ENTROPY AND BAYESIAN LEARNING

Jose H. Guardiola, Texas A&M University Corpus Christi, Department of Mathematics and Statistics.

This paper discusses an example of Bayesian statistical inference and its relationship with entropy, 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 and provides a geometric interpretation for Riemannian distances and pseudo-distances. 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.

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.

1.1 Frequentist Methods and Measures of Goodness of Fit

1.1.1 Maximum Likelihood Estimation

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

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

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

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

in practice we minimize the log-likelihood function as:

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

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 :

$$\lambda(x) = \frac{\sup\limits_{\Theta_0} L(\theta \mid x)}{\sup\limits_{\Theta} L(\theta \mid x)}$$

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:

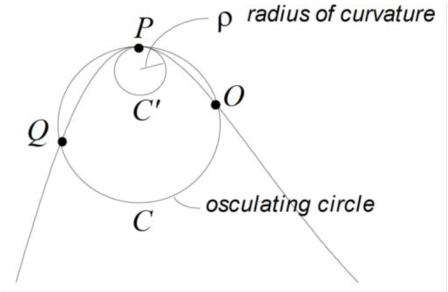
$$I(\theta) = -n \cdot E\left[\frac{\partial^2 \ln f(x)}{\partial \theta^2}\right] = n \cdot E\left[\left(\frac{\partial^2 \ln f(x)}{\partial \theta^2}\right)^2\right]$$

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:



The ratio of curvature can be expressed as:

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

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:

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

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

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:

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

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 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

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

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

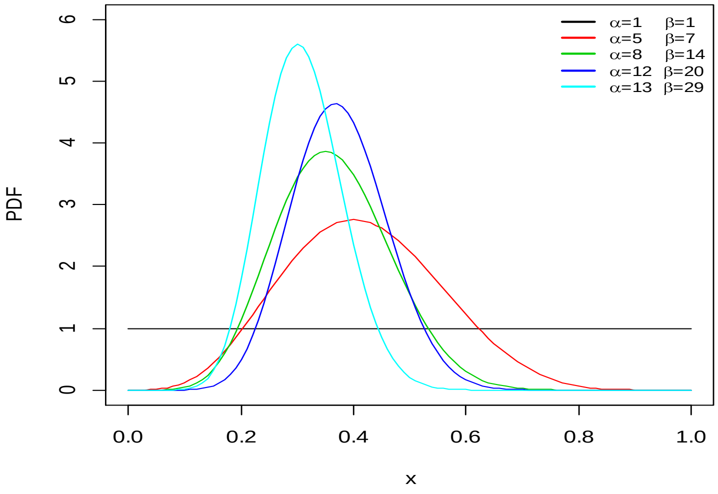


Figure 1. Sequence of Bayesian posteriors

2.2 Bayesian Factors for Two Hypotheses

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

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

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:

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

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:

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

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

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

where ψ is the digamma function:

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

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

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

Symmetry can be forced by adding the transpose to this matrix, *J-divergence = KLD+KLD'*:

$$Jdivergence = \begin{bmatrix} 0.000 & 2.835 & 6.585 & 9.808 & 15.082 \\ 2.835 & 0.000 & 0.385 & 0.679 & 2.165 \\ 6.585 & 0.385 & 0.000 & .095 & 0.606 \\ 9.808 & 0.679 & .095 & 0.000 & 0.741 \\ 15.082 & 2.165 & 0.606 & 0.741 & 0.000 \end{bmatrix}$$

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. RELATIONSHIP AMONG MEASUREMENTS

3.1 Relationship Among Fisher Information and KL Divergence

Using Taylor series we can express the KL divergence as:

$$D_{KL}\left(P_{\theta^*}\|P_\theta\right)=D_{KL}\left(P_{\theta^*}\|P_\theta\right)\Big|_{\theta^*}+\frac{\Delta\theta_i}{1!}\frac{\partial}{\partial\theta_i}D_{KL}\left(P_{\theta^*}\|P_\theta\right)\Big|_{\theta^*}+\frac{\Delta\theta_i\Delta\theta_j}{2!}\frac{\partial^2}{\partial\theta_i\partial\theta_j}D_{KL}\left(P_{\theta^*}\|P_\theta\right)\Big|_{\theta^*}+\ldots$$

