Maximum Entropy and Bayesian Learning

Jose H. Guardiola[1], Hassan Elsalloukh[2]
[1]Texas A&M University Corpus Christi, Department of Mathematics and Statistics, CI-309, 6300 Ocean Drive, Corpus Christi, TX 78412
[2]University of Arkansas at Little Rock, Department of Mathematics and Statistics, DKSN 617, 2801 S. University, Little Rock, AR 72204

Abstract
This paper discusses the relationship between the maximum entropy approach and Bayesian statistical inference, the Kullback-Leibler divergence, and Fisher's information. Using an example from science it shows how the information gain in Bayesian updating can be measured using the Kullback-Leibler divergence or cross entropy, change in entropy, and Fisher's information. This example discusses the relationship between these measurements. A numeric example is developed and detailed results are discussed under information theory and statistical point of views by comparing related quantities. Bayesian inference results and theory are interpreted using information concepts, entropy and statistical measurements, finally some conclusions are drawn regarding the information gain and relationships with other statistical procedures.

Key Words: Bayesian estimation, Maximum entropy estimation, Maximum likelihood estimation, Information theory, Information gain, Posterior distribution

1. Basic Concepts in Estimation

Frequentist statistics and Bayesian statistics have different approaches for model selection and parameter estimation. As it is well known, the classical approach to statistics is widely used and it has readily available software but the incorporation of prior information is not possible, the parameters of a distribution are considered fixed but unknown and the estimation methods such as maximum likelihood estimation, integrate over the data, even the unobserved data, and the interpretation of results is more difficult as it has to refer to the classical mantra "under repeated sampling…". The popular maximum likelihood estimation can lead to results that are inconsistent with the likelihood principle and occasionally can lead to some non-sense results such as negative variances.

On the other hand, Bayesian statistics overcomes some of the difficulties mentioned in the frequentist approach, as it can easily incorporate prior information and the interpretation of results is very straightforward as the parameter estimation can be expressed in terms of probabilities, without having to use the repeated sampling mantra. In Bayesian statistics, data are considered fixed and the estimation method integrates over the parameter space as the latter can be considered random variables with a probability distribution that is going to be determined. All these process is consistent with the probability laws. The estimation process is difficult and can only be done explicitly in very simple cases, but for more practical problems we have to use Markov chain Monte Carlo simulation to be able to estimate the posterior distribution. Critics of Bayesian

statistics argue that the prior estimation is subjective, mainly because using different priors can lead to different results, however, sometimes even that different priors can be used, a general pattern can emerge. In summary, the Bayesian philosophy is based on learning and information gain, and it depends on two things: the prior and the posterior. You can only say how much you learned if you know what your prior belief is.

1.1 Frequentist Methods and Measures of Goodness of Fit

This section summarizes the frequentist estimation methods that will be used in the following sections.

1.1.1 Maximum Likelihood Estimation

Given a random sample from a distribution $\mathbf{X} = \left(x_1, x_2 ..... x_n\right) = x_i^n$ and given the likelihood function as the joint distribution:

$$L\left(\theta \mid x_1, ..., x_n\right) = L\left(\theta \mid x_i^n\right) = f\left(x_i^n \mid \theta\right) = \prod_{i=1}^{n} f\left(x_i \mid \theta\right)$$

For maximum likelihood estimation we choose the value of the parameter that maximizes the likelihood function:

$$\hat{\theta} = \arg \max_{\theta} \mathrm{L}\left(\theta; x_i^n\right)$$

in practice we minimize the log-likelihood function as:

$$\hat{\theta} = \arg \max_{\theta} \log \mathrm{L}\left(\theta; x_i^n\right)$$

Or we minimize the negative of the log likelihood function. $\hat{\theta}$ is the maximum likelihood estimate (MLE).

1.1.2 Likelihood Ratio Test

A related criteria for testing hypothesis regarding the value of a parameter is the likelihood ratio test whose null and alternative hypothesis that can be expressed as $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^C$, then the likelihood ratio test can be expressed as :

$$\lambda\left(x\right) = \frac{\sup_{\Theta_0} L\left(\theta \mid x\right)}{\sup_{\Theta} L\left(\theta \mid x\right)}$$

A large value is in favor of the null hypothesis, while a small value is in favor of the alternative hypothesis (Casella 1990).

