Tina Memo No. 1997-001
Presented at: TIPR'97, Prague 9-11 June, 1997 (prize paper).
and publiseed in Kybernetika, 34, 4, 363-368, 1997.

The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data.

N. A. Thacker, F. J. Aherne and P. I. Rockett

Last updated
1 / 12 / 1998

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.

1 Abstract

A recurring problem that arises throughout the sciences is that of deciding whether two statistical distributions differ or are consistent - currently the chi-squared statistic is the most commonly used technique for addressing this problem. This paper explains the drawbacks of the chi-squared statistic for comparing measurements over large distances in pattern space and suggests that the Bhattacharyya measure can avoid such difficulties. The original interpretation of the Bhattacharyya metric as a geometric similarity measure is reviewed and it is pointed out that this derivation is independent of the use of the Bhattacharyya measure as an upper bound on misclassification in a two-class problem. The affinity between the Bhattacharyya and Matusita measures is described and we show that the measure is applicable to any distribution of data. We explain that the Bhattacharyya measure is consistent with an assumption of a Poisson generation mechanism for individual measurements in a distribution which is applicable to a frequency (histogram) or probabilistic data set and suggest application of the Bhattacharyya measure to the field of system identification.

Keywords: Bhattacharyya, distance, Matusita, similarity, Poisson mechanism.

Nomenclature

Symbol	Definition
S	sample space of an experiment
$P(A)$	probability of event A where $A \subseteq S$
$X(a)$	observed value of random variable X for outcome $a \in S$
∇g	derivative of g
$P(B\|A)$	probability of observing event B given that event A has occurred
$f_A(x)$	probability density function of the random variable A
$N(\mu, \sigma^2)$	Normal or Gaussian distribution with mean μ and variance σ^2
$P(\omega_I\|x)$	a posteriori probability of class I given sample x has occurred

2 Introduction

Proving that two distributions differ or are consistent is a problem that arises in many areas of scientific research. When we wish to compare one data set to a known distribution, or to compare two equally unknown distributions, the most commonly used technique is the chi-squared test. In this paper we propose the Bhattacharyya metric as an alternative similarity measure and we demonstrate the advantages of this measure over the chi-squared method. There are various forms of the chi-squared statistic depending on whether we want to compare two unknown distributions, or to compare a known distribution with an unknown distribution. For example, suppose our data is binned, that N_i is the number of events observed in the ith bin and that n_i is the number expected according to some known distribution. The chi-squared statistic is:

$$\chi^2 = \sum_i (N_i - n_i)^2 / n_i$$

Similarly, for the case of comparing two unknown distributions, suppose R_i is the frequency-coded quantity contained in bin i for the first data set and S_i the frequency coded quantity for the same bin of the second data set. Here the chi-squared statistic is:

$$\chi^2 = \sum_i (R_i - S_i)^2 / (R_i + S_i)$$

The Euclidean distance term of the numerator determines the similarity and squaring ensures no cancelling occurs when summing over all bins. Also, as the mean and variance of a Poisson distribution are equal, the denominator contains the estimate of the variances of the binned data of unknown distribution thus normalising the comparison. Hence the above definitions assume the content of each bin to be a Poisson-distributed random variable. We can immediately see that singularities will arise from the chi-squared statistic whenever empty bins are compared. We will show that the Bhattacharyya measure has no such problem as it forces all Poisson errors to be constant, thus ensuring that the denominator can never take on a zero value. Also we show that the chi-squared statistic is not an accurate comparison measure over large distances in a statistical pattern space and that in such cases the

Bhattacharyya measure is a more meaningful measure. Further we show the Bhattacharyya measure to be: Self consistent, unbiased and applicable to any distribution. We also suggest the measure can be applied to the field of system identification.

