22

Maximum Likelihood and Clustering

Rather than enumerate all hypotheses – which may be exponential in number – we can save a lot of time by homing in on one good hypothesis that fits the data well. This is the philosophy behind the maximum likelihood method, which identifies the setting of the parameter vector $\boldsymbol{\theta}$ that maximizes the likelihood, $P(\text{Data} \,|\, \boldsymbol{\theta}, \mathcal{H})$.

For some models the maximum likelihood parameters can be identified instantly from the data; for more complex models, finding the maximum likelihood parameters may require an iterative algorithm.

For any model, it is usually easiest to work with the *logarithm* of the likelihood rather than the likelihood, since likelihoods, being products of the probabilities of many data points, tend to be very small. Likelihoods multiply; log likelihoods add.

▶ 22.1 Maximum likelihood for one Gaussian

We return to the Gaussian for our first examples. Assume we have data $\{x_n\}_{n=1}^N$. The log likelihood is:

$$\ln P(\{x_n\}_{n=1}^N \,|\, \mu, \sigma) = -N \ln(\sqrt{2\pi}\sigma) - \sum_n (x_n - \mu)^2/(2\sigma^2). \quad (22.1)$$

The likelihood can be expressed in terms of two functions of the data, the sample mean

$$\bar{x} \equiv \sum_{n=1}^N x_n/N, \quad (22.2)$$

and the sum of square deviations

$$S \equiv \sum_n (x_n - \bar{x})^2 : \quad (22.3)$$

$$\ln P(\{x_n\}_{n=1}^N \,|\, \mu, \sigma) = -N \ln(\sqrt{2\pi}\sigma) - [N(\mu - \bar{x})^2 + S]/(2\sigma^2). \quad (22.4)$$

Because the likelihood depends on the data only through $\bar{x}$ and S, these two quantities are known as *sufficient statistics*.

Example 22.1. Differentiate the log likelihood with respect to μ and show that, if the standard deviation is known to be σ, the maximum likelihood mean μ of a Gaussian is equal to the sample mean $\bar{x}$, for any value of σ.

Solution.

$$\frac{\partial}{\partial \mu} \ln P = -\frac{N(\mu - \bar{x})}{\sigma^2} \quad (22.5)$$

$$= 0 \quad \text{when } \mu = \bar{x}. \qquad \square \quad (22.6)$$

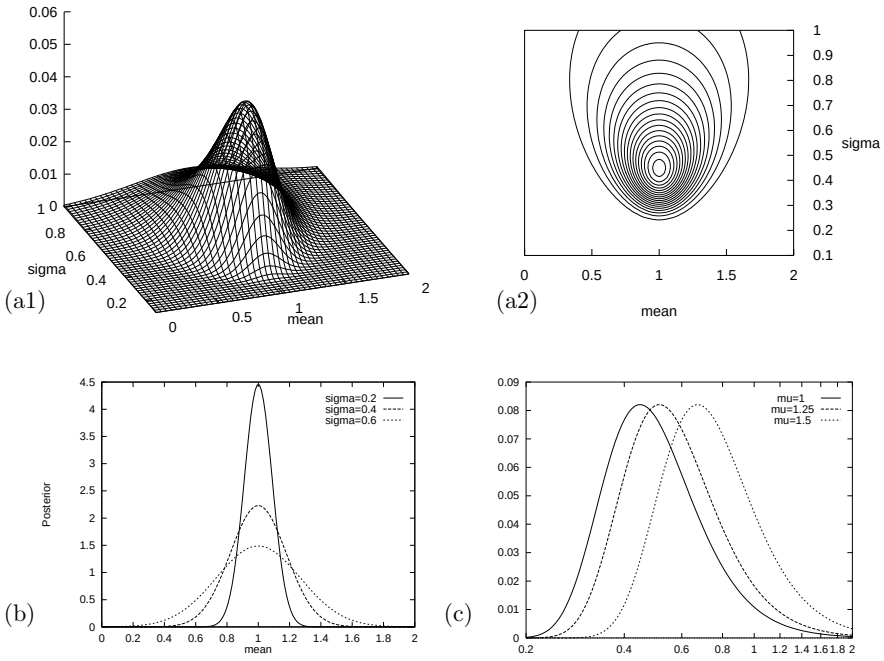


Figure 22.1. The likelihood function for the parameters of a Gaussian distribution.
(a1, a2) Surface plot and contour plot of the log likelihood as a function of μ and σ. The data set of $N = 5$ points had mean $\bar{x} = 1.0$ and $S = \sum(x - \bar{x})^2 = 1.0$.
(b) The posterior probability of μ for various values of σ.
(c) The posterior probability of σ for various fixed values of μ (shown as a density over $\ln \sigma$).

If we Taylor-expand the log likelihood about the maximum, we can define approximate error bars on the maximum likelihood parameter: we use a quadratic approximation to estimate how far from the maximum-likelihood parameter setting we can go before the likelihood falls by some standard factor, for example $e^{1/2}$, or $e^{4/2}$. In the special case of a likelihood that is a Gaussian function of the parameters, the quadratic approximation is exact.

Example 22.2. Find the second derivative of the log likelihood with respect to μ, and find the error bars on μ, given the data and σ.

Solution.

$$\frac{\partial^2}{\partial \mu^2} \ln P = -\frac{N}{\sigma^2}. \qquad \square \quad (22.7)$$

Comparing this curvature with the curvature of the log of a Gaussian distribution over μ of standard deviation σ_μ, $\exp(-\mu^2/(2\sigma_\mu^2))$, which is $-1/\sigma_\mu^2$, we can deduce that the error bars on μ (derived from the likelihood function) are

$$\sigma_\mu = \frac{\sigma}{\sqrt{N}}. \qquad (22.8)$$

The error bars have this property: at the two points $\mu = \bar{x} \pm \sigma_\mu$, the likelihood is smaller than its maximum value by a factor of $e^{1/2}$.

Example 22.3. Find the maximum likelihood standard deviation σ of a Gaussian, whose mean is known to be μ, in the light of data $\{x_n\}_{n=1}^N$. Find the second derivative of the log likelihood with respect to $\ln \sigma$, and error bars on $\ln \sigma$.

Solution. The likelihood's dependence on σ is

$$\ln P(\{x_n\}_{n=1}^N \,|\, \mu, \sigma) = -N \ln(\sqrt{2\pi}\sigma) - \frac{S_{\rm tot}}{(2\sigma^2)}, \qquad (22.9)$$

where $S_{\rm tot} = \sum_n (x_n - \mu)^2$. To find the maximum of the likelihood, we can differentiate with respect to $\ln \sigma$. [It's often most hygienic to differentiate with

respect to $\ln u$ rather than u, when u is a scale variable; we use $\mathrm{d}u^n/\mathrm{d}(\ln u) = nu^n$.]

$$\frac{\partial \ln P(\{x_n\}_{n=1}^N \,|\, \mu, \sigma)}{\partial \ln \sigma} = -N + \frac{S_{\rm tot}}{\sigma^2} \tag{22.10}$$

This derivative is zero when

$$\sigma^2 = \frac{S_{\rm tot}}{N}, \tag{22.11}$$

i.e.,

$$\sigma = \sqrt{\frac{\sum_{n=1}^N (x_n - \mu)^2}{N}}. \tag{22.12}$$

The second derivative is

$$\frac{\partial^2 \ln P(\{x_n\}_{n=1}^N \,|\, \mu, \sigma)}{\partial (\ln \sigma)^2} = -2\frac{S_{\rm tot}}{\sigma^2}, \tag{22.13}$$

and at the maximum-likelihood value of σ^2, this equals $-2N$. So error bars on $\ln \sigma$ are

$$\sigma_{\ln \sigma} = \frac{1}{\sqrt{2N}}. \qquad \square \tag{22.14}$$

▷ Exercise 22.4.[1] Show that the values of μ and $\ln \sigma$ that jointly maximize the likelihood are: $\{\mu, \sigma\}_{\rm ML} = \left\{\bar{x}, \sigma_N = \sqrt{S/N}\right\}$, where

$$\sigma_N \equiv \sqrt{\frac{\sum_{n=1}^N (x_n - \bar{x})^2}{N}}. \tag{22.15}$$

▶ 22.2 Maximum likelihood for a mixture of Gaussians

We now derive an algorithm for fitting a mixture of Gaussians to one-dimensional data. In fact, this algorithm is so important to understand that, *you*, gentle reader, get to derive the algorithm. Please work through the following exercise.

