

The Bhattacharyya Measure Requires No Bias Correction.

N.A.Thacker.

Last updated
20 / 11 / 2007

This document forms part of the **Statistics and Segmentation Series (2008-001)**
available from www.tina-vision.net.

- 2007-008 Tutorial: Defining Probability for Science.
- 2001-007 Performance Characterisation in Computer Vision:
The Role of Statistics in Testing and Design.
- 2002-007 The Effects of an Arcsin Square Root Transform on a Binomial Distributed Quantity.
- 2001-010 The Effects of a Square Root Transform on a Poisson Distributed Quantity.
- 2004-004 Shannon Entropy, Renyi Entropy, and Information.
- 2002-002 Validating MRI Field Homogeneity Correction Using Image Information Measures.
- 2004-001 Empirical Validation of Covariance Estimates for Mutual Information Coregistration.
- 2004-005 The Equal Variance Domain: Issues Surrounding the Use of Probability Densities in
Algorithm Design.
- 2009-008 Avoiding Zero and Infinity in Sample Based Algorithms.
- 2001-008 Derivation of the Renormalisation Formula for the Product of Uniform Probability
Distributions and Extension to Non-Integer Dimensionality.
- 2001-005 Model Selection and Convergence of the EM Algorithm.
- 2003-007 Noise Filtering and Testing for MR Using a Multi-Dimensional Partial Volume Model.
- 2002-004 A Novel Method for Non-Parametric Image Subtraction:
Identification of Enhancing Lesions in Multiple Sclerosis from MR Images.
- 2001-014 Bayesian and Non-Bayesian Probabilistic Models for Image Analysis.
- 1997-001 The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data.
- 1999-001 The Bhattacharyya Measure requires no Bias Correction.
- 1999-004 B-Fitting: An Estimation Technique With Automatic Parameter Selection.
- 2005-008 Tutorial: Beyond Likelihood.



Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

The Bhattacharyya Measure Requires No Bias Correction.

Abstract

This paper aims to show that the Bhattacharyya measure [2]

$$L_B = \int_a^b \sqrt{p(x|f)p(x|h)} dx$$

for comparing probability density distributions $p(x|f)$ and $p(x|h)$ has zero bias in the sense of the standard problems associated with maximum likelihood measures and estimated using the Akaike Information Criterion (AIC [1]). We start by reviewing the origins of this measure, when derived as the limit of a Likelihood measure for large samples. We then show how the assumptions used in the derivation of AIC are satisfied exactly for the L_B measure, which has a predicted bias of exactly zero¹.

1 Introduction

It is known that the square-root transform is the variance normalising transform for Poisson data [7]. This produces a close approximation to a Gaussian distribution for small sample sizes, which becomes exact for infinitely large samples [3]. In a recent paper we evaluated the suitability of using the variance normalising square-root transform as the basis of a measure for comparison of Poisson samples (such as histograms) [13]. Differences between transformed values can be interpreted as a Maximum Likelihood (effectively least squares) measure. The analysis shows that this space is thus a natural domain in which to construct a Euclidean distance measure as (unlike simple weighted differences between Poisson variables) minimum cost paths between locations are straight lines. This is the equal variance domain for such data [12]. The logical extension of this work is to apply the measure in the limit of an infinite number of samples, ie: a frequentist definition of probability density. In the process we regenerate the Bhattacharyya measure as a measure of similarity between two probability distributions. The Likelihood origins of this derivation imply the possibility that the measure might require correction due to bias (in the AIC sense [1]). In this paper we show that the AIC measure is indeed applicable to this measure but when evaluated is found to be zero for L_B . Thus showing that the Bhattacharyya measure is unbiased and can therefore be used for model selection, by comparing the degree of overlap between observed and predicted probability densities. This problem is directly related to the so called “bias-variance dilemma” in the area of artificial neural networks [5] and is covered in [11] with further examples in [8, 10].

