COMPARING THE BAYESIAN AND LIKELIHOOD APPROACHES TO INFERENCE: A GRAPHICAL APPROACH

William Bolstad
University of Waikato, New Zealand
bolstad@waikato.ac.nz

Both likelihood inference and Bayesian inference arise from a surface defined on the inference universe which is the Cartesian product of the parameter space and the sample space. Likelihood inference uses the sampling surface which is a probability distribution in the sampling dimension only. Bayesian inference uses the joint probability distribution defined on the inference universe The likelihood function and the Bayesian posterior distribution come from cutting the respective surfaces with a (hyper)plane parallel to the parameter space and through the observed sample values. Unlike the likelihood function, the posterior distribution always will be a probability distribution. This is responsible for the different choices of estimators, and the different way the two approaches have of dealing with nuisance parameters. In this paper we present a graphical approach for teaching the difference between the two approaches.

INTRODUCTION

In this paper we graphically illustrate the differences and similarities between the maximum likelihood and Bayesian approaches to inference. We will see that under the two approaches (1) The estimators come from different surfaces, (2) Even when the surfaces are the same shape (flat priors) the estimators are chosen to satisfy different criteria, and (3) The two approaches have different ways of dealing with nuisance parameters. The observation(s) come from the sampling distribution $f(y|\theta)$ where θ is the fixed parameter value. It gives the probability distribution over all possible observation values for the given value of the parameter. The parameter space, Θ is the set of all possible parameter values. It ordinarily has the same dimension as the total number of parameters, p. The sample space S, is the set of all possible values of the observation(s). Its dimension is the number of observations n. When we are in the exponential family of distributions, the dimension of the sample space may be reduced to the number of sufficient statistics. We define the inference universe of the problem to be the Cartesian product of the parameter space and the sample space. See Bolstad (2010). It is the $p+n$ dimensional space where the first p dimensions are the parameter space, and the remaining n dimensions are the sample space. We do not ever observe the parameter, so the position in those coordinates are always unknown. However, we do observe the sample, so we know the last n coordinates.

SINGLE PARAMETER CASE

We will let the dimensions be $p=1$ and $n=1$ for illustrative purposes. This is the case for a single parameter and a single observation (or observations from a one-dimensional exponential family). Figures 1, 2, and 3 are exact in this case. When we have $p \geq 2$ the same ideas hold, however we cannot project the surface defined on the inference universe down to a two dimensional graph. With multiple parameters, Figures 1, 2, and 3 can be considered to be schematic diagrams that represent the ideas rather than exact representations.

Maximum Likelihood Estimation

We are trying to choose an estimator to represent the unknown value of the parameter. The sampling distribution $f(y|\theta)$ is a function of both the value of the observation and the parameter value. Given the value θ, it gives the probability distribution of the observation y. It is defined for all points in the inference universe. Thus it forms a surface defined on the inference universe. It is a probability distribution in the observation dimension for each particular value of the parameter. However it is not a probability distribution in the parameter dimension. The first panel of Figure 1

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

shows the sampling distribution surface in 3D perspective. The likelihood function has the same functional form as the sampling distribution, only y is held at the observed value, and θ is allowed to vary over all possible values. It is found by cutting the sampling distribution surface with a vertical plane parallel to the θ axis through the observed value, as shown in the second panel of Figure 1. Likelihood inference is based on the likelihood function. Since it is not a probability density, Fisher decided that the best estimator of the parameter is the value that has the highest value of the likelihood function. He named this the maximum likelihood estimator, (MLE). The MLE is invariant under any reparamaterization.