Exercise 22.5.[2, p.310] A random variable x is assumed to have a probability distribution that is a *mixture of two Gaussians*,

$$P(x \,|\, \mu_1, \mu_2, \sigma) = \left[\sum_{k=1}^{2} p_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right)\right], \tag{22.16}$$

where the two Gaussians are given the labels $k = 1$ and $k = 2$; the prior probability of the class label k is $\{p_1 = 1/2,\, p_2 = 1/2\}$; $\{\mu_k\}$ are the means of the two Gaussians; and both have standard deviation σ. For brevity, we denote these parameters by $\boldsymbol{\theta} \equiv \{\{\mu_k\}, \sigma\}$.

A data set consists of N points $\{x_n\}_{n=1}^N$ which are assumed to be independent samples from this distribution. Let k_n denote the unknown class label of the nth point.

Assuming that $\{\mu_k\}$ and σ are known, show that the posterior probability of the class label k_n of the nth point can be written as

$$\begin{aligned} P(k_n = 1 \,|\, x_n, \boldsymbol{\theta}) &= \frac{1}{1 + \exp[-(w_1 x_n + w_0)]} \\ P(k_n = 2 \,|\, x_n, \boldsymbol{\theta}) &= \frac{1}{1 + \exp[+(w_1 x_n + w_0)]}, \end{aligned} \tag{22.17}$$

and give expressions for w_1 and w_0.

Assume now that the means $\{\mu_k\}$ are *not* known, and that we wish to infer them from the data $\{x_n\}_{n=1}^N$. (The standard deviation σ is known.) In the remainder of this question we will derive an iterative algorithm for finding values for $\{\mu_k\}$ that maximize the likelihood,

$$P(\{x_n\}_{n=1}^N \,|\, \{\mu_k\}, \sigma) = \prod_n P(x_n \,|\, \{\mu_k\}, \sigma). \tag{22.18}$$

Let L denote the natural log of the likelihood. Show that the derivative of the log likelihood with respect to μ_k is given by

$$\frac{\partial}{\partial \mu_k} L = \sum_n p_{k|n} \frac{(x_n - \mu_k)}{\sigma^2}, \tag{22.19}$$

where $p_{k|n} \equiv P(k_n = k \,|\, x_n, \boldsymbol{\theta})$ appeared above at equation (22.17).

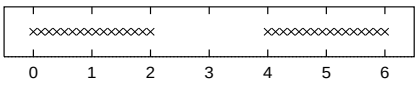
Show, neglecting terms in $\frac{\partial}{\partial \mu_k} P(k_n = k \,|\, x_n, \boldsymbol{\theta})$, that the second derivative is approximately given by

$$\frac{\partial^2}{\partial \mu_k^2} L = -\sum_n p_{k|n} \frac{1}{\sigma^2}. \tag{22.20}$$

Hence show that from an initial state μ_1, μ_2, an approximate Newton–Raphson step updates these parameters to μ_1', μ_2', where

$$\mu_k' = \frac{\sum_n p_{k|n} x_n}{\sum_n p_{k|n}}. \tag{22.21}$$

[The Newton–Raphson method for maximizing $L(\mu)$ updates μ to $\mu' = \mu - \left[\frac{\partial L}{\partial \mu} \Big/ \frac{\partial^2 L}{\partial \mu^2} \right]$.]



Assuming that $\sigma = 1$, sketch a contour plot of the likelihood function as a function of μ_1 and μ_2 for the data set shown above. The data set consists of 32 points. Describe the peaks in your sketch and indicate their widths.

Notice that the algorithm you have derived for maximizing the likelihood is identical to the soft K-means algorithm of section 20.4. Now that it is clear that clustering can be viewed as mixture-density-modelling, we are able to derive enhancements to the K-means algorithm, which rectify the problems we noted earlier.

▶ 22.3 Enhancements to soft K-means

Algorithm 22.2 shows a version of the soft-K-means algorithm corresponding to a modelling assumption that each cluster is a spherical Gaussian having its own width (each cluster has its own $\beta^{(k)} = 1/\sigma_k^2$). The algorithm updates the lengthscales σ_k for itself. The algorithm also includes cluster weight parameters $\pi_1, \pi_2, \ldots, \pi_K$ which also update themselves, allowing accurate modelling of data from clusters of unequal weights. This algorithm is demonstrated in figure 22.3 for two data sets that we've seen before. The second example shows

Assignment step. The responsibilities are

$$r_k^{(n)} = \frac{\pi_k \frac{1}{(\sqrt{2\pi}\sigma_k)^I} \exp\left(-\frac{1}{\sigma_k^2} d(\mathbf{m}^{(k)}, \mathbf{x}^{(n)})\right)}{\sum_{k'} \pi_k \frac{1}{(\sqrt{2\pi}\sigma_{k'})^I} \exp\left(-\frac{1}{\sigma_{k'}^2} d(\mathbf{m}^{(k')}, \mathbf{x}^{(n)})\right)} \tag{22.22}$$

where I is the dimensionality of $\mathbf{x}$.

Update step. Each cluster's parameters, $\mathbf{m}^{(k)}$, π_k, and σ_k^2, are adjusted to match the data points that it is responsible for.

$$\mathbf{m}^{(k)} = \frac{\sum_n r_k^{(n)} \mathbf{x}^{(n)}}{R^{(k)}} \tag{22.23}$$

$$\sigma_k^2 = \frac{\sum_n r_k^{(n)} (\mathbf{x}^{(n)} - \mathbf{m}^{(k)})^2}{I R^{(k)}} \tag{22.24}$$

$$\pi_k = \frac{R^{(k)}}{\sum_k R^{(k)}} \tag{22.25}$$

where $R^{(k)}$ is the total responsibility of mean k,

$$R^{(k)} = \sum_n r_k^{(n)}. \tag{22.26}$$

Algorithm 22.2. The soft K-means algorithm, version 2.

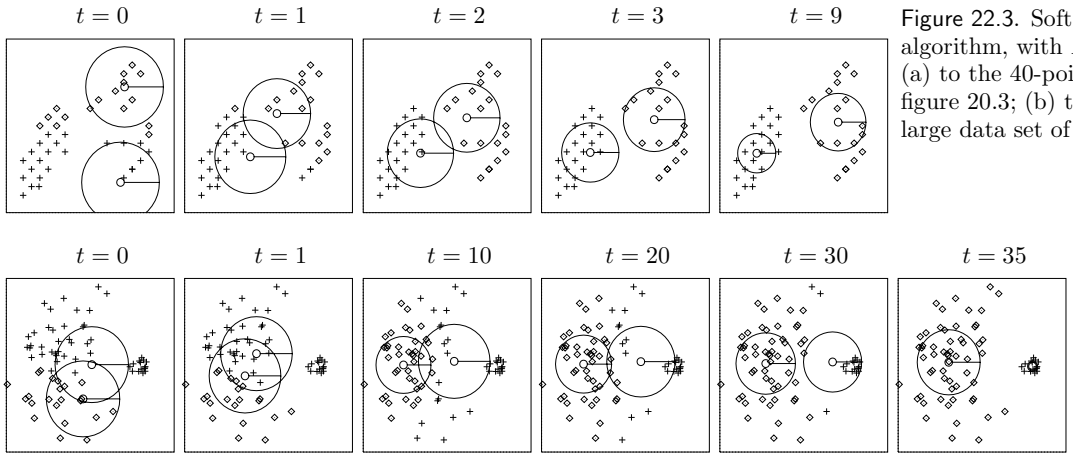


Figure 22.3. Soft K-means algorithm, with $K = 2$, applied (a) to the 40-point data set of figure 20.3; (b) to the little 'n' large data set of figure 20.5.

$$r_k^{(n)} = \frac{\pi_k \frac{1}{\prod_{i=1}^I \sqrt{2\pi}\sigma_i^{(k)}} \exp\left(-\sum_{i=1}^I (m_i^{(k)} - x_i^{(n)})^2 \Big/ 2(\sigma_i^{(k)})^2\right)}{\sum_{k'} \text{ (numerator, with } k' \text{ in place of } k)} \tag{22.27}$$

$$\sigma_i^{2(k)} = \frac{\sum_n r_k^{(n)} (x_i^{(n)} - m_i^{(k)})^2}{R^{(k)}} \tag{22.28}$$

Algorithm 22.4. The soft K-means algorithm, version 3, which corresponds to a model of axis-aligned Gaussians.

