

Tina Memo No. 2005-008  
Presented as a tutorial at VIE 2008, Xi' An, China.

# Tutorial: Statistical Design of Quantitative Vision Systems; Beyond Likelihood

N.A.Thacker

Last updated  
25 / 9 / 2011

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

# Beyond Likelihood.

N.A.Thacker. 18/5/2005, last modified 25/9/11

## Preface

*This document summarises my own views regarding attempts to apply quantitative concepts of probability in data analysis and algorithm design. Originally intended for my RA's and students, as a stand against arbitrariness, it has now been made publicly available, and used as the basis for tutorials. However, it remains a working document which is speculative in places and evolves with my own understanding. For example, in previous versions of the document it was claimed that maximum Likelihood generated results which varied with non-linear transformation of the data. As a consequence, my description of the use of intervals used to related Likelihood to probability resulted in terms which were the inverse of a Jeffreys prior and did not have the equivalence claimed. These errors have now been corrected. Other modifications will continue to be made subject to the progress of future research. Although I hope that the document will become more authoritative with time, the opinions contained, and particularly criticism of convention, should be read with this in mind. I expect readers to form their own opinions regarding the validity of conclusions based upon their own knowledge and arguments presented.*

## Abstract.

Probability densities are not probabilities and expressions constructed using them must be used with care in order to maintain a link to probability theory. In this document we will adopt the convention of upper case  $P$  for probability and lower case  $p$  for probability density, in order to avoid confusion. This lack of equivalence is relevant to Likelihood estimation.

The purpose of this document is to summarise general aspects of this issue, and in particular to assess the limits of Likelihood approaches and the possible need for a theoretical generalisation. I have presented this topic in the form of a tutorial which summarises theoretical findings from 20 years of published work. It is my hope that this will raise the general level of awareness regarding scientific validity in the context of algorithm design. The issues addressed are relevant to anyone trying to understand algorithms based upon statistical principles, and the multiple (often contradictory) papers which claim some level of practical utility. The development of the analysis, based upon identification of methods which have the best predictive capabilities, leads in a direction which is dismissive of several approaches currently seen as hot topics. In particular, an attempt is made to relate data analysis to quantitative (and testable) use of probability in order to identify unique methods for the design of a statistical analysis. This breaks with the tradition that use of probability theory is somehow inevitably based upon arbitrary choices. However, it provides a basis for what we regard as the nature of scientific progress, and for an approach to the solution of problems in artificial intelligence.

## Maximum Likelihood

This document will concern itself with the use of Maximum Likelihood (ML<sup>1</sup>) for parameter estimation. The conventional definition of this process can be taken as that presented in [11] where the joint probability of  $n$  independent observations is regarded as a function of an unknown parameter  $A$  and is written as;

$$\ell(\mathbf{x}|A) = f(x_1|A)f(x_2|A)f(x_3|A)\dots f(x_n|A) \quad (0)$$

Where the various  $f(x_i|A)$  are normalised to have unit integral. The ML principle involves taking our estimate of  $A$  such that  $\ell(x|A)$  is a maximum. It is the general validity of this approach, and in particular the consequences for specific ways of defining  $f(x_i|A)$  which we wish to address.

## Is there something wrong with Likelihood?

The most common attitude to Likelihood as a methodology for algorithm design appears to be that there must be something wrong with it. In algorithmic research areas such as neural networks, computer vision, machine learning and medical image analysis (which are those with which I am most familiar) most excitement is always caused by novel methods which are not derived from Likelihood. Then, when someone eventually gets around to relating the algorithm to Likelihood (as has happened for Hough Transforms, the Back Propagation neural network

---

<sup>1</sup>I will refer to such applications and the underlying machinery using the real noun *Likelihood*, and the function itself as lower case *likelihood*.

training algorithm, Mutual Information, the Kalman filter, Support Vector Machines, Boosting... need I go on?) the interest seems to wane and a new novel approach is selected for mass implementation. It is almost as if the use of Likelihood cannot be novel and is therefore not interesting. So what is behind this allergy to Likelihood, is the method theoretically flawed in some way, and does this flaw justify the continual search for something different?

Over the years I have heard many comments relating to Likelihood. In particular I have heard claims that Likelihood generates algorithms which have bias, though equally (in comparison to Bayes approaches for example) I have heard researchers say that the likelihood term is unbiased. So what is actually going on? To answer this question we must first define bias. Statisticians have several definitions for bias (depending upon the task) but the only ones we are interested in are those relating to parameter estimation. Here there are two definitions which are relevant; "bias" is defined as the difference between the true parameter value (that which would be used to generate the data) and the expectation of the distribution of estimates. The other is related to the definition of a "consistent" estimator, which simply states that a consistent estimator is one which estimates the true parameters for infinite samples of data. In this document we will derive the likelihood expression from probability theory and then use that to analyse these biases and possible source of such bias from Likelihood estimates.

Should we really worry about bias in parameter estimation tasks? The standard approach to solving this problem in the context of an analysis would be to estimate the source and magnitude of bias using a simulation approach. Scientific results are often quoted with statistical and systematic errors for this purpose. This is a perfectly adequate way of dealing with bias, but it does have its limitations. In particular simulated data must be an equivalent sample to that on which the estimate is to be made. This is a reasonable requirement for one off analyses, but may be regarded as strictly unworkable in the context of a general analysis system which needs to produce estimates from arbitrary samples of data. Though bias in some applications may actually be seen as helpful (particularly if the task was not properly specified to begin with), it generally results in unreliability. The problem here is one of characterisation, an algorithm which produces a bias which is sample dependant is not characterisable in any meaningful way for arbitrary data sets. This issue may be recognised as one of the main criticisms levelled against the body of algorithms suggested in computer/machine vision over the last few decades. That is, algorithms have performance which cannot be predicted on data sets which are different to those in the original publication, thus limiting the scientific value of the work. It should therefore be regarded as important to design algorithms so that they are expected to be valid (and therefore unbiased) to begin with.

In addition Likelihood can be criticised for being a biased statistic, and as a consequence being unable to identify the correct model. This limitation has consequences for any problem in machine learning. This problem will also be briefly discussed, along with the expected characteristics of any candidate solution.

## Deriving Likelihood from Probability

Standard reference texts regularly state that likelihood is not a probability. However, Fisher's intention when suggesting this approach included the requirement that the ratio of likelihoods for two sets of parameters should equal the ratio of the corresponding probabilities. Fisher also said that he wanted likelihood to be a unique solution to any given data analysis tasks (so that it conforms to notions of the scientific method).

The likelihood function (LF) is often motivated via Bayes Theorem, and in addition the individual terms ( $f(x_i|A)$ ) are often directly associated with probability densities  $p(x_i|A)$  rather than probabilities  $P(x_i|A)$ . To get a probability from a probability density we must integrate the density over a specific interval. Blurring the distinction between probability and probability density and interchanging them leads to non-quantitative (possibly invalid) formulations. Starting from a definition of LF makes it difficult to appreciate the various subtleties of this issue. Here, we will start by deriving the Likelihood method (see [16] for more details). The aim is to define the probability of observing a set of sample data  $\mathbf{x}_i$  in a multi-dimensional (continuum) space based upon a specific density function  $\lambda(\mathbf{x}_i)$ . The conventional notation for probability applies to discrete events, not continuous variables. So to achieve our aims we define a finite set of sample cells ( $\mathbf{X}_i$ ) and compute the probability of observing data within each and then take the continuum limit. The correct statistical model for this is the Poisson distribution, so that the probability of observing a given number of samples within each discrete region within the non-zero portion of the probability density is;

$$P(n_i, \lambda) = \frac{\exp(-\lambda)\lambda^{n_i}}{n_i!}$$

We can therefore write down the log probability for the model in terms of the data as the sum of independent terms for each quantity of observation  $n_i = 0, n_i = 1, n_i = 2 \dots$ ;

$$\log(P(\lambda)) = \sum_i^{N_0} \log P(0, \lambda(\mathbf{X}_i)) + \sum_i^{N_1} \log P(1, \lambda(\mathbf{X}_i)) + \sum_i^{N_2} \log P(2, \lambda(\mathbf{X}_i)) + \dots$$

the first being  $N_0$  empty cells, the second being  $N_1$  cells containing one sample and  $N_2$  cells containing two samples, etc. Equally we can write this as;

$$\begin{aligned} \log(P(\lambda)) &= \sum_i^{N_0+N_1+N_2+\dots} \log P(0, \lambda(\mathbf{X}_i)) + \\ &\sum_i^{N_1} \log P(1, \lambda(\mathbf{X}_i)) - \log P(0, \lambda(\mathbf{X}_i)) + \sum_i^{N_2} \log P(2, \lambda(\mathbf{X}_i)) - \log P(0, \lambda(\mathbf{X}_i)) + \dots \\ &= \sum_i^{N_1+N_2+\dots} n_i \log(\lambda(\mathbf{X}_i)) - \sum_i^{N_0+N_1+N_2+\dots} \lambda(\mathbf{X}_i) + k_1 \end{aligned} \quad (1)$$

where  $k_1$  is a constant. This measure (first described by Fermi) is referred to as Extended Maximum Likelihood (EML) [3]. It is extended in the sense that, rather than simply computing the likelihood for a fixed density distribution it allows parameters describing the density distribution to also be determined. This expression can therefore be taken as the probability of obtaining a particular sample of data from a set of finite states. We would therefore expect this to be the basis of standard fitting routines (such as regression using least squares and its relatives) and also density estimation (such as Expectation Maximisation). We can see that the second term (corresponding to the probability of a zero entry in every location) is the integrated density, which we can make constant as part of the model definition. Therefore, only the first term is data dependant.

