

The Equal Variance Domain: Issues Surrounding the Use of Probability Densities in Algorithm Design

Neil A. Thacker, Paul A. Bromiley and Emanuele Trucco

Last updated
22 / 6 / 2010

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 Equal Variance Domain: Issues Surrounding the Use of Probability Densities in Algorithm Design

Neil A. Thacker^a, Paul A. Bromiley^a and Emanuele Trucco^b
a Division of Imaging Science and Biomedical Engineering,
School of Medicine, University of Manchester, M13 9PT, U.K.
`neil.thacker;paul.bromiley@manchester.ac.uk`
b Vision, Image and Signal Processing Group,
School of Engineering and Physical Sciences,
Heriot Watt University, Riccarton, Edinburgh, EH14 4AS, UK.
`e.trucco@hw.ac.uk`

Abstract

We discuss various statistical procedures, including the construction of likelihood functions, entropy or information measures, probability similarity measures, and hypothesis tests, which are frequently used as the basis for machine vision algorithms. In each case we demonstrate that incautious transformation between the discrete and continuous cases when using these measures can introduce problems such as arbitrary scalings, and therefore bias. We also demonstrate that, by working in a homoscedastic space, the *equal variance domain*, many of these problems can be avoided. We conclude with some observations on the need for statistical rigour in the construction of machine vision algorithms.

1 Introduction

1.1 Topic and motivation

Image analysis researchers are familiar with probability, in particular Bayesian, MAP (maximum a-posteriori probabilities) and likelihood-based constructs. However, the use of these techniques is not always as straightforward as it might at first appear. This paper draws attention to common sources of potential error at the design level and, importantly, their consequences. We proceed from two basic assumptions; first, that there is in general only one theoretically correct way to analyse data, and second, that in cases where multiple approaches seem possible, we must look to their origins in probability theory for guidance.

The concept of a probability density function for any parameter, or arbitrary function of that parameter, is a familiar one. However, we intend to show that care must be taken in the choice of measurement domain, since probabilistic expressions defined as continuous functions cannot be expected to be valid for arbitrary definitions of a problem: continuous probability expressions do not necessarily define true (invariant) statistics.

A second issue concerns the generalisation of definitions of the probability of discrete events to the continuous case. Probability values always refer to *single*, discrete events; extending definitions to continuous variables requires particular care. This is true for any definition of probability, including both the Bayesian scheme, in which probability represents a degree of belief, and the frequentist scheme, in which the probability of an event reflects the frequency with which that event occurs.

The aim of this document is to describe the implications of taking the limit of a set of discrete events (in terms of number and density) in order to define a probability density for a continuous variable. Many of the observations described here result from attempts by the authors, over the course of the last decade, to find the relationships between different, successful approaches to the analysis of image data. In general, it has proven difficult to find theoretical grounding for these observations in the statistical literature.

The issues discussed in this paper are important for the correct design of any algorithm relying on probabilistic models. We stress that our discussion is not purely theoretical; we have tried to illustrate the practical consequences of incorrect design with examples drawn from common pattern recognition and computer vision problems.

The remainder of this paper is organised into two main parts. In the first, several theoretical issues related to the correct use of statistical constructs are discussed. These issues are used to motivate the development of a unifying model for the correct design of likelihood functions, the *equal variance domain*. The use of this

principle is then demonstrated in various contexts, such as hypothesis testing and the definition of information measures. The second part of this paper presents examples illustrating appropriate definitions of probability densities and likelihood functions. The examples are based on two general problems from computer vision and pattern recognition, namely fitting curve models to data and comparing probability density functions. Concluding remarks and discussion are given in Section 4.

We adopt the following notational conventions. Probabilities are represented by upper case P , whilst probability densities are represented by lower case p . Similarly, discrete variables are represented by upper case X , whilst continuous variables are represented by lower case x . Vector quantities are represented in boldface e.g. \mathbf{A} .

1.2 Related literature

This document covers a range of theoretical principles for the construction of computer vision algorithms, including likelihood, hypothesis tests, and entropy. It is generally accepted that algorithms should be based upon statistical methods derived from probability theory. For example, many common algorithms that may once have been interpreted as original have ultimately been reconciled with likelihood. This includes the Hough transform [11,24], mutual information [2,3], and least-squares techniques [13] such as the Kalman filter.

Hypothesis tests [14] cover a large area of statistical methods. Despite their prevalence in virtually every area of science, the concept has been rather neglected in computer vision, implying that much algorithmic potential remains unexploited. In previous work we have shown that hypothesis tests can be automatically generated from sample data in order to construct statistically valid measures of probability for arbitrary distributions of data [4,6]. Hypothesis tests are strongly associated with likelihood, as likelihood methods are often used to estimate model parameters and covariances for use in a hypothesis test.

One of the most common approaches to algorithm construction is based on information theory. The growth of telecommunications in the early twentieth century led several researchers to study the information content of signals. The seminal work of Shannon [22], building on papers by Nyquist [15,16] and Hartley [10], rationalised early efforts into a coherent mathematical theory of communication and initiated information theory as an area of research. Shannon stated that a measure of the amount of information $H(\mathbf{P})$ contained in a series of events $P_1 \dots P_N$ should satisfy three requirements: H should be continuous in the P_i ; if all the P_i are equal, so $P_i = 1/N$, then H should be a monotonic increasing function of N ; H should be additive. He then proved that the only H satisfying these three requirements is

$$H(\mathbf{P}) = -K \sum_{i=1}^N P_i \ln P_i$$

where K is a positive constant. This quantity has since become known as the *Shannon entropy*: for systems of discrete variables, it is identical to the expectation value of the likelihood (see below). Shannon entropy has been used in a variety of applications, and is often taken to be the theoretical origin of the mutual information measure used in multi-modality medical image coregistration [29].

The general issue addressed in this paper is the correct use of concepts from probability and statistics in algorithmic design. Background materials for the discussion include [14] (hypothesis testing), [22] (definition of entropy measures), [25] (histogram similarity measures), and [23] (assumptions underpinning likelihood). [27] draws on these sources to specify a general approach to the construction of machine vision algorithms.

2 Part I: Theoretical issues

2.1 The correct use of likelihood

2.1.1 Background

Suppose that we have a set of n data $X_{i=1 \dots n}$ and some hypothesised model of the data generation process that depends on a vector of parameters \mathbf{A} . The probability that a given datum X_i could be generated by the model can be written as $P(X_i|\mathbf{A})$. If the data are independent, then the joint probability of generating the entire data set can be written as

$$P(\mathbf{X}|\mathbf{A}) = \prod_i P(X_i|\mathbf{A}).$$

When considered as a function of the data with some specified (i.e. fixed) \mathbf{A} this is a simple joint probability. However, when considered as a continuous function of \mathbf{A} for fixed $x_{i=1 \dots n}$, the equation is no longer a joint probability and does not obey the axioms of probability. Such a quantity is called a likelihood, following Fisher [8].