3 Original Derivation of the Bhattacharyya Measure

Bhattacharyya's original interpretation of the measure was geometric [1]. He considered two multinomial populations each consisting of k categories classes with associated probabilities $p_1, p_2, ..., p_k$ and $p'_1, p'_2, ..., p'_k$ respectively. Then, as $\sum_i^k p_i = 1$ and $\sum_i^k p'_i = 1$, he noted that $(\sqrt{p}_1, ..., \sqrt{p}_k)$ and $(\sqrt{p}'_1, ..., \sqrt{p}'_k)$ could be considered as the direction cosines of two vectors in k-dimensional space referred to a system of orthogonal co-ordinate axes. As a measure of divergence between the two populations Bhattacharyya used the square of the angle between the two position vectors. If θ is the angle between the vectors then:

$$cos\theta = \sum_i^k \sqrt{p_i p'_i}$$

Thus if the two populations are identical:

$$cos\theta = 1$$

corresponding to $\theta = 0$, hence we see the intuitive motivation behind the definition as the vectors are co-linear. Bhattacharyya further showed that by passing to the limiting case a measure of divergence could be obtained between two populations defined in any way given that the two populations have the same number of variates.

4 The Bhattacharyya Bound

In this section we consider a two-class problem where each sample belongs to one of two mutually exclusive classes (the conditional density functions and the a priori probabilities are assumed known). The sample serves as input to a decision rule whereby we classify each sample to one of the two classes. In general, decision rules do not lead to perfect classification and in order to evaluate the performance of a decision rule we must calculate the probability of error - that is, the probability that the sample is assigned to the wrong class. If we define the a posteriori probability of class I given$x$ as $P(\omega_I|x)$ and similarly $P(\omega_{II}|x)$ for class II then the conditional error $r(x)$, given x, is either $P(\omega_I|x)$ or $P(\omega_{II}|x)$ (whichever is smaller), as described by Fukunaga [2]. That is:

$$r(x) = min[P(\omega_I|x), P(\omega_{II}|x)]$$

The total error, which is called the Bayes error is computed by the expectation of $r(x)$, $E[r(x)]$:

$$E[r(x)] = \int r(x)P(x)dx$$

where $P(x)$ is the probability of observing the pattern . An upper bound on the above integrand can be obtained by making use of the fact that $min[a, b] \leq a^s b^{1-s}$ for $0 \leq s \leq 1,\ a, b \geq 0$. This is commonly known as the Chernoff bound and taking the case of $s = 0.5$ gives the Bhattacharyya bound:

$$\int P(\omega_I|x)^{0.5} P(\omega_{II}|x)^{0.5} P(x)dx$$

or equivalently:

$$\int f_I(x)^{0.5} f_{II}^{0.5} dx$$

where $f_I(x) = P(\omega_I|x)P(x)$ and $f_{II} = P(\omega_{II}|x)P(x)$. Thus the Bhattacharyya bound integrates over all positions in the domain and assumes that the sample belongs to only one of the two classes. This assumption is a major restriction on the scope of the method as it should strictly only be applied to simple two class problems where this is known to be the case. It is therefore not an absolute similarity measure but rather a relative separation measure. Below we propose an alternative interpretation of the Bhattacharyya measure which has far wider potential for application.

5 Relationship Between Matusita and Bhattacharyya Distances

The Matusita [3] distance between two probability density functions is defined by [1]:

$$\int_{-\infty}^{\infty} (\sqrt{f}_A(x) - \sqrt{f}_B(x))^2 dx$$

and the Bhattacharyya distance by:

$$\int_{-\infty}^{\infty} \sqrt{f}_A(x)\sqrt{f}_B(x) dx$$

They are related as follows:

$$\int_{-\infty}^{\infty} (\sqrt{f}_A(x) - \sqrt{f}_B(x))^2 dx \;=\; 2 - 2\int_{-\infty}^{\infty} \sqrt{f}_A(x)\sqrt{f}_B(x) dx$$

thus minimising the Matusita distance is equivalent to maximising the Bhattacharyya distance. Matusita [3] originally noted this relationship and referred to the 'affinity' between the two measures.