Generalisation of (1) to continuous variables ( $x$ ) is now straight forward, but requires us to define the way in which quantities such as  $\lambda(\mathbf{X})$  (for discrete values) relate to  $\lambda(\mathbf{x})$  (for continuous values) and so probability densities from a parametric model  $p(x|A)$ . It was this step in Fisher's original work which was skipped over, with Fisher opting instead to state the required properties of the estimator and then try to work backwards. What we require is the probability of observing a measurement like the one we observe located at  $x$ , this quantity is proportional to  $\lambda(x)$ . For this we need a consistent way of defining an interval around this value with which to compute a probability ( $P(x|A)$ ). Such a process necessarily requires us to define what we mean by "like". For a measurement, this requires knowledge of the statistical location of the measurement  $\sigma(x)$  (that is the measurement or repeatability error not necessarily just the width of the probability density distribution). There are several circumstances which relate to this. Often the error may be provided for us as a fixed value for each measurement. Also, our Likelihood model may be capable of uniquely specifying the error on a measurement, once the parameters  $A$  are known. However, if we are constructing a Likelihood model and know that the expected error on a measurement depends upon the true (generator) value of the observed data (ie: the space is heteroscedastic), we can never uniquely determine the true variance on each measurement. We will return to this point below, and for now assume that the variance term required can be constructed from a combination of the measurement value and the model parameters ( $\sigma(x, A)$ ). We therefore define the probability of observing the measurement  $x$  within an interval of  $\pm\kappa\sigma$  ( $\kappa \ll 1$ ) according to

$$P(x|A) = \int_{x-\kappa\sigma}^{x+\kappa\sigma} p(x|A) dx \approx 2\kappa\sigma(x, A)p(x|A)$$

where  $p(x|A)$  is the integral normalised probability density (see Appendix A). The approximation becomes exact in the limit of  $\kappa \rightarrow 0$  and

$$\lambda(x) \propto P(x|A) \propto \sigma_x(A)p(x|A)$$

This result implies that for a finite sample ( $N = N_1 + N_2 + \dots$ ) of data ( $x_n$ ), minimising the quantity;

$$L = - \frac{1}{N} \sum_n^N \log(\sigma_{x_n}(A)p(x_n|A)) \quad (2)$$

subject to the constraint that the integrated density distribution is fixed for each measurement, will maximise the probability of generating the data with the assumed model  $A$ <sup>2</sup>. The normalisation requirement for probability densities is a standard part of the Likelihood definition, and our expression for  $L$  is now consistent with that for the logarithm of  $\ell(\mathbf{x}|A)$  in equation 0. Notice that the  $\sigma_x(A)$  term will have no effect on this process provided that it is a constant for each measurement (i.e. avoiding the possibility of expected errors varying as a function of the model parameters, see above), and therefore may be dropped from some definitions of likelihood. The one suggested by Fisher, and found in text books for integral normalised probability densities is the following

$$L = - \frac{1}{N} \sum_n^N \log(p(x_n|A)) \quad (3)$$

---

<sup>2</sup>The use of the normalisation term  $N$  is strictly unnecessary at this point but will become important below.

Indeed, we must fix our assumed value of  $\sigma_x(A)$  if we wish the information in our data to remain fixed, so that we can make meaningful comparisons between likelihoods during parameter estimation<sup>3</sup>. With this constraint in place the two methods will give identical parameter estimates, though it has been noticed previously that there are subtle differences in what can be done with the resulting test statistics [4].

Take for example the problem of fitting a theoretical curve  $x(A)$  to measurements with varying Gaussian errors. Enforcing the normalisation of each likelihood term and applying equation (3) will produce

$$L = \frac{1}{N} \sum_n \log\left(\frac{1}{\sqrt{2\pi}\sigma_{x_n}}\right) + (x_n - x_n(A))^2/2\sigma_{x_n}^2$$

The sample and measurement domain dependency, due to  $\sigma_{x_n}$ , of the first term is cancelled if we derive this from equation (2), and it is this form which permits comparison with standard statistical tables (with suitable adjustment of  $N$  for bias, eg: a chi-square per degree of freedom). Thus maintaining the link between probability and Likelihood as attempted in equation (2) would appear to be the required approach if we need to use the resulting goodness of fit measures in a quantitative manner<sup>4</sup>. We can see this process in operation if we transform the data. Conventional likelihood relies on the different resulting likelihoods being related by a constant. Equation (2) on the other hand, is invariant to non-linear data transformation, ie:  $P(x|A) = P(y|A)$  for any  $y = f(x)$  (at least on the basis of error propagation). This is also mathematically equivalent to normalising probability densities to the peak of the measurement distribution [2], rather than the area. Arguments supporting these approaches as the fundamental basis for Likelihood are presented in [4]<sup>5</sup>.

It is often said that Likelihood is not a probability, though our definition using equation (2) shows that it easily interpreted as probability with regard to the data, we just need to understand what the definition is with regard to the parameters. Certainly it is not a probability density for the estimated parameters. This can best be understood by observing that the Likelihood function is invariant to re-definition of the parameters. Whereas, probability densities should change upon transformation. For example, if we are estimating mass or mass squared from a set of data the likelihood function will be the same for corresponding values. We can recognise this invariance as being the equivalent behaviour as for data transformations in equation (2). This implies that likelihood is proportional to the conditional probability (as defined) not the conditional density. This observation allows us to recover the density, and for a single parameter

$$L \propto P(x|A) \propto \sigma_A(x)p(A|x)$$

so that  $p(A|x)$  is obtained by dividing with  $\sigma_A(x)$ . Multi-dimensional cases should follow when using the volume of the parameter covariance.

We can conclude that equation (2) is proportional to the probability of the data being generated by the parameters, defined using an interval proportional to the expected errors on the parameters (implied by the data). We have arrived at this result simply by paying heed to the differences that exist between densities over continuous variables and probabilities. One might think that the above derivation of Likelihood is the complete justification for the approach. However, we will provide arguments below which explain why we need to push the derivation back further. We will suggest that the true origins of Likelihood really lie in the concept of optimal prediction.

## Bias on Likelihood based parameter Estimates.

We now need to assess these definitions for the possibility of bias on the parameters  $A$ . In order to estimate the bias for the case of infinite data we can use the true probability distribution of the data  $p(x)$  and compute the expectation of the likelihood (excluding the  $\sigma_x(A)$  term).

$$\langle L \rangle = - \int p(x) \log(p(x|A)) dx / \int p(x) dx \quad (4)$$

For fixed integrated probability density.

However, we can understand the effects of excluding these terms better if we use the full expression. The expectation of the probability of the data given the model, for continuous valued observations  $x$  follows from (1) and is written as

$$Q = \int p(x) \log(\sigma_x(A)p(x|A)) dx - \beta \int p(x|A) dx + k_2 \quad (5)$$

<sup>3</sup>Any attempt to estimate  $\sigma_x(A)$  during optimisation will likely result in infinities or parameter bias.

<sup>4</sup>Such observations validate our choice for  $P(x|A)$  even though we might expect that this may only be an approximation (Appendix A).

<sup>5</sup>In the cited paper the approach is shown to have desirable properties, rather than being derived. The mathematical equivalence occurs because for an area normalised probability density  $p(x_{max}|A) \propto 1/\sigma_x$ .

where  $\beta$  is an arbitrary constant (needed to relate probability densities to probabilities). Consider the assumed probability density  $p(x|A)$  as a variable function. We can then ask what distribution  $p(x|A)$  must take in order to maximise  $Q$ . Differentiating (5) with respect to a specific  $p(x|A)$  (at a single value of  $x$ ) and setting to zero ( $\sigma_x(A)$  terms are assumed to be constant), we get

$$\frac{p(x)}{p(x|A)} = \beta$$

In order for the technique to be unbiased, the observed distribution  $p(x)$  must match the one assumed in the Likelihood model  $p(x|A)$ , so that all parameters of the data generation process are regenerated by the estimation.