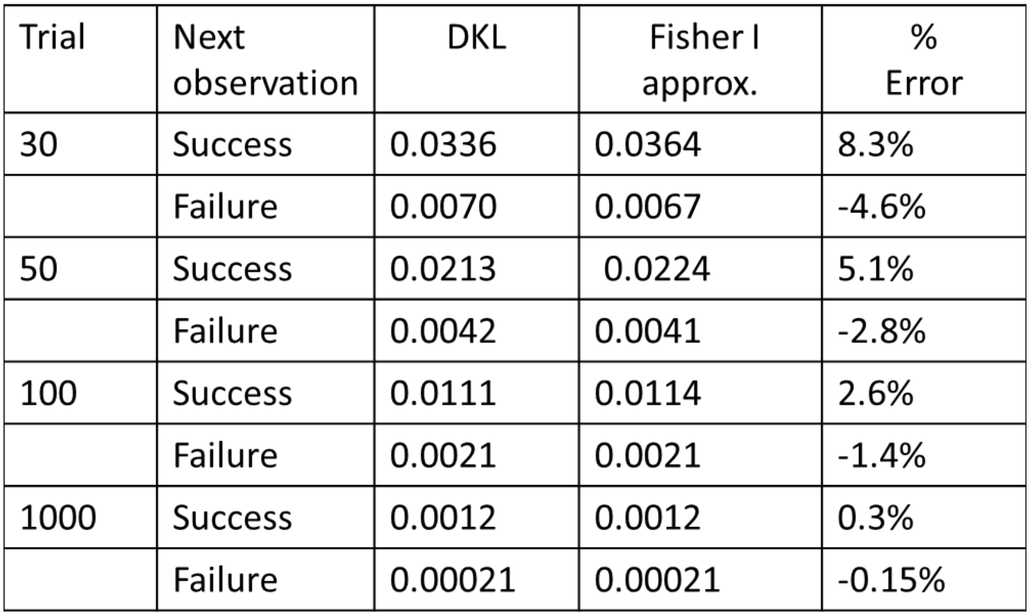
we can see that the first and second term vanishes at the maximum, and this result can be written as:

$$D_{KL}\left(P_{\theta^*}\|P_\theta\right)\approx\frac{\Delta\theta_i\Delta\theta_j}{2!}\frac{\partial^2}{\partial\theta_i\partial\theta_j}D_{KL}\left(P_{\theta^*}\|P_\theta\right)\Big|_{\theta^*}$$

where we can write an approximation that is valid when the two distribution are close:

$$D_{KL}\left(P_{\theta^*}\|P_\theta\right)\approx\frac{1}{2}\left(\theta-\theta^*\right)^t I\left(\theta^*\right)\left(\theta-\theta^*\right)$$

Therefore, we can see that the KL divergence can be approximated using the Fisher information. Next, we can summarize a numeric demonstration for this approximation:



Trial	Next observation	DKL	Fisher I approx.	% Error
30	Success	0.0336	0.0364	8.3%
	Failure	0.0070	0.0067	-4.6%
50	Success	0.0213	0.0224	5.1%
	Failure	0.0042	0.0041	-2.8%
100	Success	0.0111	0.0114	2.6%
	Failure	0.0021	0.0021	-1.4%
1000	Success	0.0012	0.0012	0.3%
	Failure	0.00021	0.00021	-0.15%

Table 1. KLD approximation using the Fisher information.