Estimators for the model parameters can be generated by maximising the likelihood or, since any monotonic transformation of the likelihood itself does not change the position of the peak, the log of the likelihood. Therefore,

$$\left. \frac{\partial \ln L}{\partial A} \right|_{A=E(A)} = 0$$

is commonly referred to as the likelihood equation, where $E(A)$ is the desired estimator of the parameter A . For example, assuming Gaussian distributions for the residuals (the differences between the model predictions $m_i(\mathbf{A})$ and the data) gives the familiar expression for weighted least-squares

$$\ln L = - \sum_i (x_i - m_i(\mathbf{A}))^2 / 2\sigma_i^2$$

A simple generalisation of this approach is to assume a non-Gaussian form, which is particularly successful when the true residual distribution of the data has long tails [20].

It could be suggested that likelihood can be derived from Bayes theory, so that a more principled approach should involve maximum a-posteriori probabilities (MAP). However, maximum likelihood estimators have several convenient properties [18]. They are invariant under transformations of the parameter space, so that

$$f[E(A)] = E[f(A)]$$

where f is some arbitrary transformation function. In addition, they are in general consistent (i.e. unbiased in the limit of infinite statistics)¹, although they are generally biased in the small sample limit as an inevitable consequence of the parameter space transformation invariance. Finally, likelihood also provides a framework for the analysis of the information content of the data, including quantifiable constraints on the parameter estimates e.g. error bars. This cannot be delivered directly by MAP.

The concept of likelihood extends directly to probability densities. Since

$$dP = p(x)dx$$

the individual probabilities required can be generated by integrating over a suitable interval, giving

$$L = \prod_i p(x_i|\mathbf{A})dx \quad \text{or} \quad \ln L = \sum_i \ln p(x_i|\mathbf{A})dx,$$

However, the dx terms are often taken to be implicit in such equations and omitted, disguising their significance. This causes complications if we think dx should ever change as a function of the parameters. Redefinition of the intervals dx is also equivalent to a transformation of the data space i.e. a redefinition of the probability terms from which the likelihood is calculated. Even though likelihood is invariant to transformations of the *parameter* space, it is not invariant to transformations of the *data* space. Conventional use relies on the property that the function differs only by a constant, so that optima are preserved. Therefore, for quantitative tasks, a principled method for specifying the data space to be used is required.

The key point is that the construction of likelihood cannot be applied indiscriminately. Each assumption on which it relies must be valid (or nearly so) in order to construct a valid algorithm. The following sections discuss several aspects of the correct use of likelihood.

2.1.2 Choosing correct distributions

In order to construct a valid likelihood *the assumed distribution must actually match the true distribution of the data*. In practice, we might expect that this can be checked by observing the residuals of the data around the estimated solution. We will investigate this below and show complications which arise due to the choice of data space. When applied appropriately, such tests facilitate the selection of a suitable distribution, but raise a question; when is the assumed distribution ‘good enough’?

This question can be approached by applying either qualitative checks or statistical tests to the shape of the data distributions. If, for the quantity of data typically under analysis, the residual distributions differ significantly from those expected, then the algorithm could be improved. This allows the iterative development of an appropriate likelihood definition even without prior information regarding the data generation process.

¹Note that the lack of bias applies to the parameter estimates, not to the likelihood itself: this point will be expanded upon in the later sections.

2.1.3 Statistical data independence

The construction of a simple additive log-likelihood assumes statistical independence of the data, i.e., the residual of one datum conveys no information about the residual of any other: the residuals are uncorrelated. It is partly the task of appropriate model selection to achieve this independence; the correct model will de-correlate the residuals.

In practice, this can be checked by plotting joint distributions of suspected correlations. For example, if temporal correlation between adjacent measurements in a 1-D time signal s_t is suspected, the scatter plot of s_t vs s_{t+1} can be examined for unwanted structure. In the case of complete independence, the outer product of the marginal distributions must completely describe the observed structure of the 2-D distribution

$$P(x_1, x_2 | \mathbf{A}) = P(x_1 | \mathbf{A})P(x_2 | \mathbf{A}).$$

As with residual shape (see the previous section), we do not have to accept correlation in the data should we find it. It is possible to model the correlation and ‘whiten’ the distribution by, for example, rotating to the principal axes of the correlation. Re-projection of the data into the new, rotated space allows us to continue with the likelihood formulation. Clearly, some correlations may be more complex for this strategy, and it is this issue that has led to the investigation of non-linear PCA and associated approaches [9].

The issues discussed so far are generally well-known, and whitening of data spaces is commonly found in the literature (e.g. [20]). However, there is a more subtle issue relating to likelihood which identifies a common problem with expressions related to probability densities.

2.1.4 The importance of the correct likelihood definition

Consider a situation in which the above approach, based on empirical validation, is used to generate two different likelihoods for the same problem. The first uses the original data set x_i , and the second models the residuals in some other domain, via a monotonic, non-linear transformation $f(x_i)$. In such a case *there will be two different solutions* to the maximum-likelihood estimate of the model parameters.

Consider a simple example. We want to compare a set of independent measurements to see if the size of some circular objects follow a given model. We can choose to represent the circles with probability densities either for measured area, $p(a_i | \boldsymbol{\theta})$, or measured radius, $p(r_i | \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of model parameters, and $a_i = 2\pi r_i^2$. We decide to perform a set of experiments (say repeatability of measurements) in order to calibrate the noise process for both. Due to the quadratic relationship between the parameters, the likelihoods computed from the empirical distributions of radius and area *will be different*, that is,

$$\sum_i \ln p(r_i | \boldsymbol{\theta}) \neq \sum_i \ln p(a_i | \boldsymbol{\theta}).$$

At first sight we might think that this has no effect on any estimates of $\boldsymbol{\theta}$ as the difference between the spaces is due to a derivative term which is fixed for each data point, and therefore contributes only an offset to the likelihood function.

$$\frac{p(r_i | \boldsymbol{\theta})}{2\pi r_i} = p(a_i | \boldsymbol{\theta}) \rightarrow - \sum_i \log(p(r_i | \boldsymbol{\theta})) = - \sum_i \log p(a_i | \boldsymbol{\theta}) + \text{const}$$

This is true if we have a theoretical definition of the distributions which can be consistently applied between domains. However, in many practical situations we must estimate these distributions from data. As we will show below, inappropriate choice of domain can then lead to different estimates.

Moreover, *the problem arose because we used continuous variables*. If we had formulated probabilities from a set of discrete events X_i rather than continuous measurements x_i , then the non-linear transformation of the data space would not change the grouping of all measurements corresponding to a particular event. Of course, the transformation may still have changed our idea of how the discrete events are distributed across the space ($f(X_i)$), but this does not change their respective probabilities (Fig. 1).

