Inferring a Gaussian distribution

Thomas P. Minka

July 25, 2001

Abstract

A common question in statistical modeling is "which out of a continuum of models are likely to have generated this data?" For the Gaussian class of models, this question can be answered completely and exactly. This paper derives the exact posterior distribution over the mean and variance of the generating distribution, i.e. $p(\mathbf{m}, \mathbf{V}|\mathbf{X})$, as well as the marginals $p(\mathbf{m}|\mathbf{X})$ and $p(\mathbf{V}|\mathbf{X})$. It also derives $p(\mathbf{X}|\text{Gaussian})$, the probability that the data came from any Gaussian whatsoever. From this we can get the posterior predictive density $p(\mathbf{x}|\mathbf{X})$, which has the most practical importance. The analysis is done for noninformative priors and for arbitrary conjugate priors. The presentation borrows from MacKay (1995). The paper concludes with a simulated classification experiment demonstrating the advantage of the Bayesian method over maximum-likelihood and unbiased estimation.

1 Introduction

We have observed N independent data points $\mathbf{X} = [\mathbf{x}_1..\mathbf{x}_N]$ from the same density parameterized by θ. This situation can be concisely expressed by the factoring

$$p(\mathbf{x}_1..\mathbf{x}_N, \theta) = p(\mathbf{x}_1|\theta)..p(\mathbf{x}_N|\theta)p(\theta)$$

Furthermore, we restrict θ to the class of Gaussian densities, i.e. $\theta = (\mathbf{m}, \mathbf{V})$:

$$p(\mathbf{x}|\mathbf{m}, \mathbf{V}) = \mathcal{N}(\mathbf{x}; \mathbf{m}, \mathbf{V}) = \frac{1}{|2\pi\mathbf{V}|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{x}-\mathbf{m}))$$

For now, we assume the absence of information about $\mathbf{m}$ and $\mathbf{V}$ other than the data $\mathbf{x}_1..\mathbf{x}_N$. How can we choose the prior $p(\mathbf{m}, \mathbf{V})$ to be noninformative? If we want to be fair to all Gaussians, then we should at least be fair to all translations and scalings of the measurement space. In other words, no Gaussian should be favored just because of our choice of origin and measurement units. For the prior to be invariant to all translations of the measurement space, we must have

$$\begin{aligned} p_{\tilde{\mathbf{m}}\mathbf{V}}(\hat{\mathbf{m}}, \hat{\mathbf{V}}) &= p_{\mathbf{m}\mathbf{V}}(\hat{\mathbf{m}}, \hat{\mathbf{V}}) \\ \text{if } \tilde{\mathbf{m}} &= \mathbf{m} + \mathbf{a} \end{aligned}$$

For the prior to be invariant to all scalings of the measurement space, we must have

$$\begin{aligned} p_{\tilde{\mathbf{m}}\tilde{\mathbf{V}}}(\hat{\mathbf{m}}, \hat{\mathbf{V}}) &= p_{\mathbf{m}\mathbf{V}}(\hat{\mathbf{m}}, \hat{\mathbf{V}}) \\ \text{if } \tilde{\mathbf{m}} &= c\mathbf{m} \\ \tilde{\mathbf{V}} &= c^2\mathbf{V} \end{aligned}$$

These two conditions are satisfied by *exactly one* distribution:

$$\begin{aligned}
p(\mathbf{m}, \mathbf{V}) &= p(\mathbf{m}|\mathbf{V})p(\mathbf{V}) && (1)\\
&= \lim_{k\to 0} \mathcal{N}(\mathbf{m}; \mathbf{m}_0, \mathbf{V}/k)\ \mathcal{IW}(\mathbf{V}; k\mathbf{V}_0, k) \qquad \text{(see appendix A)} && (2)\\
&= \alpha\, |2\pi\mathbf{V}|^{-1/2}\, |\mathbf{V}|^{-(d+1)/2} && (3)
\end{aligned}$$

where d is the dimensionality of $\mathbf{m}$. The first term is $p(\mathbf{m}|\mathbf{V})$ and the second term is $p(\mathbf{V})$. Note that $\mathbf{m}$ is not independent of $\mathbf{V}$ under this prior.

Another way to motivate this density is that it is the only density which makes the Fisher information for the parameters invariant to all possible reparameterizations of the Gaussian (e.g. using $\mathbf{W} = \mathbf{V}^{-1}$ instead of $\mathbf{V}$) (Jeffreys, 1961).

Since this density does not integrate to 1, it is improper and the symbol α is used to denote a normalizing constant which approaches zero. Throughout the paper α will serve to flag densities which are improper. Note that using an improper density in a calculation does not necessarily mean that the result will be improper, as we shall see.

Strictly speaking, one should always use the limit formula in the calculations and then take the limit at the end of the inference process. For example, the integral $\int_{\mathbf{m}} p(\mathbf{m}, \mathbf{V}) = p(\mathbf{V})$ can only be done this way. However, in the interest of clarity, this paper only uses (2) when necessary, preferring (3) whenever the result is the same.

It is worth repeating that this prior, and the resulting posterior, must change if we have any side information, such as $\mathbf{m} > 0$ or $|\mathbf{V}| < \mathbf{m}$. For example, if we know $\mathbf{m} > 0$ then $p(\mathbf{m}, \mathbf{V})$ would be 0 for $\mathbf{m} \leq 0$ and 1 otherwise.

All of the uses of $p(\cdot)$ which follow have implicit these assumptions. When other assumptions are considered, they will be made explicit, as in $p(\cdot|$ other assumption).

2 The joint posterior

Given a data set, the most honest inference about $(\mathbf{m}, \mathbf{V})$ we can make is to give a probability density for it. This density is unique, given what we have assumed so far.

Using Bayes' rule, the data generates a distribution over possible parameters, just as the parameters generate a distribution over possible data. This reversal is perfectly natural in probability theory, which has no notion of time or causation anyway. The joint posterior is readily computed to be

$$p(\mathbf{m}, \mathbf{V}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{m}, \mathbf{V})p(\mathbf{m}, \mathbf{V})}{p(\mathbf{X})} \tag{4}$$

$$= \frac{p(\mathbf{m},\mathbf{V})}{p(\mathbf{X})} \frac{1}{|2\pi\mathbf{V}|^{N/2}} \exp(-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{m})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{x}_i - \mathbf{m})) \tag{5}$$

which can be more conveniently written as

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_i \mathbf{x}_i = \mathbf{X1}/N \tag{6}$$

$$\mathbf{S} = \sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} = (\sum_i \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}) - N\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}} = \mathbf{X}(\mathbf{I} - \mathbf{11}^{\mathrm{T}}/N)\mathbf{X}^{\mathrm{T}} \tag{7}$$

$$p(\mathbf{m},\mathbf{V}|\mathbf{X}) = \frac{p(\mathbf{m},\mathbf{V})}{p(\mathbf{X})} \frac{1}{|2\pi\mathbf{V}|^{N/2}} \exp(-\frac{N}{2}(\mathbf{m} - \bar{\mathbf{x}})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{m} - \bar{\mathbf{x}}))\exp(-\frac{1}{2}\mathrm{tr}(\mathbf{S}\mathbf{V}^{-1})) \tag{8}$$

The symbol $\mathbf{1}$ denotes a column vector of 1's. The matrix $(\mathbf{I} - \mathbf{11}^{\mathrm{T}}/N)$ above is called the *centering matrix*, since it removes the mean from any matrix of data it multiplies. The matrix $\mathbf{S}$ is called the *scatter matrix* of the data.

