24

Exact Marginalization

How can we avoid the exponentially large cost of complete enumeration of all hypotheses? Before we stoop to approximate methods, we explore two approaches to exact marginalization: first, marginalization over continuous variables (sometimes known as nuisance parameters) by doing *integrals*; and second, summation over discrete variables by message-passing.

Exact marginalization over continuous parameters is a macho activity enjoyed by those who are fluent in definite integration. This chapter uses gamma distributions; as was explained in the previous chapter, gamma distributions are a lot like Gaussian distributions, except that whereas the Gaussian goes from $-\infty$ to ∞, gamma distributions go from 0 to ∞.

▶ 24.1 Inferring the mean and variance of a Gaussian distribution

We discuss again the one-dimensional Gaussian distribution, parameterized by a mean μ and a standard deviation σ:

$$P(x \,|\, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \equiv \text{Normal}(x; \mu, \sigma^2). \tag{24.1}$$

When inferring these parameters, we must specify their prior distribution. The prior gives us the opportunity to include specific knowledge that we have about μ and σ (from independent experiments, or on theoretical grounds, for example). If we have no such knowledge, then we can construct an appropriate prior that embodies our supposed ignorance. In section 21.2, we assumed a uniform prior over the range of parameters plotted. If we wish to be able to perform exact marginalizations, it may be useful to consider *conjugate priors*; these are priors whose functional form combines naturally with the likelihood such that the inferences have a convenient form.

Conjugate priors for μ and σ

The conjugate prior for a mean μ is a Gaussian: we introduce two 'hyperparameters', μ_0 and σ_μ, which parameterize the prior on μ, and write $P(\mu \,|\, \mu_0, \sigma_\mu) = \text{Normal}(\mu; \mu_0, \sigma_\mu^2)$. In the limit $\mu_0 = 0$, $\sigma_\mu \to \infty$, we obtain the *noninformative prior* for a location parameter, the flat prior. This is *noninformative* because it is *invariant* under the natural reparameterization $\mu' = \mu + c$. The prior $P(\mu) =$ const. is also an *improper* prior, that is, it is not normalizable.

The conjugate prior for a standard deviation σ is a gamma distribution, which has two parameters b_β and c_β. It is most convenient to define the prior

density of the inverse variance (the *precision* parameter) $\beta = 1/\sigma^2$:

$$P(\beta) = \Gamma(\beta; b_\beta, c_\beta) = \frac{1}{\Gamma(c_\beta)} \frac{\beta^{c_\beta - 1}}{b_\beta^{c_\beta}} \exp\left(-\frac{\beta}{b_\beta}\right), \quad 0 \leq \beta < \infty. \tag{24.2}$$

This is a simple peaked distribution with mean $b_\beta c_\beta$ and variance $b_\beta^2 c_\beta$. In the limit $b_\beta c_\beta = 1, c_\beta \to 0$, we obtain the noninformative prior for a scale parameter, the $1/\sigma$ prior. This is 'noninformative' because it is invariant under the reparameterization $\sigma' = c\sigma$. The $1/\sigma$ prior is less strange-looking if we examine the resulting density over $\ln \sigma$, or $\ln \beta$, which is flat. This is the prior that expresses ignorance about σ by saying 'well, it could be 10, or it could be 1, or it could be 0.1, ...' Scale variables such as σ are usually best represented in terms of their logarithm. Again, this noninformative $1/\sigma$ prior is improper.

Reminder: when we change variables from σ to $l(\sigma)$, a one-to-one function of σ, the probability density transforms from $P_\sigma(\sigma)$ to

$$P_l(l) = P_\sigma(\sigma) \left| \frac{\partial \sigma}{\partial l} \right|.$$

Here, the Jacobian is

$$\left| \frac{\partial \sigma}{\partial \ln \sigma} \right| = \sigma.$$

In the following examples, I will use the improper noninformative priors for μ and σ. Using improper priors is viewed as distasteful in some circles, so let me excuse myself by saying it's for the sake of readability; if I included proper priors, the calculations could still be done but the key points would be obscured by the flood of extra parameters.

Maximum likelihood and marginalization: σ_N and σ_{N-1}

The task of inferring the mean and standard deviation of a Gaussian distribution from N samples is a familiar one, though maybe not everyone understands the difference between the σ_N and σ_{N-1} buttons on their calculator. Let us recap the formulae, then derive them.