2 Comparison of Probability Densities.

We define probability densities as a Poisson sample of data N_i distributed on some continuous variable x_i , in the infinite sample limit. A similarity measure for probability densities can be derived in the following manner. We start in the discrete space X_i of finite samples N_i . We can therefore construct a statistical similarity measure for two samples \mathbf{N} and \mathbf{M} with different but known total sample sizes

$$N_{tot} = \sum_i N(X_i) \quad \text{and} \quad M_{tot} = \sum_i M(X_i)$$

in the form of the probability of generating distribution M from distribution N given Poisson perturbations, as

$$\begin{aligned} \ln(P) &\approx \sum (\sqrt{M(X_i)} - \sqrt{N(X_i)})^2 \\ &= M_{tot} + N_{tot} - 2 \sum_i \sqrt{M(X_i)} \sqrt{N(X_i)}. \end{aligned}$$

Notice that this similarity measure is valid for all N_{tot} and M_{tot} , and that the line of the minimum-cost path takes us through distributions of varying normalisation. Also, the approximation to Likelihood improves with increasing sample sizes and is exact in the limit.

¹This bias is not related to the bias obtained when using the measure as an upper limit on the Bayes classification error between two distributions.

Now consider two distributions $D_N(X_i) = N(X_i)/N_{tot}$ and $D_M(X_i) = M(X_i)/M_{tot}$, in the absence of the normalisation terms. What can we say about the difference between the vectors of densities \mathbf{D}_N and \mathbf{D}_M ? We know from the above analysis that we strictly needed N_{tot} and M_{tot} in order to construct a valid statistical similarity, as the quantity of data in each sample defines the information content. However, consider the following measure:

$$\begin{aligned} L_D &= \sum_i (\sqrt{D_M(X_i)} - \sqrt{D_N(X_i)})^2 \\ &= 2 - 2 \sum_i \sqrt{D_M(X_i)} \sqrt{D_N(X_i)} \\ &= 2 - 2 \frac{\sum_i \sqrt{M(X_i)} \sqrt{N(X_i)}}{\sqrt{M_{tot} N_{tot}}}. \end{aligned}$$

We can see that it differs from the log probability expression only by offset and normalisation terms. We can therefore use this measure to compare densities, regarded as frequency ratios, for fixed but arbitrary normalisations of \mathbf{N} and \mathbf{M} . Although we have chosen to normalise to N_{tot} and M_{tot} , the same analysis would hold regardless of what we might use.

To apply this to probability densities, we observe that the above analysis is still valid when taking the limit $N_{tot} \rightarrow \infty$ and $M_{tot} \rightarrow \infty$. Now densities ratios become probabilities, so that

$$L_P = 2 - 2 \sum_i \sqrt{P(X_i|M)} \sqrt{P(X_i|N)}$$

is a valid way of comparing vectors of probabilities. Equally, as we approach a continuum limit $X_i \rightarrow x_i$ with $P(X_i) = p(x) \Delta X_i \rightarrow p(x) dx$, the ratios become densities, and for unit integral normalised probability densities we get the Matusita measure (L_M),

$$L_M = \int_a^b (\sqrt{p(x|M)} - \sqrt{p(x|N)})^2 dx$$

which is monotonically related to the Bhattacharyya measure (L_B),

$$= 2 - 2 \int_a^b \sqrt{p(x|M)} \sqrt{p(x|N)} dx = 2 - 2L_B$$

Notice that nothing in the above construction *required* that the integral of a *probability density* must be unity; this has only been done for convenience. Moreover, the freedom to be able to specify an interval a, b implies that there cannot be a unique normalisation, ie: the density is defined in the interval we choose to observe it. This interpretation of probability density is therefore consistent with the conventional observation. *Probability densities are not probabilities and do not obey the standard laws of probability.*

3 The Akaike Information Criterion (AIC)

Conventional approaches to parameter estimation are often developed from maximum likelihood. In particular many approaches are based on weighted least squares fitting;

$$\chi^2 = \sum_{i=1}^N (y_i - f(x_i, \theta))^2 / \sigma_i^2$$

where θ is an estimate of the vector of parameters for the function f and y_i is the data set with expected error σ . It is now well known that this function gives a biased result, in that as more model parameters are added the χ^2 will reduce, eventually to zero. Such a statistic can therefore not be used directly for model comparison and selection.