In one dimension, this is just

$$p(m, v|\mathbf{X}) = \frac{1}{p(\mathbf{X})}\frac{\alpha}{v}\frac{1}{(2\pi v)^{(N+1)/2}}\exp(-\frac{1}{2v}\left[N(m - \bar{x})^2 + S\right]) \tag{9}$$

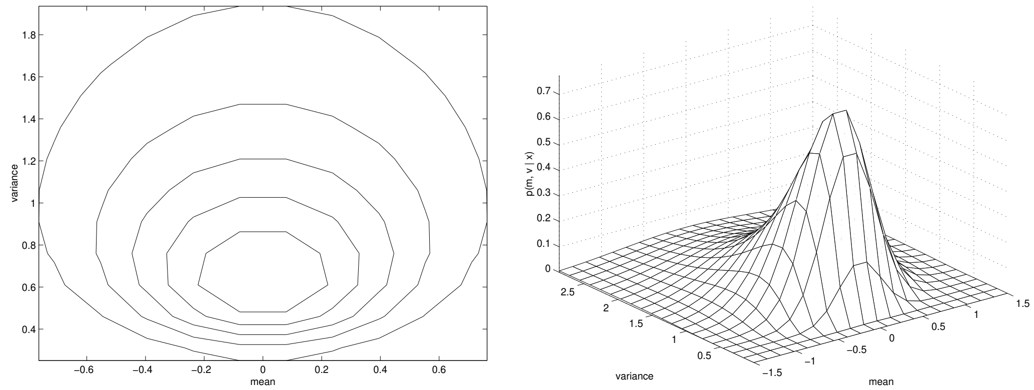

Figure 1: Contour and mesh plots of the joint posterior density. The data set of $N = 5$ points had $\bar{x} = 0$ and $S = N$. The maximum is at $(0, 5/8)$. Notice the asymmetry and long tail in v.

Figure 1 plots the contours of this density. The m dimension is symmetric about $\bar{x}$ with Gaussian falloff, but the v dimension is not symmetric, with a maximum at $S/(N + 3)$ and a long tail. The noninformative prior for v causes the mode for the variance to be smaller than would be found using the mode of the likelihood (which is S/N). There is a fairly wide region of highly probable parameter values. Keep this picture in mind whenever someone claims to have "fit a Gaussian" to their data.

The joint posterior density answers the question, "given the data and the noninformative priors, what might $\mathbf{m}$ and $\mathbf{V}$ be?" Another question we can ask is "given the data, the noninformative priors, and assuming a particular value of $\mathbf{V}$, what might $\mathbf{m}$ be?" The answer is the density $p(\mathbf{m}|\mathbf{V},\mathbf{X})$, which can be computed as

$$p(\mathbf{m}|\mathbf{V},\mathbf{X}) = \frac{p(\mathbf{m},\mathbf{V}|\mathbf{X})}{\int_{\mathbf{m}} p(\mathbf{m},\mathbf{V}|\mathbf{X})} \tag{10}$$

The integral is solved in the next section, but as a shortcut, we can observe that the dependence on $\mathbf{m}$ has the form of a Gaussian, ergo:

$$p(\mathbf{m}|\mathbf{V},\mathbf{X}) = \mathcal{N}(\mathbf{m};\bar{\mathbf{x}},\mathbf{V}/N) = \frac{N^{d/2}}{|2\pi\mathbf{V}|^{1/2}}\exp(-\frac{N}{2}(\mathbf{m}-\bar{\mathbf{x}})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{m}-\bar{\mathbf{x}})) \tag{11}$$

All questions about the parameters can be answered by performing integrals over the joint posterior, as will be demonstrated further in the next two sections.

3 The marginal posterior for the variance

Let us now ask the question "given the data, and the noninformative priors, what might $\mathbf{V}$ be?" This question is different since we are not interested in $\mathbf{m}$, and therefore must integrate it out:

$$p(\mathbf{V}|\mathbf{X}) = \int_{\mathbf{m}} p(\mathbf{m},\mathbf{V}|\mathbf{X})$$

The result is not the same as setting $\mathbf{m}$ to a particular value, such as $\bar{\mathbf{x}}$; the integration accounts for *every* possible value of $\mathbf{m}$ at once.

3.1 Univariate case

In one dimension, we proceed by factoring the joint density to create a solvable Gaussian integral:

$$\begin{aligned}
p(v|\mathbf{X}) &= \int_m p(m,v|\mathbf{X}) && (12)\\
&= \frac{1}{p(\mathbf{X})}\frac{\alpha}{v}\frac{1}{(2\pi v)^{(N+1)/2}}\int_m \exp(-\frac{1}{2v}\left[N(m-\bar{x})^2+S\right]) && (13)\\
&= \frac{1}{p(\mathbf{X})}\frac{\alpha}{v}\frac{1}{\sqrt{N}(2\pi v)^{N/2}}\exp(-\frac{S}{2v})\int_m \frac{\sqrt{N}}{\sqrt{2\pi v}}\exp(-\frac{N}{2v}(m-\bar{x})^2) && (14)\\
&= \frac{1}{p(\mathbf{X})}\frac{\alpha}{v}\frac{1}{\sqrt{N}(2\pi v)^{N/2}}\exp(-\frac{S}{2v}) && (15)\\
&= \frac{1}{p(\mathbf{X})}p(v)p(\mathbf{X}|v) && (16)
\end{aligned}$$

By matching up terms, this gives us a formula for the marginalized likelihood $p(\mathbf{X}|v)$. If we used a uniform prior over v, then this would be proportional to the posterior for v. The maximum of the marginalized likelihood occurs at S/N, the sample variance of the data.