The difference between the use of probabilities and PDF’s arises because the former are derived from the latter through integration. In the limit of small intervals Δx_i we can write

$$P(X_i | \mathbf{A}) = p(x_i | \mathbf{A}) \Delta x_i.$$

If we apply a transformation, then preserving the probability requires adjustment of the interval (Fig. 2)

$$P(X_i | \mathbf{A}) = p(f(x_i) | \mathbf{A}) \Delta f(x_i)$$

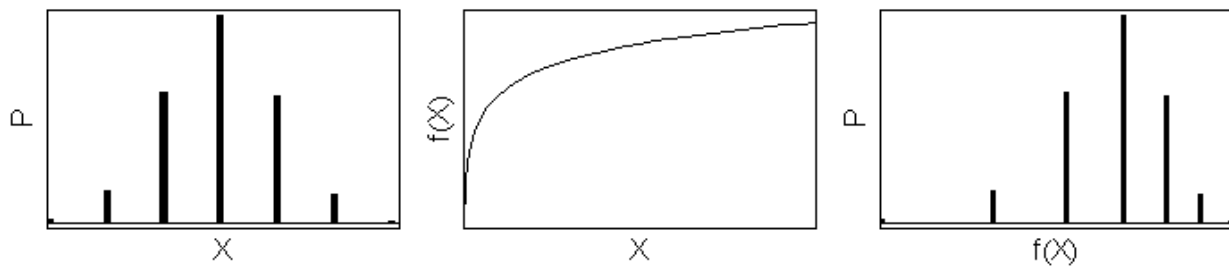


Figure 1: Transformation of discrete samples.

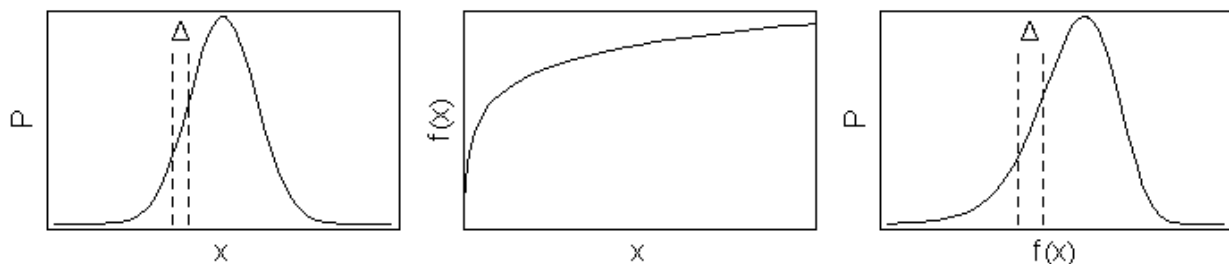


Figure 2: Transformation of probability densities demonstrating conservation of probability within a consistent interval.

Early work on likelihood [8] tended to make this point explicit, drawing a distinction between a probability, which depends on the integration interval, and a likelihood, which is independent of intervals and thus does not obey the laws of probability. The standard notation in current use blurs this distinction, considering the interval to be implicit in relationships between probabilities and PDFs.

2.1.5 The equal variance domain

In many applications, as in the case of circle size measurement, it may be difficult to define a discrete system from which to derive the continuous formulation. In such cases, we suggest that the appropriate equivalent discrete event can be generalised on the basis of statistical separability of the data: the correct space in which to formulate the likelihood (construct the probability density) is the one in which the expected localisation of the measurement within the continuous variable is the same everywhere, i.e., *equal variance*. In the statistical literature, this property is described as ‘homoscedasticity’, and is often referred to as the ‘natural representation’ of a problem, but we introduce the more descriptive term *equal variance domain*. Without this the link between probability and estimation is broken, resulting in bias. In the example of the simple circle measurement, we should aim to construct the likelihood in a space observed to have equal error (see Section 3.1). In some cases, we may observe a functional dependency on the error ($x \pm \sigma(x)$), and apply a transformation to remove this dependency [18, 23]. We will show below how the use of transformations addresses problems with the construction of some hypothesis tests. However, properly constructed, maximum likelihood is already invariant to data transformation, once a given likelihood estimator is found to have bias this will not remove it.

The ‘equal variance’ property therefore constrains the data which can be correctly represented for use in algorithms, reducing the possible number of strictly valid methods that can be constructed. The issue is important for theoretical and practical reasons, and needs to be addressed on a case-by-case basis². Fortunately, in many data analysis problems the raw data is already in the equal variance domain, and so defining likelihoods and hypothesis tests without awareness of the issue results in appropriate formulations. In image processing, for instance, assuming random, uniform and equal additive noise on grey level data is often a very good approximation. Equally, however, it is possible to formulate approaches that do not work as well as expected.

²One way to finesse many of the issues associated with equal variance is to design algorithms based upon likelihood ratios.

2.2 Likelihood and statistical algorithms

2.2.1 Hypothesis Testing

Approaches based on hypothesis tests contrast with Bayesian methods of data analysis in that they require only one distribution model in order to compute a statistic [4]. Specifically, rather than computing the most likely interpretation of a set of data given a set of alternative models, they compute the probability that data which was ‘less like’ that observed could be selected from the model. The definition of ‘less like’ is often couched in terms of a simple similarity measure, such as the distance to the mean (or mode) of the expected distribution. At first sight this may appear to be independent of distribution definitions: we elaborate on this point below.

The computation of hypothesis probabilities involves integration of the probability density beyond the point defined by the data being considered, and so any non-linear monotonic rescaling of the domain has no effect on the result. However, it is common to invoke the so-called *ordering principle* to obtain a principled definition of ‘less like’. Here, all values along the measurement axis are re-ordered according to the expected frequency of occurrence [14]. This can be taken to be the more formal definition of the application of hypothesis testing, since it utilises the only available information with which to define the concept of ‘less like’. For simple (monotonic) distributions, this generates exactly the ordering methods generally applied. However, non-linear rescaling of the measurement domain can result in a new ordering of ‘less like’ for the data. Specifically, the integral beyond a specific test value can result in different probabilities. An example is given in Section 3. From a purely empirical point of view, this new hypothesis probability still has the same quantitative prediction capability as the original measure. A probability of 30% still means that 30% of the time data which was less like the test data would be drawn from the sample. But the definition of ‘less like’ has changed, and so has the specific 30% portion of the distribution we are referring to. A unique ordering is therefore required, consistent with the application of probability theory, and this can be provided by the equal variance principle.

The only way to make definite statements regarding the differences between likely frequencies of events along the measurement axis would be to start from the definition of a discrete system and take the limit of the continuous problem. In addition, the ordering principle is often justified as the choice for the variable which minimises the ‘length’ in the measurement domain of any interval. It is therefore consistent to attach some concept of statistical similarity to this definition of length. Thus, *in the equal variance domain, the ordering principle is the choice resulting in the most compact form of any statistical interval based on the characteristic variability of the measurement.* Such a statistic is by construction the most informative way of summarising any data, as it puts the tightest bounds on any inference. It would be consistent to speculate on the possibility of this benefit for the application of the equal variance approach to likelihood, which may equally correspond to the tightest predictive (Cramer-Rao) bounds in quantitative analysis.

2.2.2 Information Measures

No general discussion of the legitimacy of using information measures in algorithmic construction will be attempted here (see [5]). We limit the discussion to several observations concerning Shannon entropy, the most commonly used information measure in algorithm design. Shannon entropy attempts to define the information content in a signal composed of a string of symbols drawn from a discrete alphabet, and is constructed as a sum over individual data terms.