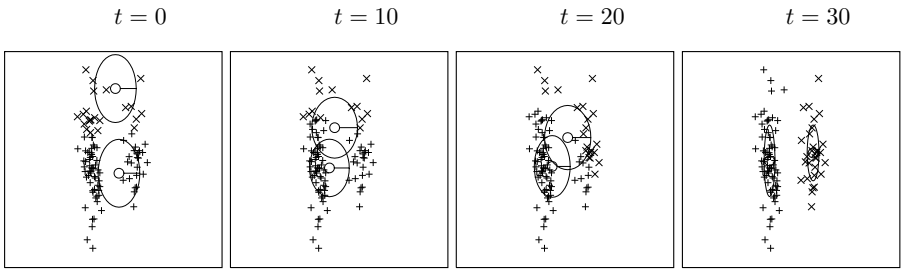


Figure 22.5. Soft K-means algorithm, version 3, applied to the data consisting of two cigar-shaped clusters. $K = 2$ (cf. figure 20.6).

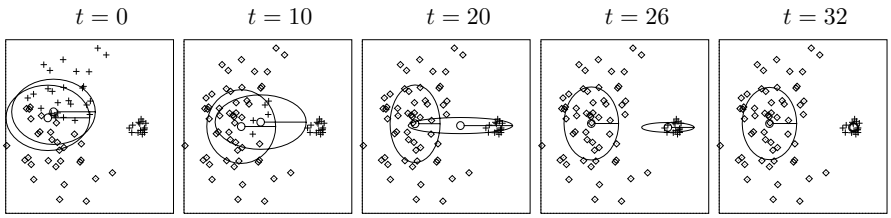


Figure 22.6. Soft K-means algorithm, version 3, applied to the little 'n' large data set. $K = 2$.

that convergence can take a long time, but eventually the algorithm identifies the small cluster and the large cluster.

Soft K-means, version 2, is a maximum-likelihood algorithm for fitting a mixture of *spherical Gaussians* to data – 'spherical' meaning that the variance of the Gaussian is the same in all directions. This algorithm is still no good at modelling the cigar-shaped clusters of figure 20.6. If we wish to model the clusters by axis-aligned Gaussians with possibly-unequal variances, we replace the assignment rule (22.22) and the variance update rule (22.24) by the rules (22.27) and (22.28) displayed in algorithm 22.4.

A proof that the algorithm does indeed maximize the likelihood is deferred to section 33.7.

This third version of soft K-means is demonstrated in figure 22.5 on the 'two cigars' data set of figure 20.6. After 30 iterations, the algorithm correctly locates the two clusters. Figure 22.6 shows the same algorithm applied to the little 'n' large data set; again, the correct cluster locations are found.

▶ 22.4 A fatal flaw of maximum likelihood

Finally, figure 22.7 sounds a cautionary note: when we fit $K = 4$ means to our first toy data set, we sometimes find that very small clusters form, covering just one or two data points. This is a pathological property of soft K-means clustering, versions 2 and 3.

▷ Exercise 22.6.[2] Investigate what happens if one mean $\mathbf{m}^{(k)}$ sits exactly on top of one data point; show that if the variance σ_k^2 is sufficiently small, then no return is possible: σ_k^2 becomes ever smaller.

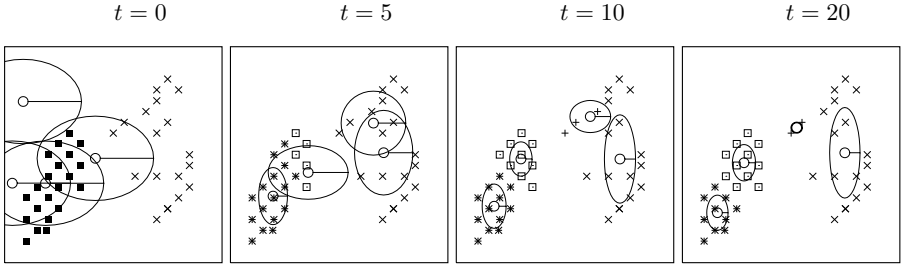


Figure 22.7. Soft K-means algorithm applied to a data set of 40 points. $K = 4$. Notice that at convergence, one very small cluster has formed between two data points.

KABOOM!

Soft K-means can blow up. Put one cluster exactly on one data point and let its variance go to zero – you can obtain an arbitrarily large likelihood! Maximum likelihood methods can break down by finding highly tuned models that fit part of the data perfectly. This phenomenon is known as overfitting. The reason we are not interested in these solutions with enormous likelihood is this: sure, these parameter-settings may have enormous posterior probability *density*, but the density is large over only a very small *volume* of parameter space. So the probability *mass* associated with these likelihood spikes is usually tiny.

We conclude that maximum likelihood methods are not a satisfactory general solution to data-modelling problems: the likelihood may be infinitely large at certain parameter settings. Even if the likelihood does not have infinitely-large spikes, the maximum of the likelihood is often unrepresentative, in high-dimensional problems.

Even in low-dimensional problems, maximum likelihood solutions can be unrepresentative. As you may know from basic statistics, the maximum likelihood estimator (22.15) for a Gaussian's standard deviation, σ_N, is a *biased* estimator, a topic that we'll take up in Chapter 24.

The maximum a posteriori (MAP) method

A popular replacement for maximizing the likelihood is maximizing the Bayesian posterior probability density of the parameters instead. However, multiplying the likelihood by a prior and maximizing the posterior does not make the above problems go away; the posterior density often also has infinitely-large spikes, and the maximum of the posterior probability density is often unrepresentative of the whole posterior distribution. Think back to the concept of typicality, which we encountered in Chapter 4: in high dimensions, most of the probability mass is in a typical set whose properties are quite different from the points that have the maximum probability density. Maxima are atypical.

A further reason for disliking the maximum *a posteriori* is that it is *basis-dependent*. If we make a nonlinear change of basis from the parameter θ to the parameter $u = f(\theta)$ then the probability density of θ is transformed to

$$P(u) = P(\theta) \left| \frac{\partial \theta}{\partial u} \right| . \tag{22.29}$$

The maximum of the density $P(u)$ will usually not coincide with the maximum of the density $P(\theta)$. (For figures illustrating such nonlinear changes of basis, see the next chapter.) It seems undesirable to use a method whose answers change when we change representation.

Further reading

The soft K-means algorithm is at the heart of the automatic classification package, AutoClass (Hanson *et al.*, 1991b; Hanson *et al.*, 1991a).

▶ 22.5 Further exercises

Exercises where maximum likelihood may be useful

Exercise 22.7.[3] Make a version of the K-means algorithm that models the data as a mixture of K arbitrary Gaussians, i.e., Gaussians that are not constrained to be axis-aligned.

▷ Exercise 22.8.[2] (a) A photon counter is pointed at a remote star for one minute, in order to infer the brightness, i.e., the rate of photons arriving at the counter per minute, λ. Assuming the number of photons collected r has a Poisson distribution with mean λ,

$$P(r \mid \lambda) = \exp(-\lambda)\frac{\lambda^r}{r!}, \tag{22.30}$$

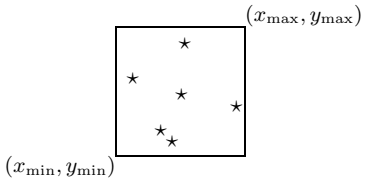
what is the maximum likelihood estimate for λ, given $r = 9$? Find error bars on $\ln \lambda$.

(b) Same situation, but now we assume that the counter detects not only photons from the star but also 'background' photons. The background rate of photons is known to be $b = 13$ photons per minute. We assume the number of photons collected, r, has a Poisson distribution with mean $\lambda + b$. Now, given $r = 9$ detected photons, what is the maximum likelihood estimate for λ? Comment on this answer, discussing also the Bayesian posterior distribution, and the 'unbiased estimator' of sampling theory, $\hat{\lambda} \equiv r - b$.