However, we've assumed the absence of side information about v, so to continue with the non-informative prior we must compute $p(\mathbf{X})$. Since $p(v|\mathbf{X})$ must sum to one over v, we know that

$$p(\mathbf{X}) = \int_0^\infty \frac{\alpha}{v} \frac{1}{\sqrt{N}(2\pi v)^{N/2}} \exp(-\frac{S}{2v})\ dv \tag{17}$$

To solve this, we invoke the formula

$$\int_0^\infty v^{-k-1} \exp(-a/v)\ dv = \int_0^\infty w^{k-1} \exp(-aw)\ dw = \frac{\Gamma(k)}{a^k} \qquad \text{if } k > 0 \text{ and } a > 0 \tag{18}$$

where $\Gamma(k)$ is the Gamma function, equal to $(k-1)!$ for natural k but valid for all real k. Letting $a = S/2$ and $k = N/2$ gives

$$p(\mathbf{X}) = \begin{cases} \dfrac{\alpha\Gamma(N/2)}{\sqrt{N}(\pi S)^{N/2}} & \text{if } N > 1 \\ \alpha((x - m_0)^2 + v_0)^{-1/2} \qquad \text{(see appendix B)} & \text{if } N = 1 \end{cases} \tag{19}$$

This formula is significant since it is the probability of the data set averaged over all possible Gaussians, hence can be used to measure the Gaussianity of a data set. For example, if $N = 2$ then

$$p(x_1, x_2) = \frac{\alpha\sqrt{2}}{\pi(x_1 - x_2)^2} \tag{20}$$

Note that $p(\mathbf{X})$ is indeed invariant to translation and scaling of $\mathbf{X}$ (scaling $\mathbf{X}$ by c scales the density by $1/c^N$ which cancels the Jacobian). It is also invariant to all affine transformations $\mathbf{X}' = \mathbf{AX}$ (the proof is left as an exercise).

With a uniform prior ($p(m, v) = \alpha$ instead of $p(m, v) = \alpha(2\pi)^{-1}v^{-3/2}$) we would have obtained

$$p(\mathbf{X}|\text{uniform prior}) = \frac{\alpha\Gamma((N-3)/2)}{2\pi\sqrt{N}(\pi S)^{(N-3)/2}} \qquad \text{if } N > 3 \tag{21}$$

which is *not* invariant to scaling of $\mathbf{X}$, i.e. data will be considered more or less Gaussian just from changing the measurement units.

We can plug (19) into (9) to completely specify the joint density for m and v:

$$p(m, v|\mathbf{X}) = \begin{cases} \dfrac{1}{\Gamma(N/2)v} \left(\dfrac{N}{\pi S}\right)^{1/2} \left(\dfrac{S}{2v}\right)^{(N+1)/2} \exp(-\dfrac{1}{2v}\left[N(m - \bar{x})^2 + S\right]) & \text{if } N > 1 \\ \dfrac{\alpha}{2\pi v^2} \exp(-\dfrac{(m - x)^2}{2v}) & \text{if } N = 1 \end{cases} \tag{22}$$

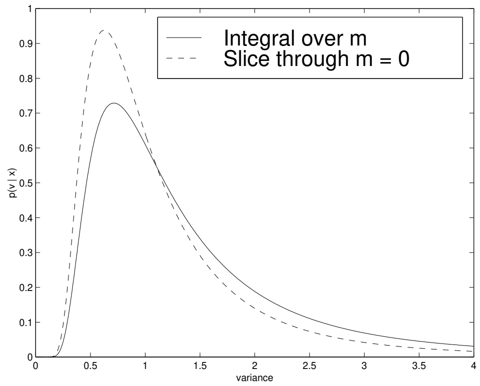

Figure 2: The posterior density for the variance parameter is smoother than a slice through the joint posterior at $m = \bar{x}$. The data set of $N = 5$ points had $\bar{x} = 0$ and $S = N$.

And by substituting (19) into (15) the variance posterior becomes

$$p(v|\mathbf{X}) = \begin{cases} \frac{1}{\Gamma(N/2)v} \left(\frac{S}{2v}\right)^{N/2} \exp(-\frac{S}{2v}) & \text{if } N > 1 \\ \alpha(2\pi)^{-1} v^{-3/2} & \text{if } N = 1 \end{cases} \tag{23}$$

$$\sim \chi^{-2}(S, N) \tag{24}$$

This formula can also be derived by observing that the dependence on v in (15) has the form of an inverse Chi-square distribution. See appendix A for the definition of this and other standard distributions used in this paper. The mode of $p(v|\mathbf{X})$ is $S/(N+2)$.

Had we simply substituted $m = \bar{x}$ in the joint density, we would have

$$p(v|m = \bar{x}, \mathbf{X}) = \frac{1}{\Gamma((N+1)/2)v} \left(\frac{S}{2v}\right)^{(N+1)/2} \exp(-\frac{S}{2v}) \tag{25}$$

See figure 2 for a comparison.

3.2 Multivariate case

In higher dimensions, we proceed as before, factoring the dependence on $\mathbf{m}$ out of the joint density:

$$p(\mathbf{V}|\mathbf{X}) = \int_{\mathbf{m}} p(\mathbf{m}, \mathbf{V}|\mathbf{X}) \tag{26}$$

$$= \frac{1}{p(\mathbf{X})} \frac{\alpha}{|\mathbf{V}|^{(d+1)/2}} \frac{1}{N^{d/2}\, |2\pi\mathbf{V}|^{N/2}} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{S}\mathbf{V}^{-1})) \tag{27}$$

$$
\begin{aligned}
&\int_{\mathbf{m}} \frac{N^{d/2}}{|2\pi\mathbf{V}|^{1/2}} \exp(-\frac{N}{2}(\mathbf{m}-\bar{\mathbf{x}})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{m}-\bar{\mathbf{x}})) \\
&= \frac{1}{p(\mathbf{X})} \frac{\alpha}{|\mathbf{V}|^{(d+1)/2}} \frac{1}{N^{d/2}\,|2\pi\mathbf{V}|^{N/2}} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{S}\mathbf{V}^{-1})) && (28) \\
&= \frac{1}{p(\mathbf{X})} p(\mathbf{V})p(\mathbf{X}|\mathbf{V}) && (29)
\end{aligned}
$$

Now we invoke the more general integral formula

$$
\begin{aligned}
\int_{\mathbf{V}\geq 0} |\mathbf{V}|^{-k-(d+1)/2} \exp(-\mathrm{tr}(\mathbf{A}\mathbf{V}^{-1})) &= \int_{\mathbf{W}\geq 0} |\mathbf{W}|^{k-(d+1)/2} \exp(-\mathrm{tr}(\mathbf{A}\mathbf{W})) && (30) \\
&= |\mathbf{A}|^{-k}\, \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma(k+1/2-i/2) && (31) \\
&\quad \text{if } k > (d-1)/2 \text{ and } |\mathbf{A}| > 0
\end{aligned}
$$

to get

$$
\begin{aligned}
p(\mathbf{X}) &= \int_{\mathbf{V}} \frac{\alpha}{|\mathbf{V}|^{(d+1)/2}} \frac{1}{N^{d/2}\,|2\pi\mathbf{V}|^{N/2}} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{S}\mathbf{V}^{-1})) && (32) \\
p(\mathbf{X}) &= \begin{cases} \alpha \dfrac{\pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma((N+1-i)/2)}{N^{d/2}\,|\pi\mathbf{S}|^{N/2}} & \text{if } N > d \\ \alpha \left(\pi \left|\mathbf{S}_0^{-1}\mathbf{S}\right|_{+} |\mathbf{S}_0|\right)^{-N/2} \quad \text{(see appendix B)} & \text{if } N \leq d \end{cases} && (33) \\
\text{where } \mathbf{S}_0 &= (\bar{\mathbf{x}}-\mathbf{m}_0)(\bar{\mathbf{x}}-\mathbf{m}_0)^{\mathrm{T}} + \mathbf{V}_0
\end{aligned}
$$