1.1.3 Fisher Information

The information about θ in a random sample of size n is given by:

$$I\left(\theta\right) = -n \cdot E\left[\frac{\partial^2 \ln f\left(x\right)}{\partial \theta^2}\right] = n \cdot E\left[\left(\frac{\partial^2 \ln f\left(x\right)}{\partial \theta^2}\right)^2\right]$$

This expression also provides a bound of the variance of the best unbiased estimator of θ (Cramer-Rao Inequality) Freund (2014)

The observed Fisher information that we can compute from our sample is:

$$\kappa(\theta) = \frac{1}{\rho(\theta)} = \left| \frac{\dfrac{d^2 \log L(\theta)}{d\theta^2}}{\left\{1 + \left[\dfrac{d \log L(\theta)}{d\theta}\right]^2\right\}^{3/2}} \right|$$

That expression can be interpreted geometrically as:

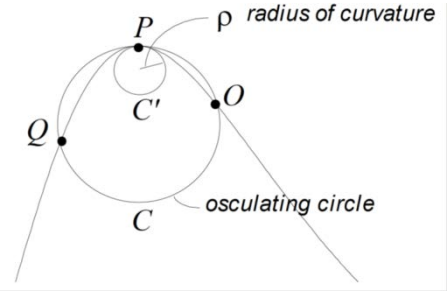

Figure 1: Geometric interpretation of Fisher information.

The ratio of curvature can be expressed as:

$$I\left(\hat{\theta}\right) = \kappa\left(\hat{\theta}\right) = -\left.\frac{d^2 \log L(\theta)}{d\theta^2}\right|_{\theta=\hat{\theta}}$$

And the curvature evaluated at the MLE, is known as the observed Fisher information (Huzurbazar, 1949)

1.2 Bayesian Estimation

Initially the parameters can be assigned a prior distribution that describe what we know about these parameters, and that combined with the new information from a sample it allows us to update our knowledge and obtain a posterior distribution using the Bayes rule:

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\int P(x|\theta)P(\theta)d\theta}$$

Then we can use last expression to make a probability statement about the parameter θ.

1.2.1 Bayes Factors
Bayes factors allow us to compare two models (Bernardo, 1994) as follows:

$$K = \frac{P(data|M_1)}{P(data|M_2)} = \frac{\int P(\theta_1|M_1)P(data|\theta_1, M_1)d\theta_1}{\int P(\theta_2|M_2)P(data|\theta_2, M_2)d\theta_2}$$

If instead of the Bayes factors integral the maximum likelihood estimators are used, the test becomes the likelihood ratio test.

1.2.2 Deviance information Criteria
Spiegelhalter, Best, Carlin and Van Der Linde (2001) developed a generalization of the Akaike information criteria (AIC), known as Deviance Information Criteria (DIC) and they showed that it is asymptotically equivalent to AIC.

1.3 Entropy and Information Theory
This section summarizes the basics concepts of entropy and information theory.

1.3.1 Entropy
Boltzmann (1872) quantifies the entropy of a thermodynamic system as:

$$S = K \log W$$

where, S = Entropy, K = Boltzmann constant, W = number of microstates in the system.

1.3.2 Information Theory
Shannon (1948) defines entropy of a discrete random variable and probability mass function:

$$H(X) = E\left[-\ln\left(P(X)\right)\right] = -\sum_{i=1}^{n} P(x_i)\log_b(x_i)$$

When the distribution is continuous, the sum is replaced with an integral:

$$H(X) = -\int P(x)\log_b P(x)\,dx$$

1.3.3 Principle of Maximum Entropy
This principle was stated by Jaynes in 1957[1],[2], where he emphasized a natural correspondence between statistical mechanics and information theory. Statistical mechanics can be seen as an application of logical inference and information theory.