Given data $D = \{x_n\}_{n=1}^N$, an 'estimator' of μ is

$$\bar{x} \equiv \textstyle\sum_{n=1}^N x_n / N, \tag{24.3}$$

and two estimators of σ are:

$$\sigma_N \equiv \sqrt{\frac{\sum_{n=1}^N (x_n - \bar{x})^2}{N}} \quad \text{and} \quad \sigma_{N-1} \equiv \sqrt{\frac{\sum_{n=1}^N (x_n - \bar{x})^2}{N-1}}. \tag{24.4}$$

There are two principal paradigms for statistics: sampling theory and Bayesian inference. In sampling theory (also known as 'frequentist' or orthodox statistics), one invents *estimators* of quantities of interest and then chooses between those estimators using some criterion measuring their sampling properties; there is no clear principle for deciding which criterion to use to measure the performance of an estimator; nor, for most criteria, is there any systematic procedure for the construction of optimal estimators. In Bayesian inference, in contrast, once we have made explicit all our assumptions about the model and the data, our inferences are mechanical. Whatever question we wish to pose, the rules of probability theory give a unique answer which consistently takes into account all the given information. Human-designed estimators and confidence intervals have no role in Bayesian inference; human input only enters into the important tasks of designing the hypothesis space (that is, the specification of the model and all its probability distributions), and figuring out how to do the computations that implement inference in that space. The answers to our questions are probability distributions over the quantities of interest. We often find that the estimators of sampling theory emerge automatically as modes or means of these posterior distributions when we choose a simple hypothesis space and turn the handle of Bayesian inference.

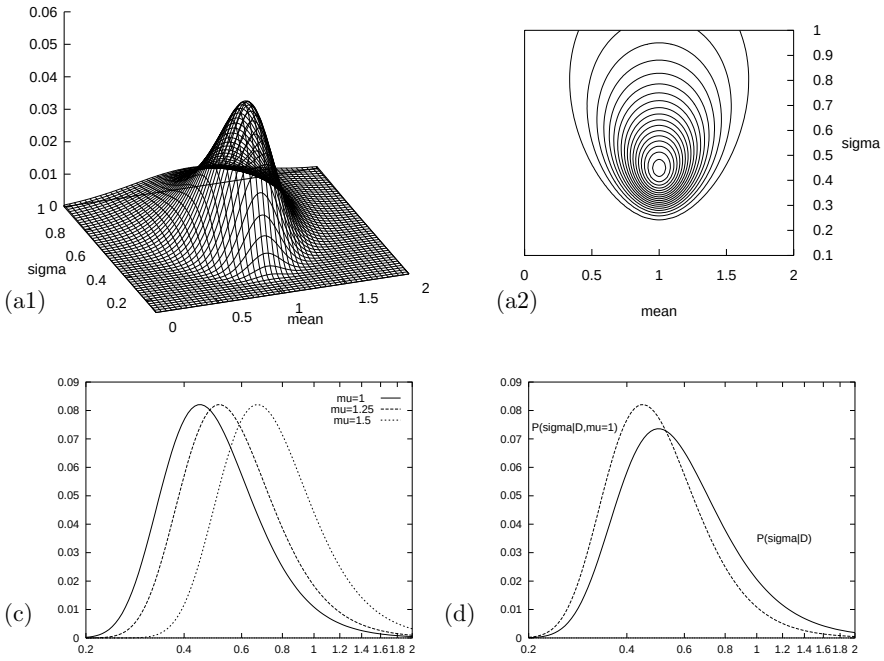


Figure 24.1. The likelihood function for the parameters of a Gaussian distribution, repeated from figure 21.5.
(a1, a2) Surface plot and contour plot of the log likelihood as a function of μ and σ. The data set of $N = 5$ points had mean $\bar{x} = 1.0$ and $S = \sum(x - \bar{x})^2 = 1.0$. Notice that the maximum is skew in σ. The two estimators of standard deviation have values $\sigma_N = 0.45$ and $\sigma_{N-1} = 0.50$.
(c) The posterior probability of σ for various fixed values of μ (shown as a density over $\ln \sigma$).
(d) The posterior probability of σ, $P(\sigma \mid D)$, assuming a flat prior on μ, obtained by projecting the probability mass in (a) onto the σ axis. The maximum of $P(\sigma \mid D)$ is at σ_{N-1}. By contrast, the maximum of $P(\sigma \mid D, \mu = \bar{x})$ is at σ_N. (Both probabilities are shows as densities over $\ln \sigma$.)