Exercise 22.9.[2] A bent coin is tossed N times, giving N_a heads and N_b tails. Assume a beta distribution prior for the probability of heads, p, for example the uniform distribution. Find the maximum likelihood and maximum *a posteriori* values of p, then find the maximum likelihood and maximum *a posteriori* values of the logit $a \equiv \ln[p/(1-p)]$. Compare with the predictive distribution, i.e., the probability that the next toss will come up heads.

▷ Exercise 22.10.[2] *Two men looked through prison bars; one saw stars, the other tried to infer where the window frame was.*

From the other side of a room, you look through a window and see stars at locations $\{(x_n, y_n)\}$. You can't see the window edges because it is totally dark apart from the stars. Assuming the window is rectangular and that the visible stars' locations are independently randomly distributed, what are the inferred values of $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$, according to maximum likelihood? Sketch the likelihood as a function of $x_{\max}$, for fixed $x_{\min}$, $y_{\min}$, and $y_{\max}$.



▷ Exercise 22.11.[3] A sailor infers his location (x, y) by measuring the bearings of three buoys whose locations (x_n, y_n) are given on his chart. Let the true bearings of the buoys be θ_n. Assuming that his measurement $\tilde{\theta}_n$ of each bearing is subject to Gaussian noise of small standard deviation σ, what is his inferred location, by maximum likelihood?

The sailor's rule of thumb says that the boat's position can be taken to be the centre of the cocked hat, the triangle produced by the intersection of the three measured bearings (figure 22.8). Can you persuade him that the maximum likelihood answer is better?

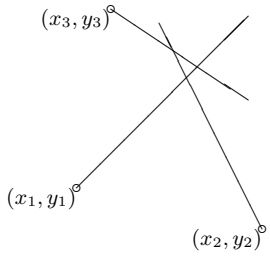


Figure 22.8. The standard way of drawing three slightly inconsistent bearings on a chart produces a triangle called a cocked hat. Where is the sailor?

▷ Exercise 22.12.[3, p.310] Maximum likelihood fitting of an exponential-family model.

Assume that a variable $\mathbf{x}$ comes from a probability distribution of the form

$$P(\mathbf{x} \mid \mathbf{w}) = \frac{1}{Z(\mathbf{w})} \exp\left(\sum_k w_k f_k(\mathbf{x})\right), \tag{22.31}$$

where the functions $f_k(\mathbf{x})$ are given, and the parameters $\mathbf{w} = \{w_k\}$ are not known. A data set $\{\mathbf{x}^{(n)}\}$ of N points is supplied.

Show by differentiating the log likelihood that the maximum-likelihood parameters $\mathbf{w}_{\rm ML}$ satisfy

$$\sum_{\mathbf{x}} P(\mathbf{x} \,|\, \mathbf{w}_{\rm ML}) f_k(\mathbf{x}) = \frac{1}{N} \sum_n f_k(\mathbf{x}^{(n)}), \tag{22.32}$$

where the left-hand sum is over *all* $\mathbf{x}$, and the right-hand sum is over the data points. A shorthand for this result is that each function-average under the fitted model must equal the function-average found in the data:

$$\langle f_k \rangle_{P(\mathbf{x} \,|\, \mathbf{w}_{\rm ML})} = \langle f_k \rangle_{\rm Data}. \tag{22.33}$$

▷ Exercise 22.13.[3] 'Maximum entropy' fitting of models to constraints.

When confronted by a probability distribution $P(\mathbf{x})$ about which only a few facts are known, the *maximum entropy principle* (maxent) offers a rule for *choosing* a distribution that satisfies those constraints. According to maxent, you should select the $P(\mathbf{x})$ that maximizes the entropy

$$H = \sum_{\mathbf{x}} P(\mathbf{x}) \log 1/P(\mathbf{x}), \tag{22.34}$$

subject to the constraints. Assuming the constraints assert that the *averages* of certain functions $f_k(\mathbf{x})$ are known, i.e.,

$$\langle f_k \rangle_{P(\mathbf{x})} = F_k, \tag{22.35}$$

show, by introducing Lagrange multipliers (one for each constraint, including normalization), that the maximum-entropy distribution has the form

$$P(\mathbf{x})_{\rm Maxent} = \frac{1}{Z} \exp\left(\sum_k w_k f_k(\mathbf{x}) \right), \tag{22.36}$$

where the parameters Z and $\{w_k\}$ are set such that the constraints (22.35) are satisfied.

And hence the maximum entropy method gives identical results to maximum likelihood fitting of an exponential-family model (previous exercise).

The maximum entropy method has sometimes been recommended as a method for assigning prior distributions in Bayesian modelling. While the outcomes of the maximum entropy method are sometimes interesting and thought-provoking, I do not advocate maxent as *the* approach to assigning priors.

Maximum entropy is also sometimes proposed as a method for solving inference problems – for example, 'given that the mean score of this unfair six-sided die is 2.5, what is its probability distribution $(p_1, p_2, p_3, p_4, p_5, p_6)$?' I think it is a bad idea to use maximum entropy in this way; it can give silly answers. The correct way to solve inference problems is to use Bayes' theorem.

Exercises where maximum likelihood and MAP have difficulties

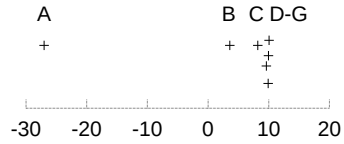
▷ Exercise 22.14.[2] This exercise explores the idea that maximizing a probability density is a poor way to find a point that is representative of the density. Consider a Gaussian distribution in a k-dimensional space, $P(\mathbf{w}) = (1/\sqrt{2\pi}\,\sigma_W)^k \exp(-\sum_1^k w_i^2/2\sigma_W^2)$. Show that nearly all of the probability mass of a Gaussian is in a thin shell of radius $r = \sqrt{k}\sigma_W$ and of thickness proportional to $r/\sqrt{k}$. For example, in 1000 dimensions, 90% of the mass of a Gaussian with $\sigma_W = 1$ is in a shell of radius 31.6 and thickness 2.8. However, the probability *density* at the origin is $e^{k/2} \simeq 10^{217}$ times bigger than the density at this shell where most of the probability mass is.

Now consider two Gaussian densities in 1000 dimensions that differ in radius σ_W by just 1%, and that contain equal total probability mass. Show that the maximum probability density is greater at the centre of the Gaussian with smaller σ_W by a factor of $\sim\exp(0.01k) \simeq 20\,000$.

In ill-posed problems, a typical posterior distribution is often a weighted superposition of Gaussians with varying means and standard deviations, so the true posterior has a skew peak, with the maximum of the probability density located near the mean of the Gaussian distribution that has the smallest standard deviation, not the Gaussian with the greatest weight.

▷ Exercise 22.15.[3] The seven scientists. N datapoints $\{x_n\}$ are drawn from N distributions, all of which are Gaussian with a common mean μ but with different unknown standard deviations σ_n. What are the maximum likelihood parameters $\mu, \{\sigma_n\}$ given the data? For example, seven scientists (A, B, C, D, E, F, G) with wildly-differing experimental skills measure μ. You expect some of them to do accurate work (i.e., to have small σ_n), and some of them to turn in wildly inaccurate answers (i.e., to have enormous σ_n). Figure 22.9 shows their seven results. What is μ, and how reliable is each scientist?

I hope you agree that, intuitively, it looks pretty certain that A and B are both inept measurers, that D–G are better, and that the true value of μ is somewhere close to 10. But what does maximizing the likelihood tell you?



Scientist	x_n
A	−27.020
B	3.570
C	8.191
D	9.898
E	9.603
F	9.945
G	10.056

Figure 22.9. Seven measurements $\{x_n\}$ of a parameter μ by seven scientists each having his own noise-level σ_n.

Exercise 22.16.[3] Problems with MAP method. A collection of widgets $i = 1, \ldots, k$ have a property called 'wodge', w_i, which we measure, widget by widget, in noisy experiments with a known noise level $\sigma_\nu = 1.0$. Our model for these quantities is that they come from a Gaussian prior $P(w_i \,|\, \alpha) = \text{Normal}(0, 1/\alpha)$, where $\alpha = 1/\sigma_W^2$ is not known. Our prior for this variance is flat over $\log \sigma_W$ from $\sigma_W = 0.1$ to $\sigma_W = 10$.

Scenario 1. Suppose four widgets have been measured and give the following data: $\{d_1, d_2, d_3, d_4\} = \{2.2, -2.2, 2.8, -2.8\}$. We are interested in inferring the wodges of these four widgets.