This can be rephrased as saying that for the correct model, in the limit of infinite sample data, a Likelihood estimate of parameters will have no bias. **Likelihood can be applied as a consistent (unbiased) estimator, provided that an effort is made to maintain a link between probabilities and probability densities.** Note that likelihood formulations which allow the distributions to vary during optimisation will violate this requirement. I believe that this is the origin of the bias found when estimating variances of distributions. Note that if we are unaware of this issue we can only expect to get an unbiased estimate by good fortune. We must then do more work to understand the results of the estimation process, by using a Monte-Carlo for example to assess these issues with a statistically equivalent sample.

## Statistical Bias

We now need to ask if Likelihood is “biased” in the statistical sense of the word. The expectation of a parameter estimate is;

$$\langle A \rangle = \int P(A)A \, dA / \int P(A) \, dA \quad (6)$$

Does Likelihood generate a biased estimate of  $A$  based upon this definition, and how does this relate to the concept of “consistency”?

The first thing to consider is our definition of  $P(A)$ , which in the above expression appears to be simply the probability of the parameters, clearly if we are using data to estimate  $A$  then  $P(A) \rightarrow P(A|\mathbf{x})$ , where  $x$  is the data. As  $P(A|\mathbf{x}) \propto P(\mathbf{x}|A)$  for frequentist priors the above expression then becomes;

$$\langle A \rangle = \int P(\mathbf{x}|A)A \, dA / \int P(\mathbf{x}|A) \, dA$$

If this expression is expected to identify a poor estimation process then the implication is that  $\langle A \rangle$  is the best estimate of the parameters given all of the data. Rather than attempting to estimate (6) we can therefore consider how (6) compares to the process of Likelihood estimation, which we already know can be used as a framework for consistent estimation of  $A$ . One way to think about this is to consider the process of taking the limit of an infinite sample as the following extension to (2)

$$L = \frac{-1}{NM} \sum_m^M \sum_n^N \log(\sigma_{x_n}(A)p(x_n|A))$$

where  $L$  is now the combined likelihood for a set of  $M$  samples used to estimate multiple estimates  $A_m$ . Clearly, in the limit of  $M \rightarrow \infty$ ,  $L \rightarrow \langle L \rangle$ . We can write this as

$$L = \frac{-1}{M} \sum_m^M \log(P(\mathbf{x}_m|A))$$

where  $\mathbf{x}_m$  is the set of measurements used for each estimate  $A_m$ . What this illustrates is that combining multiple estimates of  $A$  into one “optimal” value should strictly be done by making full use of not only the parameter estimate, but also the shape of the resulting Likelihood distribution about that estimate. In addition, the best estimate is obtained by taking the minimum, this is data fusion in the standard sense.

The statistical definition of bias takes no account of the distribution of the Likelihood around individual estimates and computes an average not a maximum. Standard statistical texts emphasise appropriate use of the likelihood profile [11] rather than simply the ML estimate for this reason<sup>6</sup>. The expectation of a parameter, though superficially an obvious quantity to define, is not consistent with the definition of Likelihood. Therefore, we cannot

<sup>6</sup>The standard criticism that Maximum Likelihood is an arbitrary way to estimate parameters is true, but summary of the information pertaining to the parameters found in the data as this estimate **plus** covariance can be quantitatively valid and as good as any other way of parameterising this distribution.

expect to get a consistent estimate (the true value) of  $A$  if we choose simply to average. As Likelihood can be used as the basis for the specification of quantitatively valid confidence limits, the existence of such bias certainly can not invalidate Likelihood as a concept. The likelihood function seems valid, if Maximum Likelihood is defined as producing biased parameter estimates on this basis then it is perhaps something that we should not be unduly worried about.

## Bias on Likelihood Estimates

Although the parameters estimated using Likelihood may well be consistent, it is well known that the Likelihood value itself is not unbiased. More complex models always give better likelihood scores and as a consequence Likelihood is incapable of selecting between model hypotheses. Criticism of Likelihood is often followed by the claim that its various problems are avoided if analysis techniques are designed using likelihood ratios. In particular, by subtracting the LF values for two competing model hypotheses we can eliminate the effect of working with probability densities. As the likelihood score has a complexity dependent bias, this claim is simply not true when the competing models have differing numbers of parameters. This problem should not necessarily lead us to abandon Likelihood as a design technique, but it does tell us that any task which embodies the model selection problem cannot be approached in this way. Clearly, this limitation to Likelihood is likely to have a larger impact in applications where the best model to describe the data is not known a-priori (such as scene interpretation tasks in computer vision, but also architecture development in neural networks, and selecting ways of storing data and deriving decision boundaries in machine learning). This then becomes our main motivation for developing a technique which goes beyond Likelihood.

We need to consider if this problem could be associated with the approximations used to relate probability densities such as  $p(x|A)$  to the probability of observing a continuous variable with an arbitrary localisation distribution. Our choice here to scale to the measurement accuracy may be a only first order approximation, as it takes no specific account of the shape of the measurement localisation distribution. However, some basic reasoning illustrates that this is not the cause of the problem. If it were then the discrete form of the joint probability would be capable of model selection and would not just give the best match for the most complex model. Whereas, when matching a model which predicts discrete probabilities computed from a model to a set of observed samples, the best model would be the one which had enough degrees of freedom to exactly match the model to the data. Thus the approximation of the measurement distribution may be a factor but it is not the cause of the bias in Likelihood formulations for continuous random variables. **Likelihood cannot be applied directly to the task of model selection and the reason would appear to arise from the use of joint probability as a measure of similarity.** Thus we must consider the possibility that we need a different way to assess agreement between data and a model.

## Bayesian and Information Based Extensions to Likelihood.

Likelihood is often embedded within a Maximum A-posteriori Probability (MAP) framework, and is almost universally accepted as a technique for incorporating prior knowledge into estimation processes thereby regularising the solution. However, the published literature contains many contradictory solutions, often for the same problem. In addition, multiple alternatives to Likelihood have been suggested based upon the use of ‘information theory’ and in particular the ideas presented by Shannon [10]. Such multiplicity of approaches is not an issue we should simply ignore, instead we should look for the reasons behind this potential source of confusion. My intention in writing this section is to try to provide the reader with a basis for some critical thinking.

### A Critical Analysis of MAP

MAP is often presented as a “point estimation” technique, as is Likelihood, and more general Bayesian methods are advertised as distribution based. Bayesian methods are often justified by the observation that the likelihood term (eg:  $P(X|A)$ ) is a function of the data, and what we would really like is a quantitative understanding of the probability as a function of the parameters (ie:  $P(A|X) = P(X|A)P(A)/P(X)$ ). Under normal circumstances Bayes theorem is applied to problems where there are a fixed number of interpretations, but if we choose to make this categorisation across the parameters (say  $P(A_i)$ ) we can consider this as a distribution over  $A$  (more will be said regarding this step below). Under these circumstances we can write

$$P(A|X) \propto P(X|A)P(A)$$

to get a distribution over our parameter. Obtaining such distributions is claimed to be more informative and superior to using a point estimator.

However, practitioners who use Likelihood quantitatively are interested not just in the solution, but also the distribution around it and in particular the parameter error covariance which characterises the 2nd order shape as a function of the estimated parameters. Such knowledge is essential for any scientific use of derived results. Thus determining the constraint imposed by the data in the parameter space is not solely the province of a Bayesian approach. Estimation of covariance is a standard and quantitatively valid procedure which is related to the concepts of the Cramer-Rao bound and confidence limits. So how is the Bayesian approach related and is it valid?

In comparison to  $P(X|A)$ ,  $P(A|X)$  is conditional on a lot of additional assumptions. This makes it far more difficult to interpret  $P(A|X)$  in the context of other scientific work. It can also inhibit direct comparison of these values with sample data and has led in some instances to observations that Bayesian methods do not have quantitative validity (such as the issue of ‘coverage’).