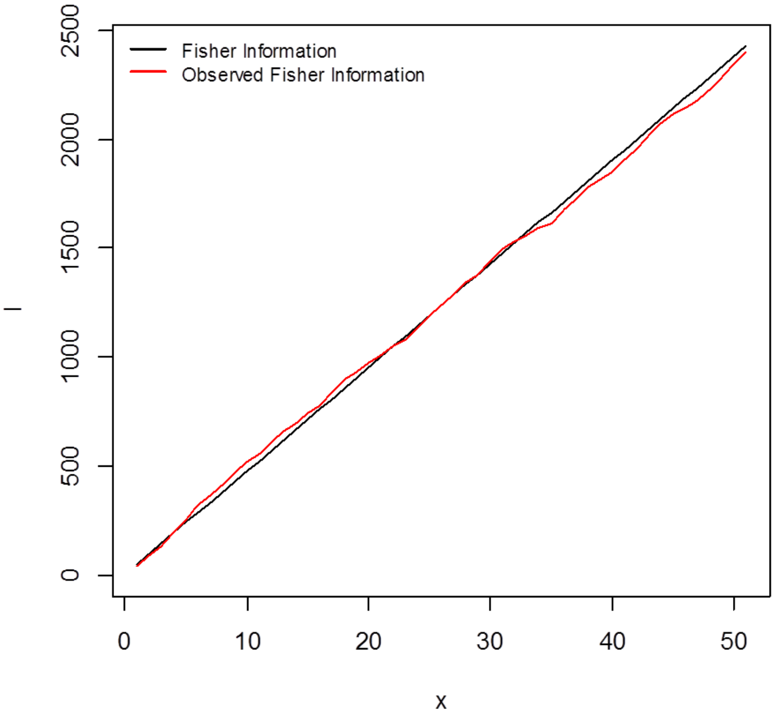


Figure 2. Fisher information and observed Fisher information for 50 updates.

3.2 Relationship Between KLD and Akaike Information Criteria

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

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

Akaike (1974) observed that:

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

that can be written as:

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

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 $n \rightarrow \infty$ with probability approaching 1, the model with the minimum AIC score will possess the smallest Kullback-Leibler divergence (Schmidt, 2008).

3.3 Deviance Information Criteria (DIC)

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.

3.4 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). In this case the cross entropy, the entropy and the AIC can be computed as:

Model	KLD	H(f)	Cross Entropy
α= 1, β=1	NA	0.000	NA
α= 5, β=7	0.580	-0.580	0.000
α= 8, β=14	0.144	-0.887	-0.744
α= 12, β=20	0.041	-1.058	-1.018
α= 13, β=29	0.326	-1.239	-0.914

Table 2. KLD, Entropy and Cross Entropy.

Developing a numeric example for 120 updates of the posterior.

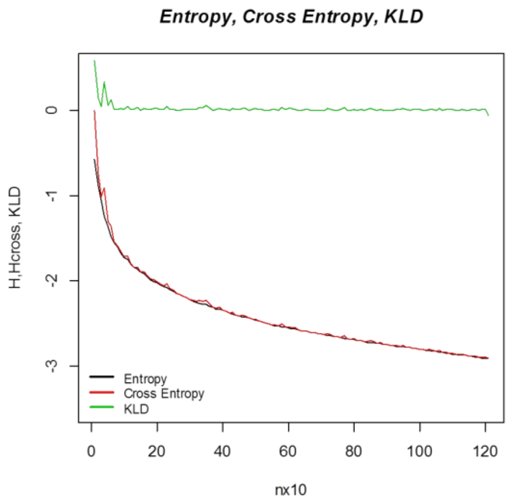


Figure 3. Entropy, Cross Entropy, and KLD.

4. DISSIMILARITY MATRICES

4.1 Multidimensional Scaling

Multidimensional scaling can be used for transforming the original distance matrix to two dimensions. The procedure for multidimensional scaling can be summarized as follows:

Set up the distance matrix D and apply double centering,

$$\mathbf{B} = -\frac{1}{2}\mathbf{JDJ} \qquad \text{where} \qquad \mathbf{J} = \mathbf{I} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{1'1}$$

then, extract the *m* largest eigenvalues (2 in this case) and

$$\mathbf{X} = \mathbf{E}_m \mathbf{\Lambda}_m^{1/2}$$

Where $\mathbf{E_m}$ is the matrix of m eigenvectors and $\mathbf{\Lambda_m}$ is the diagonal matrix of *m* eigenvalues of **B.**

Then for four Bayesian updates of the posterior (see Figure 1) we have the following sequence of Bayesian updates for the first coordinate.