6 Maximum Likelihood Derivation of the Chi-squared Statistic

In this section we obtain the chi-squared statistic using least squares as a maximum likelihood estimator, enabling us to produce a result which will be used later. Suppose we are fitting M data points (x_i, y_i) $[i = 1, ..., M]$ to a model that has P parameters a_j $[j = 1, ..., P]$. The model predicts a functional relationship between the measurements and a set of parameters :

$$y(x) \;=\; y(x; a_1, ..., a_P)$$

The dependence of the model on the set of parameters is given explicitly by the right-hand side form. Maximum likelihood estimation identifies the probability of the data given a particular set of parameters as the likelihood of the parameters given the data. This likelihood is then maximised with respect to the parameters, thus giving the set of parameters that was most likely to produce the observed data. Now suppose each data point yi has a measurement error that is independently random and distributed as a Gaussian around the "true" model $y(x)$ with standard deviation σ_i. The probability of observing the data set is the product of the probabilities of each point and is known as the likelihood, i.e.:

$$L \;=\; \prod_{i=1}^{M} exp(-(y_i - y(x_i; a_1...a_p))^2/2\sigma^2)/(\sqrt{2}\pi\sigma_i)$$

maximising the above is equivalent to maximising its logarithm or minimising the negative of its logarithm:

$$+0.5Mln(2\pi) \;+\; \sum_{i=1}^{M} ln\sigma_i \;+\; \sum_{i=1}^{M} (y_i - y(x_i; a_1...a_p))^2/2\sigma_i^2 \qquad (1)$$

Now the above is a chi-squared statistic but since M and $\sum_{i=1}^{M} ln\sigma_i$ are all constants, minimising this equation is equivalent to minimising:

$$\sum_{i=1}^{M} (y_i - y(x_i; a_1...a_p))^2/2\sigma_i^2$$

This is the more familiar form of the chi-squared statistic for comparing a known and unknown distribution assuming independent Gaussian errors. Frequency measures from sampled data are of course distributed in accordance with the Poisson distribution and in Appendix 1 we show how the Poisson distribution approximates a Gaussian for a large number of independent samples. Thus by substituting Poisson errors into the above we obtain the definition of the chi-squared for comparing known and unknown distributions as given in the introduction.

If we now substitute $M = 1$, $\sigma_1 \;=\; \sqrt{\sigma_A^2 + \sigma_b^2}$ $y_1 \;=\; \mu_A$ and $y(x_1; \mu, \sigma^2) \;=\; \mu_B$, the chi-squared statistic of (1) becomes:

$$+0.5ln(2\pi) + 0.5ln(\sigma_A^2 + \sigma_B^2) + (\mu_A - \mu_b)^2/2(\sigma_A^2 + \sigma_B^2)$$

(This is equivalent to comparing two points observations and from distributions respectively).This result will be used later.

[1] Thanks to Vasu Parameswaran for pointing out a long standing typographical error.

7 Drawbacks of Chi-squared Statistics

In the field of pattern recognition we often need to determine the similarity between two points observations in a high dimensional space. In such domains the nature of how the errors vary over the space can influence the shortest path between two points observations. For example consider the case of Poisson distributed measurements in two dimensions. Since the mean and variance of a Poisson distribution are equal the space is a region of smoothly changing variance as illustrated described by the ellipses of Figure 1. In this space the shortest distance between two observations points that are close to each other can be approximated by a straight line. However, for two distant observationspoints the shortest distance between them is not necessarily a straight line but a curved path as shown by the dotted line in Figure 2. This presents obvious difficulties when trying to construct a simple similarity measure in such a space. Moreover, if a chi-squared statistic was used in this example the Euclidean distance term in the numerator of the statistic implicitly assumes that a straight line distance is the shortest path between the observationspoints - in this case clearly it is not. Thus the chi-squared statistic is not meaningful when comparing distributions with Poisson errors over large distances.