Properties of Entropy:

- Entropy is nonnegative (discrete case)
- When the distribution is uniform entropy increases with the cardinality of support
- If $\pi(\theta) = 1$ for some θ, entropy vanishes.
- The more "uniform", the greater the entropy
- Entropy is invariant with respect to permutations in the support of θ

- Entropy is a continuous function of $\pi(\theta)$

Maximum Entropy Distribution:

Suppose Θ is discrete and P be all the probability distributions on Θ.
A distribution $\pi \in P$ is a maximum entropy distribution if:

$$En(\pi^*) = \sup_{\pi \in \mathrm{P}} En(\pi)$$

Restricted Maximum Entropy Distribution:
Suppose Θ is discrete and let π be a probability distributions on Θ. Let g_k be a function defined on Θ such as $E[g_k(\theta)]$ exists, for k =1,2,…n. Suppose we know that:

$$E^{\pi}\left[g_k(\theta)\right] = \sum_{\Theta} \pi(\theta_i) g_k(\theta_i) = \mu_k$$

Then the restricted maximum entropy distribution, subject to constraints is:

$$\pi^*(\theta_i) = \frac{\exp\left\{\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\right\}}{\sum_{\Theta} \exp\left\{\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\right\}} = \frac{\exp\left\{\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\right\}}{Z(\lambda)}$$

Where $\lambda=(\lambda_1, \lambda_2, \ldots\ldots \lambda_m)$ is determined from:

$$\mu_k = \frac{\partial}{\partial \lambda_k} \log Z(\lambda) \quad ,k=1,2,\ldots..m$$

1.3.4 Kullback-Leibler Divergence
In information theory and probability the Kullback-Leibler divergence (KLD) is a measure of the difference between two probability distributions:

$$D_{KL}(P \| Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

For continuous random variables the KLD can be expressed with integrals as follows:

$$D_{KL}(P \| Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

Properties for the KLD:

$$D_{KL}(P \| Q) \geq 0$$

$$D_{KL}(P \| P) = 0$$

It is a pseudo-distance :

$$D_{KL}\left(P\|Q\right) \neq D_{KL}\left(Q\|P\right)$$

The asymmetry issue was addressed by Jeffreys as:

$$J-divergence = D_{KL}\left(P\|Q\right) + D_{KL}\left(Q\|P\right)$$

2. Bayesian and Frequentist Models

This section develops a numeric example and the corresponding inferences for frequentist and Bayesian estimation.

2.1 Bayesian Example

Suppose we want to determine the sex ratio θ for a certain kind of animal. Because we don't have previous information we can start with a uniform distribution as a prior for θ. We are going to take 4 sets of 10 observations each and we update our prior after every 10 samples, we repeat this process 4 times.

$$P\left(\theta|x\right) = \frac{P\left(x|\theta\right)P\left(\theta\right)}{\int P\left(x|\theta\right)P\left(\theta\right)d\theta}$$

Using the gamma distribution as the prior

$$h\left(\theta\right) = \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\theta^{\alpha-1}\left(1-\theta\right)^{\beta-1} \qquad 0<\theta<1$$

The likelihood is a binomial

$$f\left(x|\theta\right) = \binom{n}{x}\theta^{x}\left(1-\theta\right)^{n-x}$$

The posterior is the well-known result

$$\varphi\left(\theta|x\right) = \Gamma\left(x+\alpha, n-x+\beta\right)$$

Table 1: Sequence of generated data and updated parameters

Iteration	Sequence of generated data	Updated parameter (mode)
1st	0 1 0 0 1 1 0 0 1 0	$p = 0.4$
2nd	0 0 0 1 0 1 1 0 0 0	$p = 0.35$
3rd	1 0 1 0 1 1 0 0 0 0	$p = 0.3667$
4th	0 0 1 0 0 0 0 0 0 0	$p = 0.3$