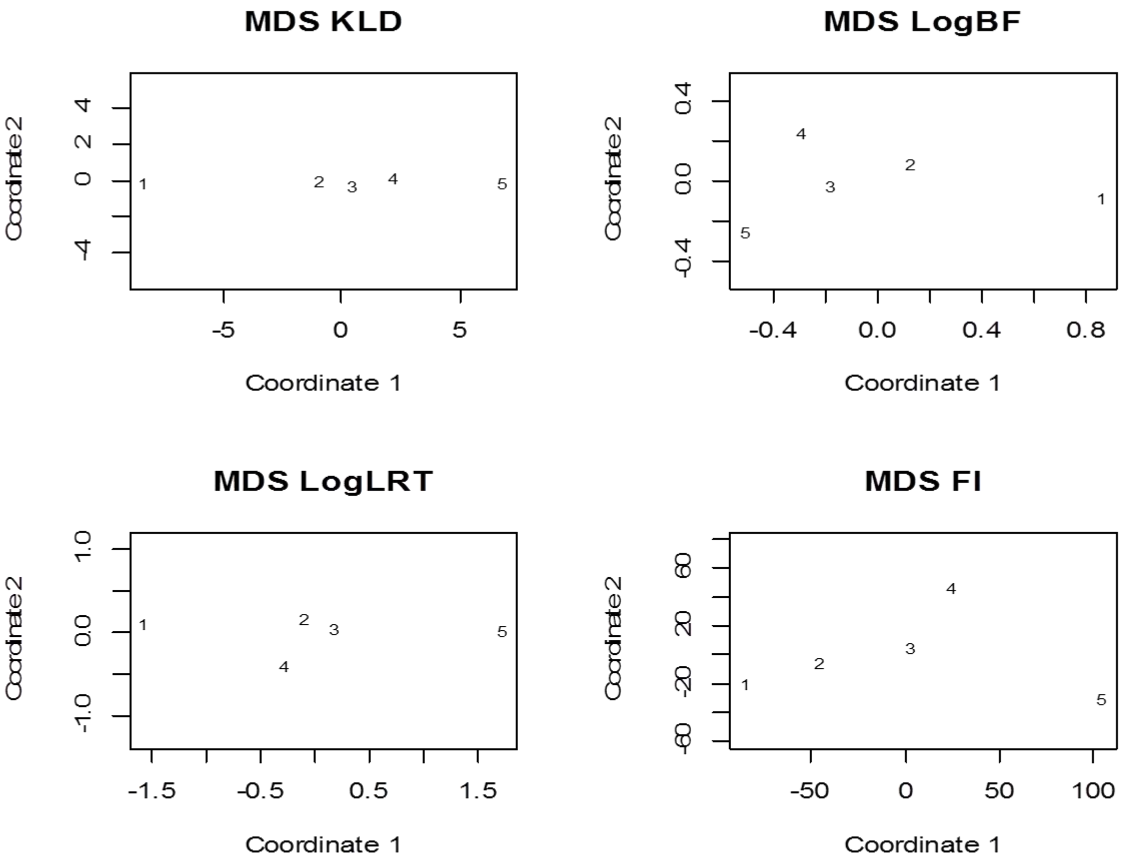


Figure 4. Bayesian updates sequence for the first coordinate of MDS.

Similarly, for 14 Bayesian updates of the posterior as shown in Figure 5.