(a) Find the values of $\mathbf{w}$ and α that maximize the posterior probability $P(\mathbf{w}, \log\alpha \,|\, \mathbf{d})$.

(b) Marginalize over α and find the posterior probability density of $\mathbf{w}$ given the data. [Integration skills required. See MacKay (1999a) for solution.] Find maxima of $P(\mathbf{w} \,|\, \mathbf{d})$. [Answer: two maxima – one at $\mathbf{w}_{\rm MP} = \{1.8, -1.8, 2.2, -2.2\}$, with error bars on all four parameters

(obtained from Gaussian approximation to the posterior) ± 0.9; and one at $\mathbf{w}'_{\rm MP} = \{0.03, -0.03, 0.04, -0.04\}$ with error bars ± 0.1.]

Scenario 2. Suppose in addition to the four measurements above we are now informed that there are four more widgets that have been measured with a much less accurate instrument, having $\sigma'_\nu = 100.0$. Thus we now have both well-determined and ill-determined parameters, as in a typical ill-posed problem. The data from these measurements were a string of uninformative values, $\{d_5, d_6, d_7, d_8\} = \{100, -100, 100, -100\}$.

We are again asked to infer the wodges of the widgets. Intuitively, our inferences about the well-measured widgets should be negligibly affected by this vacuous information about the poorly-measured widgets. But what happens to the MAP method?

(a) Find the values of $\mathbf{w}$ and α that maximize the posterior probability $P(\mathbf{w}, \log\alpha \,|\, \mathbf{d})$.

(b) Find maxima of $P(\mathbf{w} \,|\, \mathbf{d})$. [Answer: only one maximum, $\mathbf{w}_{\rm MP} = \{0.03, -0.03, 0.03, -0.03, 0.0001, -0.0001, 0.0001, -0.0001\}$, with error bars on all eight parameters ± 0.11.]

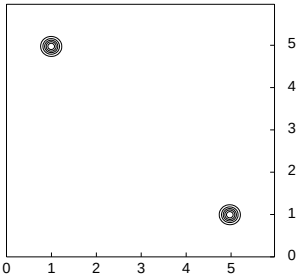


Figure 22.10. The likelihood as a function of μ_1 and μ_2.

▶ 22.6 Solutions

Solution to exercise 22.5 (p.302). Figure 22.10 shows a contour plot of the likelihood function for the 32 data points. The peaks are pretty-near centred on the points $(1, 5)$ and $(5, 1)$, and are pretty-near circular in their contours. The width of each of the peaks is a standard deviation of $\sigma/\sqrt{16} = 1/4$. The peaks are roughly Gaussian in shape.

Solution to exercise 22.12 (p.307). The log likelihood is:

$$\ln P(\{\mathbf{x}^{(n)}\} \,|\, \mathbf{w}) = -N \ln Z(\mathbf{w}) + \sum_n \sum_k w_k f_k(\mathbf{x}^{(n)}). \tag{22.37}$$

$$\frac{\partial}{\partial w_k} \ln P(\{\mathbf{x}^{(n)}\} \,|\, \mathbf{w}) = -N \frac{\partial}{\partial w_k} \ln Z(\mathbf{w}) + \sum_n f_k(\mathbf{x}). \tag{22.38}$$

Now, the fun part is what happens when we differentiate the log of the normalizing constant:

$$\frac{\partial}{\partial w_k} \ln Z(\mathbf{w}) \;=\; \frac{1}{Z(\mathbf{w})} \sum_{\mathbf{x}} \frac{\partial}{\partial w_k} \exp\left(\sum_{k'} w_{k'} f_{k'}(\mathbf{x}) \right)$$

$$= \frac{1}{Z(\mathbf{w})} \sum_{\mathbf{x}} \exp\left(\sum_{k'} w_{k'} f_{k'}(\mathbf{x}) \right) f_k(\mathbf{x}) \;=\; \sum_{\mathbf{x}} P(\mathbf{x} \,|\, \mathbf{w}) f_k(\mathbf{x}), \tag{22.39}$$

so

$$\frac{\partial}{\partial w_k} \ln P(\{\mathbf{x}^{(n)}\} \,|\, \mathbf{w}) = -N \sum_{\mathbf{x}} P(\mathbf{x} \,|\, \mathbf{w}) f_k(\mathbf{x}) + \sum_n f_k(\mathbf{x}), \tag{22.40}$$

and at the maximum of the likelihood,

$$\sum_{\mathbf{x}} P(\mathbf{x} \,|\, \mathbf{w}_{\rm ML}) f_k(\mathbf{x}) = \frac{1}{N} \sum_n f_k(\mathbf{x}^{(n)}). \tag{22.41}$$

23

Useful Probability Distributions

In Bayesian data modelling, there's a small collection of probability distributions that come up again and again. The purpose of this chapter is to introduce these distributions so that they won't be intimidating when encountered in combat situations.

There is no need to memorize any of them, except perhaps the Gaussian; if a distribution is important enough, it will memorize itself, and otherwise, it can easily be looked up.

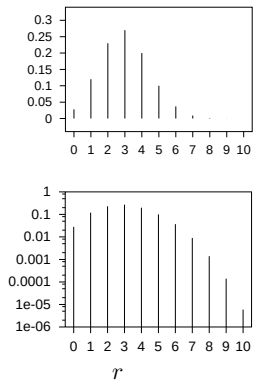


Figure 23.1. The binomial distribution $P(r \mid f = 0.3, N = 10)$, on a linear scale (top) and a logarithmic scale (bottom).

▶ 23.1 Distributions over integers

Binomial, Poisson, exponential

We already encountered the binomial distribution and the Poisson distribution on page 2.

The *binomial distribution* for an integer r with parameters f (the bias, $f \in [0, 1]$) and N (the number of trials) is:

$$P(r \mid f, N) = \binom{N}{r} f^r (1-f)^{N-r} \quad r \in \{0, 1, 2, \ldots, N\}. \tag{23.1}$$

The binomial distribution arises, for example, when we flip a bent coin, with bias f, N times, and observe the number of heads, r.

The *Poisson distribution* with parameter $\lambda > 0$ is:

$$P(r \mid \lambda) = e^{-\lambda} \frac{\lambda^r}{r!} \quad r \in \{0, 1, 2, \ldots\}. \tag{23.2}$$

The Poisson distribution arises, for example, when we count the number of photons r that arrive in a pixel during a fixed interval, given that the mean intensity on the pixel corresponds to an average number of photons λ.

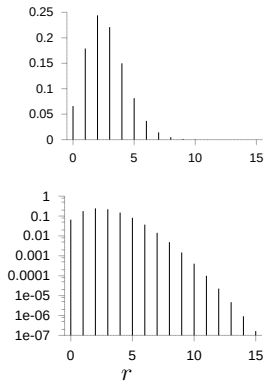


Figure 23.2. The Poisson distribution $P(r \mid \lambda = 2.7)$, on a linear scale (top) and a logarithmic scale (bottom).

The *exponential distribution on integers*,,

$$P(r \mid f) = f^r (1-f) \quad r \in (0, 1, 2, \ldots, \infty), \tag{23.3}$$

arises in waiting problems. How long will you have to wait until a six is rolled, if a fair six-sided dice is rolled? Answer: the probability distribution of the number of rolls, r, is exponential over integers with parameter $f = 5/6$. The distribution may also be written

$$P(r \mid f) = (1-f)\, e^{-\lambda r} \quad r \in (0, 1, 2, \ldots, \infty), \tag{23.4}$$

where $\lambda = \ln(1/f)$.

▶ 23.2 Distributions over unbounded real numbers

Gaussian, Student, Cauchy, biexponential, inverse-cosh.

The *Gaussian distribution* or normal distribution with mean μ and standard deviation σ is

$$P(x \mid \mu, \sigma) = \frac{1}{Z} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad x \in (-\infty, \infty), \tag{23.5}$$

where

$$Z = \sqrt{2\pi\sigma^2}. \tag{23.6}$$

It is sometimes useful to work with the quantity $\tau \equiv 1/\sigma^2$, which is called the *precision* parameter of the Gaussian.