However, as Shannon described, this can only be valid if the individual terms are independent i.e. the symbols in the signal are uncorrelated. The goal of compression algorithms can be stated as the removal of such correlations, and so it would seem to be legitimate to apply compression to any data set prior to computation of the entropy. Therefore, in the absence of a definition of the optimal compression scheme, Shannon entropy must be an upper bound on the information content of a data set unless the data are independent by construction. One particular manifestation of this issue in image processing is the use of entropy measures computed from image histograms: the entropy of an *image histogram* is not the entropy of the *image*. A grey-level histogram contains none of the spatial information in the data. The computation of the true entropy of an image would require the construction of a multi-dimensional probability distribution describing local grey-level structure. Using histograms to construct entropy estimates can only result in an upper bound on the image entropy.

In addition, again as described by Shannon, the entropy becomes scale-dependent when applied to continuous distributions. Kendall [17] describes the same problem in generalised information measures. This is fundamentally the same issue as that affecting likelihood; when entropy and related measures are expressed as continuous integrals,

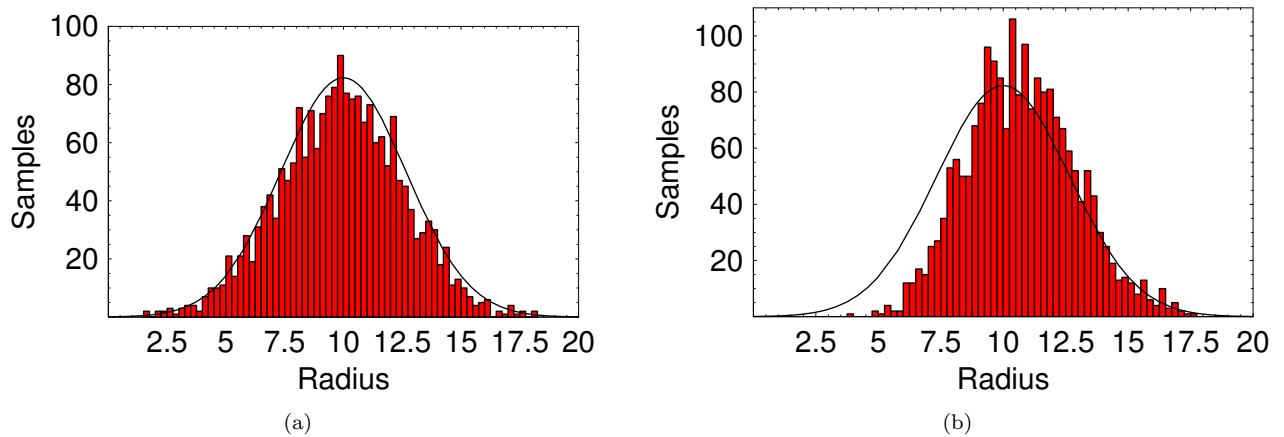


Figure 3: Monte-Carlo distributions for circle radius based upon (a) a radial likelihood and (b) an area based likelihood. The theoretical curves are shown in both cases, demonstrating bias in the area-based approach.

it is possible to obtain *any* numerical value via a rescaling transformation $x \rightarrow y = f(x)$:

$$\int p(x)\log(p(x))dx \neq \int p(y)\log(p(y))dy.$$

Such a measure cannot therefore be afforded the status of a true statistic. For example, the use of entropy measures has been suggested for determining non-linear multiplicative corrections for MR images acquired with a field inhomogeneity [12]. However, simply scanning the solution around the expected answer is sufficient to demonstrate that the measure cannot be used for this task [26]. The potential solution to this problem is similar to the previous one for hypothesis tests. If the continuous entropy were constructed in an equal variance domain then it would be consistent, up to a scale factor, with any underlying discrete definition. Notice that this procedure, although necessary to identify a unique entropy, does not ensure data independence.

Another problem is the lack of shift invariance when histograms of samples from continuous distributions are used to calculate entropy or information measures. The computed quantities may vary with the boundary locations of the binning process, as demonstrated in Section 3.3. This problem is not solved by the equal-variance approach alone, and thus undermines the validity of these measures as cost functions. The standard theoretical interpretation of these approaches does not suggest why this problem should arise or how it can be solved. There would appear to be two possible interpretations: that these approaches are strictly only applicable to uniquely specifiable discrete systems (as in statistical physics); or that they are applicable to purely continuous functions in conjunction with the equal-variance approach (or similar) to set the probability scale. The first of these possibilities is certainly valid; the second appears to require more evidence. We have recently related likelihood to a popular form of the mutual information construct used for image registration [2, 3]. In doing so we have then been able to apply the usual formulation for the estimation of error covariances and demonstrate its validity. We believe that this is sufficient to show that the true origins of this measure (at least in this case) are likelihood and not information theory. In general, we must accept that entropy is defined for discrete systems, and application to continuous variables can result in an infinite number of possible measures, including trivial ones.

3 Part II: Examples

This section presents examples demonstrating the practical importance of correct formulations. First, three specific problems with the formulation of measures based on probability density, as described above, are demonstrated. Second, the application of the equal-variance approach is demonstrated using two problems from the field of computer vision and pattern recognition, namely curve fitting and measurement of probability density similarity.

3.1 Bias from Likelihood Distributions

The assertion that empirical approaches to the construction of maximum likelihood estimates formulated in domains other than the equal variance domain suffer from bias can be demonstrated using the circle area measurement example described in Section 2.1. 2,000 sets of ten radii were generated with additive, uniform Gaussian noise, around a central value of $\mu = 10$ and with a standard deviation of $\sigma = 8$. The maximum likelihood estimate of

the mean was then found for each set. The results are shown in Fig. 3, along with the theoretical distribution computed for 10 measurements and 1 degree of freedom. The same set of data were then transformed to areas ($a = \pi r^2$). The maximum likelihood estimate for the mean was again found, using a likelihood constructed using the empirically observed distribution for area. The results, converted back to equivalent radii, are also shown in Fig. 3. A systematic overestimate of the mean, inconsistent with the original central value, can be clearly observed even though the empirically observed distribution for the area was used in the construction of the likelihood. Thus the output estimate of the parameters describing the model are not the same as the ones we used to generate the data: there is a bias.

The reason this has happened is that the empirical distribution estimated from the observed samples of area for one fixed μ has one fixed shape, which is assumed appropriate for all μ . However, transformation of the specific likelihood used for radius to area would provide densities which vary according to

$$p(a) = \frac{1}{\pi\sigma\sqrt{8a}} e^{-\frac{(\sqrt{a/\pi} - \mu)^2}{2\sigma^2}}.$$

In order to resolve this problem we would need to estimate likelihood distributions from sample data for every possible value of μ , and use these varying functions instead. The assumption of a fixed likelihood function (common to all measurements) cannot be made outside of the equal variance space.

This simple example illustrates one reason why, for example, 3-D data from a stereo vision system should be analysed in disparity space and not in 3-D space. The problem here is often identified with skewed measurement distributions, a separate but related problem. Distribution skewing is not in itself a problem, provided that the *same* skewed distribution is used in the construction of the likelihood. However, combining for example a Gaussian assumption with skewed data will produce parameter bias.