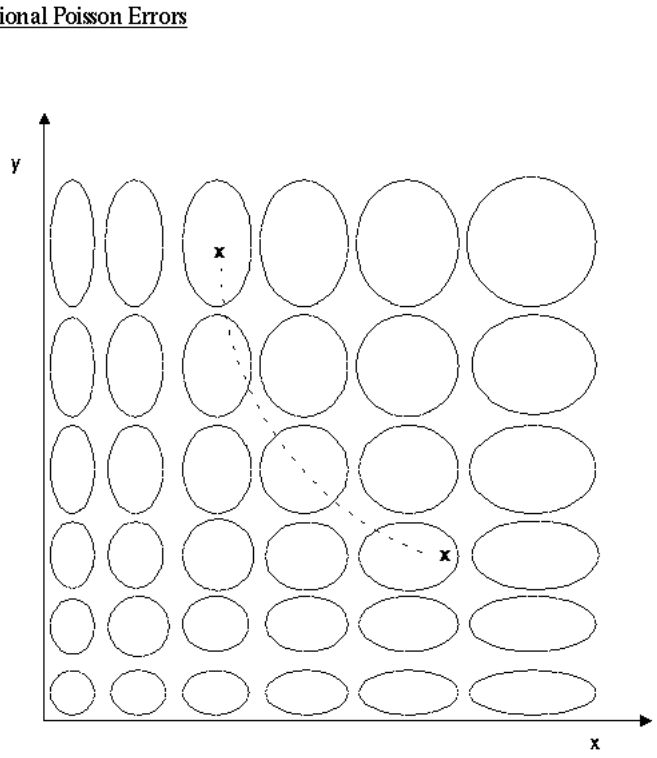

Figure 1: Illustration of the Shortest Distance Between Two Observations Points w ith Two-dimensional Poisson Errors

8 Advantage of the Bhattacharyya Measure

We now show that by using the Bhattacharyya measure all Poisson errors are forced to be constant therefore ensuring the minimum distance between two observationspoints is indeed a straight line. Suppose we have data $\{y_i \pm \delta y_i\}_{i=1}^{M}$ where δy_i represents measurement error (or standard deviation). Further suppose this error can be expressed as some function of y_i , i.e. $\delta y_i = f(y_i)$. Then we seek a function g that maps all errors to a constant, that is for all i. We can approximate the derivative of g by:

$$\nabla g \approx \delta g/\delta y \Rightarrow \delta g = \nabla g \delta y$$

therefore

$$k \approx \nabla g \delta y \Rightarrow \nabla g = k/\delta y$$

therefore

$$g \approx k \int dy/f(y)$$

Now for a Poisson distribution with mean y the variance is also y and the standard deviation, thus for Poisson measurements we have $f(y) = \sqrt{y}$. Hence:

$$g = k \int dy/\sqrt{y}$$

$$g = k\sqrt{y} + C$$

So our solution set is the family of square-root functions. We can select $k = 1$ and the boundary conditions tell us that $C = 0$ thus $g = \sqrt{y}$. So we have shown that for a measurement domain of Poisson variables the square-root function of the Bhattacharyya measure transforms into a domain where all errors are constant. Moreover, by mapping to a domain where all errors to be are constant we have avoided the problem of evaluating the minimum of a curved path integral by ensuring that a straight line measure is always the minimum distance between two observationspoints, as shown in Figure 2.

Application of the Bhattacharyya Measure

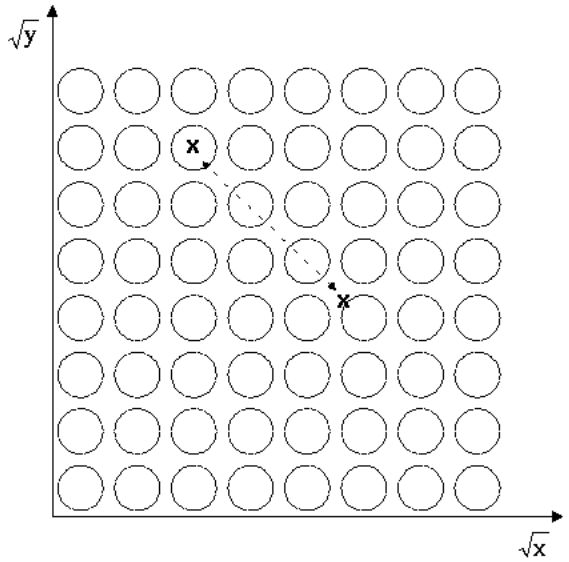

Figure 2: Illustration of the Shortest Distance Between Two Observations Points After Application of the Bhattacharyya Measure