A sample z from a standard univariate Gaussian can be generated by computing

$$z = \cos(2\pi u_1)\sqrt{2\ln(1/u_2)}, \tag{23.7}$$

where u_1 and u_2 are uniformly distributed in $(0, 1)$. A second sample $z_2 = \sin(2\pi u_1)\sqrt{2\ln(1/u_2)}$, independent of the first, can then be obtained for free.

The Gaussian distribution is widely used and often asserted to be a very common distribution in the real world, but I am sceptical about this assertion. Yes, *unimodal* distributions may be common; but a Gaussian is a special, rather extreme, unimodal distribution. It has very light tails: the log-probability-density decreases quadratically. The typical deviation of x from μ is σ, but the respective probabilities that x deviates from μ by more than 2σ, 3σ, 4σ, and 5σ, are 0.046, 0.003, 6×10^{-5}, and 6×10^{-7}. In my experience, deviations from a mean four or five times greater than the typical deviation may be rare, but not as rare as 6×10^{-5}! I therefore urge caution in the use of Gaussian distributions: if a variable that is modelled with a Gaussian actually has a heavier-tailed distribution, the rest of the model will contort itself to reduce the deviations of the outliers, like a sheet of paper being crushed by a rubber band.

▷ Exercise 23.1.[1] Pick a variable that is supposedly bell-shaped in probability distribution, gather data, and make a plot of the variable's empirical distribution. Show the distribution as a histogram on a log scale and investigate whether the tails are well-modelled by a Gaussian distribution. [One example of a variable to study is the amplitude of an audio signal.]

One distribution with heavier tails than a Gaussian is a *mixture of Gaussians*. A mixture of two Gaussians, for example, is defined by two means, two standard deviations, and two *mixing coefficients* π_1 and π_2, satisfying $\pi_1 + \pi_2 = 1$, $\pi_i \geq 0$.

$$P(x \mid \mu_1, \sigma_1, \pi_1, \mu_2, \sigma_2, \pi_2) = \frac{\pi_1}{\sqrt{2\pi}\sigma_1} \exp\left(-\tfrac{(x-\mu_1)^2}{2\sigma_1^2}\right) + \frac{\pi_2}{\sqrt{2\pi}\sigma_2} \exp\left(-\tfrac{(x-\mu_2)^2}{2\sigma_2^2}\right).$$

If we take an appropriately weighted mixture of an infinite number of Gaussians, all having mean μ, we obtain a *Student-t distribution*,

$$P(x \mid \mu, s, n) = \frac{1}{Z} \frac{1}{(1 + (x-\mu)^2/(ns^2))^{(n+1)/2}}, \tag{23.8}$$

where

$$Z = \sqrt{\pi n s^2} \frac{\Gamma(n/2)}{\Gamma((n+1)/2)} \tag{23.9}$$

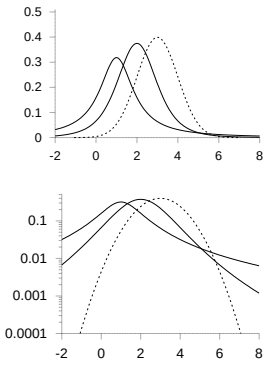


Figure 23.3. Three unimodal distributions. Two Student distributions, with parameters $(m, s) = (1, 1)$ (heavy line) (a Cauchy distribution) and $(2, 4)$ (light line), and a Gaussian distribution with mean $\mu = 3$ and standard deviation $\sigma = 3$ (dashed line), shown on linear vertical scales (top) and logarithmic vertical scales (bottom). Notice that the heavy tails of the Cauchy distribution are scarcely evident in the upper 'bell-shaped curve'.

and n is called the number of degrees of freedom and Γ is the gamma function. If $n > 1$ then the Student distribution (23.8) has a mean and that mean is μ. If $n > 2$ the distribution also has a finite variance, $\sigma^2 = ns^2/(n-2)$. As $n \to \infty$, the Student distribution approaches the normal distribution with mean μ and standard deviation s. The Student distribution arises both in classical statistics (as the sampling-theoretic distribution of certain statistics) and in Bayesian inference (as the probability distribution of a variable coming from a Gaussian distribution whose standard deviation we aren't sure of).

In the special case $n = 1$, the Student distribution is called the *Cauchy distribution*.

A distribution whose tails are intermediate in heaviness between Student and Gaussian is the *biexponential distribution*,

$$P(x \mid \mu, s) = \frac{1}{Z} \exp\left(-\frac{|x-\mu|}{s}\right) \quad x \in (-\infty, \infty) \tag{23.10}$$

where

$$Z = 2s. \tag{23.11}$$

The *inverse-cosh distribution*

$$P(x \mid \beta) \propto \frac{1}{[\cosh(\beta x)]^{1/\beta}} \tag{23.12}$$

is a popular model in independent component analysis. In the limit of large β, the probability distribution $P(x \mid \beta)$ becomes a biexponential distribution. In the limit $\beta \to 0$ $P(x \mid \beta)$ approaches a Gaussian with mean zero and variance $1/\beta$.

▶ 23.3 Distributions over positive real numbers

Exponential, gamma, inverse-gamma, and log-normal.

The *exponential distribution*,

$$P(x \mid s) = \frac{1}{Z} \exp\left(-\frac{x}{s}\right) \quad x \in (0, \infty), \tag{23.13}$$

where

$$Z = s, \tag{23.14}$$

arises in waiting problems. How long will you have to wait for a bus in Poissonville, given that buses arrive independently at random with one every s minutes on average? Answer: the probability distribution of your wait, x, is exponential with mean s.

The *gamma distribution* is like a Gaussian distribution, except whereas the Gaussian goes from $-\infty$ to ∞, gamma distributions go from 0 to ∞. Just as the Gaussian distribution has two parameters μ and σ which control the mean and width of the distribution, the gamma distribution has two parameters. It is the product of the one-parameter exponential distribution (23.13) with a polynomial, x^{c-1}. The exponent c in the polynomial is the second parameter.

$$P(x \mid s, c) = \Gamma(x; s, c) = \frac{1}{Z}\left(\frac{x}{s}\right)^{c-1} \exp\left(-\frac{x}{s}\right), \quad 0 \leq x < \infty \tag{23.15}$$

where

$$Z = \Gamma(c)s. \tag{23.16}$$

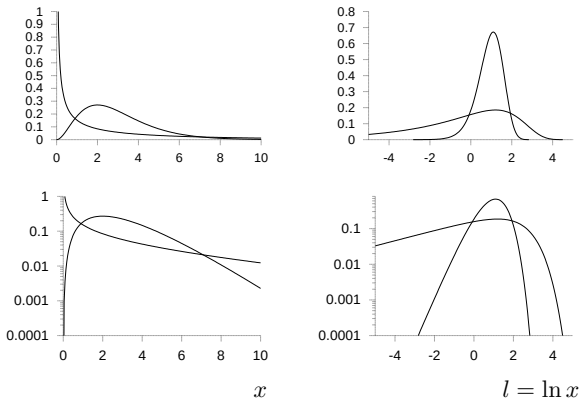


Figure 23.4. Two gamma distributions, with parameters $(s, c) = (1, 3)$ (heavy lines) and $10, 0.3$ (light lines), shown on linear vertical scales (top) and logarithmic vertical scales (bottom); and shown as a function of x on the left (23.15) and $l = \ln x$ on the right (23.18).

This is a simple peaked distribution with mean sc and variance s^2c.

It is often natural to represent a positive real variable x in terms of its logarithm $l = \ln x$. The probability density of l is

$$P(l) = P(x(l)) \left| \frac{\partial x}{\partial l} \right| = P(x(l))x(l) \tag{23.17}$$

$$= \frac{1}{Z_l} \left(\frac{x(l)}{s} \right)^c \exp\left(-\frac{x(l)}{s} \right), \tag{23.18}$$

where

$$Z_l = \Gamma(c). \tag{23.19}$$

[The gamma distribution is named after its normalizing constant – an odd convention, it seems to me!]

Figure 23.4 shows a couple of gamma distributions as a function of x and of l. Notice that where the original gamma distribution (23.15) may have a 'spike' at $x = 0$, the distribution over l never has such a spike. The spike is an artefact of a bad choice of basis.

In the limit $sc = 1, c \to 0$, we obtain the noninformative prior for a scale parameter, the $1/x$ prior. This improper prior is called noninformative because it has no associated length scale, no characteristic value of x, so it prefers all values of x equally. It is invariant under the reparameterization $x = mx$. If we transform the $1/x$ probability density into a density over $l = \ln x$ we find the latter density is uniform.