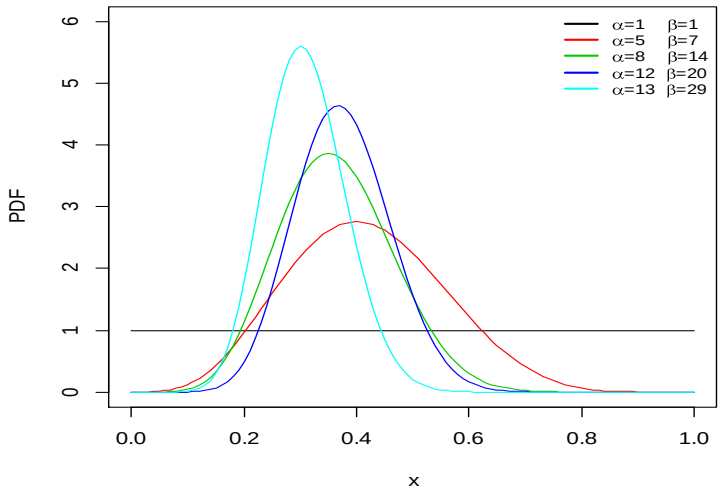

Figure 2: Sequence of Bayesian posteriors for four updates of 10 trials each.

2.2 Bayesian Factors for Two Hypotheses

In general the Bayesian factor for comparing two competing hypotheses can be written as:

$$K = \frac{P\left(data\middle|M_1\right)}{P\left(data\middle|M_2\right)} = \frac{\int P\left(\theta_1\middle|M_1\right)P\left(data\middle|\theta_1,M_1\right)d\theta_1}{\int P\left(\theta_2\middle|M_2\right)P\left(data\middle|\theta_2,M_2\right)d\theta_2}$$

In particular, solving the integrals for two betas:

$$K = \frac{B\left(\alpha_1+x,\beta_1+n-x\right)B\left(\alpha_2,\beta_2\right)}{B\left(\alpha_2+x,\beta_2+n-x\right)B\left(\alpha_1,\beta_1\right)}$$

Dissimilarity matrices can be computed for Bayes factors showing every possible combination for the posterior distributions for the four updates:

$$BF = \begin{bmatrix} 1.000 & 2.017 & 2.785 & 3.321 & 3.993 \\ 0.496 & 1.000 & 1.232 & 1.348 & 1.976 \\ 0.359 & 0.812 & 1.000 & 1.068 & 1.365 \\ 0.301 & 0.742 & 0.936 & 1.000 & 1.712 \\ 0.250 & 0.506 & 0.733 & 0.584 & 1.000 \end{bmatrix}$$

Computing the natural log of the previous matrix in order to be able to express these factors as a distance:

$$\ln(BF) = \begin{bmatrix} 0.000 & 0.702 & 1.024 & 1.200 & 1.384 \\ -0.702 & 0.000 & 0.208 & 0.299 & 0.681 \\ -1.024 & -0.208 & 0.000 & .066 & 0.311 \\ -1.200 & -0.208 & -.066 & 0.000 & 0.538 \\ -1.384 & -0.681 & -0.311 & -0.538 & 0.000 \end{bmatrix}$$

Table 2: Jeffreys interpretation of Bayes factors

K	dHart	bits	Strength of evidence
$< 10^0$	< 0		negative (supports M_2)
10^0 to $10^{1/2}$	0 to 5	0 to 1.6	barely worth mentioning
$10^{1/2}$ to 10^1	5 to 10	1.6 to 3.3	substantial
10^1 to $10^{3/2}$	10 to 15	3.3 to 5.0	strong
$10^{3/2}$ to 10^2	15 to 20	5.0 to 6.6	very strong
$> 10^2$	> 20	> 6.6	decisive

2.3 Likelihood Ratio Test Results

The likelihood ratio test for the sequence of updated posteriors can be computed in a similar manner that we did to compute the Bayes factors, and transforming to logarithms we can obtain the following dissimilarity matrix:

$$LogLRT = \begin{bmatrix} 0.000 & 0.201 & 0.914 & 1.080 & 3.291 \\ -0.201 & 0.000 & 0.216 & 0.106 & 1.171 \\ -0.914 & -0.216 & 0.000 & 0.054 & 0.464 \\ -1.080 & -0.106 & -0.054 & 0.000 & 1.863 \\ -3.291 & -1.171 & -0.464 & -1.863 & 0.000 \end{bmatrix}$$

2.4 Kullback-Leibler Divergence Between Two Betas

From the general form for KLD is:

$$D_{KL}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

Solving the integrals for two betas (Schumitzky, 2014):

$$D_{KL}(B_1, B_2) = \psi(\alpha_1)(\alpha_1 - \alpha_2) + \psi(\beta_1)(\beta_1 - \beta_2) + \psi(\alpha_1 + \beta_1)(\alpha_2 + \beta_2 - (\alpha_1 + \beta_1)) + \log\left[\frac{B(\alpha_2, \beta_2)}{B(\alpha_1, \beta_1)}\right]$$

where ψ is the digamma function:

$$\psi(x) = \frac{d}{dx}\ln(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)}$$