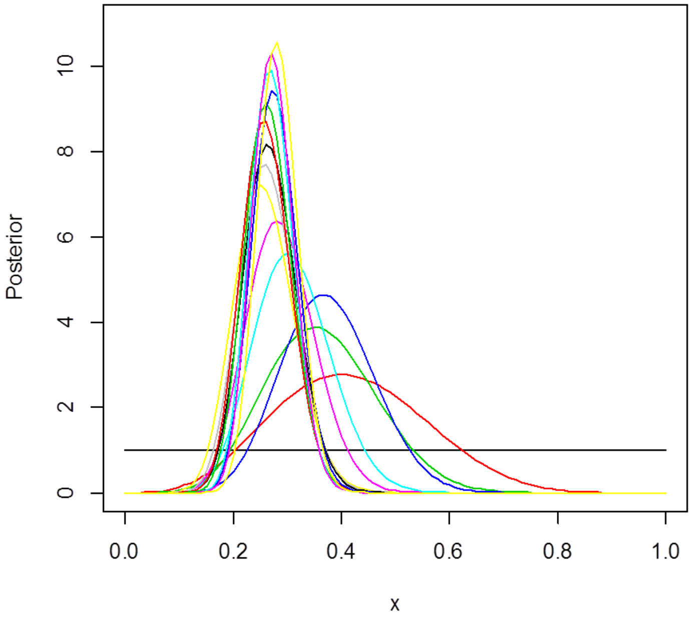


Figure 5. 14 Bayesian updates of the posterior.

Then, computing the corresponding MDS for the previous updates of the posterior, the change in the first two coordinates can be shown in two dimensions as:

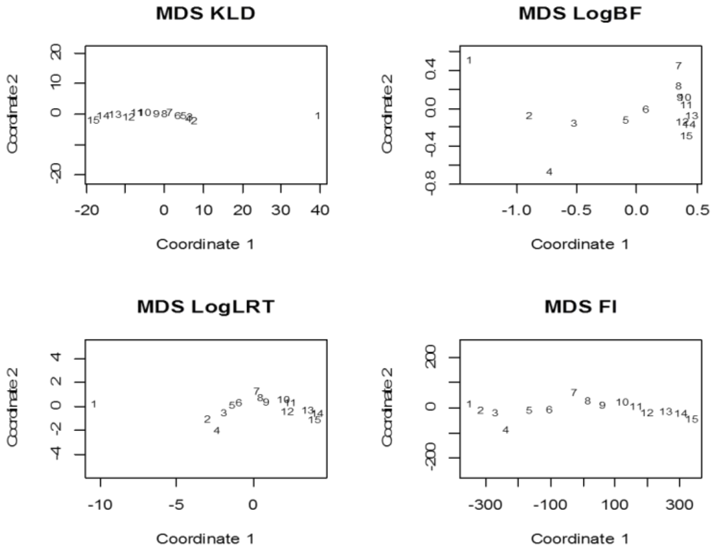


Figure 6. Bayesian updates sequence for the first coordinate of MDS using 14 iterations.

Similarly, for 29 Bayesian updates of the posterior as shown in Figure 6.

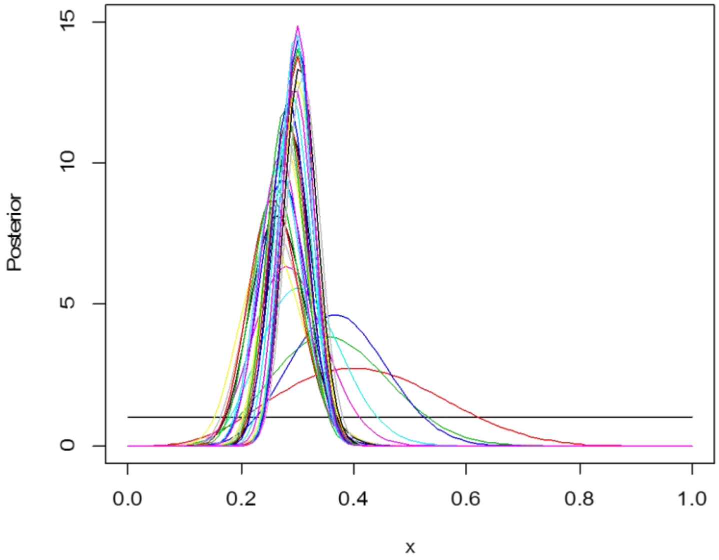


Figure 7. 29 Bayesian updates of the posterior.