The equivalent probabilistic form of the χ^2 is written as follows;

$$\chi^2 = -2 \sum_{i=1}^N \log(p(x_i, \theta))$$

The limit of the bias is estimated directly as ;

$$q = \left\langle 2 \sum_{i=1}^N \log(p(x_i, \theta)) \right\rangle - \left\langle 2 \sum_{i=1}^N \log(p(x_i)) \right\rangle$$

where $p(x_i)$ is the true probability from the correct model and $\langle X \rangle$ denotes the expectation operation. We can expand this about the true solution θ_0 as;

$$q = \langle 2 \sum_{i=1}^N [\log(p(x_i, \theta_0)) + (\theta - \theta_0) \partial \log(p(x_i, \theta_0)) / \partial \theta + \frac{1}{2} (\theta - \theta_0)^T H(x_i, \theta_0) (\theta - \theta_0) + h.o.t] \rangle - \langle 2 \sum_{i=1}^N \log(p(x_i)) \rangle$$

where $H(x_i, \theta_0)$ is the Hessian of the log probability for a single data point. The second term has an expectation value of zero and excluding the higher orders the remaining terms can be re-written as;

$$q' = \langle 2 \sum_{i=1}^N \log(p(x_i, \theta_0)) - 2 \sum_{i=1}^N \log(p(x_i)) \rangle + \langle \sum_{i=1}^N (\theta - \theta_0)^T H(x_i, \theta_0) (\theta - \theta_0) \rangle$$

The first expectation term is $2n$ independent estimates of the Kullback-Liebler distance $L_{KL}(p, p_{\theta_0})$ and the second term can be re-written using the matrix *trace* identity such that

$$q' = 2nL_{KL}(p, p_{\theta_0}) + \text{trace}(\langle \sum_{i=1}^N H(x_i, \theta_0) (\theta - \theta_0) (\theta - \theta_0)^T \rangle)$$

which can be reduced further to

$$q' = 2nL_{KL}(p, p_{\theta_0}) + \text{trace}(\langle \sum_{i=1}^N H(x_i, \theta_0) \rangle \langle (\theta - \theta_0) (\theta - \theta_0)^T \rangle)$$

This result is now directly interpretable, as for the correct model the Kullback-Liebler distance is expected to be zero². The remaining term contains the information matrix for the data, more frequently used to approximate the inverse covariance of the parameters and the covariance on the parameters. For a well determined system we would expect the trace of the product of these matrices to be the rank of the parameter covariance. This is simply the number of model parameters k and leads to the standard form of the AIC measure used for model selection

$$AIC = \chi^2 + k$$

For badly determined parameters the information matrix may not be full rank and the trace will be the number of linearly independent parameters determined in the model with this data set.

In fact this derivation (based upon that presented in [15]) is over simplistic, and the more accurate estimate gives a correction term of $2k$ [4]. This difference does not however, invalidate what follows. For the benefit of the following sections it is worth pointing out one feature of a key assumptions in this derivation. This is; the quadratic expansion of the expectation will be exact for Gaussian distributed errors and a linear function model, or equivalently in the limit of very small errors.

4 Estimating bias of the Bhattacharyya Measure.

In this section we aim to show that the AIC measure is directly applicable to the Bhattacharyya measure and that the expected value of the bias is zero.

We start by considering the distance measure

$$L = 1 - \frac{(b-a)}{(4mN)} \sum_{i=1}^N \frac{(\sqrt{f_i} - \sqrt{h_i})^2}{\text{var}(\sqrt{f_i}) + \text{var}(\sqrt{h_i})}$$

²We are not claiming here that this measure is the correct way of performing this comparison, only that KL is zero for identical PDFs.

where f_i and h_i represent frequency measures from m samples in the range a to b of a quantized variable x . The term in the sum is a maximum likelihood estimator (least squares) which assumes that $\sqrt{f_i}$ and $\sqrt{h_i}$ have Gaussian distributions and from the argument above the expected bias is N as each frequency measurement also represents an independent model parameter. The results on estimation of bias using the Akaike approach presented in the previous section are therefore directly relevant and the total bias on L is given by;

$$q' = -(b-a)/(4m)$$

as we have N independent variables. In the limit that the number of samples m becomes infinite the estimated frequency ratios tend to conditional probabilities $P(i|f)$ and $P(i|h)$. Moreover, using the law of large numbers and error propagation it is clear that the variance terms become equal and constant [13]

$$\text{var}(\sqrt{f_i}) = \text{var}(\sqrt{h_i}) = 1/4$$

while in this limit the distributions for $\sqrt{f_i}$ and $\sqrt{h_i}$ become exactly Gaussian. In this limit we can now rewrite L as

$$L_p = 1 - \frac{(b-a)}{2N} \sum_{i=1}^N (\sqrt{P(i|f)} - \sqrt{P(i|h)})^2$$

The bias q' on L_p will clearly tend to zero in this limit. Thus we can use such measures as estimates of similarity between vectors of probability values without worrying about bias. In other words they are suitable for model selection.