The Entropy and cross entropy for two Betas is as follows:

$$H(p) = InB(\alpha_1, \beta) - (\alpha_1 - 1)\psi(\alpha_1) - (\beta_1 - 1)\psi\beta_1 + (\alpha_1 + \beta_1 - 2)\psi(\alpha_1 + \beta_1)$$

And the cross entropy is:

$$H(p,q) = InB(\alpha_2, \beta_2) - (\alpha_2 - 1)\psi(\alpha_1) - (\beta_2 - 1)\psi\beta_1 + (\alpha_2 + \beta_2 - 2)\psi(\alpha_1 + \beta_1)$$

The relationship between them is:

$$D_{KL} = H(p,q) - H(p)$$

The change in entropy due to Bayesian learning for four updates of 10 trials each is as follows:

Table 3: Change in Entropy due to Bayesian learning, four updates of 10 trials at a time.

Model	KLD	H(p)	Cross Entropy
α= 1, β=1	----	0	-------
α= 5, β=7	2.255	0	2.255
α= 8, β=14	0.241	-0.580	-0.339
α= 12, β=20	0.054	-0.887	-0.833
α= 13, β=29	0.416	-1.058	-0.643

Then, the numeric results for the dissimilarity matrix for KL divergence:

$$KLD = \begin{bmatrix} 0.000 & 2.255 & 5.698 & 8.750 & 13.843 \\ 0.580 & 0.000 & 0.241 & 0.459 & 1.617 \\ 0.887 & 0.144 & 0.000 & 0.054 & 0.386 \\ 1.058 & 0.220 & 0.041 & 0.000 & 0.416 \\ 1.239 & 0.548 & 0.219 & 0.326 & 0.000 \end{bmatrix}$$

And the change in entropy, Kullback-Leibler divergence and cross entropy can be shown as follows:

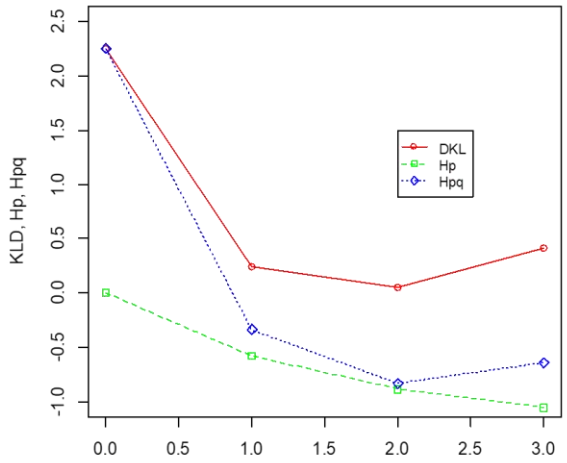

Figure 3: Relationship among KLD, Entropy and Cross Entropy for four updates of 10 trials each.

The Bayesian learning in the numeric discrete case that we are discussing can be computed as:

$$D_{KL}\left(P_i \| P_{i+1}\right) = \sum_{\forall i} p_i(x) \log \frac{p_i(x)}{p_{i+1}(x)}$$

And the numeric results for the learning process can be summarized in the following table:

Table 3: KLD, Entropy and Cross Entropy.

Mode	KLD	Mean	KLD
$p = no\ mode$	----	$p = 0.5$	-----
$p = 0.4$	----	$p = 0.4167$	0.0141
$p = 0.35$	0.0054	$p = 0.3636$	0.0060
$p = 0.3667$	0.0006	$p = 0.375$	0.0003
$p = 0.30$	0.0102	$p = 0.3095$	0.0097

The Bayesian updates on the posterior can also be computed for a sequence of 120 updates of 10 trails each, and the updated posterior can be shown as follows:

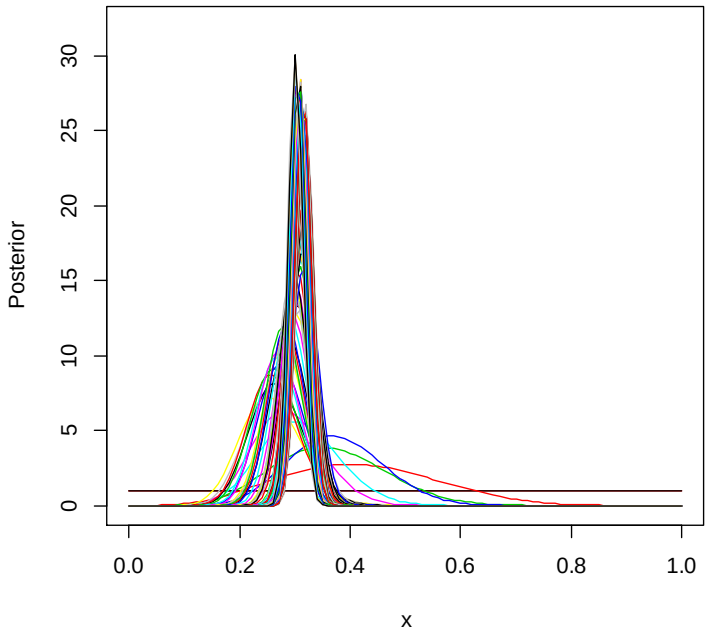

Figure 4: Sequence of Bayesian posteriors for 120 updates of 10 trials each.

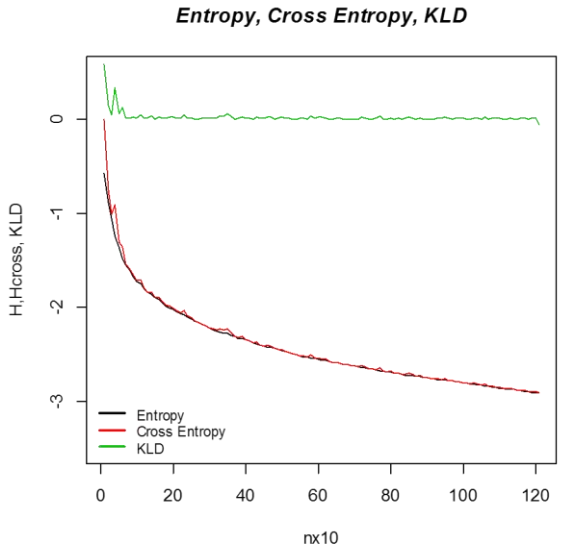

Figure 5: Relationship among Entropy, Cross Entropy, and KLD for 120 updates of 10 trials each.

2.5 Observed Fisher Information

Following the same kind of pairwise comparison among the different posteriors we can compute a dissimilarity matrix for the observed Fisher's information using the expression:

$$I\left(\hat{\theta}\right) = \kappa\left(\hat{\theta}\right) = -\left.\frac{d^2 \log L(\theta)}{d\theta^2}\right|_{\theta=\hat{\theta}}$$

Then, the corresponding dissimilarity matrix is:

$$FI(observed) = \begin{bmatrix} 0.00 & 41.67 & 87.91 & 129.19 & 190.48 \\ 41.67 & 0.00 & 47.62 & 87.91 & 153.41 \\ 87.91 & 47.62 & 0.000 & 41.67 & 106.67 \\ 129.19 & 87.91 & 41.67 & 0.000 & 111.11 \\ 190.48 & 153.41 & 106.67 & 111.11 & 0.000 \end{bmatrix}$$

3. Relationships Among Measurements

This section discusses the relationships among the most important measurements discussed before.

3.1 Relationships between Kullback-Liebler Divergence and Akaike Information Criteria.

Using the KLD between two distributions we can write the expression:

$$D_{KL}\left(f \| f^* | \theta\right) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{f^*(x|\theta)} dx$$

Akaike (1974) observed that:

$$D_{KL}\left(f \| f^* | \theta\right) = \int_{-\infty}^{\infty} f(x) \log f(x) dx - \int_{-\infty}^{\infty} f(x) \log f^*(x|\theta) dx$$

that can be written as:

$$D_{KL}\left(f \| f^* | \theta\right) = -H(f) - E\left[\log f^*(x|\theta)\right]$$

Since the entropy is free of parameters it can be ignored, and the minimization of the second term provides a basis for model comparison, this is not a measure of goodness of fit. As $n \to \infty$ with probability approaching 1, the model with the minimum AIC score will possess the smallest Kullback-Leibler divergence (Schmidt, 2008).