Then, computing the corresponding MDS for the previous updates of the posterior, the change in the first two coordinates can be shown in two dimensions as:

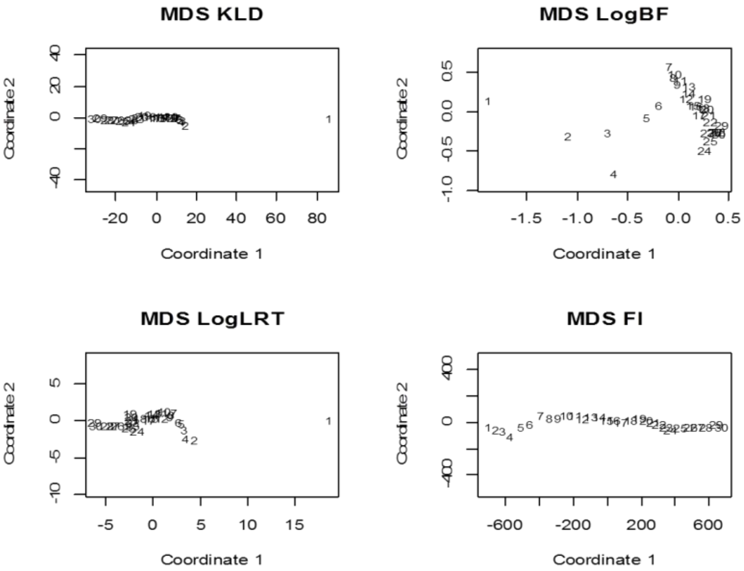


Figure 8. MDS for the first two coordinates of 29 Bayesian updates.

4.2 Correlations Among Measurements

A second approach that can be used here is to compute the correlations between first coordinates of the MDS, and then correlations for second coordinates of the MDS, and computing the correlation matrices among those measurements as follows:

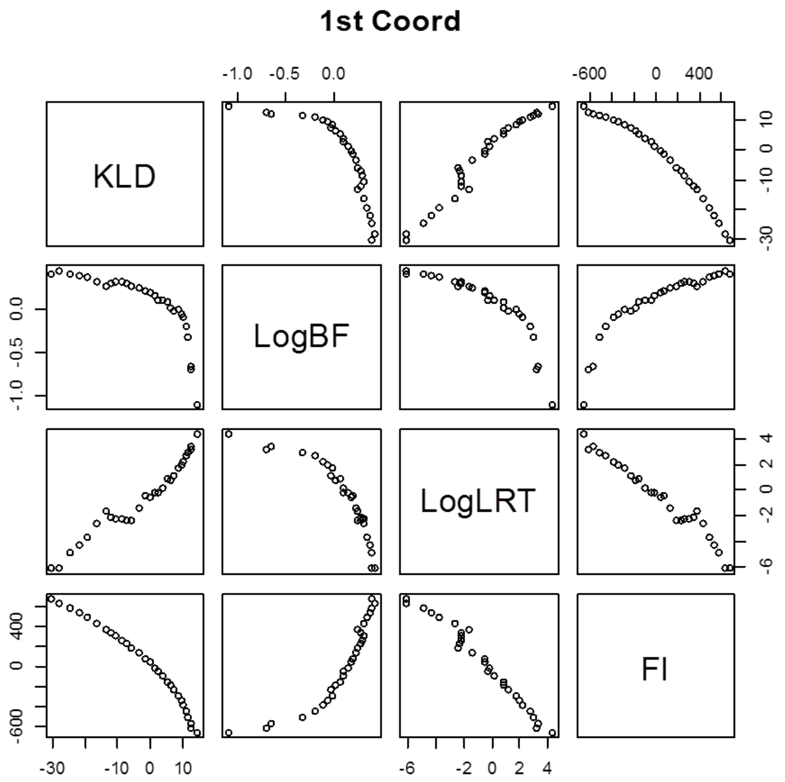


Figure 9. Correlations among first coordinates of multidimensional scaling.