We have already seen one way which priors can be constructed, as the term which allows consistent estimation of probability density. We could argue that these are not really priors at all (see below). The priors in Bayes theorem originate as a way of generating the data from a set of discrete competing interpretations and not as prior knowledge regarding expected values of real parameters (per-se). If we choose to have a distribution of competing models distinguished by parameter values, then the only mechanism we have to determine this distribution is from examples<sup>7</sup>. However, although the likelihood term  $P(X|A)$  is invariant to transformation of parameters  $A \rightarrow B$ , the sample distribution is not. If this is the way we choose to define our priors then ( $P(A) \neq P(B)$ ). As parameterization (for example; deciding to work with mass  $m$  or  $m^2$ ) is a choice which is available to the experimenter this gives us a degree of arbitrariness which should be judged intolerable in any scientific context. Consequently, using MAP we can produce any answer we like from a sample of data and an assumed prior by applying the appropriate non-linear transformation to the parameters (thereby changing the prior), unless we have taken specific steps in the methodology (absent in conventional usage) to avoid this. The steps which are taken generally focus not upon issues of invariance but instead upon trying to select priors in such a way that changing them causes a minimal effect upon estimated parameters. This addresses the criticism of the use of arbitrary choices in scientific analysis, but may succeed only in replacing the description “completely arbitrary” with “slightly arbitrary”. It does not address the fundamental problem of non-invariance under parameter transformation and it provides another opportunity for the generation of multiple contradictory publications.

This issue is analogous to the transformation of the parameter space in Likelihood, and would presumably require an equivalent solution. Take for example the case of a non-informative uniform prior  $P(A)$ . To recover a density  $p(A)$  we must divide by an appropriate interval which quantifies the level of uncertainty we have in the parameters. Having obtained the density we can apply a redefinition in the form of a transformation of the parameter space  $A \rightarrow B$  to construct the corresponding density for the new parameters  $p(B)$ . To obtain our corresponding prior  $P(B)$  we now integrate over the equivalent (transformed) interval, regaining once again a uniform uninformative prior ie.  $P(A) = P(B)$ . By staying within a logical framework for use of probability notation there are no contradictions, and a uniform uninformative prior for one definition of parameters is a uniform uninformative prior for all, destroying the myth that the uninformative prior is an arbitrary choice. The uninformative uniform prior is the one which has no impact on the estimation, and lets the data speak. The use of any other form of prior is no longer consistent with summarising the information present in data.

There is nothing intrinsically wrong with Bayes theorem and in situations other than competing model parameters the arbitrary intervals will generally cancel ensuring that there is no problem. However, by not defining the interval over which we need to integrate our observed distributions we are effectively replacing probability with probability density ( $p(A)$ ). The common criticism; Where do we get the prior distributions from?, is not just a problem of finding the right set of data. It is also; How do we define intervals for the calculation of consistent probabilities from sample distributions? This observation poses real problems for the frequentist interpretation of prior information when using Bayes theorem<sup>8</sup>.

Those who advocate Bayesian methods for parameter estimation will often show evidence that it is superior to Likelihood. However, if the approach used for likelihood made the error of missing the requirement of homoscedasticity (or equivalently using an appropriate measurement interval) then the Bayesian approach (such as Jeffreys Priors, see below) is improving parameter estimates only as a first order correction for this error<sup>9</sup>. It is the approximate nature of this process which introduces one way of generating multiple solutions (publications). What we might generally observe is that for carefully constructed Likelihoods (which maintain a link to quantitative probability) the inclusion of prior terms in MAP estimation not only increases the chance of getting parameter estimates close to those we would like, but also introduces bias and prevents valid estimates of error covariance. In accordance

<sup>7</sup>I exclude here the idea that priors follow from simple mathematical rules, which would otherwise avoid the criticism, on the grounds that our theories must be descriptive of the real world, and if the world changes the priors should also change.

<sup>8</sup>This could be one of the reasons why strong Bayesian’s refuse to accept a frequentist interpretation of probability.

<sup>9</sup>This correction must be first order and not exact as the issue of heteroscedacity is a function of the data and Bayesian priors are defined as a function of the parameters.

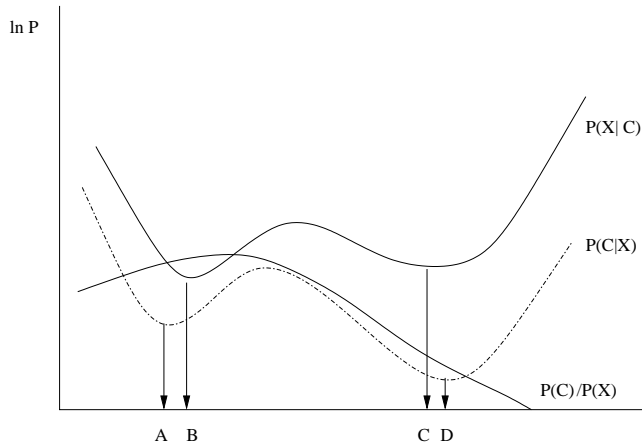


Figure 1: The figure shows a log Likelihood before ( $P(X|C)$ ) and after ( $P(C|X)$ ) modification with a prior ( $P(C)/P(X)$ ) which varies smoothly across the parameter space. It is legitimate to use prior information in the selection of candidate minima from Likelihood (solution C chosen instead of B). However, use of prior information during optimisation (MAP), results in parameter bias (solution D instead of C), and prevents ‘honest’ estimation of measurement covariance.

with Popper’s ideas, regarding the use of falsification and induction in science, we cannot interpret individual cases of improved estimation results as validation of the Bayesian methodology, particularly when reporting bias ensures that unsuccessful results do not get published.

This analysis contradicts popular opinion, which generally regards MAP as the correct way to approach Likelihood. One way to avoid these issues, and get the best from our estimation process, is to consider selection from candidate local likelihood optima using a-posterior probability (Figure 1) rather than optimising MAP. This provides a quantitative mechanism for selection of the ‘priors’, as the process which maximises the correct choice of local minimum in a sample of representative data. We might even reasonably suggest that this is what MAP is approximating and the origins of favourable evaluations. Perhaps, if we wish to incorporate prior knowledge into our estimation process we should consider doing this by using additional likelihood terms, which satisfy the same statistical restrictions (such as independence and invariance) as the rest of the data, rather than re-defining these terms as priors and simply ignoring these issues.

### Example; When are Priors Priors?

Jeffreys priors were defined in order that Likelihood could be used to construct probability densities which transforms correctly between parameter spaces. An idea which is equivalent to Fisher’s scientific idea of uniqueness. We can therefore interpret Jeffreys result as the form of prior necessary to regenerate a meaningful estimation process when using a Bayesian framework.

Specifically a prior probability is chosen according to

$$p(A) = \sqrt{\int_{-\infty}^{\infty} p(x|A) \partial^2 L / \partial A^2 dx}$$

where  $L$  is the log likelihood. This can be interpreted as the square root of the expectation of the Fisher information associated with the measurement  $x$ . When the likelihood term is multiplied by this prior it is therefore equivalent to using  $\sigma_x$  to scale the probability density from each independent measurement, as advocated in the previous sections.

Philosophically, we should take note of Popper’s analysis of the role of priors. One could argue that as the Likelihood estimation bias is sample dependent there is no simple function of the parameters which will remove bias for all data sets. It may therefore be wrong to refer to such correction terms as ‘priors’. Under the Bayesian interpretation use of this prior results in  $P(A|\mathbf{x})$ , whereas the frequentist approach tells us that this  $\sigma$  term was always necessary in order to compute a quantitative  $p(\mathbf{x}|A)$ . The prior here is uniform, and so the use of Jeffreys priors is uninformative, but there is no need to conclude that it results in a calculation of  $P(A|\mathbf{x})$ <sup>10</sup>.

<sup>10</sup>Strictly this requires an additional step to logically constrain allowable ranges of parameters using the axioms of the parametric theory. For example for mass  $m \geq 0$ .

Though I think the use of this approach has validity for the quantitative analysis of data, persisting in the Bayesian interpretation as the theoretical basis for use of Jeffreys priors is unnecessary and can only lead to confusion. Rather, it should now be possible for the reader to see the approach for what it is, a solution to the problems caused by defining Likelihood with probability densities rather than probabilities. An independent re-construction of the quantitative (frequentist) approach with a Bayesian disguise. Further, conventional maximum likelihood is invariant non-linear re-definition of the parameters. Use of a Jeffreys prior results in an estimate of the density which changes depending upon the parameters chosen. So, although it delivers self-consistent densities, these are not something we should simply maximise in order to estimate parameters. You could say that this approach would indeed be MAD (Maximisation of A-posteriori Densities) rather than MAP.

### **Critical Comments Regarding *Information*.**

We must now consider if ideas from information theory are capable of providing an extension to theoretical interpretation of Likelihood as joint probability. Importantly, Shannon's approach was never theoretically derived, but was suggested as a mathematical form which satisfied certain computational requirements. As such, the approach abandons quantitative use of probability in favour of a new theoretical interpretation, which borrows terms like entropy from statistical physics for added credibility. These ideas are shown to have real value for analysis of tasks such as data transmission and compression, with Shannon's Theorem able to make predictions regarding the theoretical efficiency of a compression process.