3.2 Non-Unique Hypothesis Tests

To illustrate the issue of non-unique confidence intervals from hypothesis tests, consider the application of the transformation

$$f(x) = x + \frac{1}{2} \cos 2x.$$

to an angular random variate x with a Gaussian distribution. The effects of this transformation are illustrated in Fig. 4. Confidence regions containing the same proportion of the probability density are shown for both the original and transformed data; the ordering principle results in different boundaries of confidence regions. For this specific and extreme case we can see that, although the definition of the confidence interval is still quantitatively useful as a description of the likely measurement values, the unimodal distribution is transformed to a bimodal distribution, which can no longer be concisely described simply by mean and variance. The non-equivalence of confidence intervals causes a change in the result of a hypothesis test for any measured data. The equal-variance domain resolves the potential ambiguity of construction.

3.3 Scale Dependence of Entropy Calculations

The problem of scale-dependence in Shannon entropy when applied to probability densities can be illustrated with a simple example. Figure 5(a) shows the entropy of a Gaussian distribution calculated using histograms of various bin sizes, given on the plot as multiples of the standard deviation of the original Gaussian. The contents of each bin were calculated as definite integrals of the Gaussian in order to avoid sampling problems. The ordinate shows the displacement between the position of the peak of the Gaussian and the central bin boundary of the histogram, in percentages of the bin size. The plots illustrate two points which will be discussed in turn.

First, for large bin sizes, the entropy changes as the Gaussian is moved across the histogram. The difference in entropy between the point at which the peak of the Gaussian and the central bin boundary of the histogram coincide and the point at which they are displaced by 50% of the bin width is shown in Fig. 5(b). The change illustrates that the additivity property of entropy is lost when applied to samples from probability densities. Consider two extreme examples: for infinitesimally small histogram bins, any infinitesimal shift in the position of the Gaussian will result in an identical histogram, offset by one bin. If the bin size is large with respect to the standard deviation of the Gaussian, then when the peak lies on a bin boundary, the Gaussian is split across two bins giving an entropy of -0.693 : when the peak lies at the centre of a bin, the whole distribution is contained within that bin giving an entropy of 0. The differences become significant for any bin size larger than $\approx 1\sigma$.

Second, working with continuous distributions can avoid the effect described above (at the possible cost of additional computational complexity), but does not solve the problem of scale dependency. This can be seen in the changes of

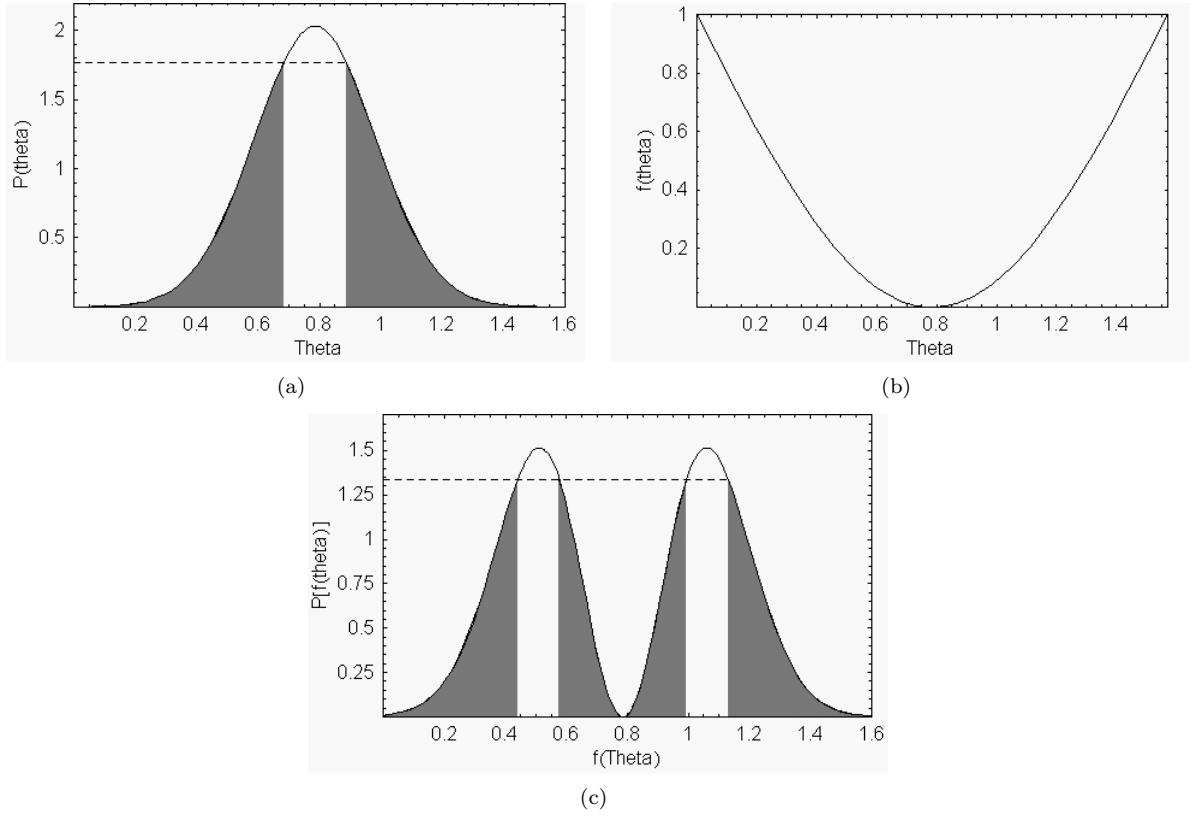


Figure 4: Application of the ordering principle for transformed probability densities showing (a) a Gaussian distribution, (b) a transformation function and (c) the distribution of the transformed variable and the new confidence interval following the transformation.

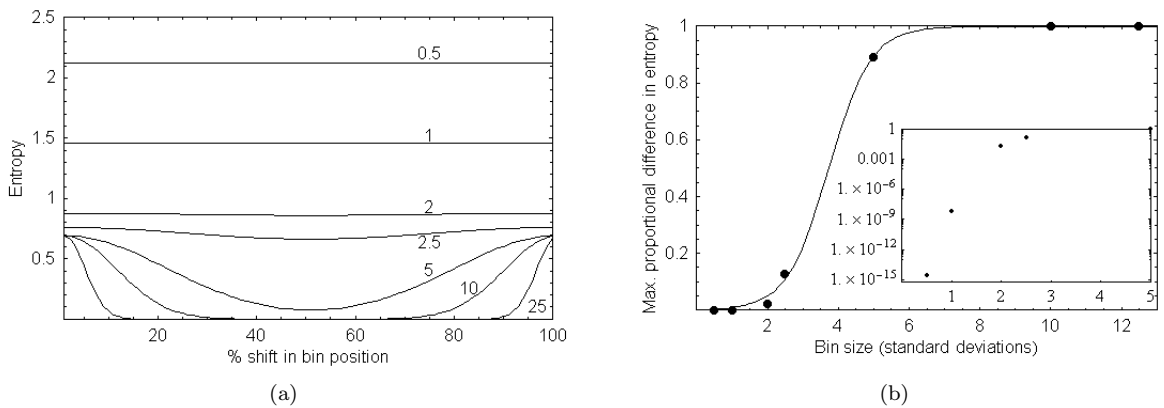


Figure 5: The Shannon entropy of a Gaussian distribution. Each curve in (a) shows the result from a histograms with different bin sizes (given in standard deviations). The x axis shows the distance between the central bin boundary and the peak of the Gaussian in percentages of the bin size. The proportional difference in the Shannon entropies of a Gaussian distribution calculated when the central bin boundary of the histogram is aligned to the peak of the Gaussian, compared to when it is away by 50% of the bin width, is shown in (b) for varying bin sizes (given in standard deviations of the Gaussian). The inset figure shows a logarithmic plot of the lower points.