And the corresponding correlation matrix, following the order; KLD, LogBF, LogLRT, FI:

$$\begin{bmatrix} 1.000 & -0.778 & 0.977 & -0.973 \\ -0.778 & 1.000 & -0.858 & 0.883 \\ 0.977 & -0.858 & 1.000 & -0.984 \\ -0.973 & 0.883 & -0.984 & 1.000 \end{bmatrix}$$

Similarly for the second coordinate of the MDS the following scatterplots can be developed:

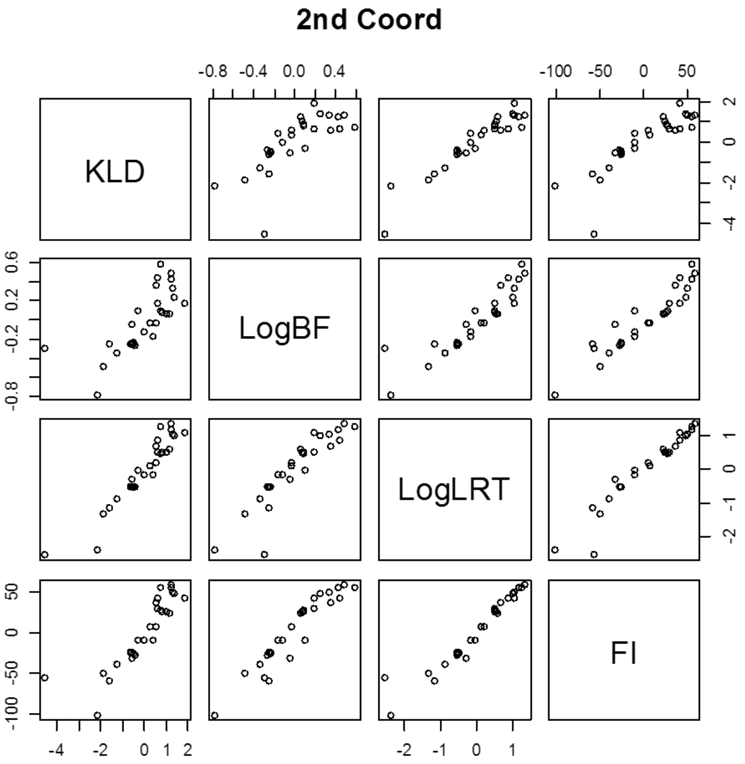


Figure 10. Correlations among second coordinates of multidimensional scaling.

And the corresponding correlation matrix for second coordinates of the MDS, following the order; KLD, LogBF, LogLRT, FI:

$$\begin{bmatrix} 1.000 & 0.742 & 0.947 & 0.864 \\ 0.742 & 1.000 & 0.909 & 0.945 \\ 0.947 & 0.909 & 1.000 & -0.984 \\ 0.864 & 0.945 & -0.984 & 1.000 \end{bmatrix}$$

4.3 Multidimensional Scaling Results

- KLD, Bayes factors and LRT show a large step for the first iteration and smaller steps for the following iterations
- KLD and FI showed a clear uni-directional sequence in Bayesian Learning
- All model measurements showed some non-linear behavior for correlations for the first coordinate and a more linear behavior for the second coordinate

4.4 Angles Between Dissimilarity Matrices

Since the Frobenius inner product ⟨A,⟩_F="tr"(A^T B) is a generalization of the dot product to matrices, the expression for the measure of the angle between two matrices A and B becomes:

$$\theta = \cos^{-1}\left(\frac{\operatorname{tr}(AB)}{\|A\|_F\|B\|_F}\right)$$

The last expression is a measure for the proximity of matrices in the *n*-dimensional space. Computing the angles between all dissimilarity matrices we have:

$$angles = \phi = \begin{bmatrix} 0 & 0.730 & 0.247 & 1.052 \\ 0.730 & 0 & 0.551 & 0.535 \\ 0.247 & 0.551 & 0 & 0.919 \\ 1.052 & 0.535 & 0.919 & 0 \end{bmatrix}$$

And the corresponding cosines or correlations for those angles:

$$\cos(\phi) = \begin{bmatrix} 0 & 0.745 & 0.970 & 0.496 \\ 0.745 & 0 & 0.852 & 0.860 \\ 0.970 & 0.852 & 0 & 0.607 \\ 0.496 & 0.860 & 0.607 & 0 \end{bmatrix}$$

From these results it can be seen that, the closest matrices in the *n*-dimensional space are between KLD and LRT with an angle of 0.247 and corresponding correlation of 0.97. The matrices that are more dissimilar are KLD and FI with an angle of 1.052 and corresponding correlation of 0.496.