In addition however, applications have arisen in areas such as appearance modelling, MR bias field correction and image alignment (registration). In these applications, and examples suggested in Shannon's original paper, these measures are directly equivalent to the expectation of the likelihood  $\langle L \rangle$ , subject to various assumptions (such as data independence). We must therefore be careful to attribute any success of these approaches to the correct theoretical interpretation.

There is some evidence that use of Shannon entropy for estimation tasks is not theoretically justified. Firstly, quantitative understanding of the results from such algorithms, such as error estimates, requires the Likelihood interpretation and associated methods for the calculation of covariance [7]. Secondly, limitations of the quantitative use of entropy are often interpreted in information terms as putting an upper bound on the 'true' value. You can describe this situation as the consequence of an assumption of data independence. This may be sufficient for analysis of data transmission but ignoring the correlations in data is not a good basis for an estimation task. In fact, we can say that the 'true' entropy in these cases could be computed using a definition of joint probability which takes account of these correlations, and not simply Shannon Entropy. In cases where the additional (independence) assumption required to construct the entropy measure from joint probability do not apply, our understanding of Likelihood, and in particular its relationship to the Cramer-Rao bound, tells us that this will result in inferior algorithms.

Issues of limits implied by correlation are avoided in the most popular work on entropy (that associated with trying to incorporate prior knowledge using the so called MAX ENT), by only considering sets of independent uniquely distinguishable states. This idea is often illustrated using Jaynes "Dice Problem", which tries to estimate a useful distribution of probabilities over dice outcomes based upon an observation of mean "dot count". It has been noted however, that as this solution is independent of the size of sample from which this mean is assessed the solution is non-statistical. Once the problem has been correctly specified, so as to clearly elucidate all of the implied assumptions hidden in the original formulation, it becomes apparent that Jaynes' formulation is arbitrary [15]. Any justification for such an approach, such as it is, is found to lie with joint probability (ie: MAX PROB). In seeking an underlying concept for Likelihood we have only succeeded in coming full circle.

Based upon this analysis, use of entropy measures for estimation can therefore not be expected to have either practical advantage or theoretical superiority over the appropriate use of Likelihood<sup>11</sup>. Although ignoring data correlation might result in highly convenient mathematical forms, suitable for simplified algorithm designs, we can eliminate these ideas as a valid theoretical generalisation of Likelihood.

### **Example; Feature Selection**

An issue which is related to the problem of model selection involves recent attempts to construct optimisation measures for use in feature selection. This is another of those situations where there are multiple and contradictory suggestions now in the literature (each claiming validity through some level of empirical testing). In these papers, methods are presented as a combination of mutual information terms, eg: from data tuples. However, any

---

<sup>11</sup>except for perhaps the opportunity to publish another paper.

mathematical expression needs to be considered with regard to its properties, such as invariance under transformation of variables, and its computational form as a construct of probability before we can assume any theoretical legitimacy. As not only can different terms assume different data independence properties, (ie: making contradictory assumptions), any subtraction of terms (each only a bound) is likely to be quantitatively meaningless. In my opinion, it is therefore wrong to treat Shannon based ‘information’ terms as fundamental quantities which can be combined arbitrarily (ie: as often done in ‘information theory’). A series of unknown factors need to be true before the differences between information measures associated with alternative models become meaningful. The existence of contradictory theories is explained for me by the lack of mathematical self-consistency within the approach being taken.

In summary of this section, we can say that a theory of data analysis which is intended to provide an optimal analysis must be based upon quantitative use of probability. However, the popular approaches to use of Bayes Theory and Information are non-quantitative, and used as approximate solutions to problems with Likelihood formulations (eg: homoscedasticity, correlation, numerical stability, lack of data). Before we conclude that the vast quantity of publications applying such methods constitute validation we have to be careful. As well as the publication bias mentioned previously, the introduction of any additional degree of freedom to a technique which already has a flaw will allow an improved performance when tuned to a specific data set<sup>12</sup>. In this way even invalid approaches can obtain the empirical evidence needed for publication. Below we will also explain how the same process occurs with the issue of model selection.

If these arguments are valid it demands that we re-interpret the empirical utility of these methods within the context of quantitative probability rather than assuming theoretical validity. **Neither Bayes Theory nor Information Theory offer a quantitatively valid generalisation to the theory of Likelihood.**

## Optimising Generalisation.

Any extension to Likelihood must accord with Fisher’s original motivation and requirements. Specifically, it must be consistent with its intended use in a scientific experiment, by being unique (so that the methodology precludes multiple interpretations) and testable (ie: related to predictions via a frequentist definition of probability). In previous work we have asserted that the best interpretation of data is not the most likely generator of the data, but the model which would be most likely to generalise to unseen data drawn from the same distribution. This is not just an arbitrary choice for the replacement of joint probability. Not only does this accord with scientific practice, we can also argue that optimising the quantitative predictive capabilities of any model to future data must be seen as **the** best way of summarising any data set<sup>13</sup>.

We must also ensure that any new approach satisfies the criteria we have used to evaluate the theoretical limits of Likelihood. We can demand that any formulation must produce “consistent” estimates. A technique which genuinely optimises generalisation meets this requirement by definition, as ultimately the best model parameters for predicting data are the true parameters. **The use of generalisation as the basis for similarity not only accords with Fisher’s original aims but permits the construction of an unbiased (“consistent”) estimator.**

If we optimise generalisation, we must expect the selected model order to be a function of the information content of the data. That is more data, or more accurate data will be necessary to justify the use of a more complex model. Logically, this must apply to both the quantity of data and how the distribution of data constrains model parameters<sup>14</sup>. The required modification is therefore not something which could be specified a-priori, without knowledge of the distribution (or accuracy) of the data sample. This one observation eliminates any method of model selection that attempts to bias towards simpler models but takes no explicit account of data accuracy, from further consideration. Such methods can only be viewed as empirical corrections which are valid for constrained data sampling processes (where the information content of the data is either fixed or only a function of the parameters used to determine the prior terms). This observation explains another rather large set of contradictory approaches which have been suggested in the literature, which generally involve adding some simple bias correction to a conventional likelihood and justifying this from a ‘Bayes’ perspective. As with the issue of heteroscedacity,

---

<sup>12</sup>As a simple example, imagine the task is to add two numbers  $x + y$  and have empirical data of what the result  $d$  should be. Because of a flaw in our theory we have decided to use the function  $x + \sqrt{y}$ , if we test on the data  $d$  we will then find that on average, for a fixed data set  $x + \sqrt{y} + c$ , works better. This does not prove that  $x + \sqrt{y} + c$  is the correct formula but only that  $x + \sqrt{y}$  is not correct. What is worse, publishing our result has little academic value, as the required  $c$  is data set dependant.

<sup>13</sup>Generalisation is a well defined concept in machine learning and artificial neural networks and is related directly to the universally agreed method of testing a learning architecture, that is; How well will this specific choice perform on unseen data? Data is often split into training and testing data sets for this purpose.

<sup>14</sup>For example, when performing a curve fit a large group of data in a localised region is of less statistical value than data distributed more uniformly along the curve.

such solutions can only be expected to be an approximate solution, though in this case this approach might be expected to be less satisfactory, due to the more general nature of the model selection task in practical situations. Other authors have suggested simple modifications to Likelihood in order to correct for bias. The most famous (and relevant) here being the Akaike Information Criteria (AIC) [1] and the Neural Information Criteria (NIC) [9] (Appendix B). The latter has the ultimate result of maximising the generalisation capabilities of the selected model, as discussed above. Unfortunately, this approach does not provide anything other than conventional Likelihood as the basis for parameter estimation. **Comparison of continuous valued data, in a way which permits both the unbiased estimation of parameters and model selection requires a new definition of similarity based upon a quantitative use of probability for the prediction of generalisation.**

Quantitative estimation of generalisation should be straightforward. We can use conventional quantitative approaches to assess a specific model description of data in order to compute a probability distribution associated with the prediction of new data. Evaluation of this prediction should then be a simple case of making a comparison with the original probability distribution describing the uncertainty in the measured data. However, this requires a method for comparing the similarity between probability densities.