As the Bhattacharyya measure is equivalent to the Matusita measure we see that the Bhattacharyya metric is a chi-squared type statistic in the sense of squared Euclidean distance, although unlike the chi-squared statistic, the Bhattacharyya metric measures similarity in a domain where all errors are constant. We can show how the Bhattacharyya metric approximates the chi-squared for small distances as follows. For an arbitrary function, f acting on binned, frequency coded data of unknown distribution we can approximate the chi-squared as:

$$\chi^2 = \sum_i (R_i - S_i)^2/(R_i + S_i) \approx \sum_i (f(R_i) - f(S_i))^2/((\partial f/\partial R_i)^2 R_i + (\partial f/\partial S_i)^2 S_i)$$

The denominator of the right-hand side results from error propagation as:

$$var(f(R_i) - f(S_i)) = var(f(R_i)) + var(f(S_i))$$

$$\approx var(constant + R_i(\partial f/\partial R_i)) + var(constant + S_i(\partial f/\partial S_i))$$

$$(\partial f/\partial R_i)^2 var\ R_i + (\partial f/\partial S_i)^2 var\ S_i$$

(since the variance of a constant term is zero). Now substituting into the above gives:

$$\sum_i (\sqrt{R_i} - \sqrt{S_i})^2/((1/\sqrt{4R_i})^2 R_i + (1/\sqrt{4S_i})^2 S_i)$$

$$= 2\sum_i (\sqrt{R_i} - \sqrt{S_i})^2$$

which is a scaled Matusita distance measure and rearranging this gives:

$$2\sum_i(\sqrt{R_i}-\sqrt{S_i})^2 \;=\; constant - 4\sum_i \sqrt{R_i}\sqrt{S_i}$$

Thus the Bhattacharyya measure approximates the chi-squared similarity function. It should also be noted that by transforming all variances to be constant the Bhattacharyya measure avoids the singularity problem of the chi-squared statistic when comparing empty bins.

9 Bhattacharyya Measure Applied to Histograms

The Bhattacharyya measure can be used to compare the similarity between two histograms as follows: If we let R_i be the frequency coded quantity in bin i (normalised such that $\sum_i R_i \;=\; 1$) for the first histogram and S_i a similar quantity for the second histogram. Then we can assume R_i to be a Poisson distributed random variable and similarly for S_i. We propose the Bhattacharyya statistic $\sum_i \sqrt{R_i}\sqrt{S_i}$ as a measure of similarity between the two histograms. For the case of two identical histograms we obtain $\sum_i R_i \;=\; 1$ indicating a perfect match.

10 Self-Consistency Property of the Bhattacharyya Measure

We have derived the Bhattacharyya measure for arbitrary frequency distributions and can therefore use the measure to compare probability distributions assuming that each probability estimate is the limiting estimate of a Poisson distributed value. This measure should be applicable to any data distribution and it is instructive to apply the measure to the Poisson distribution itself. If R_i , the observed quantity of bin i for the first histogram, has a Poisson distribution with mean r_i the probability of observing R_i in bin i is:

$$P(R_i) \;=\; exp(-r_i)r_i^{R_i}/R_i!$$

If we assume similar properties for the second histogram where S_i is the quantity in bin i and that this has Poisson distribution with mean s_i we can apply the Bhattacharyya measure over all values to give:

$$\sum_{x=0}^{\infty}(exp(-r_i)r_i^{x_i}/x_i!)^{0.5}(exp(-s_i)s_i^{x_i}/x_i!)^{0.5}$$

taking the natural logarithm (any monotonically increasing function would be legitimate) gives:

$$\begin{aligned} &-r_i/2 \;-\; s_i/2 \;+\; \sqrt{r_i s_i} \\ &= -(\sqrt{r}_i - \sqrt{s}_i)^2 \end{aligned}$$

So we have applied the Bhattacharyya measure to Poissonly distributed data and manipulated to give the Matusita measure. As the Matusita measure leads us to the Bhattacharyya metric we have shown a self-consistency property of the measure. Most importantly this implies that the similarity measure statistic is unbiased.