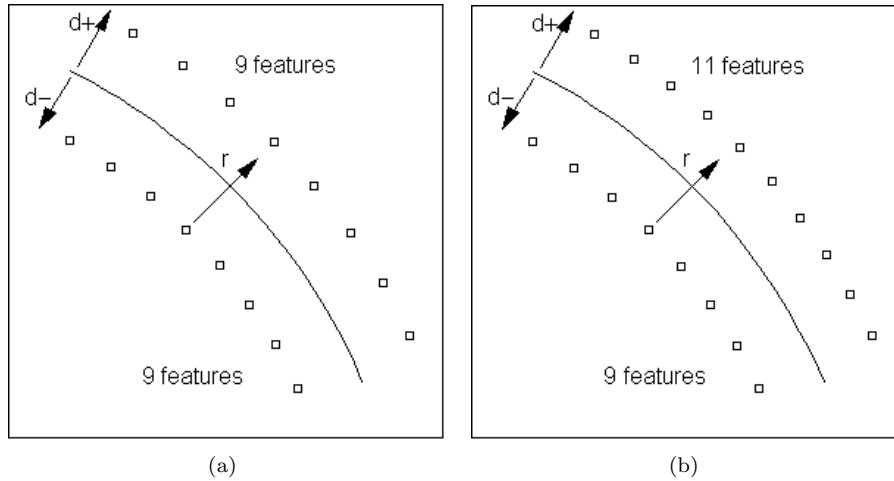


Figure 6: Edge density generation for the equivalent of a nearest point likelihood (a) and the relative increase in edge density needed on the outside of the curve in order to generate a connected string (b).

the intercept of the curves. The entropy is dependent on the bin size of the histogram, even for bin sizes less than 1σ . Changing the bin size is equivalent to changing the scale at which the underlying generating PDF is sampled. Many algorithms in the literature are designed so that they are sensitive to either or both of these effects. Working in the equal variance domain and using a bin size consistent with the resolution of features within the distribution mitigates the effects of these problems.

3.4 A Computer Vision Example: Fitting Edge Strings.

We now apply the concept of equal variance to the very common problem of fitting models to edge strings detected from image data. For simplicity, we assume that the accuracy of feature location is the same for all edges, so that the equal variance domain is the image co-ordinate system. Therefore, to construct the likelihood we have only to estimate the data density distribution for edges perturbed by noise. Strictly, we should start from the noise in the image, but we assume here, again for simplicity, that feature locators have approximately fixed spatial accuracy. In addition we will ignore the complication of quantisation of an image into pixels, and assume that such effects average out in the statistics of large samples, particularly if an effort is made to work with sub-pixel feature locations.

One obvious approach would be to define the data density as a 2D perturbation from multiple locations along a curve. This might be reasonable for fitting isolated features such as corners, but it would not be appropriate for matching a curve to a set of edge features. In this case, the direction along the detected feature string conveys little or no information. Moreover, as edge detection algorithms are designed to produce connected strings, the density of detected features along the string is fixed. Our definition of likelihood must incorporate this property.

Let us start by assuming the 1-D nearest-point likelihood model, giving the radial density model

$$p(x, y|\mathbf{s}) = \exp(-d^2/2\sigma^2).$$

This implies a constant probability of generating a point at a distance d from the curve, inside or outside (Fig. 6). The problem with applying a 1-D density model to 2-D data is that the consequences of changing orientation for the sampling process are easily neglected. As shown in the figure, a generation process with this property will appear to increase the separation between data moved by the noise process to the outside of the curve, relative to those on the inside. Therefore, it will not produce a fixed average interval edge string, and the change in interval will be a function of local radius of curvature (r), resulting in a systematic distortion of the curve shape.

In order to get the same quantity of edges inside and outside of the fitted curve, we must correct the expected probability density in proportion to the expected increase in spacing as a function of local radius of curvature, r , and distance from the curve, d . The density model for this case is

$$p(x, y|\mathbf{s}) = \frac{r+d}{r} \exp(-d^2/2\sigma^2),$$

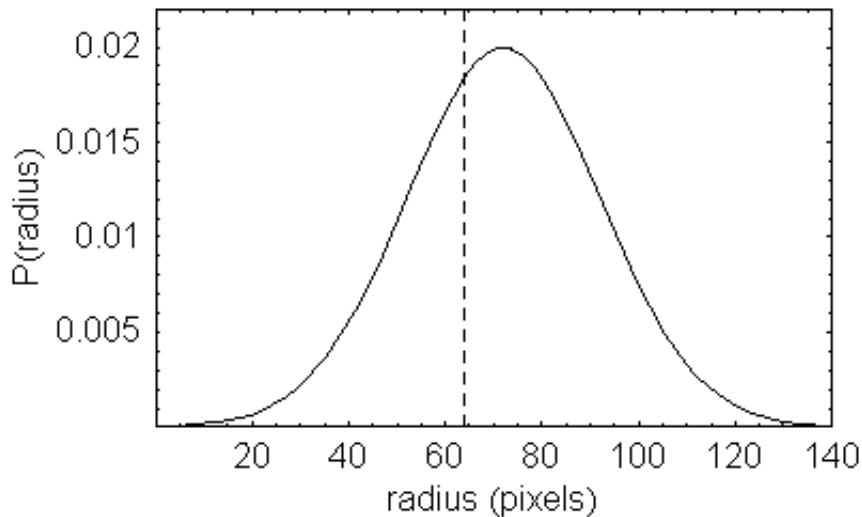


Figure 7: Distribution of the estimated radius of curvature of an edge string. The dashed line shows the true mean, demonstrating the potential for bias.

so that the corresponding likelihood is

$$L = \sum_{t=0}^T d_t^2 / 2\sigma^2 - \ln\left(\frac{r_t + d_t}{r_t}\right).$$

Figure 7 shows the result of a numerical simulation of the process, for the rather extreme case of a circle of radius 64 pixels with a feature location accuracy of 20 pixels. These parameters were chosen to make the bias visible in the distribution for single measurements. The bias scales to more realistic measurements in proportion to the ratio of radius to localisation error.