There are many contradictory methods suggested for such a comparison in the pattern recognition literature (Kullback-Liebler Divergence, Matusita, etc.), with the prevalent view being that the task necessarily involves some degree of arbitrariness. If we choose to use KL-Divergence then we will have the expectation of the log Likelihood, which fits with the entropy/information based interpretation. We note here that this cannot be taken as a theoretical justification in terms as a fundamental principle, precisely because this is a ‘‘Divergence’’ and not a ‘‘Measure’’<sup>15</sup>. By contrast, in statistics any notion of similarity must be firmly rooted in a quantitative understanding of allowable variation, and this is often based upon characteristic sampling models, such as Poisson, Binomial and Gaussian. Unfortunately, the pattern recognition literature seems to have made no attempt to be consistent with conventional statistics and distribution assumptions. Partly this is due to the acceptance of a subjective definition of probability, which we conclude is unscientific in [19]. In particular there is no concept which relates to the variation (or error) in a probability, the very notion if considered at all is quickly dismissed as meaningless. However, the multitude of possible similarity measures can be seen as resulting from the inability to address this issue.

Definitions for various forms of frequentist probability follow directly from the large sample limits of conventional models. This allows us to interpret variation in probability as we approach this limit. In particular, probability density is the large sample limit of a Poisson process. From this definition, the correct way to compare two probability density distributions ( $d_1(x)$  and  $d_2(x)$ ) is to use the Matusita (or equivalent Bhattacharyya) measure [14].

$$M = \int_{-\infty}^{\infty} (\sqrt{d_1(x)} - \sqrt{d_2(x)})^2 dx \quad (7)$$

We can observe that this approach also has the invariance characteristics we require<sup>16</sup>. Note also that for similar distributions, such a density comparison will be very similar to a KL-Divergence (Appendix D), so that optimising the expectation of log Likelihood will approximate the process of maximising generalisation. This provides a justification for such approaches, as an approximate form, without having to appeal to notions of Entropy or Information for the underlying principle. Note also that the idea of optimal generalisation is quantitatively testable, while the concept of information can only ever be relative, and is always in the form of an approximate correction. Thus, in the latter case, there seems to be no experiment we can do which will quantitatively validate the approach, and those experiments which are done are generally based upon accuracy of prediction (i.e. generalisation). In this context, the ideas of information theory may be nothing more than a retrospective justification for a convenient computational form which shows empirical merit. If true, this constitutes misuse and distortion of the physics concepts to which they are related and may lead to spurious ideas in future. It seems more straight forward to suggest that minimising KL approximately maximises the similarity between the uncertainty of prediction of the model and the data distribution, i.e. quantitative optimisation of prediction. This has nothing to do with classical notions of entropy, and need not even be related to Shannon information. As a physicist by training, it is my opinion that the KL divergence and entropy functions are sufficiently simple as computational forms that they happen to look the same merely by chance.

However, for our new (generalisation) approach there appears to be two ways that we can introduce multiple pieces

<sup>15</sup>Texts which discuss this issue [18] will generally make a distinction between ‘‘truth’’ and ‘‘model’’ in order to explain the asymmetry found in KLD, while forgetting that in general use the roles of these two distributions (what we see as model and what data) can be interchanged for some definitions of comparison. For example, take the prediction of a single value (theoretical model) and a single measurement, prior to the experiment we can compute the degree of similarity between distributions for any hypothesised measurement. In so doing we have completely reversed conventional definitions of model and measurement. Our model is a prediction of possible measurements, and our data is our prior knowledge obtained from theory.

<sup>16</sup>Invariance under interchange of distributions  $P_1 \leftrightarrow P_2$  and non-linear transformation of the measurement space  $x \leftrightarrow y = f(x)$ .

of data, it is legitimate to increase the dimensional character of  $x$ , and also to sum individual integrals. We can choose to treat the set of measurements either as a set of individual predictions  $x_i$  ( $x_1$  or  $x_2$  or  $x_3$  ..) or as a vector  $\mathbf{x}$  ( $x_1$  and  $x_2$  and  $x_3$  ..). For example a comparison between a model and a set of data ( $x_i$ ) localised by measurement kernels  $k_i(x)$  would be

$$L'(\mathbf{A}) = \sum_i^N \int_{-\infty}^{\infty} (\sqrt{k_i(x)} - \sqrt{p_i(x|\mathbf{A})})^2 dx \quad (8)$$

or

$$L'(\mathbf{A}) = \int_{\mathbf{x}} (\sqrt{k(\mathbf{x})} - \sqrt{p(\mathbf{x}|\mathbf{A})})^2 dx_1 \dots dx_N \quad (9)$$

The terms in these equations represent quantitatively valid ('honest') estimates of distributions, in particular  $p_i(x|\mathbf{A})$  is the distribution of unseen data as predicted by the estimated model parameters<sup>17</sup>.

If we choose to use a vector measurement and compute the probability overlap in a single multi-dimensional data space (equation 9), this corresponds (as for Likelihood) to demanding that all data were simultaneously generated by the assumed model (*and's*). Equation 8 on the other hand allows each piece of data to be addressed separately (*or's*). We could also have mixed data composed of sets of vectors. This framework is therefore broader in scope for it's assessment of evidence than a joint probability, with equations 8 and 9 being the logical extremes of this process.

This measure is expected to decrease with increasing quantities of data and will approach zero in the limit of infinite quantities. Therefore, as for Likelihood scores, it cannot be compared directly between non-equivalent samples without an additional normalisation. If we start from equation 9 and choose to ignore the effects of parameter stability, for linear models there will be a monotonic relationship between the likelihood (the distance between the point defining the location of the data and the model constraint manifold) and the new measure (the multi-dimensional distribution overlap integral). In this situation any optimisation will generate the same parameter estimates as Likelihood. However, both inclusion of the effects of parameter stability and non-linear model constraints are expected to give different results which genuinely optimise generalisation. We have shown in previous work how this approach solves the key problem of model selection [13, 6, 8, 12].

Conventional understanding of Likelihood holds that the assumption of a Gaussian distributed measurement results in a quadratic form of likelihood. This measure is ubiquitous in the computer vision literature, as it holds the key to constructing closed form solutions to systems of linear equations. We might ask the question; What alternative measure could we use to compare real valued (Gaussian distributed) measurements with a fixed model? This question is investigated in Appendix C starting from equation 8. We see that with each data point contributing to the overall score from its own overlap integral, we obtain something equivalent to a robust kernel.

Optimisation of generalisation can therefore be said to be a genuine extension to, and justification for, the Likelihood method. While also offering not only a solution for model selection but a method for dealing with non-linearity and issues of robustness for cases where data is not necessarily generated by one process. In such cases, the assessment of data does not appear to require prior knowledge of alternative data generators. Each measurement is assessed according to its level of agreement with the assumed model on the basis of the known errors, thereby (as per Fisher's requirements) providing a unique method for assessment of data, even when contaminated by an unknown process.

## Conclusions

We suggest here, in accordance with our previous published work, that in order to gain agreement with probability theory we must go back to first principles and derive the correct way to compare a measurement of a continuous variable with a function using the definition of probability density comparison based upon the limit of Poisson samples. From this approach we would conclude that Likelihood can be applied as a "consistent" approach for the estimation of parameters in a fixed (known) model, but that it may well produce "bias" in the statistical sense of the word. In general however, the concept of "consistency" appears to be the basis for the most appropriate theoretical definition of bias. This apparent contradiction perhaps explains the conflicting statements made regarding the use of Likelihood for parameter estimation. The existence of bias depends upon how you define it and the specific methodology used for likelihood construction.

Our preferred definition of bias appears to be valid only for infinite samples of data. How then should we evaluate the performance of Likelihood for finite samples if we cannot use the expectation of the parameter estimate? In fact,

<sup>17</sup>Derivation of such expressions must therefore take appropriate account of how accurately the model parameters can be determined for a given choice of model and are (despite the notation) distinctly different to conventional likelihood terms. In our work, obtaining an 'honest' prediction of generalisation was achieved using an analytic form of cross validation.

for all quantitative statistical methods, the only thing which really matters is the issue of “honesty”, that is, does the estimation technique deliver estimates of probability which match the world data samples. Here all we need to say is that a technique which has been derived from probability with distributions chosen carefully to match real world samples must be “honest” by construction, if probability theory is (as advertised) a self consistent method of data analysis. On this basis of a definition of Likelihood formulated carefully, so that the model quantitatively predicts the data distribution, will have no bias. It should therefore be possible to use the theoretical definition of Likelihood to investigate possible causes of bias in naive implementations. An example is given in our recent paper [17]. This makes Likelihood a very powerful theoretical tool in algorithmic research.