▷ Exercise 23.2.[1] Imagine that we reparameterize a positive variable x in terms of its cube root, $u = x^{1/3}$. If the probability density of x is the improper distribution $1/x$, what is the probability density of u?

The gamma distribution is always a unimodal density over $l = \ln x$, and, as can be seen in the figures, it is asymmetric. If x has a gamma distribution, and we decide to work in terms of the inverse of x, $v = 1/x$, we obtain a new distribution, in which the density over l is flipped left-for-right: the probability density of v is called an *inverse-gamma distribution*,

$$P(v \mid s, c) = \frac{1}{Z_v} \left(\frac{1}{sv} \right)^{c+1} \exp\left(-\frac{1}{sv} \right), \quad 0 \leq v < \infty \tag{23.20}$$

where

$$Z_v = \Gamma(c)/s. \tag{23.21}$$

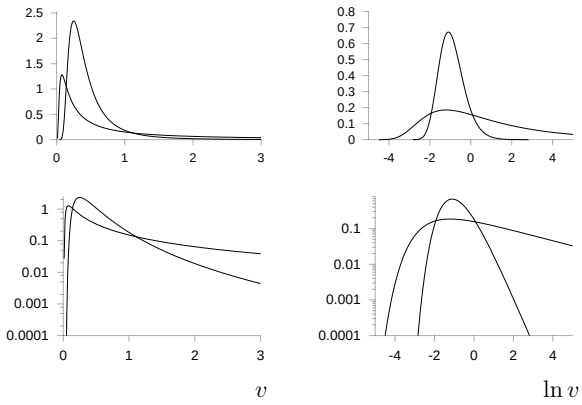


Figure 23.5. Two inverse gamma distributions, with parameters $(s, c) = (1, 3)$ (heavy lines) and $10, 0.3$ (light lines), shown on linear vertical scales (top) and logarithmic vertical scales (bottom); and shown as a function of x on the left and $l = \ln x$ on the right.

Gamma and inverse gamma distributions crop up in many inference problems in which a positive quantity is inferred from data. Examples include inferring the variance of Gaussian noise from some noise samples, and inferring the rate parameter of a Poisson distribution from the count.

Gamma distributions also arise naturally in the distributions of waiting times between Poisson-distributed events. Given a Poisson process with rate λ, the probability density of the arrival time x of the mth event is

$$\frac{\lambda(\lambda x)^{m-1}}{(m-1)!}\, e^{-\lambda x}. \tag{23.22}$$

Log-normal distribution

Another distribution over a positive real number x is the *log-normal* distribution, which is the distribution that results when $l = \ln x$ has a normal distribution. We define m to be the median value of x, and s to be the standard deviation of $\ln x$.

$$P(l \,|\, m, s) = \frac{1}{Z} \exp\left(-\frac{(l - \ln m)^2}{2s^2}\right) \quad l \in (-\infty, \infty), \tag{23.23}$$

where

$$Z = \sqrt{2\pi s^2}, \tag{23.24}$$

implies

$$P(x \,|\, m, s) = \frac{1}{x}\frac{1}{Z} \exp\left(-\frac{(\ln x - \ln m)^2}{2s^2}\right) \quad x \in (0, \infty). \tag{23.25}$$

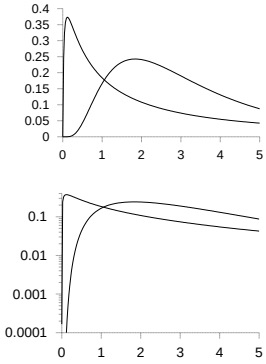


Figure 23.6. Two log-normal distributions, with parameters $(m, s) = (3, 1.8)$ (heavy line) and $(3, 0.7)$ (light line), shown on linear vertical scales (top) and logarithmic vertical scales (bottom). [Yes, they really do have the same value of the median, $m = 3$.]

▶ 23.4 Distributions over periodic variables

A periodic variable θ is a real number $\in [0, 2\pi]$ having the property that $\theta = 0$ and $\theta = 2\pi$ are equivalent.

A distribution that plays for periodic variables the role played by the Gaussian distribution for real variables is the *Von Mises distribution*:

$$P(\theta \,|\, \mu, \beta) = \frac{1}{Z} \exp\left(\beta \cos(\theta - \mu)\right) \quad \theta \in (0, 2\pi). \tag{23.26}$$

The normalizing constant is $Z = 2\pi I_0(\beta)$, where $I_0(x)$ is a modified Bessel function.

A distribution that arises from Brownian diffusion around the circle is the wrapped Gaussian distribution,

$$P(\theta \,|\, \mu, \sigma) = \sum_{n=-\infty}^{\infty} \text{Normal}(\theta; (\mu + 2\pi n), \sigma^2) \quad \theta \in (0, 2\pi). \tag{23.27}$$

▶ 23.5 Distributions over probabilities

Beta distribution, Dirichlet distribution, entropic distribution

The *beta distribution* is a probability density over a variable p that is a probability, $p \in (0, 1)$:

$$P(p \,|\, u_1, u_2) = \frac{1}{Z(u_1, u_2)} p^{u_1 - 1} (1 - p)^{u_2 - 1}. \tag{23.28}$$

The parameters u_1, u_2 may take any positive value. The normalizing constant is the beta function,

$$Z(u_1, u_2) = \frac{\Gamma(u_1)\Gamma(u_2)}{\Gamma(u_1 + u_2)}. \tag{23.29}$$

Special cases include the uniform distribution – $u_1 = 1, u_2 = 1$; the Jeffreys prior – $u_1 = 0.5, u_2 = 0.5$; and the improper Laplace prior – $u_1 = 0, u_2 = 0$. If we transform the beta distribution to the corresponding density over the logit $l \equiv \ln p/(1 - p)$, we find it is always a pleasant bell-shaped density over l, while the density over p may have singularities at $p = 0$ and $p = 1$ (figure 23.7).

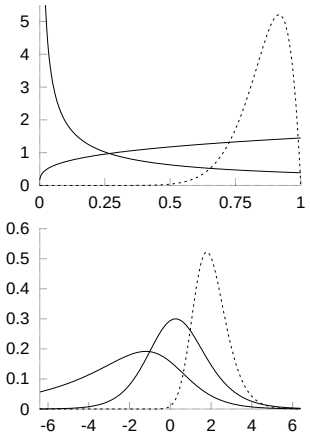


Figure 23.7. Three beta distributions, with $(u_1, u_2) = (0.3, 1)$, $(1.3, 1)$, and $(12, 2)$. The upper figure shows $P(p \,|\, u_1, u_2)$ as a function of p; the lower shows the corresponding density over the *logit*,

$$\ln \frac{p}{1 - p}.$$

Notice how well-behaved the densities are as a function of the logit.

More dimensions

The *Dirichlet distribution* is a density over an I-dimensional vector $\mathbf{p}$ whose I components are positive and sum to 1. The beta distribution is a special case of a Dirichlet distribution with $I = 2$. The Dirichlet distribution is parameterized by a measure $\mathbf{u}$ (a vector with all coefficients $u_i > 0$) which I will write here as $\mathbf{u} = \alpha\mathbf{m}$, where $\mathbf{m}$ is a normalized measure over the I components ($\sum m_i = 1$), and α is positive:

$$P(\mathbf{p} \,|\, \alpha\mathbf{m}) = \frac{1}{Z(\alpha\mathbf{m})} \prod_{i=1}^{I} p_i^{\alpha m_i - 1} \delta\left(\textstyle\sum_i p_i - 1\right) \equiv \text{Dirichlet}^{(I)}(\mathbf{p} \,|\, \alpha\mathbf{m}). \tag{23.30}$$

The function $\delta(x)$ is the Dirac delta function, which restricts the distribution to the simplex such that $\mathbf{p}$ is normalized, i.e., $\sum_i p_i = 1$. The normalizing constant of the Dirichlet distribution is:

$$Z(\alpha\mathbf{m}) = \prod_i \Gamma(\alpha m_i) \,/\, \Gamma(\alpha) \,. \tag{23.31}$$