11 Independence Between Bhattacharyya Measure and Bin Widths

Given that we are to use a similarity measure of the form $\sum_i g_i g_i'$ as a basis for comparing histograms, we now show that for the Bhattacharyya metric the contribution to the measure is the same irrespective of how the quantities are divided between bins. Suppose that the quantity contained in bin i for the first histogram can be described by a function of some variable R , i.e. $g_i \;=\; f(R)$. The contribution to the similarity measure attributed by bin i is given by $g_i g_i'$. Now imagine that this quantity is translated so that it falls across the boundary of two bins j and k with the relative significance of each component being a and b (such that$a + b = 1$) respectively then:

$$g_j \;=\; f(R_0 a) \quad and \quad g_k \;=\; f(R_0 b)$$

and where R_0 denotes a fixed bin quantity. In order that the contributions to the similarity measure are equal irrespective of how the quantity has been apportioned between the bins we require:

$$f(R_0)g_i' \;=\; f(R_0 a)g_j' \;+\; f(R_0 b)g_k'$$

As the best possible match we could obtain would be that of two identical histograms, in this case we can write that we require:

$$f^2(R_0) \;=\; f^2(R_0 a) \;+\; f^2(R_0 b)$$

Since $f^2(R)$ must be linear the only choice of the function we can make is:

$$f(R) \;=\; KR^{0.5}$$

which gives the Bhattacharyya measure. Therefore the Bhattacharyya statistic is unaffected by the distribution of data across the histogram and is the only form of sum-of-product functions with this property. It is interesting to note that the chi-squared statistic performs equally well if R_i is split into two bins containing aR_i and bR_i respectively. This is because the original chi-squared statistic:

$$(R_i - S_i)^2/(R_i + S_i)$$

becomes:

$$\begin{aligned}
&(aR_i - aS_i)^2/(aR_i + aS_i) \;+\; (bR_i - bS_i)^2/(bR_i + bS_i) \\
&= \; a(R_i - S_i)^2/(R_i + S_i) \;+\; b(R_i - S_i)^2/(R_i + S_i) \\
&= \; (R_i - S_i)^2/(R_i + S_i)
\end{aligned}$$

12 Bhattacharyya Measure Applied to Univariate Gaussian Distributions

In this section we show how the Bhattacharyya measure behaves in comparison with a chi-squared statistic for the case of univariate Gaussian distributions with different means and variances.

For example consider two univariate Gaussian probability density functions:

$$\begin{aligned}
f_A(x) \;&=\; exp(-(x - \mu_A)^2/a\sigma_A^2)/\sqrt{2\pi\sigma_A} \\
f_B(x) \;&=\; exp(-(x - \mu_B)^2/a\sigma_B^2)/\sqrt{2\pi\sigma_B}
\end{aligned}$$

Consider the integral below as a general similarity measure:

$$\int_{-\infty}^{\infty} (f_A(x))^n (f_B(x))^n dx$$

This has a solution given by:

$$exp(-n(\mu_A - \mu_B)^2/2(\sigma_A^2 + \sigma_B^2))\sqrt{2\pi(\sigma_A^2\sigma_B^2)}/(n(2\pi\sigma_A\sigma_B)^n\sqrt{\sigma_A^2 + \sigma_B^2}))$$

and therefore the case of n = 0.5 gives a solution:

$$exp(-(\mu_A - \mu_B)^2/4(\sigma_A^2 + \sigma_B^2))\sqrt{2\sigma_A\sigma_B}/\sqrt{\sigma_A^2 + \sigma_B^2} \qquad (3)$$

and n = 1 gives the solution:

$$exp(-(\mu_A - \mu_B)^2/2(\sigma_A^2 + \sigma_B^2))\sqrt{2\pi(\sigma_A^2 + \sigma_B^2)} \qquad (4)$$

Now by comparing the negative natural logarithm of (4) with (2) we see that (4) represents a chi-squared statistic. By considering the dimension of the constantnormalisation term in (4) we see that this quantity is not dimensionless, thus the statistic will depend upon the measurement scale used. Moreover for this statistic to have a value of unity we require both, $\mu_A \;=\; \mu_B \;\; \sigma_A^2 + \sigma_B^2 = 1/(2\pi)$. As there are many solutions to the second constraint many equivalent measures can be obtained by using different variance quantities; and thus the chi-squared measure does not draw a good comparison between the two distributions but simply their means.