Likelihood, (even applied in homoscedastic spaces where there is a more direct link to probability), is incapable of solving the problem of model selection. This then would appear to be the main area of concern for algorithmic research. We have suggested a method for comparing data distributions using a probability overlap, in order to maximise the generalisation of the model to unseen data. We have explained how this concept generates not one but two forms for the quantitative assessment of data, one of which being related to Likelihood. We have then compared the other approach on the assumption of Gaussian distributions, to standard least squares, which is the conventional Likelihood based method. From this we can see two things, least squares approaches are consistent up to 3rd order with this new measure, with the next correction term not being important until around  $\Delta/\sigma \approx 2$ . In addition the new approach will saturate at a finite value for large  $\Delta$ . **Application of this new definition of similarity, in the limit of large samples (where model selection is not an issue) does not always regenerate Likelihood. For the case of Gaussian distributed data it can also be used to generate a robust statistic.** In comparison to conventional robust Likelihood this approach has one very important characteristic. As the similarity metric converges to a constant (finite) value for large deviations from the model, data sampled in these regions have no effect on the optimal parameter estimates. In contrast, and as sensible as this a behaviour would seem, strict adherence to the definition of Likelihood, as it appears in reference texts (see above), would normally prevent us from defining such a distribution, as the functions corresponding to our  $f(x_i|A)$  terms (having arbitrary and potentially infinite extent) are non-integrable<sup>18</sup>.

I appreciate that the logical flow of this document will be difficult to follow for many readers, particularly if you are not acquainted with the idea of quantitative use of probability. I will therefore finish with an attempt to summarise the key conclusions of the analysis presented in this document, with the hope that the reader will begin to appreciate what is being suggested and will go back to see where these statements have originated.

- Likelihood can be applied as a consistent (unbiased) estimator, provided that an effort is made to maintain a link between probabilities and probability densities.
- Likelihood cannot be applied directly to the task of model selection and the reason would appear to arise from the use of joint probability as a measure of similarity.
- Neither Bayes Theory nor Information Theory offer a quantitatively valid generalisation to the theory of Likelihood.
- The use of generalisation as the basis for similarity not only accords with Fisher’s original aims but permits the construction of an unbiased (“consistent”) estimator.
- Optimising the expectation of a log likelihood (where the expectation is taken over the parameters) is an approximation to this principle.
- Comparison of continuous valued data, in a way which permits both the unbiased estimation of parameters and model selection requires a new definition of similarity based upon a quantitative use of probability for the prediction of generalisation.
- Application of this new definition of similarity, in the limit of large samples (where model selection is not an issue) does not always regenerate Likelihood. For the case of Gaussian distributed data it generates something akin to a robust statistic.

---

<sup>18</sup>This raises the interesting possibility that the well known sensitivity of least squares approaches to outlier data is actually due to the definition of Likelihood (the joint probability of the observed data rather than an attempt to maximise generalisation capability) and not an inevitable consequence of an assumption of Gaussian errors. This therefore casts the conventional theoretical justification for “robust” methods in a new light :ie the problem was caused by assuming that all data was generated by one set of model parameters, rather than the shape of the assumed Likelihood distribution.

## Appendix A: Relating probabilities and probability densities in quantitative analysis.

The probability of observing a values  $x$  within an interval  $\delta$  from a probability density  $p(x)$  is simply;

$$P(x, \delta) = \int_{x-\delta/2}^{x+\delta/2} p(x') dx'$$

The idea of an interval is consistent with assuming a rectangular distribution for the measurement process (ie: a quantised (or integer) variable  $x$ ), which is saying something specific regarding our measurement accuracy. In this case, for slowly varying (or approximately linear)  $p(x)$

$$P(x, \delta) \approx \delta p(x)$$

However, we would also like to know how to compute the probability  $P(x, k(x))$  of observing a measurement  $x$  for other measurement processes  $k(x)$ , such as Gaussian errors.

We can see that the **probability density** must be proportional to the convolution of the measurement distribution with the model distribution ie:

$$p(x, k(x)) = \int_{-\infty}^{\infty} k(x' - x)p(x') dx'$$

for normalised densities. Some refer to this as integrating over the unknown, and others may recognise it as a common step in many papers which are motivated from a Bayesian perspective. However, it can be described in words as: computing the total number of ways that  $x$  could have been generated from  $p(x)$  given a measurement process with distribution  $k(x)$ .

Of course generating a probability requires us to once again specify an interval (we can replace  $p(x)$  with  $p(x, k(x))$ ) and continue as above. This is then entirely consistent with a quantitative (frequentist) application of probability and therefore also for use as the required joint probability in Likelihood formulations. The question is; How do we define the interval?

Strictly, for a quantitative task, we need any statistical quantities we calculate to give meaningful (absolute) values. Probabilities (upper case  $P$  here ) can be quantitatively tested. However, probability densities (lower case  $p$  here) are not absolute, for example they vary according to the definition of measurement variable  $x$ , even simple rescaling of the variable changes the probability density corresponding to a particular physical event. Obtaining a quantitative value requires us to choose one way of using the probability density which can be consistently applied.

Taking our cue from the start of this appendix we can choose to define the interval using the expected width (or standard deviation)  $\sigma$  of the measurement process  $k(x)$  so that;

$$P(x, k(x)) \propto \sqrt{\text{var}(k(x))} p(x, k(x)) = \sigma p(x, k(x))$$

This way, although we still have an arbitrary constant, we can at least be sure that we treat measurements with different errors consistently and independently of the chosen measurement domain.

Assuming that this is the correct scaling measure <sup>19</sup> we can go slightly further, as a rectangular distribution has a standard deviation of  $\delta/\sqrt{12}$ . Which implies

$$P(x, k(x)) \approx \sqrt{12} \sigma p(x, k(x))$$

in order to obtain quantitative agreement between different distributions.

## Appendix B: AIC and NIC

The original derivation of AIC was motivated by information arguments [1], but the aim here is to show that the same result can be obtained purely by considering the expected bias on the Likelihood score, without apparently any need for information theory.

---

<sup>19</sup>There is still some element of doubt as to whether this is the appropriate way to characterise the ‘width’ of all distributions, but for similar shaped distributions this should at least be proportional to the required factor.

Conventional approaches to parameter estimation are often developed from Maximum Likelihood. It is now well known that this approach gives a biased result, in that as more model parameters  $\theta$  are added the log-likelihood (or  $\chi^2$ ) of the model generating the data  $x_i$  approaches 0. For a set  $N$  of independent data  $i$  we can write;

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

The limit of the bias is defined directly as ;

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

where  $p(x_i)$  is the true probability density from the correct model and  $\langle X \rangle$  denotes the expectation operation. We can expand this about the true solution  $\theta_0$  as;

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

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

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

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

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

which can be reduced further to

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

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

$$AIC = -2 \log(p(x_i, \theta_0)) + 2K$$

which for Gaussian distributed variables with known variance is

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

For unknown variance (setting the variable  $\sigma$  as an estimated parameter) fitting with  $n$  data samples we get<sup>20</sup>

$$AIC = n \log(\sigma^2) + 2K$$

---

<sup>20</sup>Though it might be argued that changing the value of  $\sigma$  between models is allowing the assumed information to change, so that observed differences in AIC are meaningless. This would be in accordance with my own belief that for situations where there is a true model to find, it is impossible to perform model selection reliably in the absence of knowledge of the measurement process.

For badly determined parameters the information matrix may not be full rank and the trace will be the number of linearly independent parameters determined in the model with this data set. This analysis does not alter our idea of the best estimate of the parameters given the data, but it does change our idea of the conformity of the model for a particular number of degrees of freedom. Notice therefore that an unbiased estimate of the test statistic at the parameters actually estimated ( $\theta$  and not their true value  $\theta_0$ ) for an equivalent second sample of data, requires an additional correction of  $k$  as suggested by the AIC approach (Figure 2). This is also equivalent to the NIC except for an overall factor of two which appears in some texts.

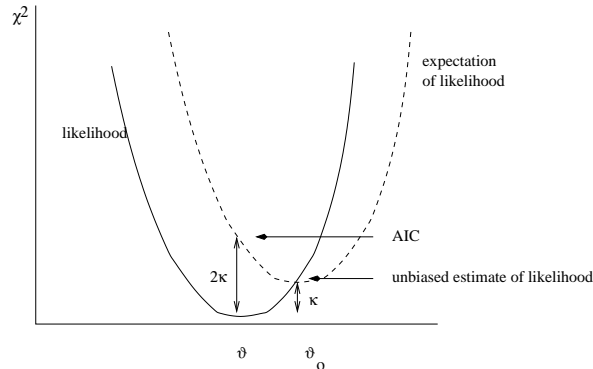


Figure 2: The relationship between AIC ( defined on the expectation of the likelihood around the true solution  $\theta_0$  ), NIC ( defined at the Likelihood estimate of the parameters  $\theta$  ) and the original likelihood function.