5. CONCLUSIONS

The relationship among the measurements can be summarized on the following figure:

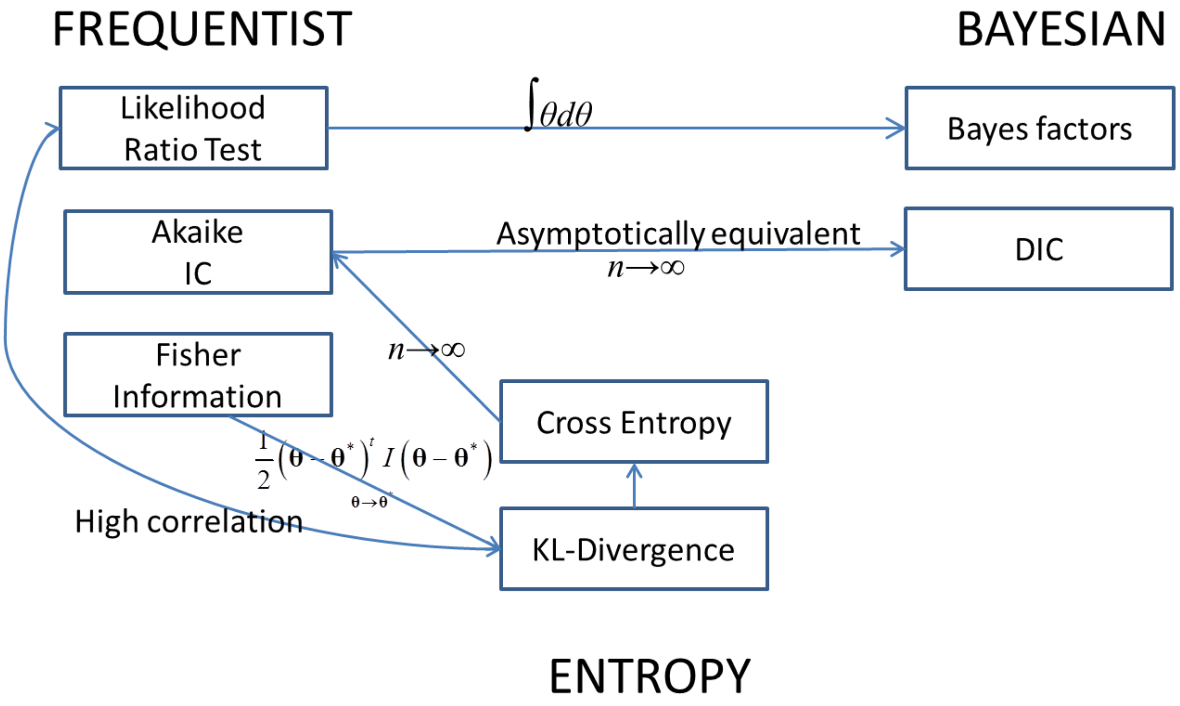


Figure 11. Relationships among measurements.

- Fisher Information despite its name is not considered a measure of information (in information theory) but rather its reciprocal defines a bound of the variance of an unbiased estimator (Cramer-Rao lower bound)
- All model selection measurements are correlated among them, the highest correlation was between KLD and LRT
- 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 is shown in every step
- The smaller angles between matrices in the *n*-dimensional space is between KLD and LRT and thus they showed the smaller corresponding correlation
- The computational effort for KLD is a lot greater than for computing the AIC or Fisher Information
- The entropy and cross entropy tend to be equal as Bayesian learning is achieved
- The KLD tends to zero for smaller steps between consecutive iterations as 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

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