The vector $\mathbf{m}$ is the mean of the probability distribution:

$$\int \text{Dirichlet}^{(I)}(\mathbf{p} \,|\, \alpha\mathbf{m}) \, \mathbf{p} \, \mathrm{d}^I\mathbf{p} = \mathbf{m}. \tag{23.32}$$

When working with a probability vector $\mathbf{p}$, it is often helpful to work in the 'softmax basis', in which, for example, a three-dimensional probability $\mathbf{p} = (p_1, p_2, p_3)$ is represented by three numbers a_1, a_2, a_3 satisfying $a_1 + a_2 + a_3 = 0$ and

$$p_i = \frac{1}{Z} e^{a_i}, \text{ where } Z = \textstyle\sum_i e^{a_i}. \tag{23.33}$$

This nonlinear transformation is analogous to the $\sigma \rightarrow \ln \sigma$ transformation for a scale variable and the logit transformation for a single probability, $p \rightarrow$

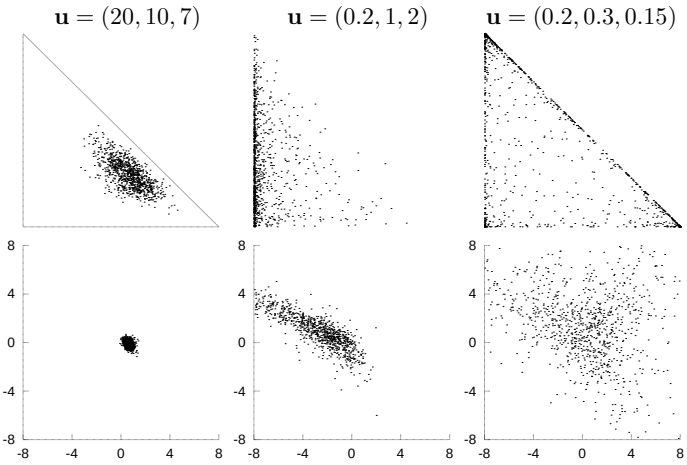


Figure 23.8. Three Dirichlet distributions over a three-dimensional probability vector (p_1, p_2, p_3). The upper figures show 1000 random draws from each distribution, showing the values of p_1 and p_2 on the two axes. $p_3 = 1 - (p_1 + p_2)$. The triangle in the first figure is the simplex of legal probability distributions.
The lower figures show the same points in the 'softmax' basis (equation (23.33)). The two axes show a_1 and a_2. $a_3 = -a_1 - a_2$.

$\ln \frac{p}{1-p}$. In the softmax basis, the ugly minus-ones in the exponents in the Dirichlet distribution (23.30) disappear, and the density is given by:

$$P(\mathbf{a} \,|\, \alpha\mathbf{m}) \propto \frac{1}{Z(\alpha\mathbf{m})} \prod_{i=1}^{I} p_i^{\alpha m_i} \delta\left(\textstyle\sum_i a_i\right). \tag{23.34}$$

The role of the parameter α can be characterized in two ways. First, α measures the sharpness of the distribution (figure 23.8); it measures how different we expect typical samples $\mathbf{p}$ from the distribution to be from the mean $\mathbf{m}$, just as the precision $\tau = 1/\sigma^2$ of a Gaussian measures how far samples stray from its mean. A large value of α produces a distribution over $\mathbf{p}$ that is sharply peaked around $\mathbf{m}$. The effect of α in higher-dimensional situations can be visualized by drawing a typical sample from the distribution $\text{Dirichlet}^{(I)}(\mathbf{p} \,|\, \alpha\mathbf{m})$, with $\mathbf{m}$ set to the uniform vector $m_i = 1/I$, and making a Zipf plot, that is, a ranked plot of the values of the components p_i. It is traditional to plot both p_i (vertical axis) and the rank (horizontal axis) on logarithmic scales so that power law relationships appear as straight lines. Figure 23.9 shows these plots for a single sample from ensembles with $I = 100$ and $I = 1000$ and with α from 0.1 to 1000. For large α, the plot is shallow with many components having similar values. For small α, typically one component p_i receives an overwhelming share of the probability, and of the small probability that remains to be shared among the other components, another component $p_{i'}$ receives a similarly large share. In the limit as α goes to zero, the plot tends to an increasingly steep power law.

Second, we can characterize the role of α in terms of the predictive distribution that results when we observe samples from $\mathbf{p}$ and obtain counts $\mathbf{F} = (F_1, F_2, \ldots, F_I)$ of the possible outcomes. The value of α defines the number of samples from $\mathbf{p}$ that are required in order that the data dominate over the prior in predictions.

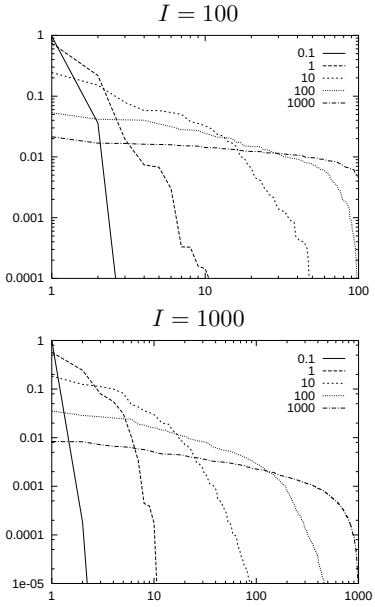


Figure 23.9. Zipf plots for random samples from Dirichlet distributions with various values of $\alpha = 0.1 \ldots 1000$. For each value of $I = 100$ or 1000 and each α, one sample $\mathbf{p}$ from the Dirichlet distribution was generated. The Zipf plot shows the probabilities p_i, ranked by magnitude, versus their rank.

Exercise 23.3.[3] The Dirichlet distribution satisfies a nice additivity property. Imagine that a biased six-sided die has two red faces and four blue faces. The die is rolled N times and two Bayesians examine the outcomes in order to infer the bias of the die and make predictions. One Bayesian has access to the red/blue colour outcomes only, and he infers a two-component probability vector (p_R, p_B). The other Bayesian has access to each full outcome: he can see which of the six faces came up, and he infers a six-component probability vector $(p_1, p_2, p_3, p_4, p_5, p_6)$, where

$p_{\mathrm{R}} = p_1 + p_2$ and $p_{\mathrm{B}} = p_3 + p_4 + p_5 + p_6$. Assuming that the second Bayesian assigns a Dirichlet distribution to $(p_1, p_2, p_3, p_4, p_5, p_6)$ with hyperparameters $(u_1, u_2, u_3, u_4, u_5, u_6)$, show that, in order for the first Bayesian's inferences to be consistent with those of the second Bayesian, the first Bayesian's prior should be a Dirichlet distribution with hyperparameters $((u_1 + u_2), (u_3 + u_4 + u_5 + u_6))$.

Hint: a brute-force approach is to compute the integral $P(p_{\mathrm{R}}, p_{\mathrm{B}}) = \int \mathrm{d}^6\mathbf{p}\, P(\mathbf{p} \,|\, \mathbf{u})\, \delta(p_{\mathrm{R}} - (p_1 + p_2))\, \delta(p_{\mathrm{B}} - (p_3 + p_4 + p_5 + p_6))$. A cheaper approach is to compute the predictive distributions, given arbitrary data $(F_1, F_2, F_3, F_4, F_5, F_6)$, and find the condition for the two predictive distributions to match for all data.

The *entropic distribution* for a probability vector $\mathbf{p}$ is sometimes used in the 'maximum entropy' image reconstruction community.

$$P(\mathbf{p} \,|\, \alpha, \mathbf{m}) = \frac{1}{Z(\alpha, \mathbf{m})} \exp[-\alpha D_{\mathrm{KL}}(\mathbf{p}||\mathbf{m})]\, \delta(\textstyle\sum_i p_i - 1)\,, \tag{23.35}$$

where $\mathbf{m}$, the measure, is a positive vector, and $D_{\mathrm{KL}}(\mathbf{p}||\mathbf{m}) = \sum_i p_i \log p_i/m_i$.

Further reading

See (MacKay and Peto, 1995) for fun with Dirichlets.

▶ 23.6 Further exercises

Exercise 23.4.[2] N datapoints $\{x_n\}$ are drawn from a gamma distribution $P(x \,|\, s, c) = \Gamma(x; s, c)$ with unknown parameters s and c. What are the maximum likelihood parameters s and c?