3.2 Deviance Information Criteria

Spiegelhalter, Best, Carlin and Van Der Linde (2001) developed a generalization of AIC known as Deviance Information Criteria (DIC) and showed that it is asymptotically equivalent to AIC.

3.3 Cross Entropy, KLD and Entropy

The AIC score is an asymptotically unbiased estimate of the cross-entropy (Schmidt, 2008), where the *Cross entropy = KL divergence +Entropy* (Murphy, 2012).

3.5 Entropy and Bayesian Statistics

- The maximum entropy principle (MEP) makes a claim of maximum ignorance, as the selected distribution is the one that makes the claim of least information.
- For the Bayesian case sometimes it can be represented as a uniform prior probability density.
-Jaynes claimed that Bayes theorem was a method to compute probabilities, while maximum entropy was a way to assign prior probability distributions (maximum entropy priors)

4. Conclusions

The conclusions of this study can be summarized as follows:

- The Bayesian philosophy is based on learning and information gain, and it depends completely on the prior and the posterior.
- Bayesian learning is proportional to the Kullback Leibler divergence (also called relative entropy)
- The KLD is larger in the direction of decreasing entropy
- The smaller steps for KLD in the direction of decreasing entropy, the closer we are to the correct model and the smaller learning that we have in every step
- The computational effort for KLD is greater than for computing AIC or DIC
- The entropy and cross entropy tend to be equal as Bayesian learning is achieved.
- The KLD tends to zero for smaller steps when Bayesian learning is achieved

References

Akaike H (1974). “A New Look at the Statistical Model Identification”, IEEE Transactions on Automatic Control, 19, 716-723.

Bernardo, J.; Smith, A. F. M. (1994). Bayesian Theory. John Wiley. ISBN 0-471-92416-4.

Boltzman, L. (1872), Neitere Studien uber das Warmegleichgewicht unter Gasmolekulen. K. Akad. Wiss. (Wein) Sitzb. 66, 275.

Casella G., Berger R. (1990). “Statistical Inference”, Duxbury Press, Belmont CA

Freund J,. Miller I., Miller M., (2014). “Mathematical Statistics with applications”, 8th Ed., Pearson, Boston, MA

Huzurbazar V.S. , (1949) “On A Property Of Distributions Admitting Sufficient Statistics” Biometrika (1949) 36 (1-2): 71-74 doi:10.1093/biomet/36.1-2.71

Jaynes, E. T. (1957). "Information Theory and Statistical Mechanics" (PDF). Physical Review. Series II. 106 (4): 620–630. Bibcode:1957PhRv..106..620J. MR 87305. doi:10.1103/PhysRev.106.620.

Jaynes, E. T. (1957). "Information Theory and Statistical Mechanics II" (PDF). Physical Review. Series II. 108 (2): 171–190. Bibcode:1957PhRv..108..171J. MR 96414. doi:10.1103/PhysRev.108.171.

Jeffreys, H., Theory of Probability, Oxford University Press; 3 ed.,ISBN-13: 978-0198503682

Mazzuchi, T.,(2006) “Bayes Estimate and Inference for Entropy and Information Index of Fit”, CiteSeerX, Pennsylvania State University

Schumitzky A., Tatarinova T.(2014), Nonlinear Mixture Models: A Bayesian Approach, Imperial College Press, (2014

Schmidt, D., Makalic, E. (2008), Model Selection Tutorial, Monash University, Melbourne Au.

Spiegelhalter, D.J., Best, N.G., Carlin B.P., and van der Linde, A. (2001), “Bayesian Measures of Model Complexity and Fit”, Journal of the Royal Statistical Society B, 38, 54-59.

Ullah A., Entropy, divergence and distance measures with econometrics applications, Journal of statistical planning and inference, 1996, 137-162