In contrast the constant term of the Bhattacharyya measure in (3) is dimensionless therefore the measure is not affected by the measurement scale used. An overlap integral function with $n = 0.5$ is the only one with this property. Further when Bhattacharyya is used to compare two identical distributions we see that the term is maximised to a value of one. Thus the Bhattacharyya similarity measure is superior to the chi-squared statistic in the sense that it can be used to compare two distribution shapes rather than just their means.

13 Relevance to System Identification

A commonly encountered problem in the field of system identification is that of fitting models to a given set of data. Here the problem we address is that of how to make the best model selection. A method often applied to the problem involves fitting the model whose prediction is closest to the measurement, however, more complex models having more free parameters, are able to predict any one data set as likely as a simpler model. Therefore we need to consider the complexity of the model, preferring the simplest model capable of accurately predicting the data in order to ensure we have the most probable hypotheses for the data. There exist many algorithms for estimating the parameters of a model of known order from noise-contaminated input-output data. In practice the model order is rarely known a priori.

The Akaike [4] Information Criteria (AIC) is a common method used for system identification. In practice it requires the estimation of the parameters of models of different order using maximum likelihood methods and then selecting the model which minimises:

$$-2ln(L) + 2p'$$

where L is the likelihood function of section 2 and p' is the number of independent parameters in the model. The AIC embodies Occam's razor since if two models are equally likely the one with the fewer parameters is chosen. The AIC test is most commonly formulated as a modification to the standard χ^2 test statistic (there are several other similar statistics). The second term effectively compensates for the systematic underestimate of the χ^2 test statistic due to a finite sample of data. However there are major limitations to the AIC test:

(1) According to Norton [5], a major limitation of the AIC test in this form arises from its assumption of Gaussian errors. We have shown the Bhattacharyya measure to have no such assumption.

(2) Another major drawback of the AIC method is that it is independent of the distribution of the data and clearly a method incorporating the data distribution, such as the Bhattacharyya measure, would be superior.

(3) Also the AIC commonly assumes the number of independent parameters to be constant and known a-priori. This is not always the case as the number of independent parameters of the model is data and model parameter dependent.

The Matusita and Bhattacharyya measures are themselves a form of test statistic but because of their construction they require no systematic error correction. The use of these measures for system identification is thus perfectly consistent with current practice.

The question we should be asking when selecting a model is: Which model will give the best prediction of unseen data based on the current data? As one example, the Kalman filter is often used to predict the value of the next data point given all the information up to that instant and can be regarded as a least-squares estimator. Many researchers have used switchable Kalman filters in order to select the most appropriate form from a set of models but we believe that the prediction power of the Kalman filter has as much to do with the width of the distribution of the prediction of the filter as the central value (i.e. the mean). and have suggested Iin previous work[6] we have suggested the Bhattacharyya measure as the appropriate selection statistic. We emphasise this point in Figure 3 for the analogous problem of function interpolation. More accurate data justifies more complex model descriptions (Occam's razor).

14 Summary

In this work we have presented the original geometric interpretation of the Bhattacharyya similarity measure. A derivation of the chi-squared statistic by maximum likelihood estimation has been given and the shortcomings of this statistic has been explained when applied over large distances in pattern space. We have shown that the Bhattacharyya measure is applicable to any data set irrespective of the distribution from which the data is sampled; moreover we have shown the measure to have all the properties that could be expected of a principled probabilistic similarity function including self-consistency and lack of bias. Arguments favouring the measure have been described and we have published previous work on the successful practical application of the measure : Lacey, Thacker and Seed [6] , Evans, Thacker and Mayhew [7] and finally Thacker, Abraham and Courtney [8]. We also describe the suitability of the measure to the field of system identification where we describe the link between prediction power and errors on data measurements used to define the model. The measure takes direct account of this correlation and can be seen to be consistent with standard approaches.