We can plug (33) into (8) to completely specify the joint density for $\mathbf{m}$ and $\mathbf{V}$:

$$
\begin{aligned}
p(\mathbf{m},\mathbf{V}|\mathbf{X}) &= \frac{1}{Z_{Nd}\,|\mathbf{V}|^{(d+1)/2}} \left|\frac{N}{\pi\mathbf{S}}\right|^{1/2} \left|\frac{\mathbf{S}\mathbf{V}^{-1}}{2}\right|^{(N+1)/2} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{S}\mathbf{V}^{-1})) && (34) \\
&\quad \exp(-\frac{N}{2}(\mathbf{m}-\bar{\mathbf{x}})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{m}-\bar{\mathbf{x}})) \qquad \text{if } N > d \\
\text{where } Z_{Nd} &= \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma((N+1-i)/2)
\end{aligned}
$$

And we can plug (33) into (28) to get an inverse Wishart distribution for the variance:

$$
\begin{aligned}
p(\mathbf{V}|\mathbf{X}) &= \frac{1}{Z_{Nd}\,|\mathbf{V}|^{(d+1)/2}} \left|\frac{\mathbf{S}\mathbf{V}^{-1}}{2}\right|^{N/2} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{S}\mathbf{V}^{-1})) \qquad \text{if } N > d && (35) \\
&\sim \mathcal{IW}(\mathbf{S}, N) && (36)
\end{aligned}
$$

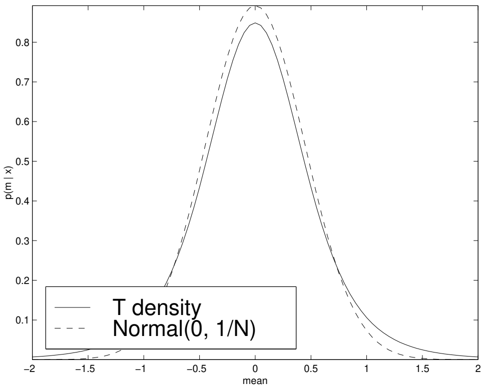

Figure 3: The posterior density for the mean parameter, a T distribution, is approximately Gaussian with mean $\bar{x}$ and variance v/N, but smoother. The data set of $N = 5$ points had $\bar{x} = 0$ and $S = N$. The curves are indistinguishable for $N > 25$.

4 The marginal posterior for the mean

Similarly, we can ask the question "given the data, and the noninformative priors, what might $\mathbf{m}$ be?" As before, we must account for every possible value of $\mathbf{V}$. The difference between integrating out $\mathbf{V}$ and fixing it to some value is especially pronounced here, since the resulting density over $\mathbf{m}$ is no longer in the Gaussian family.

Starting in one dimension, we get

$$
\begin{aligned}
p(m|\mathbf{X}) &= \int_v p(m, v|\mathbf{X}) && (37)\\
&= \frac{1}{p(\mathbf{X})} \int_0^\infty \frac{\alpha}{v} \frac{1}{(2\pi v)^{(N+1)/2}} \exp(-\frac{1}{2v}\left[N(m-\bar{x})^2 + S\right])\, dv && (38)\\
&= \frac{1}{p(\mathbf{X})} \frac{\alpha\Gamma((N+1)/2)}{\pi^{(N+1)/2}\left[N(m-\bar{x})^2 + S\right]^{(N+1)/2}} && (39)\\
&= \begin{cases} \dfrac{\Gamma((N+1)/2)}{\Gamma(N/2)} \left(\dfrac{N}{\pi S}\right)^{1/2} \left(\dfrac{N(m-\bar{x})^2}{S} + 1\right)^{-(N+1)/2} & \text{if } N > 1 \\ \dfrac{\alpha}{\pi(m-x)^2} \quad \text{(see appendix B)} & \text{if } N = 1 \end{cases} && (40)\\
&\sim \mathcal{T}(\bar{x}, S/N, N+1) && (41)
\end{aligned}
$$

This distribution has mean $\bar{x}$ and variance $\frac{1}{N(N-2)}S$ (when $N \leq 2$, the variance is infinite). For large N, it is approximately Gaussian with variance v/N, since $S \approx Nv$. See figure 3 for a comparison.

In multiple dimensions, we get

$$
p(\mathbf{m}|\mathbf{X}) = \int_{\mathbf{V}} p(\mathbf{m}, \mathbf{V}|\mathbf{X}) \quad (42)
$$

$$= \frac{1}{p(\mathbf{X})} \int_{\mathbf{V}} \frac{\alpha}{|\mathbf{V}|^{(d+1)/2}} \frac{1}{|2\pi\mathbf{V}|^{(N+1)/2}} \exp(-\frac{1}{2}\mathrm{tr}(\left[N(\mathbf{m}-\bar{\mathbf{x}})(\mathbf{m}-\bar{\mathbf{x}})^{\mathrm{T}} + \mathbf{S}\right] \mathbf{V}^{-1})) \quad (43)$$

$$= \frac{1}{p(\mathbf{X})} \frac{\alpha(2\pi)^{d(d-1)/4}}{\pi^{(N+1)d/2}} \left|N(\mathbf{m}-\bar{\mathbf{x}})(\mathbf{m}-\bar{\mathbf{x}})^{\mathrm{T}} + \mathbf{S}\right|^{-(N+1)/2} \prod_{i=1}^{d} \Gamma((N+2-i)/2) \quad (44)$$

$$= \frac{\Gamma((N+1)/2)}{\Gamma((N+1-d)/2)} \left|\frac{N\mathbf{S}^{-1}}{\pi}\right|^{1/2} (N(\mathbf{m}-\bar{\mathbf{x}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{m}-\bar{\mathbf{x}})+1)^{-(N+1)/2} \quad (45)$$

$$\sim \mathcal{T}(\bar{\mathbf{x}}, \mathbf{S}/N, N+1) \qquad \text{if } N > d \quad (46)$$

If $N \leq d$, then the distribution is improper due to our choice of an improper prior. A conjugate prior (see section 6) can remedy this.

5 The posterior predictive distribution

Finally, we ask the question "given the data, and the noninformative priors, what is the probability distribution of a new sample?" The answer, $p(\mathbf{x}|\mathbf{X})$ is called the *posterior predictive distribution*, and plays a major role in Bayesian statistics. Since it averages over all possible parameter settings for the Gaussian, it provides an optimal way to classify a new point.