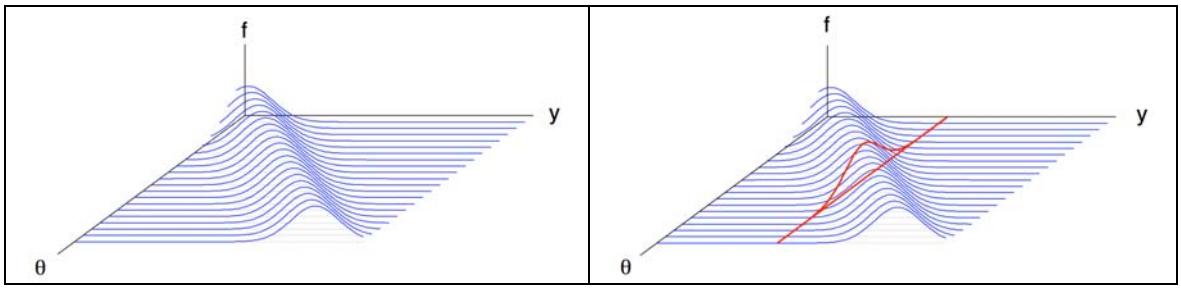

Figure 1. The observation distribution (L) and the likelihood function (R)

Bayesian Estimation

Bayesian estimation requires that we have a probability distribution defined on the parameter space before we look at the data. It is called the prior distribution. It gives our belief weights for each of the possible parameter values before we see the data. This requires that we allow a different interpretation of probability on the parameter space than on the sample space. It is measures our belief, and thus is subjective. The probability on the sample space has the usual long-run relative frequency interpretation. The prior distribution of the parameter is shown with the sampling distribution surface in the first panel of Figure 2. The joint distribution of the parameter and the observation is found by multiplying each value of the sampling distribution surface by the corresponding height of the prior distribution. This is shown in the second panel of Figure 2. To find the posterior distribution of the parameter given the observed value we cut the joint distribution of the parameter and the observation with a vertical plane parallel to the parameter axis through the observed value of y. This is shown in the first panel of Figure 3. The posterior distribution summarizes the belief we can have about all possible parameter values, given the observed data. It will always be a probability distribution, conditional on the observed data. We can use the mean of this distribution as the estimate of the parameter. See Bolstad (2007). The mean of a distribution is the value that minimizes the mean-squared deviation. Hence, the Bayesian posterior estimator minimizes the mean-squared deviation of the posterior distribution.

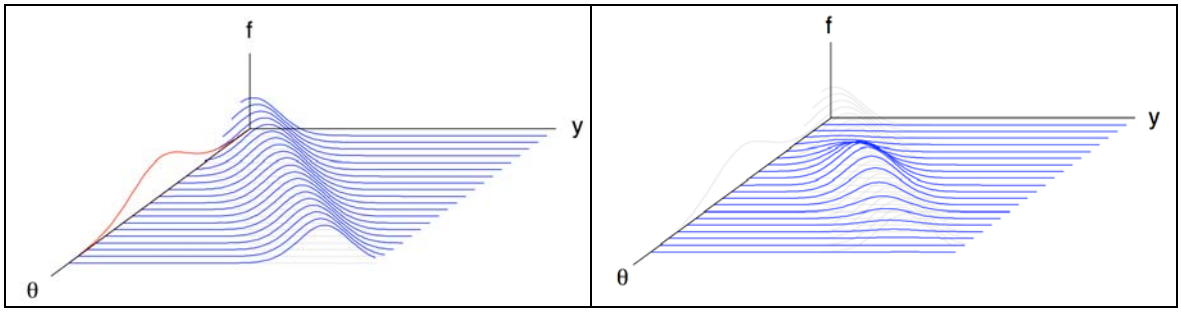

Figure 2. The prior and likelihood (L) and the joint density of θ and y (R)