Going now one step further and allowing $(b-a)/N$ to tend to zero, so that the sum becomes an integration we get.

$$L_B = 1 - \frac{1}{2} \int_a^b (\sqrt{p(x|f)} - \sqrt{p(x|h)})^2 dx$$

Where $p(x|f)$ and $p(x|h)$ are probability density functions. This again has zero bias. Rewriting this slightly by expanding the squared term and forming three separate integrations we get

$$L_B = \int_a^b \sqrt{p(x|f)} \sqrt{p(x|h)} dx$$

which is the Bhattacharyya measure, although this derivation from a likelihood measure is completely different to the original motivation [2].

It is interesting to note that on the way to generating this result, the limits ensure that all assumptions regarding the derivation of the AIC measure are satisfied exactly and $q' = q = 0$. Thus any measures for model comparison based directly on this or equivalent measures require no bias correction and can be used directly for model selection [8, 10, 14].

5 Discussion

There is nothing particularly special regarding the elimination of bias for the Bhattacharyya measure. Any probability similarity measure which can be formulated from a log-likelihood approach will have the same property, for example

$$L_{\chi^2} = \int_a^b \frac{(p(x|f) - p(x|h))^2}{(p(x|f) + p(x|h))} dx$$

which should be immediately recognisable as a generalisation of the conventional χ^2 measure, also has this property. Other probabilistic similarity measures can also be constructed which are symmetric (as required) under interchange of probability distributions. However, the Bhattacharyya measure and the related Matusita measure [6]

$$L_M = \int_a^b (\sqrt{p(x|f)} - \sqrt{p(x|h)})^2 dx$$

are the only ones which linearise the distance metric generated by the Poisson nature of the definition of probability densities. An analogous approach for the comparison of probability distributions computed using Bayes Theorem, derived for the frequentist definition as the limit of a Binomial sample, uses the *arcsin* transform and is given in [12].

References

- [1] H.Akaike, 'A new Look at Statistical Model Identification', IEEE Trans. on Automatic Control, **19**, 716, (1974).
- [2] A.Bhattacharyya. 'On a Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions',Bull, Calcutta Math Soc., **35**, 99 (1943).
- [3] N.A. Thacker and P.A. Bromiley, The Effects of a Square Root Transform on a Poisson Distributed Quantity. Tina memo, 2001-010, 2001.
- [4] J.E.Cavanaugh, Unifying the Derivations for Akaike and Corrected Akaike Information Criteria, Statistics and Probability Letters, 33, 201-208, 1997.
- [5] S.Geman, E.BienenStock and R.Doursat, 'Neural Networks and the Bias/Variance Dilemma', Neural Computation **4(1)**,1 (1992).
- [6] K.Fukunaga, 'Introduction to Statistical Pattern Recognition', 2ed, Academic Press, San Diego (1990).
- [7] K Ord and S Arnold. *Kendall's Advanced Theory of Statistics Volume 1: Distribution Theory*. Arnold, 1998.
- [8] A.J.Lacey, N.A.Thacker and N.L.Seed, 'Feature Tracking and Motion Classification Using a Switchable Model Kalman Filter.' Proc. BMVC, York, Sept. 1994.
- [9] J.Porrill, 'Fitting Ellipses and Predicting Confidence Envelopes using a Bias Corrected Kalman Filter.' Proc. 5th. Alvey Vision Conference, 175-185, Sept. 1989.
- [10] N.A.Thacker and J.E.W.Mayhew, 'Designing a Network for Context Sensitive Pattern Classification.' Neural Networks 3,3, 291-299, 1990.
- [11] N.A.Thacker, D.Prendergast and P.I.Rockett,'B-Fitting: A Statistical Estimation Technique with Automatic Parameter Selection.',Proc, BMVC, 283-292, Edinburgh, 1996.
- [12] N.A.Thacker, P.Bromiley, The Equal Variance Domain: Issues Surrounding the use of Probability Densities for Algorithm Construction. Tina memo, 2004-005, 2004.
- [13] N.A.Thacker, F.J.Aherne and P.I.Rockett, 'The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data',Kybernetika, 34, 4, 363-368, 1997.
- [14] D.J.C.MacKay, 'Bayesian Modelling and Neural Networks', Research Fellow Dissertation, Trinity College, Cambridge (1991).
- [15] B.D.Ripley, Appendix A in Pattern Recognition and Neural Networks, Cambridge University Press, 1996.