To compute the posterior predictive distribution, we consider an augmented data set $\mathbf{X}' = \{\mathbf{x}\} \cup \mathbf{X}$. Then

$$\bar{\mathbf{x}}' = \frac{\mathbf{x} + N\bar{\mathbf{x}}}{N+1} = \frac{\mathbf{x} - \bar{\mathbf{x}}}{N+1} + \bar{\mathbf{x}} \quad (47)$$

$$\mathbf{S}' = \mathbf{S} + \frac{N}{N+1}(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})^{\mathrm{T}} \quad (48)$$

So

$$p(\mathbf{x}|\mathbf{X}) = p(\mathbf{x}, \mathbf{X})/p(\mathbf{X}) = p(\mathbf{X}')/p(\mathbf{X}) \quad (49)$$

$$= \frac{\Gamma((N+1)/2)}{\Gamma((N+1-d)/2)} \left(\frac{N}{N+1}\right)^{d/2} \frac{|\pi\mathbf{S}|^{N/2}}{|\pi\mathbf{S}'|^{(N+1)/2}} \quad (50)$$

$$= \frac{\Gamma((N+1)/2)}{\Gamma((N+1-d)/2)} \left|\frac{N\mathbf{S}^{-1}}{\pi(N+1)}\right|^{1/2} \left(\frac{N(\mathbf{x}-\bar{\mathbf{x}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}})}{N+1} + 1\right)^{-(N+1)/2} \quad (51)$$

$$\sim \mathcal{T}(\bar{\mathbf{x}}, \frac{N+1}{N}\mathbf{S}, N+1) \qquad \text{if } N > d \quad (52)$$

As $N \to \infty$, this density approaches a Gaussian with mean $\bar{\mathbf{x}}$ and variance $\mathbf{S}/N$, as we should expect. See figure 4 for a comparison.

An alternative to using the exact posterior predictive distribution is to use a Gaussian with somewhat larger variance than $\mathbf{S}/N$. This is perhaps the most principled way to motivate the

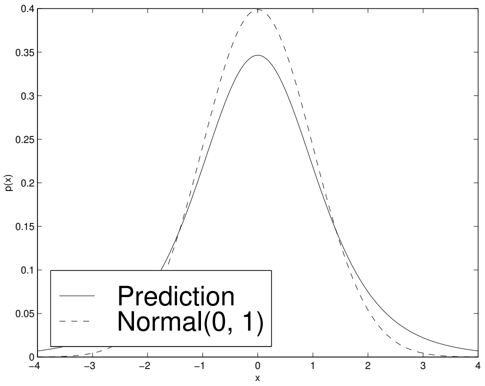

Figure 4: The posterior predictive density, a T distribution, is approximately Gaussian with mean $\bar{x}$ and variance S/N, but smoother. The data set of $N = 5$ points had $\bar{x} = 0$ and $S = N$. The curves are indistinguishable for $N > 75$.

unbiased estimator for the variance: $\mathbf{S}/(N-1)$. However, the true variance of the predictive distribution is $\frac{(N+1)}{N(N-d-1)}\mathbf{S}$ (when $N > d+1$), which may be a better choice.

6 Conjugate priors

A *conjugate prior* is a prior over the parameters which happens to have the same form as the posterior over the parameters. In the Gaussian case, this means it has the form of equation 34. This choice of prior is particularly convenient since it doesn't change the form of the posterior, as will be now shown.

The preceding analysis can be repeated for conjugate priors simply by considering an augmented data set $\mathbf{X}' = \mathbf{X} \cup \mathbf{X}_p$, where $\mathbf{X}$ has N points with sufficient statistics $\bar{\mathbf{x}}$ and $\mathbf{S}$, and $\mathbf{X}_p$ has N_p points with sufficient statistics $\bar{\mathbf{x}}_p$ and $\mathbf{S}_p$. Then

$$\bar{\mathbf{x}}' = \frac{N\bar{\mathbf{x}} + N_p\bar{\mathbf{x}}_p}{N + N_p} \tag{53}$$

$$\mathbf{S}' = \mathbf{S} + \mathbf{S}_p + \frac{N N_p}{N + N_p}(\bar{\mathbf{x}} - \bar{\mathbf{x}}_p)(\bar{\mathbf{x}} - \bar{\mathbf{x}}_p)^{\mathrm{T}} \tag{54}$$

Any conjugate prior can be interpreted as the posterior for a virtual data set $\mathbf{X}_p$. Therefore, if we are given a conjugate prior, we can recover the parameters $\bar{\mathbf{x}}_p$, $\mathbf{S}_p$, and N_p from it, then use the above formulas to get the desired posterior or predictive distribution based on the real data plus the virtual data. The noninformative priors we started with can be interpreted as the limit of a conjugate prior as $N_p \to 0$.

7 Predicting multiple samples

We can use the same data-splitting idea to predict multiple new samples. Consider an augmented data set $\mathbf{X}' = \mathbf{X} \cup \mathbf{Y}$, where $\mathbf{X}$ has N points with sufficient statistics $\bar{\mathbf{x}}$ and $\mathbf{S}$ and $\mathbf{Y}$ has K points. Then

$$\begin{aligned} \mathbf{S}' &= \mathbf{S} + (\mathbf{Y} - \bar{\mathbf{x}}\mathbf{1}^{\mathrm{T}})\mathbf{C}(\mathbf{Y} - \bar{\mathbf{x}}\mathbf{1}^{\mathrm{T}})^{\mathrm{T}} && (55) \\ \mathbf{C} &= \mathbf{I}_K - \mathbf{1}\mathbf{1}^{\mathrm{T}}/(N+K) && (56) \end{aligned}$$

so

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}) &= p(\mathbf{Y},\mathbf{X})/p(\mathbf{X}) && (57) \\ &= \frac{\prod_{i=1}^d \Gamma((N+K+1-i)/2)}{\prod_{i=1}^d \Gamma((N+1-i)/2)} \left(\frac{N}{N+K}\right)^{d/2} \frac{|\pi\mathbf{S}|^{N/2}}{|\pi\mathbf{S}'|^{(N+K)/2}} && (58) \\ &= \frac{\prod_{i=1}^d \Gamma((N+K+1-i)/2)}{\prod_{i=1}^d \Gamma((N+1-i)/2)} |\mathbf{C}|^{d/2} |\pi\mathbf{S}|^{-K/2} \left|(\mathbf{Y} - \bar{\mathbf{x}}\mathbf{1}^{\mathrm{T}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{Y} - \bar{\mathbf{x}}\mathbf{1}^{\mathrm{T}})\mathbf{C} + \mathbf{I}_K\right|^{-(N+K)/2} && (59) \\ &\sim \mathcal{T}(\bar{\mathbf{x}}\mathbf{1}^{\mathrm{T}}, \mathbf{S}, \mathbf{C}, N+K) && (60) \end{aligned}$$

This formula says that each new sample $\mathbf{y}_i$ has mean $\bar{\mathbf{x}}$, as we would expect. However, the samples are correlated by the matrix $\mathbf{C}$ since the true parameters of the Gaussian are unknown. Think of it this way: once you've seen $(K-1)$ new samples, your prediction of the Kth sample should be influenced, since you've learned more about the true parameters.