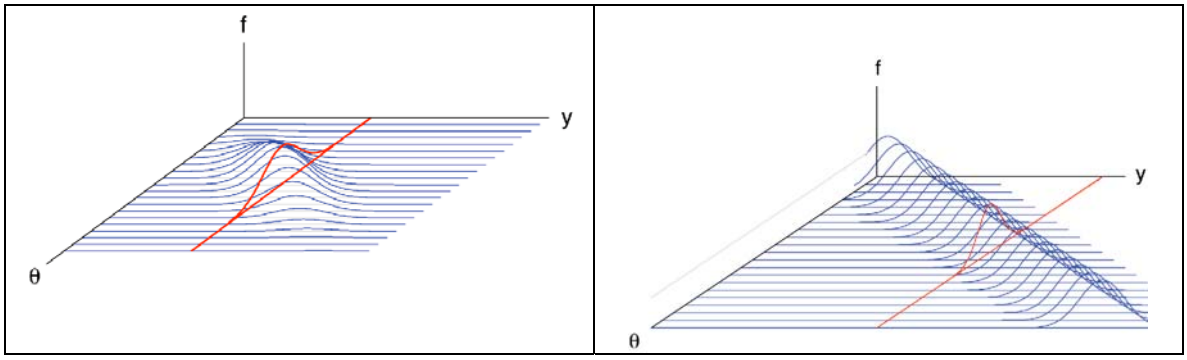

Figure 3. Posterior (L) and posterior when flat prior is used (R)

Using flat prior the posterior has same shape as the likelihood

If we decide to use a flat prior distribution that gives equal weight to all values of the parameters the joint distribution on the inference universe will be the same as the sampling distribution surface. This is shown in the second panel of Figure 3. Note that this prior distribution will be improper unless the parameter values have finite lower and upper bounds. When the prior is improper, we do not have a joint probability distribution. Nevertheless the normed likelihood function will be a probability distribution. In this case, the Bayesian posterior estimator would be the mean value (balance point) of the likelihood function. This is not generally the same value as the maximum likelihood estimator, unless the likelihood function is symmetric and unimodal such as in the normal likelihood. Figure 4 illustrates the difference between these estimators on a non-symmetric likelihood function that could also be considered a Bayesian posterior distribution with a flat prior distribution. The maximum likelihood estimator is the mode of this curve, while the Bayesian estimator is its mean. The two estimators are based on different ideas, even when the likelihood function and the posterior distribution have the same shape.

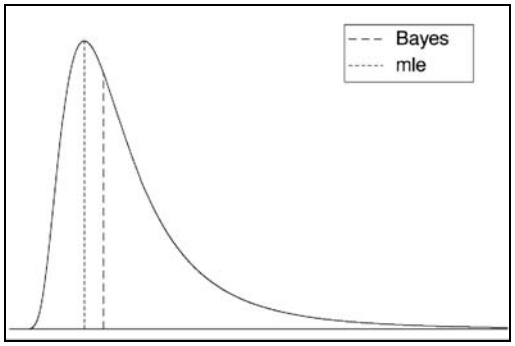

Figure 4. Maximum likelihood estimator and Bayesian estimator

MULTIPLE PARAMETER CASE

We will use the two parameter case to show what happens when there are multiple parameters. The inference universe has at least four dimensions, so we cannot graph the surface on it. The likelihood function is still found by cutting through the surface with a hyperplane parallel to the parameter space passing through the observed values. The likelihood function will be defined on the two parameter dimensions, and we graph it 3D perspective in Figure 5. In this example, we have the likelihood function where θ_1 is the mean and θ_2 is the variance for a normal random sample

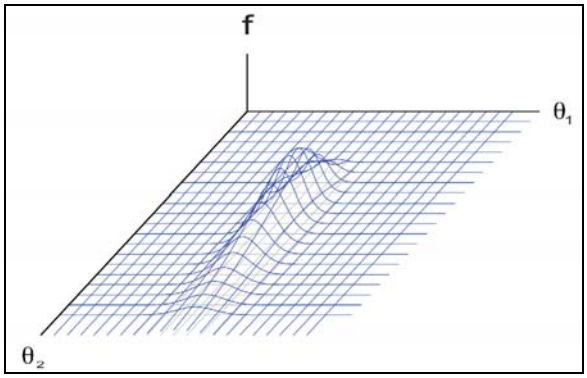

Figure 5. Joint likelihood

Nuisance parameters

Sometimes, only one of the parameters is of interest to us. We don't want to estimate the other parameters and call them "nuisance" parameters. We just want to make sure the nuisance parameters don't interfere with our inference on the parameter of interest. Because using the Bayesian approach the joint posterior distribution is a probability distribution, and using the likelihood approach the joint likelihood function is not a probability distribution, the two approaches have different ways of dealing with the nuisance parameters, even when we use independent joint flat priors and the posterior distribution and likelihood function are the same shape.

