2.8 Maximum likelihood and the exponential

Our basic distribution here (sometimes called the Laplace distribution) is

$$\text{prob}(x) = \frac{1}{2a} \exp\left[-\frac{|x|}{a}\right]$$

and so for n data X_i we have the posterior distribution for a

$$\text{prob}(a|\text{data}) \propto \left(\frac{1}{2a}\right)^n \exp\left[-\frac{\sum_{i=1}^n |X_i|}{a}\right] \text{prob}(a).$$

If we just take the prior for a as being a constant over some largish range, starting at zero, we will have a function which looks like the classical likelihood function; maximizing with respect to a we find that

$$a_1 = \frac{\sum_{i=1}^n |X_i|}{n}$$

is the most probable value for a, given the data and the prior.
There are several quite strong theoretical arguments for taking a prior

$$\text{prob}(a) \propto \frac{1}{a}$$

because it is a "scale parameter" and we want our prior ignorance to be portrayed in the same way regardless of our choice of unit. In this case we can find in the same way that

$$a_2 = \frac{\sum_{i=1}^n |X_i|}{n+1}$$

is the most probable value.
For small amounts of data, there is a large difference between a_1 and a_2 – the prior dominates the result, as we would expect. From simulations (try it yourself) a_1 seems to be better for small n.
Pragmatically, priors of the form (Lee, 1997)

$$\text{prob}(a) \propto \frac{1}{a^m} \exp\left[-\frac{R}{a}\right]$$

(m and R are parameters) are useful because the posterior will then be of the same form, which is often called an "inverse chi-square" distribution. R plays the useful rôle of preventing the divergence of the Jeffryes prior at small a. This is an example of a so-called conjugate prior.
Now suppose we have an unknown location parameter, μ:

$$\text{prob}(x) = \frac{1}{2a} \exp\left[-\frac{|x-\mu|}{a}\right].$$

For a location parameter, the appropriate prior to represent total ignorance is again theoretically well-determined

$$\text{prob}(\mu) = \text{constant}$$

over some arbitrarily wide range of μ.
The posterior, for data X_i as before, is then (including a simple Jeffryes prior on a)

$$\text{prob}(a, \mu \,|\, \text{data}) \propto \left(\frac{1}{2a}\right)^{n+1} \exp\left[-\frac{\sum_{i=1}^{n} |X_i - \mu|}{a}\right].$$

Differentiating the log of this expression, with respect to μ and a, gives two equations for their most likely values:

$$\frac{n+1}{a} - \frac{\sum_{i=1}^{n} |X_i - \mu|}{a^2} = 0$$

and (remembering carefully the definition of the derivative of the absolute value function)

$$\sum_{i=1}^{n} \text{sign}(X_i - \mu) = 0.$$

The second equation tells us that the best estimate of μ is the median, the value of x which has equal numbers of data points above and below it. We can place this value for μ into the first equation and solve for a.
Checking the results by simulation, we now find that neither the "diffuse prior" (corresponding the a_1, above) nor the Jeffryes prior (corresponding to a_2) gives unbiased results for small amounts of data. The problem arises because μ, being estimated from the data, tends artificially to reduce the scatter around itself.