8 Classification example

Now let's bring this all together. Suppose we want to classify a new data point $\mathbf{x}$ into two classes ω_1 and ω_2. All of the information that we have about these classes is (1) that their sampling distribution is Gaussian, (2) their prior probabilities are equal, and (3) some data $\mathbf{X}_1$ of size N_1 from ω_1 and data $\mathbf{X}_2$ of size N_2 from ω_2. How should we classify $\mathbf{x}$ to minimize the probability of error?

First we compute the sample mean and sample scatter of both data sets, giving $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, $\mathbf{S}_1$, and $\mathbf{S}_2$. Now if we were to use the maximum likelihood (ML) approach, then we would assume that

$$p(\mathbf{x}|\omega_i) \sim \mathcal{N}(\bar{\mathbf{x}}_i, \mathbf{S}_i/N_i) \qquad (61)$$

and classify $\mathbf{x}$ as ω_1 iff $p(\mathbf{x}|\omega_1) > p(\mathbf{x}|\omega_2)$.

If we were to use the unbiased estimator (UB) approach, then we would do the same thing except that the variances would be assumed to be $\mathbf{S}_i/(N_i - 1)$.

If we instead use the exact Bayesian (EB) approach developed in this paper, then we compute

$$p(\mathbf{x}|\mathbf{X}_i) \sim \mathcal{T}(\bar{\mathbf{x}}_i, \frac{N_i+1}{N_i}\mathbf{S}_i, N_i+1) \tag{62}$$

and classify $\mathbf{x}$ as ω_1 iff $p(\mathbf{x}|\mathbf{X}_1) > p(\mathbf{x}|\mathbf{X}_2)$.

Finally, if we use the approximate Bayesian (AB) approach, then we would assume that

$$p(\mathbf{x}|\omega_i) \sim \mathcal{N}(\bar{\mathbf{x}}_i, \frac{(N_i+1)}{N_i(N_i-d-1)}\mathbf{S}_i) \tag{63}$$

and classify $\mathbf{x}$ as ω_1 iff $p(\mathbf{x}|\omega_1) > p(\mathbf{x}|\omega_2)$.

Let's test these four approaches as follows. The true mean of ω_1 is $\mathbf{0}$ and the true mean of ω_2 is $\mathbf{1}$. The two variances are independently drawn from a $\mathcal{W}(\mathbf{I}, d)$ distribution. Let N_1, N_2, and d vary. For each choice of these, the training sets $\mathbf{X}_1$ and $\mathbf{X}_2$ are sampled from the class distributions. Then a coin is flipped 50000 times and each time an $\mathbf{x}$ is sampled from one of the two classes and classified. This gives an estimate of the error rate.

The results are reported in terms of the *advantage* of method Z:

$$\text{advantage of } Z = \frac{\text{error rate of ML} - \text{error rate of } Z}{\text{error rate of ML} - \text{optimal error rate}} \tag{64}$$

The optimal error rate is computed empirically as the error rate of using the true class distributions to classify the testing points. So a technique with an advantage of 0.5 is exceeding ML by half the distance to the optimum. The plots show the advantage averaged over 20 different choices of the class-conditional densities, with error bars to show the variation of the advantage about its mean. Of course, ML always has an advantage of zero so it is not shown on the plots.

Figure 5 plots the advantage of each of the three approaches for the case $N_2 = N_1$ as d and N_1 vary. We see a clear ordering of EB-AB-UB, with EB having the highest advantage.

Figure 6 plots the advantage of each of the three approaches for the case $N_2 = 2N_1$ as d and N_1 vary. Here the difference is even more dramatic. One can roughly conclude from this experiment that the Bayesian prediction, even with a noninformative prior, is preferable to using the maximum-likelihood estimate, especially for high dimensionality and small and uneven training sets. It should be clear from the formulas that there will be negligible difference for large training sets, however the difference in computation for the techniques is also negligible.

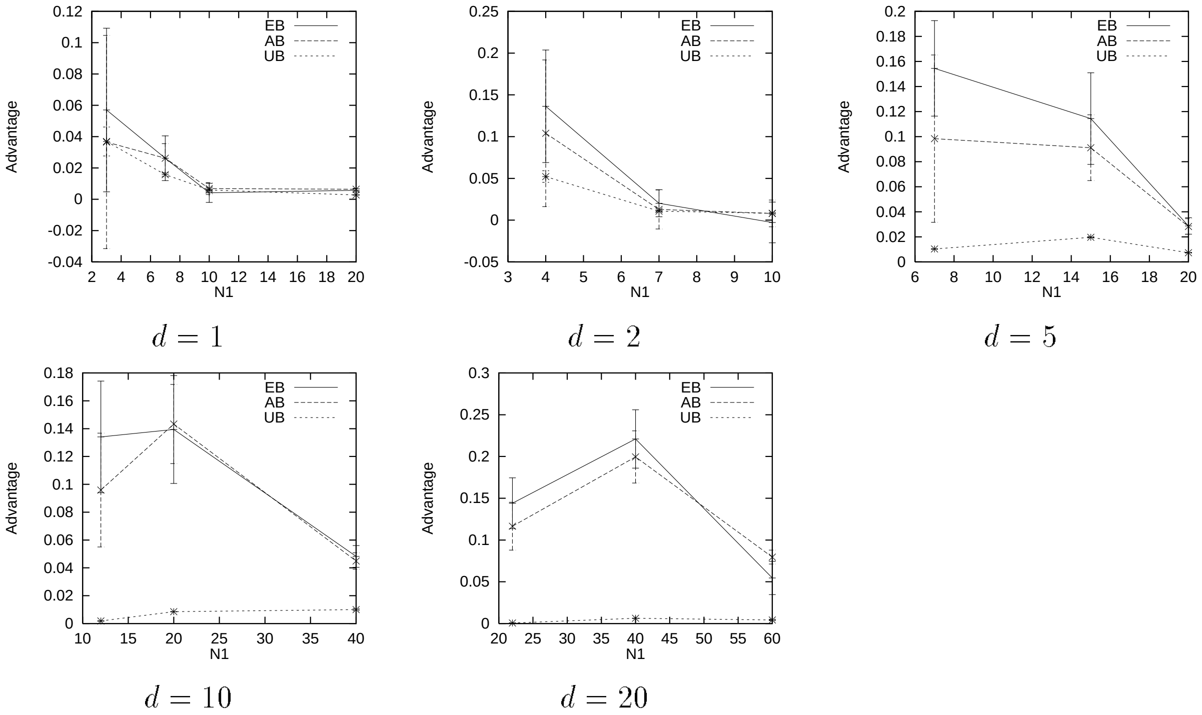

Figure 5: The average advantage over ML when $N_2 = N_1$.