Likelihodd Inference in he prsence of nuisance parameters

For instance, suppose that θ_1 is the parameter of interest, and θ_2 is a nuisance parameter. If there is a ancillary sufficient statistic, conditioning on it will give a likelihood that only depends on θ_1, the parameter of interest, and inference can be based on that conditional likelihood. This can only be true in certain exponential families, so is of limited general use when nuisance parameters are present. Instead, likelihood inference on θ_1 is often based on the profile likelihood function given by:

$$L_p(\theta_1; data) = \sup_{\theta_2}\{L(\theta_1, \theta_2; data)\}$$

where $L(\theta_1, \theta_2; data)$ is the joint likelihood function. See Kotz et. al. (1986). Essentially, the nuisance parameter has been eliminated by plugging $\hat{\theta}_2 \mid \theta_1$, the conditional maximum likelihood value of θ_2 given θ_1, into the joint likelihood. Hence

$$L_p(\theta_1; data) = L(\theta_1, \hat{\theta}_2 \mid \theta_1; data).$$

This is shown in the first panel of Figure 6. The profile likelihood function may lose some information about θ_1,compared to the joint likelihood function. Note that the maximum profile likelihood value of θ_1 will be the same as its maximum likelihood value. However confidence intervals based on profile likelihood may not be the same as those based on the joint likelihood.

Bayesian Inference in the presence of nuisance parameters

Bayesian statistics has a single way of dealing with nuisance parameters. Inference about the parameter of interest θ_1, is based on the marginal posterior $g(\theta_1 \mid data)$ which is found by integrating the nuisance parameter out of the joint posterior, a process referred to as marginalization.

$$g(\theta_1 \mid data) = \int g(\theta_1, \theta_2 \mid data)\, d\theta_2$$

Note: we are using independent flat priors for both θ_1 and θ_2 , so the joint posterior is the same shape as the joint likelihood in this example. The joint posterior distribution with the marginal

distribution is shown in the second panel of Figure 6. The marginal posterior has all the information about θ_1 that was in the joint posterior. In this example, the Bayesian posterior distribution of θ_1 found by marginalizing θ_2 out of the joint posterior, and the profile likelihood function of θ_1 turn out to have the same shape. That is not always the case. For instance, suppose we wanted to do inference on θ_2 and regarded θ_1 as the nuisance parameter. We have used independent flat priors for both parameters, so the joint posterior has the same shape as the joint likelihood. The profile likelihood and marginal posterior of θ_2 are shown in 3D perspective in the first and second panels of Figure 7, respectively. Figure 8 compares the shapes of the profile likelihood function and the marginal posterior distribution in 2D for θ_2 in this case. Clearly they have different shapes despite coming from the same two dimensional function.