Many researchers have used the Bhattacharyya similarity measure and found it advantageous. Until now the Bhattacharyya measure has been utilised by many as the result of a trial-and-error process with little understanding of why the measure works well. This work has demonstrated the reasons why the Bhattacharyya similarity measure

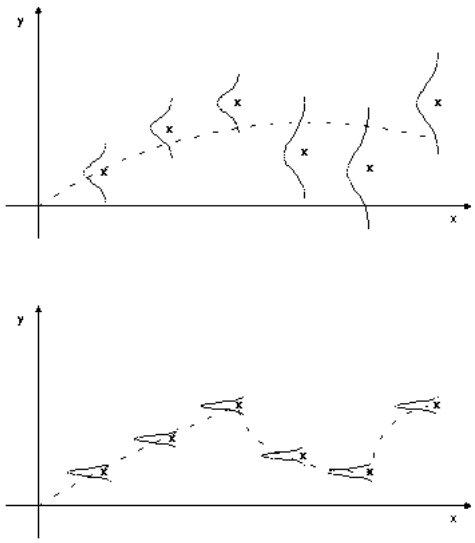

Figure 3: Diagrams Illustrating Effects of Data Accuracy on Function Complexity

should be used as an absolute similarity metric. The origin of the measure iswas independent of an upper bound on the Bayes error and we emphasise its use should not be confined by such a limiting derivation. In further work we intend to examine the robustness of the Bhattacharyya measure by empirical methods and its application to problems of automatic model selection in the field of neural networks.

Acknowledgements

F. J. Aherne would like to thank the Engineering and Physical Science Research Council (EPSRC) of the U.K. for providing a studentship. P.I. Rockett acknowledges financial support given by CEC Copernicus and N. A. Thacker would also like to thank the EPSRC (grant number GR-J10464).

Appendix 1 : The Normal Approximation to the Poisson

Suppose we have one model histogram and H candidate histograms from which we want to select the best match to the model. Then for any histogram the number of elements in bin i can be considered as a random variable having a Poisson distribution with mean r_i . We can take R_i as our estimate of r_i (the mean of the distribution) where R_i is the observed quantity in bin i. Then assuming that the bin counts are independent random variables, $X_{1i}, ..., X_{Hi}$,

across our set of histograms we can take the limiting case by letting H tend to infinity. By the additivity property of the Poisson distribution:

$$\sum_{j=1}^{H} X_{ji} \sim Po(\sum_{j=1}^{H} R_{ji})$$

By the Central Limit Theorem, the shape of this distribution tends to a Gaussian when H is large, i.e.:

$$Po(\sum_{j=1}^{H} R_{ji}) \approx N(\sum_{j=1}^{H} R_{ji}, \sum_{j=1}^{H} R_{ji})$$

Therefore in the limiting case the probabilities can be shown to have Gaussian distributions thus the probability distributions should be comparable to using a chi-squared statistic.

References

[1] A. BHATTACHARYYA, On a Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions, Bull. Calcutta Math. Soc., 35, pp. 99-110, 1943.

[2] K. FUKUNAGA, Introduction to Statistical Pattern Recognition, Second Edition, Academic Press Inc., 1990.

[3] K. MATUSITA, Decision Rules Based on Distance for Problems of Fit, Two Samples and Estimation, Ann. Mathematical Statistics, Vol. 26, pp.631-641, 1955.

[4] H. AKAIKE, A New Look at the Statistical Model Identification, IEEE Trans. on Automatic Control, Vol. 19, pp. 716-723, December 1974.

[5] J. P. NORTON, An Introduction to Identification, Academic Press, 1986.

[6] A. J. LACEY, N. A THACKER and N. L. SEED, Feature Tracking and Motion Classification Using a Switchable Model Kalman Filter, Proc. BMVC, York, September 1994.

[7] A. C. EVANS, N. A. THACKER, J. E. W. MAYHEW, The Use of Geometric Histograms for Model-Based Object Recognition, Proc. 4th BMVC, Guildford, pp.429-438, September 1993.

[8] N. A. THACKER, I. ABRAHAM and P. G. COURTNEY, Supervised Learning Extensions to the CLAM Network, Submitted to the Neural Network Journal 1994.