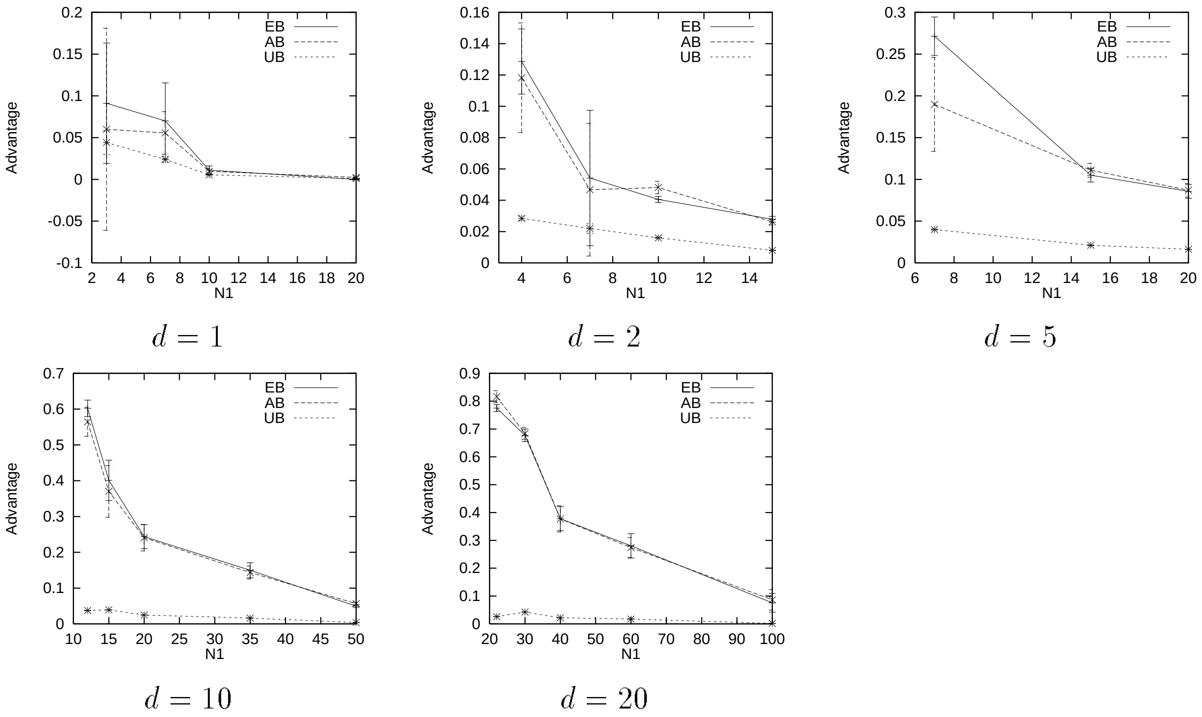

Figure 6: The average advantage over ML when $N_2 = 2N_1$.

A Standard densities

These densities and their properties are more fully described in Box and Tiao (1973).

A random vector $\mathbf{x}$ of length d is said to be T distributed with parameters $\mathbf{m}$, $\mathbf{V}$, and n if the density of $\mathbf{x}$ is

$$\begin{aligned} p(\mathbf{x}) &\sim \mathcal{T}(\mathbf{m}, \mathbf{V}, n) & (65) \\ &= \frac{\Gamma(n/2)}{\Gamma((n-d)/2)} |\pi\mathbf{V}|^{-1/2} \left((\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{x}-\mathbf{m}) + 1\right)^{-n/2} & (66) \end{aligned}$$

In the univariate case,

$$\begin{aligned} p(x) &\sim \mathcal{T}(m, v, n) & (67) \\ &= \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \frac{1}{\sqrt{\pi v}} \left(\frac{(x-m)^2}{v} + 1\right)^{-n/2} & (68) \end{aligned}$$

The mean and mode of $\mathbf{x}$ is $\mathbf{m}$. The covariance matrix of $\mathbf{x}$ is $\mathbf{V}/(n-d-2)$ when $n > d+2$ and does not exist when $n \leq d+2$. As $n \to \infty$, the density approaches $\mathcal{N}(\mathbf{m}, \mathbf{V}/n)$.

A random d by k matrix $\mathbf{X}$ is said to be T distributed with parameters $\mathbf{M}$, $\mathbf{V}$, $\mathbf{C}$, and n if the density of $\mathbf{X}$ is

$$\begin{aligned} p(\mathbf{X}) &\sim \mathcal{T}(\mathbf{M}, \mathbf{V}, \mathbf{C}, n) & (69) \\ &= \frac{\prod_{i=1}^{d} \Gamma((n+1-i)/2)}{\prod_{i=1}^{d} \Gamma((n-k+1-i)/2)} |\mathbf{C}|^{d/2} |\pi\mathbf{V}|^{-k/2} \left|(\mathbf{X}-\mathbf{M})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{X}-\mathbf{M})\mathbf{C} + \mathbf{I}_k\right|^{-n/2} & (70) \end{aligned}$$

A random d by d positive definite matrix $\mathbf{V}$ is said to be Wishart distributed with parameters $\mathbf{C}$ and n if the density of $\mathbf{V}$ is

$$\begin{aligned} p(\mathbf{V}) &\sim \mathcal{W}(\mathbf{C}, n) & (71) \\ &= \frac{1}{Z_{nd} |\mathbf{V}|^{(d+1)/2}} \left|\frac{\mathbf{V}\mathbf{C}^{-1}}{2}\right|^{n/2} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}\mathbf{C}^{-1})) & (72) \\ \text{where } Z_{nd} &= \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma((n+1-i)/2) \end{aligned}$$

In the univariate case, we would say that v is Chi-square distributed or Gamma distributed:

$$\begin{aligned} p(v) &\sim \chi^2(c, n) \sim \Gamma(n/2, 2c) & (73) \\ &= \frac{1}{\Gamma(n/2)v} \left(\frac{v}{2c}\right)^{n/2} \exp(-\frac{v}{2c}) & (74) \end{aligned}$$

For the density to be proper we must have $n > d-1$. The mean of $\mathbf{V}$ is $n\mathbf{C}$. The mode of $\mathbf{V}$ is $(n-d-1)\mathbf{C}$. In the univariate case, the variance of v is $2nc^2$.

A random d by d positive definite matrix $\mathbf{V}$ is said to be inverse Wishart distributed with parameters $\mathbf{C}$ and n if the density of $\mathbf{V}$ is

$$\begin{aligned} p(\mathbf{V}) &\sim \mathcal{IW}(\mathbf{C}, n) & (75)\\ &= \frac{1}{Z_{nd}\,|\mathbf{V}|^{(d+1)/2}}\left|\frac{\mathbf{V}^{-1}\mathbf{C}}{2}\right|^{n/2}\exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{C})) & (76)\\ \text{where } Z_{nd} &= \pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma((n+1-i)/2) \end{aligned}$$

In the univariate case, we would say that v is inverse Chi-square distributed or inverse Gamma distributed:

$$\begin{aligned} p(v) &\sim \chi^{-2}(c, n) \sim \mathcal{IG}(n/2, 2c) & (77)\\ &= \frac{1}{\Gamma(n/2)v}\left(\frac{c}{2v}\right)^{n/2}\exp(-\frac{c}{2v}) & (78) \end{aligned}$$

The mean of $\mathbf{V}$ is $\mathbf{C}/(n-d-1)$. The mode of $\mathbf{V}$ is $\mathbf{C}/(n+d+1)$. In the univariate case, the variance of v is $\dfrac{2c^2}{(n-2)^2(n-4)}$.

