$

$$x_1' = x_1^* + s, \qquad x_2' = x_2^* + s^2 \tag{34}$$

The augmented centred 2nd-order pseudostat is again defined:

$$\bar{x}_2' = x_2' - n' \left(\frac{x_1'}{n'} \right)^2 = \bar{x}_2^* + \frac{n^*}{n^*+1} \left(s - \frac{x_1^*}{n^*} \right)^2 \tag{35}$$

where we used (11). Here it is convenient to define:

$$\mu^* = \frac{x_1^*}{n^*}, \qquad v^* = \frac{\bar{x}_2^*}{n^*} \tag{36}$$

so that

$$\bar{x}_2' = n^* v^* + \frac{n^*}{n^*+1}(s-\mu^*)^2 \tag{37}$$

The predictive distribution is now:

$$\begin{aligned} P(s|x_1^*, x_2^*, m^*, n^*) &= h(s)\frac{f(x_1^*, x_2^*, m^*, n^*)}{f(x_1', x_2', m', n')} \\ &= \frac{1}{\sqrt{2\pi}}\frac{\sqrt{n^*}}{\Gamma\left(\frac{m^*}{2}\right)}\left(\frac{\bar{x}_2^*}{2}\right)^{\frac{m^*}{2}}\left(\frac{\sqrt{n'}}{\Gamma\left(\frac{m'}{2}\right)}\left(\frac{\bar{x}_2'}{2}\right)^{\frac{m'}{2}}\right)^{-1} \\ &= \frac{\Gamma\left(\frac{m^*+1}{2}\right)}{\Gamma\left(\frac{m^*}{2}\right)}\frac{1}{\sqrt{\pi v^*(n^*+1)}}\left(1+\frac{(s-\mu^*)^2}{v^*(n^*+1)}\right)^{-\frac{m^*+1}{2}} \end{aligned} \tag{38}$$

This is a student's t-distribution, with m^* degrees of freedom[3]. For $m^* \gg 1$ it becomes Gaussian. The mean and variance are[4]:

$$\langle s \rangle = \mu^* = \frac{x_1^*}{n^*} \tag{39}$$

$$\langle (s-\mu^*)^2 \rangle = \frac{n^*+1}{m^*-2}v^* = \frac{n^*-1}{m^*-2} \times \frac{\bar{x}_2^*}{n^*} \tag{40}$$

A The gamma distribution

The *gamma distribution* is defined as:

$$\mathrm{Gam}(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda) \tag{41}$$

where Γ is the gamma *function*. The gamma distribution has mean and variance:

$$\langle \lambda \rangle = \frac{a}{b}, \qquad \langle (\lambda - \langle \lambda \rangle)^2 \rangle = \frac{a}{b^2} \tag{42}$$

Since the gamma distribution integrates to 1, we find the useful result[5]:

$$\int_0^\infty \lambda^{a-1} e^{-b\lambda}\, d\lambda = \frac{\Gamma(a)}{b^a} \tag{43}$$

[3]Compare to Bishop's eq. B.64, with (his) $\frac{\lambda}{\nu} = \frac{1}{v^*(n^*+1)}$ (ours).

[4]These expectations agree with Minka's 'Inferring a Gaussian distribution', section 5, if we set our $m = n$ and his $d = 1$.

[5]which can also be derived directly from the definition of the gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\, dt$.

which we can immediately apply[6] to find for example:

$$\begin{aligned}\langle\lambda^{-1}\rangle &= \int_0^\infty \lambda^{-1}\mathrm{Gam}(\lambda|a,b)\,d\lambda \\ &= \frac{b^a}{\Gamma(a)}\int \lambda^{(a-1)-1}e^{-b\lambda}\,d\lambda \\ &= \frac{b^a}{\Gamma(a)}\frac{\Gamma(a-1)}{b^{a-1}} \\ &= \frac{b}{a-1}\end{aligned} \tag{44}$$

which is the expected value of the unknown *variance parameter* of our Gaussian. We need $a > 1$ for a well-defined expected variance. As a sanity check, notice that for $a > 1$ and $\lambda > 0$, Jensen's inequality holds: $\frac{1}{\langle\lambda\rangle} = \frac{b}{a} < \langle\frac{1}{\lambda}\rangle = \frac{b}{a-1}$.

B The Gaussian-gamma distribution

The Gaussian-gamma

$$\mathcal{GG}(\mu,\lambda|\mu_0,\beta,a,b) = \mathcal{N}(\mu|\mu_0,(\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a,b) \tag{45}$$

is important because it forms both the parameter prior and posterior. To better understand it, we can examine the means and variances of μ and λ. Since the marginal for λ is just $\mathrm{Gam}(\lambda|a,b)$, the expectations for λ are as given above in section A. To find the expectations for μ, we need to integrate over both μ and λ, which gives:[7]

$$\langle\mu\rangle = \mu_0, \qquad \langle(\mu-\mu_0)^2\rangle = \frac{\langle\lambda^{-1}\rangle}{\beta} = \frac{b}{\beta(a-1)} \tag{46}$$

The marginal for μ can be derived in closed form, because the gamma factor of the Gaussian-gamma forms a conjugate prior for the precision of the Gaussian

[6] using $\Gamma(x+1) = x\Gamma(x)$

[7] Note however that when the degrees of freedom of the marginal for μ is 1 it is the Cauchy distribution, for which the mean is undefined. As long as we have enough data, non-pathological posteriors will be obtained.

factor. The marginal is:

$$
\begin{aligned}
P(\mu|\mu_0,\beta,a,b) &= \int \mathcal{GG}(\mu,\lambda,a,b)\,d\lambda \\
&= \int \left(\frac{\beta\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(\frac{-\beta\lambda}{2}(\mu-\mu_0)^2\right) \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp\left(-b\lambda\right)\,d\lambda \\
&= \frac{\sqrt{\frac{\beta}{2\pi}}b^a}{\Gamma(a)} \int \lambda^{a+\frac{1}{2}-1}\exp\left(-\lambda\left(b+\frac{\beta}{2}(\mu-\mu_0)^2\right)\right)\,d\lambda \qquad (47) \\
&= \frac{\Gamma(a+\frac{1}{2})}{\Gamma(a)}\sqrt{\frac{\beta}{2\pi}}b^a\left(b+\frac{\beta}{2}(\mu-\mu_2)^2\right)^{-a-\frac{1}{2}} \\
&= \frac{\Gamma(a+\frac{1}{2})}{\Gamma(a)}\sqrt{\frac{\beta}{2\pi b}}\left(1+\frac{\beta}{2b}(\mu-\mu_2)^2\right)^{-a-\frac{1}{2}}
\end{aligned}
$$

where we used (43) again to solve the final integral. Comparing this to equation B.64 in Bishop's book, we see this is a student's t-distribution with $2a$ degrees of freedom and expectations as given by (46).