This shows that simple closest-point fitting routines, which do not include this second term, are likely to exhibit bias on fitted parameters. It explains the bias observed when fitting ellipses to edge strings using closest-point methods and suggests (as is observed) that fitted ellipses will generally appear to be more elongated than the true curve. The empirically determined bias correction term suggested in Porrill’s early work [19] is formally identical to the one derived here for connected features, but the suggested source was instead attributed to the linearisation of the model during the Kalman filter implementation.

3.5 A Pattern Recognition Example: Probability Density Similarity

Table 1 lists several measures that are commonly used in statistical pattern recognition for comparing probability distributions. The principle that there is generally only one correct way in which to analyse data and, that in cases where multiple approaches seem possible, we must look to the origins of the methods in probability theory, was stated above. Following this principle, the measures listed can be examined to determine which are suitable for any given task. For example, to compare two distributions with common origin, we may demand that the ‘correct’ similarity function be consistent under interchange of the two distributions, that is

$$d(\mathbf{p}_1, \mathbf{p}_2) = d(\mathbf{p}_2, \mathbf{p}_1).$$

Examination of Table 1 shows that we can eliminate Chernoff measure immediately on these grounds. In addition, the Kullback-Leibler divergence [7], often quoted as the basis of many algorithms, even in the symmetric form given in the table, does not obey the triangle inequality required of a metric (hence the term “divergence” rather than “distance”).

A further issue concerns the geometry of the underlying space in which any similarity measure is embedded. In Euclidean spaces, such as the three spatial dimensions of the real world, the appropriate distance measure is linear. However, constructs in any other parametric or data space must take account of the curvature of that space. The problem can be stated as one of finding the minimum path (and corresponding probability) along which one measurement could be perturbed to another on the basis of a known noise process. This is the basis

Measure	Function
Chernoff	$d_C = -\ln \int p^s(x W_1)p^{1-s}(x W_2)dx$
Bhattacharyya	$d_B = -\ln \int (p(x W_1)p(x W_2))^{1/2}dx$
Matusita	$d_T = \int (\sqrt{p(x W_1)} - \sqrt{p(x W_2)})^2 dx^{1/2}$
KL Divergence	$d_D = \int (p(x W_1) - p(x W_2)) \ln \frac{p(x W_1)}{p(x W_2)} dx$

Table 1: Distance measures used in statistical pattern recognition for the comparison of probability densities.

of all quantitative similarity or distance metrics in use in statistics, and is directly equivalent to the use of Fisher information [21] as a probability similarity metric.

For strict statistical measures it is accepted that great care must be taken to derive the similarity on the basis of the appropriate perturbation model. For example, independent Gaussian errors give rise to the least-squares function, which for equal variances becomes a simple Euclidean distance. Histograms, say H_i, K_i , are often compared on the basis of the χ^2 :

$$\chi^2 = \sum_i (H_i - K_i)^2 / (K_i + H_i)$$

The space of measurement defined by the standard χ^2 has some rather unusual properties. In particular, the expected variance on measured values changes as a data point moves across the space. This is a property referred to as heteroscedasticity, and creates a problem for similarity measure construction, because the space is non-Euclidean and so the shortest statistical cost path between two points is a curved path, not a straight line. We have shown in previous work [25] that the application of a square-root transformation can be used to construct a similarity space in which the variances are equal (homoscedastic): transformation to the equal variance domain produces a Euclidean space in which the appropriate similarity measure is

$$m(\mathbf{H}, \mathbf{K}) = \sum_i (\sqrt{H_i} - \sqrt{K_i})^2.$$

If, as advocated, we look at a discrete definition of probability for guidance, we can define a continuous probability distribution as the result of applying two limits: the limit of an infinite number of samples, $H_i \rightarrow P_i$, followed by the limit of an infinite number of discrete states, $P_i \rightarrow p_i$. We can then approach a correct definition for comparing probabilities from the statistically appropriate way to compare histograms. Therefore the square-root transform step, required to flatten the space for histogram similarity, is needed to flatten the space for probability density function similarity. The approximations made in this mapping become exact in the limit of large samples. This identifies the Matusita measure (Table 1), or its equivalent the Bhattacharyya measure, as the correct way to solve the problem. These two measures are also symmetric under interchange of the probabilities, as required.

It is also worth considering the effects of non-linear transformations on the continuous definition of probability similarity $x \rightarrow y = f(x)$. It is possible to show that these measures are invariant to this process. For example, for the Bhattacharyya measure,

$$\int \sqrt{p(x|W_1)}\sqrt{p(x|W_2)}dx = \int \sqrt{p(y|W_1)}\sqrt{p(y|W_2)}dy$$

The equal variance approach does not rule out the use of other measures to compare probabilities; indeed, it would be possible to define the limit of the sample in a variety of ways consistent with measures used in statistics. For example, the probability of observing a binary event would be the limit of a binomial process, $H_i/N_i \rightarrow P_i$, of N_i samples. The required variance-normalising transform in this case would be the inverse sine [18],

$$\sqrt{N_i} \sin^{-1}(\sqrt{H_i/N_i}).$$

One probability expression consistent with a binomial sampling process is a frequentist interpretation of the Bayesian formula,

$$P(W_j|x) = \frac{p(x|W_j)P(W_j)}{\sum_k p(x|W_k)P(W_k)} = \frac{p(x|W_j)P(W_j)}{p(x)}.$$

For this expression, differences between two theoretical Bayesian probability distributions (defined as a limit of data samples) with consistent prior probabilities, $P(W_j)$, would be

$$d(\mathbf{P}_1, \mathbf{P}_2)_{Bayes} = \int \frac{p_1(x)p_2(x)}{p_1(x) + p_2(x)}$$

$$(\sin^{-1}\sqrt{P_1(W_j|x)} - \sin^{-1}\sqrt{P_2(W_j|x)})^2 dx$$

4 Conclusions

This paper has discussed several issues surrounding the correct definition of likelihood and other statistical constructs, and their consequences for a range of computer vision and pattern recognition problems. These observations demonstrate that the same care taken in the design of quantitative statistical measures should be applied when designing likelihood and related constructs for vision algorithms. Specifically, probability densities of continuous random variables should be built as the limit of the discrete case. The equal variance domain provides an appropriate way to perform this task. The same principle can be applied to the construction of likelihoods, information measures, and hypothesis tests and should arguably be considered whenever a probability density must be used in an absolute way, that is, whenever a transformation of variables changes the result. We have deliberately excluded Bayesian formulations, as the issues addressed in this paper do not apply to ratios of likelihoods (but see [4] for a discussion of problems specific to Bayesian techniques).

We suggest that a similarity between computational forms has sometimes led people to regard approximations of likelihoods as information measures, arguably as the latter interpretation provides the freedom to include additional measures penalising model complexity. This is well understood in the likelihood literature and generally addressed with approaches to bias correction, such as the Akaike [1] (and related) measures. In previous work we have tried to explain how this can be addressed via the statistical approach, provided that the definition of the problem maximises generalisation capabilities of the estimated parameters [28]. A similar aim for information measures might be to seek the most compact model. This is basically a statement of Occam’s razor, which is accepted as a necessary, if not rigorous, characteristic of the model selection process.