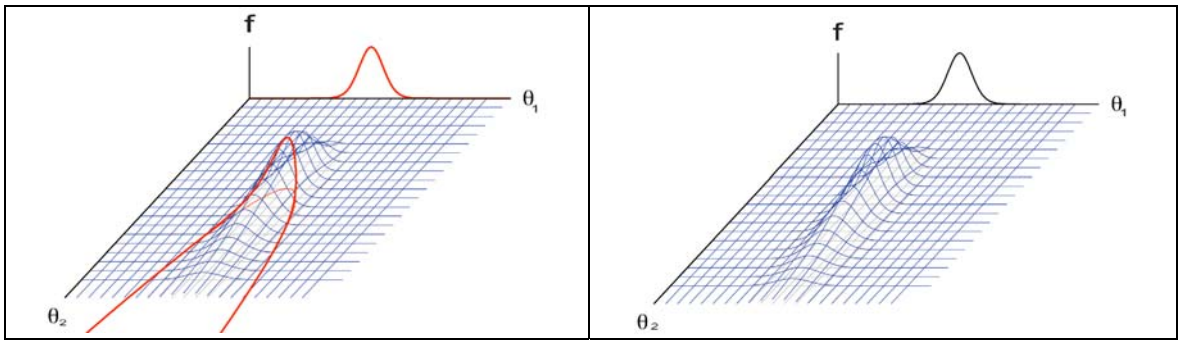

Figure 6. Profile likelihood (L) and marginal posterior (R) for θ_1

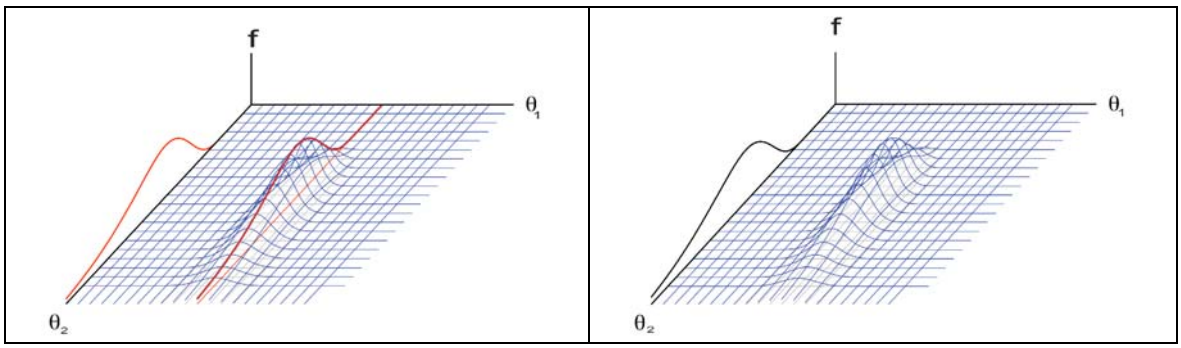

Figure 7. Profile likelihood (L) and marginal posterior (R) for θ_2

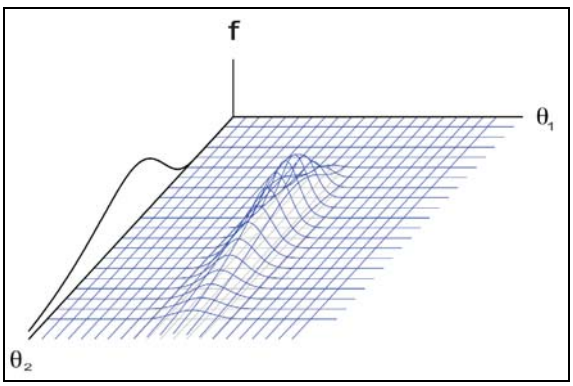

Figure 8. Profile likelihood and marginal posterior for θ_2 in 2D

CONCLUSION

Both the likelihood and Bayesian approach arise from surfaces defined on the inference universe. Cutting through these surfaces with a hyperplane that goes through the observed data yields the likelihood function and the posterior distribution which are used for likelihood inference and Bayesian inference respectively. The likelihood function is not considered a probability distribution, while the posterior distribution always is. The main differences between these two approaches stem from this difference. Certain ideas arise naturally when dealing with a probability distribution. There is no reason to use the first moment of the likelihood function without the probability interpretation, so the maximum likelihood estimator is the value that gives the highest value on the likelihood function. When a flat prior is used, the posterior distribution has the same shape as the likelihood function. Under the Bayesian approach it has a probability interpretation, so the posterior mean which minimizes the mean squared deviation will be the estimator. When there are nuisance parameters, there is no reason why they could not be integrated out of the likelihood function, and the inference be based on the marginal likelihood. However, without the probability interpretation on the joint likelihood, there is no compelling reason to do so. Instead, likelihood inference is commonly based on the profile likelihood function, where the maximum conditional likelihood values of the nuisance parameters given the parameters of interest are plugged into the joint likelihood. Under the Bayesian approach the joint posterior distribution is clearly a probability distribution. Hence Bayesian inference about the parameter of interest will be based on the marginal posterior where the nuisance parameter has been integrated out.

REFERENCES

Bolstad, W. M. (2007). *Introduction to Bayesian Statistics: Second Edition.* NY: John Wiley & Sons.

Bolstad, W. M. (2010). *Understanding Computational Bayesian Statistics*. NY: John Wiley & Sons.

Kotz, S., Johnson, N. L., & Read, C. B. (Eds.) (1986). *Encyclopedia of Statistical Sciences, Volume 7.* NY: John Wiley & Sons.