B Extreme cases

This appendix derives the special cases of the main formulas in the paper where the limit equation (2) must be used.

In the univariate case when $N = 1$, the joint density of x, m, and v is

$$p(x, m, v) = \lim_{k\to 0}\mathcal{N}(x; m, v)\mathcal{N}(m; m_0, v/k)\chi^{-2}(v; kv_0, k) \quad (79)$$

Integrating out m gives

$$\begin{aligned} p(x, v) &= \lim_{k\to 0}\chi^{-2}(v; kv_0, k)\int_m \mathcal{N}(x; m, v)\mathcal{N}(m; m_0, v/k) & (80)\\ &= \lim_{k\to 0}\chi^{-2}(v; kv_0, k)\mathcal{N}(x; m_0, v/k + v) & (81)\\ &= \lim_{k\to 0}\frac{1}{\Gamma(k/2)v}\left(\frac{kv_0}{2v}\right)^{k/2}\exp(-\frac{kv_0}{2v})\frac{k^{1/2}}{(2\pi(k+1)v)^{1/2}}\exp(-\frac{k(x-m_0)^2}{2(k+1)v}) & (82)\\ &= \lim_{k\to 0}\frac{1}{\Gamma(k/2)v}(\pi(k+1)v_0)^{-1/2}\left(\frac{kv_0}{2v}\right)^{(k+1)/2}\exp(-\frac{k(v_0 + \frac{1}{k+1}(x-m_0)^2)}{2v}) & (83) \end{aligned}$$

Integrating out v gives

$$p(x) = \lim_{k\to 0} \frac{\Gamma((k+1)/2)}{\Gamma(k/2)} (\pi(k+1)v_0)^{-1/2} \frac{v_0^{(k+1)/2}}{(\frac{1}{k+1}(x-m_0)^2 + v_0)^{(k+1)/2}} \quad (84)$$

$$= \lim_{k\to 0} \frac{\Gamma((k+1)/2)}{\Gamma(k/2)} (\pi(k+1)v_0)^{-1/2} \left(\frac{(x-m_0)^2}{(k+1)v_0} + 1\right)^{-(k+1)/2} \quad (85)$$

$$= \lim_{k\to 0} \mathcal{T}(m_0, (k+1)v_0, k+1) \quad (86)$$

$$= \alpha((x-m_0)^2 + v_0)^{-1/2} \quad (87)$$

which was used in (19). Note that this formula is the same as the limit of the predictive density (51) for x where $\bar{x} = m_0$, $S = v_0 N$, and $N \to 0$. It is also the limit of the regular formula for $p(\mathbf{X})$ where $S = v_0 k + \frac{k}{k+1}(x-m_0)^2$ for $k \to 0$, i.e. the limiting case of a conjugate prior.

In multiple dimensions, we can use this idea to get (33). The equivalent scatter matrix with a conjugate prior is $\mathbf{S} + k\mathbf{V}_0 + \frac{Nk}{N+k}(\bar{\mathbf{x}} - \mathbf{m}_0)(\bar{\mathbf{x}} - \mathbf{m}_0)^{\mathrm{T}}$. Plugging this into the regular formula for $p(\mathbf{X})$ requires computing

$$\lim_{k\to 0} |\mathbf{S} + k\mathbf{S}_0| / k^{d+1-N} \quad (88)$$

where

$$\mathbf{S}_0 = \mathbf{V}_0 + (\bar{\mathbf{x}} - \mathbf{m}_0)(\bar{\mathbf{x}} - \mathbf{m}_0)^{\mathrm{T}} \quad (89)$$

To compute this, let's define $|\mathbf{A}|_+$ to be the *pseudo-determinant* of square matrix $\mathbf{A}$:

$$|\mathbf{A}|_+ = \lim_{k\to 0} |\mathbf{A} + k\mathbf{I}| / k^{\mathrm{rows}(\mathbf{A}) - \mathrm{rank}(\mathbf{A})} \quad (90)$$

which is the product of the nonzero eigenvalues of $\mathbf{A}$ (or 1 if they are all zero). Then

$$\lim_{k\to 0} |\mathbf{S} + k\mathbf{S}_0| / k^{d+1-N} = \left|\mathbf{S}_0^{-1}\mathbf{S}\right|_+ |\mathbf{S}_0| \quad (91)$$

which gives (33).

Returning to the univariate case, if we instead integrate out v first:

$$p(x,m) = \lim_{k\to 0} \int_v \frac{1}{\Gamma(k/2)v} \left(\frac{kv_0}{2v}\right)^{k/2} \frac{\sqrt{k}}{2\pi v} \exp(-\frac{kv_0 + k(m-m_0)^2 + (x-m)^2}{2v}) \quad (92)$$

$$= \lim_{k\to 0} \int_v \frac{1}{\Gamma(k/2)\sqrt{k}v} \frac{1}{\pi v_0} \left(\frac{kv_0}{2v}\right)^{(k+2)/2} \exp(-\frac{kv_0 + k(m-m_0)^2 + (x-m)^2}{2v}) \quad (93)$$

$$= \lim_{k\to 0} \frac{\Gamma((k+2)/2)}{\Gamma(k/2)\sqrt{k}} \frac{1}{\pi v_0} \frac{(kv_0)^{(k+2)/2}}{(kv_0 + k(m-m_0)^2 + (x-m)^2)^{(k+2)/2}} \quad (94)$$

$$= \lim_{k\to 0} \frac{\Gamma((k+2)/2)}{\Gamma(k/2)\sqrt{k}} \frac{1}{\pi v_0} \left(1 + \frac{(m-m_0)^2}{v_0} + \frac{(x-m)^2}{kv_0}\right)^{-(k+2)/2} \quad (95)$$

$$= \frac{\alpha}{\pi(x-m)^2} \quad (96)$$

which was used in (40).

Acknowledgements

Thanks to Rosalind Picard for improving the presentation and Ali Rahimi for checking the equations.

References

[1] George E. P. Box and George C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, 1973.

[2] Harold Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, third edition, 1961.

[3] D. J. C. MacKay. Developments in probabilistic modelling with neural networks—ensemble learning. In *Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks, Nijmegen, Netherlands, 14-15 September 1995*, pages 191–198, Berlin, 1995. Springer. `http://www.cs.toronto.edu/~mackay/README.html`.