We have suggested the selection of variable domains in accordance with measurement accuracy, in order to maintain a solid link with probability theory. This appears to contradict an assumption frequently encountered in pattern recognition, namely that problems can be solved **uniquely** without matching the approach to the process generating the data sample used. We believe this assumption is, quite simply, false. The very information content of a data set is fundamentally limited by measurement reproducibility: any attempt to construct probabilistic decision systems ignoring information content (repeatability error) can only result in one of an infinite set of possible solutions, selected by an implicit scaling. Such an approach can never be directly reconcilable with probability theory: put another way, it cannot be expected to solve correctly the key problem of model selection [28] (the “bias variance dilemma” in the neural networks literature).

To put the above issue into historical context; when Gauss originally derived the method of least-squares, its properties were all assessed in an equal variance space. Later statisticians introduced the word ‘homoscedastic’ in order to describe data with this property. In 2003 Robert Engle received a Nobel prize for his work in the 1980’s on analysis of ‘heteroscedastic’ data. Clearly, the behaviour of measurement error within a measurement space is considered a fundamental issue. Meanwhile, the development of standard statistical practices, particularly when applied to computer vision algorithms, has managed to overlook this issue. Indeed researchers have sought to justify their work via theories other than quantitative use of probability.

We close the paper with some wide-ranging conclusions:

1. There need not always be multiple, alternative ways of approaching the analysis of data. Accepting that there are ‘correct’ ways of dealing with data, which result in the most informative interpretation, may place restrictions on the number of possible algorithms, but should lead to reliable and well-designed techniques, ultimately the very basis of reliable components of vision systems.
2. We have shown how the equal-variance domain defines the space in which to compare probabilities correctly. The underlying issue is scale-dependence. Discrete probabilities, as definite integrals of an underlying PDF, contain an *implicit* definition of scale, which must be consistent between algorithmic calculations to make optimisations meaningful. The equal-variance domain provides such a definition of scale, and therefore a solution to problems of scale dependence arising with measures such as the Shannon entropy.
3. A better theoretical understanding of the foundations of algorithms ultimately allows us to design better algorithms. Algorithmic tests are indeed needed to assess quantitative behaviour, but proving the validity of a given approach need not rely *entirely* on empirical analysis. It should be sufficient to identify the statistical characteristics of the data which must hold for the approach to be valid. Designing valid comparisons of algorithms requires attention to the expected statistical characteristics of the data. Real progress calls for equal care in experiment design and interpretation of results.

Acknowledgement

The authors would like to acknowledge the support of the MIAS (Medical Images and Signals) IRC under EPSRC grant no. GR/N14248/01 and the UK Medical Research Council Grant No. D2025/31 in funding part of this work.

References

- [1] H Akaike. A new look at statistical model identification. In *IEEE Transactions on Automatic Control*, volume 19, page 716, 1974.
- [2] P A Bromiley, M Pokric, and N A Thacker. Computing covariances for mutual information coregistration. In *Proceedings MIUA 2004*, 2004.
- [3] P A Bromiley, M Pokric, and N A Thacker. Empirical evaluation of covariance matrices for mutual information coregistration. In *Proceedings MICCAI 2004*, 2004.
- [4] P A Bromiley, M L J Scott, M. Pokrić, A J Lacey, and N A Thacker. Bayesian and non-bayesian probabilistic models for magnetic resonance image analysis. *Image and Vision Computing, Special Edition; The use of Probabilistic Models in Computer Vision*, 21:851–864, 2003.
- [5] P A Bromiley, N A Thacker, and E Bouhova-Thacker. Tina memo 2004-004: Shannon entropy, renyi entropy, and information. Technical report, www.tina-vision.net/docs/memos/2004-004, Imaging Science and Biomedical Engineering Division, Medical School, University of Manchester, 2004.
- [6] P A Bromiley, N A Thacker, and P Courtney. Non-parametric subtraction using grey level scattergrams. *Image and Vision Computing*, 20:609–617, 2002.
- [7] T M Cover and J A Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.
- [8] R A Fisher. Theory of statistical estimation. *Proc. Cambridge Philosophical Society*, 2:700–725, 1925.
- [9] K Fukenaga. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, 2nd edition, 1990.
- [10] R V L Hartley. Transmission of information. *Bell Systems Technical Journal*, page 535, July 1928.
- [11] N Kiryati, Y Eldar, and A M Bruckstein. A probabilistic hough transform. *Pattern Recognition*, 24(4):303–316, 1991.
- [12] B Likear. *Registration and Restoration of Medical Images*. PhD thesis, University of Utrecht, 2000.
- [13] P Meer, D Mintz, A Rosenfeld, and D Y Kim. Robust regression methods for computer vision: A review. *International Journal of Computer Vision*, 6(1):59–70, 1991.
- [14] J Neyman. X-outline of a theory of statistical estimation based on the classical theory of probability. *Phil. Trans. Royal Soc. London*, A236:333–380, 1937.
- [15] H Nyquist. Certain factors affecting telegraph speed. *Bell Systems Technical Journal*, page 324, April 1924.
- [16] H Nyquist. Certain topics in telegraph transmission theory. *A.I.E.E. Trans.*, page 617, April 1928.
- [17] K Ord and S Arnold. *Kendall's Advanced Theory of Statistics Volume 1: Distribution Theory*. Arnold, 1998.
- [18] K Ord and S Arnold. *Kendall's Advanced Theory of Statistics Volume 2: Classical Inference and the Linear Model*. Arnold, 1998.
- [19] J Porrill. Fitting ellipses and predicting confidence envelopes using a bias corrected kalman filter. In *Proceedings 5th. Alvey Vision Conference*, pages 175–185, Oxford, 1989.
- [20] W H Press, B P Flannery, S A Teukolsky, and W T Vetterling. *Numerical Recipes in C*. Cambridge University Press, New York, 2nd edition, 1992.
- [21] C R Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.
- [22] C E Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423 and 623–656, Jul and Oct 1948.

- [23] G W Snedecor and W G Cochran. *Statistical Methods*, chapter 15. Iowa State Press, 8 edition, 1989.
- [24] R S Stephens. A probabilistic approach to the hough transform. In *Proc. BMVC 1990*, pages 55–60, 1990.
- [25] N A Thacker, F Ahearne, and P I Rockett. The bhattacharryya metric as an absolute similarity measure for frequency coded data. *Kybernetika*, 34(4):363–368, 1997.
- [26] N A Thacker, A J Lacey, and P A Bromiley. Validating mri field homogeneity correction using information measures. In *Proceedings BMVC 2002*, pages 626–635, Cardiff, 2002.
- [27] N A Thacker, A J Lacey, P Courtney, and G S Rees. Tina memo 2002-005: An empirical design methodology for the construction of machine vision systems. Technical report, www.tina-vision.net/docs/memos/2002-005, Imaging Science and Biomedical Engineering Division, Medical School, University of Manchester, 2002.
- [28] N A Thacker, D Prendergast, and P I Rockett. B-fitting: A statistical estimation technique with automatic parameter selection. In *Proceedings BMVC 1996*, pages 283–292, Edinburgh, 1996.
- [29] P Viola and W M Wells. Alignment by maximisation of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.