In sampling theory, the estimators above can be motivated as follows. $\bar{x}$ is an unbiased estimator of μ which, out of all the possible unbiased estimators of μ, has smallest variance (where this variance is computed by averaging over an ensemble of imaginary experiments in which the data samples are assumed to come from an unknown Gaussian distribution). The estimator $(\bar{x}, \sigma_N)$ is the maximum likelihood estimator for (μ, σ). The estimator σ_N is *biased*, however: the expectation of σ_N, given σ, averaging over many imagined experiments, is not σ.

Exercise 24.1.[2, p.323] Give an intuitive explanation why the estimator σ_N is biased.

This bias motivates the invention, in sampling theory, of σ_{N-1}, which can be shown to be an unbiased estimator. Or to be precise, it is σ^2_{N-1} that is an unbiased estimator of σ^2.

We now look at some Bayesian inferences for this problem, assuming non-informative priors for μ and σ. The emphasis is thus not on the priors, but rather on (a) the likelihood function, and (b) the concept of marginalization. The joint posterior probability of μ and σ is proportional to the likelihood function illustrated by a contour plot in figure 24.1a. The log likelihood is:

$$\ln P(\{x_n\}_{n=1}^N \mid \mu, \sigma) = -N \ln(\sqrt{2\pi}\sigma) - \sum_n (x_n - \mu)^2/(2\sigma^2), \quad (24.5)$$

$$= -N \ln(\sqrt{2\pi}\sigma) - [N(\mu - \bar{x})^2 + S]/(2\sigma^2), \quad (24.6)$$

where $S \equiv \sum_n (x_n - \bar{x})^2$. Given the Gaussian model, the likelihood can be expressed in terms of the two functions of the data $\bar{x}$ and S, so these two quantities are known as 'sufficient statistics'. The posterior probability of μ and σ is, using the improper priors:

$$P(\mu, \sigma \mid \{x_n\}_{n=1}^N) = \frac{P(\{x_n\}_{n=1}^N \mid \mu, \sigma) P(\mu, \sigma)}{P(\{x_n\}_{n=1}^N)} \quad (24.7)$$

$$= \frac{\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{N(\mu-\bar{x})^2+S}{2\sigma^2}\right) \frac{1}{\sigma_\mu} \frac{1}{\sigma}}{P(\{x_n\}_{n=1}^N)}. \quad (24.8)$$

This function describes the answer to the question, 'given the data, and the noninformative priors, what might μ and σ be?' It may be of interest to find the parameter values that maximize the posterior probability, though it should be emphasized that posterior probability maxima have no fundamental status in Bayesian inference, since their location depends on the choice of basis. Here we choose the basis $(\mu, \ln\sigma)$, in which our prior is flat, so that the posterior probability maximum coincides with the maximum of the likelihood. As we saw in exercise 22.4 (p.302), the maximum likelihood solution for μ and $\ln\sigma$ is $\{\mu, \sigma\}_{\rm ML} = \left\{\bar{x}, \sigma_N = \sqrt{S/N}\right\}$.

There is more to the posterior distribution than just its mode. As can be seen in figure 24.1a, the likelihood has a skew peak. As we increase σ, the width of the conditional distribution of μ increases (figure 22.1b). And if we fix μ to a sequence of values moving away from the sample mean $\bar{x}$, we obtain a sequence of conditional distributions over σ whose maxima move to increasing values of σ (figure 24.1c).

The posterior probability of μ given σ is

$$P(\mu \,|\, \{x_n\}_{n=1}^N, \sigma) = \frac{P(\{x_n\}_{n=1}^N \,|\, \mu, \sigma) P(\mu)}{P(\{x_n\}_{n=1}^N \,|\, \sigma)} \tag{24.9}$$

$$\propto \exp(-N(\mu - \bar{x})^2/(2\sigma^2)) \tag{24.10}$$

$$= \text{Normal}(\mu; \bar{x}, \sigma^2/N). \tag{24.11}$$

We note the familiar $\sigma/\sqrt{N}$ scaling of the error bars on μ.

Let us now ask the question 'given the data, and the noninformative priors, what might σ be?' This question differs from the first one we asked in that we are now not interested in μ. This parameter must therefore be *marginalized* over. The posterior probability of σ is:

$$P(\sigma \,|\, \{x_n\}_{n=1}^N) = \frac{P(\{x_n\}_{n=1}^N \,|\, \sigma) P(\sigma)}{P(\{x_n\}_{n=1}^N)}. \tag{24.12}$$

The data-dependent term $P(\{x_n\}_{n=1}^N \,|\, \sigma)$ appeared earlier as the normalizing constant in equation (24.9); one name for this quantity is the 'evidence', or marginal likelihood, for σ. We obtain the evidence for σ by integrating out μ; a noninformative prior $P(\mu) =$ constant is assumed; we call this constant $1/\sigma_\mu$, so that we can think of the prior as a top-hat prior of width σ_μ. The Gaussian integral, $P(\{x_n\}_{n=1}^N \,|\, \sigma) = \int P(\{x_n\}_{n=1}^N \,|\, \mu, \sigma) P(\mu)\, d\mu$, yields:

$$\ln P(\{x_n\}_{n=1}^N \,|\, \sigma) = -N \ln(\sqrt{2\pi}\sigma) - \frac{S}{2\sigma^2} + \ln \frac{\sqrt{2\pi}\sigma/\sqrt{N}}{\sigma_\mu}. \tag{24.13}$$

The first two terms are the best-fit log likelihood (i.e., the log likelihood with $\mu = \bar{x}$). The last term is the log of the *Occam factor* which penalizes smaller values of σ. (We will discuss Occam factors more in Chapter 28.) When we differentiate the log evidence with respect to $\ln\sigma$, to find the most probable σ, the additional volume factor $(\sigma/\sqrt{N})$ shifts the maximum from σ_N to

$$\sigma_{N-1} = \sqrt{S/(N-1)}. \tag{24.14}$$

Intuitively, the denominator $(N-1)$ counts the number of noise measurements contained in the quantity $S = \sum_n (x_n - \bar{x})^2$. The sum contains N residuals squared, but there are only $(N-1)$ effective noise measurements because the determination of one parameter μ from the data causes one dimension of noise to be gobbled up in unavoidable overfitting. In the terminology of classical

statistics, the Bayesian's best guess for σ sets χ^2 (the measure of deviance defined by $\chi^2 \equiv \sum_n (x_n - \hat{\mu})^2/\hat{\sigma}^2$) equal to the number of degrees of freedom, $N-1$.

Figure 24.1d shows the posterior probability of σ, which is proportional to the marginal likelihood. This may be contrasted with the posterior probability of σ with μ fixed to its most probable value, $\bar{x}=1$, which is shown in figure 24.1c and d.

The final inference we might wish to make is 'given the data, what is μ?'

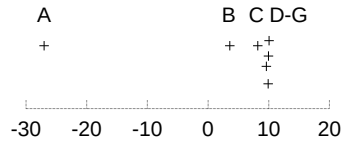
▷ Exercise 24.2.[3] Marginalize over σ and obtain the posterior marginal distribution of μ, which is a Student-t distribution:

$$P(\mu \,|\, D) \propto 1/\left(N(\mu - \bar{x})^2 + S\right)^{N/2}. \tag{24.15}$$

Further reading

A bible of exact marginalization is Bretthorst's (1988) book on Bayesian spectrum analysis and parameter estimation.

▶ 24.2 Exercises



▷ Exercise 24.3.[3] [This exercise requires macho integration capabilities.] Give a Bayesian solution to exercise 22.15 (p.309), where seven scientists of varying capabilities have measured μ with personal noise levels σ_n, and we are interested in inferring μ. Let the prior on each σ_n be a broad prior, for example a gamma distribution with parameters $(s, c) = (10, 0.1)$. Find the posterior distribution of μ. Plot it, and explore its properties for a variety of data sets such as the one given, and the data set $\{x_n\} = \{13.01, 7.39\}$.

[Hint: first find the posterior distribution of σ_n given μ and x_n, $P(\sigma_n \,|\, x_n, \mu)$. Note that the normalizing constant for this inference is $P(x_n \,|\, \mu)$. Marginalize over σ_n to find this normalizing constant, then use Bayes' theorem a second time to find $P(\mu \,|\, \{x_n\})$.]

▶ 24.3 Solutions

Solution to exercise 24.1 (p.321). 1. The data points are distributed with mean squared deviation σ^2 about the true mean. 2. The sample mean is unlikely to exactly equal the true mean. 3. The sample mean is the value of μ that minimizes the sum squared deviation of the data points from μ. Any other value of μ (in particular, the true value of μ) will have a larger value of the sum-squared deviation that $\mu = \bar{x}$.

So the expected mean squared deviation from the sample mean is necessarily smaller than the mean squared deviation σ^2 about the true mean.