## Appendix C: Consequences for Least-Squares Fitting.

Starting from our assertion that perhaps we should be optimising the generalisation capabilities of the model for data generated with the assumed perturbation process, we need a way of comparing quantitative (probabilistic) predictions of data with the observed values. The Matusita measure (7) can be derived as the log probability of the similarity of two density distributions which follows directly from a definition of probability density as a limit of a Poisson data sample. As discussed above, this measure also satisfies our other restrictions on appropriate use of Likelihood, it is invariant to measurement space definition. Also, it can be interpreted (via a limit of finite samples) as having equivalent definitions of the model complexity for all probability distributions, so we avoid the problems generally associated with model selection. In other words, we are free to use this measure (based upon Likelihood as a measure of joint probability) to derive an expression expected to replace Likelihood (now to be based upon generalisation). For standard Likelihood a (joint probability) comparison of a single data point with Gaussian errors of known distribution width ( $\sigma$ ) and the residual between the model and the data ( $\Delta$ ), would be simply  $L_G = \Delta^2/\sigma^2$ . Equation 7 would imply the following (generalisation) measure for comparison of a single data point with Gaussian errors of width  $\sigma$ <sup>21</sup>;

$$L'_G = \alpha \int [\exp(-(x - \Delta)^2/4\sigma^2) - \exp(-x^2/4\sigma^2)]^2 dx$$

where  $\alpha$  is a constant introduced by the process of defining a probability distribution as a limit of a Poisson sample, and  $\Delta$  is the residual (separation in  $x$ ) between the model and the data.

We can rewrite this as;

$$\begin{aligned} L'_G &= \alpha [2\sqrt{2\pi}\sigma - 2\exp(-\Delta^2/8\sigma^2) \int \exp(-2(x - \Delta/2)^2/4\sigma^2) dx] \\ &= \alpha 2\sqrt{2\pi}\sigma [1 - \exp(-\Delta^2/8\sigma^2)] \end{aligned}$$

Using the standard expansion for an exponential we get;

$$L'_G = \alpha 2\sqrt{2\pi}\sigma [\Delta^2/8\sigma^2 - \frac{(\Delta^2/8\sigma^2)^2}{2!} + \frac{(\Delta^2/8\sigma^2)^3}{3!} - \dots]$$

<sup>21</sup>In previous work [13] the probability distribution for the model was constructed in a way which included the effects of stability of parameter estimation, it was this which made model selection possible. This modification (which prevents construction of a similarity measure for individual data points) has been excluded here for simplicity, producing the correct expression for the large sample limit where the effects of parameter instability become small in comparison to the intrinsic noise on the data.

Thus we can get first order agreement between the standard least-squares measure and this new measure by setting  $\alpha = 4/\sqrt{2\pi}\sigma$ . We can now write the new measure in terms of the old as;

$$L'_G = L_G - \Delta^4/16\sigma^4 + \Delta^6/384\sigma^6 + h.o.t. = 8 [1 - \exp(-\Delta^2/8\sigma^2)]$$

It can be observed that for conventional least-squares estimation, computed as a *product* of Gaussians of width  $\sigma$  in order to determine **the parameters which have the highest probability of generating the data set**, there is an alternative approach computed as a *sum* of Gaussians with width  $2\sigma$ , which can be used to identify **the parameters which describe the maximum quantity of data**. This simple example illustrates the difference between using generalisation rather than joint probability as the principle for parameter estimation. In this case, the new principle will not solve the model selection problem, because we have specifically eliminated the effects of sample size in this analysis. However, we might hope that this measure would now formulate the effects of measurement distribution in a valid way, so that we avoid the approximation of probabilities using variance estimates to scale the local interval. This has implications for the theoretical interpretation of algorithms which use voting as a key component, such as the Hough Transform.

## Appendix D: KL Divergence and Matusita Density Comparison

We can show how Kullback-Liebler Divergence approximates the Matusita Measure as a density overlap estimate as follows. Defining them respectively as

$$K.L. = \int p_1(x) \log(p_2(x)/p_1(x)) dx$$

and

$$M = 2 - 2 \int \sqrt{p_1(x)p_2(x)} dx$$

where the integral term is the Bhattacharyya measure. Now defining the densities w.r.t their difference

$$p_2(x) = p_1(x) + \Delta(x)$$

we have

$$K.L. = \int p_1(x) \log[1 + \Delta(x)/p_1(x)] dx$$

and

$$M = 2 - 2 \int p_1(x) \sqrt{1 + \Delta(x)/p_1(x)} dx$$

which makes explicit the similarity in form of the two expressions. These can now be written using the usual expansions giving

$$K.L. = \int [\Delta(x) - \Delta(x)^2/(2p_1(x)) + \Delta(x)^3/(3p_1(x)) - \dots] dx$$

and

$$M = \int [\Delta(x) - \Delta(x)^2/(4p_1(x)) + \Delta(x)^3/(8p_1(x)) - \dots] dx$$

So that for small  $\Delta(x)$  (to a first order difference in densities) the two measures are approximately equivalent.

In our opinion the Matusita measure is the correct way to compare probability densities, as this measure follows directly from a Frequentist definition of density in the large sample limit of Poisson processes. One can equally attempt to derive log Likelihood from Poisson statistics and then K.L. divergence from its expectation. The asymmetry of the resulting *divergence* is direct proof that this approach has gone astray. The limit of a density has to be approached for both distributions simultaneously in order to obtain the *measure*. However, in practice the K.L. divergence is far more convenient as the basis for algorithm construction. This may lead to researchers using the K.L. divergence, and trying to argue theoretical validity based upon analogies with Entropy. One could equally point out analogies between the Bhattacharyya measure and the overlap integral in quantum mechanics (used to compute the probability of a transition between quantum states), but there really is no need if a first principles proof applicable to data sampling and analysis is already known.

## Acknowledgements

Appendix A was motivated from discussions with Tim Cootes regarding early drafts of this document. The various arguments have then been refined following discussions with Jamie Gilmore and Paul Bromiley.

## References

- [1] H.Akaike, 'A new Look at Statistical Model Identification', IEEE Trans. on Automatic Control, **19**, 716, (1974).
- [2] S. Baker and R.D. Cousins, Clarification of the use of Chi-Square and Likelihood Functions to Fit Histograms, Nucl. Inst. and Meth. in Phys. Res., 221, 437-442, 1984.
- [3] R.J. Barlow. Statistics: A Guide to the use of Statistical Methods in the Physical Sciences. John Wiley and Sons, U.K., 1989.
- [4] J. Berkson, Minimum Chi-Square, not Maximum Likelihood! The Annals of Statistics, Vol. 8, 3, 457-487, 1980.
- [5] N.A.Thacker, P.Bromiley, The Equal Variance Domain: Issues Surrounding the use of Probability Densities for Algorithm Construction. Tina memo, 2004-005, 2004.
- [6] T.F.Cootes, N.A.Thacker and C.J.Taylor, Automatic Model Selection by Modelling the Distribution of Residuals, ECCV 2002 (IV), LNCS 2353, 621-635, 2002.
- [7] N.A. Thacker, P.A. Bromiley and M. Pokric, Computing Covariances for Mutual Information Co-registration. Tina-Memo, 2001-013, 2001.
- [8] A.J.Lacey, N.A.Thacker and N.L.Seed, 'Feature Tracking and Motion Classification Using a Switchable Model Kalman Filter.' Proc. BMVC, York, Sept. 1994.
- [9] B.D.Ripley, Appendix A in Pattern Recognition and Neural Networks, Cambridge University Press, 1996.
- [10] C.E.Shannon, A Mathematical Theory of Communication, Bell Systems Technical Journal, 27, 379-423 and 623-656, 1948.
- [11] A. Stuart, K.Ord and S.Arnold,. Kendall's Advanced Theories of Statistics. Volume, 2A, Classical Inference and the Linear Model, Oxford University Press, 1999.
- [12] N.A.Thacker, P.A.Riocreux, and R.B.Yates, 'Assessing the Completeness Properties of Pairwise Geometric Histograms", Image and Vision Computing, 13, 5, 423-429, 1995.
- [13] N.A.Thacker, D.Prendergast and P.I.Rockett, 'B-Fitting: A Statistical Estimation Technique with Automatic Parameter Selection.', Proc, BMVC, 283-292, Edinburgh, 1996.
- [14] N.A.Thacker, F.Ahearne and P.I.Rockett, 'The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data.' Kybernetika, 34, 4, 363-368, 1997.
- [15] M. Grendar and M. Grendar, Maximum Entropy: Clearing up Mysteries, ISSN 1099-4300, Entropy, 3, 58-63, 2001.
- [16] N.A.Thacker., Parameter Estimation for EM Mixture Modelling and its Relationship to Likelihood and EML. Tina Mmemo 2004-006.
- [17] N.A.Thacker, Curve Fitting and Image Potentials: A Unification within the Likelihood Framework. Tina Memo 2005-006, 2005.
- [18] K.P.Burnham, D.R.Anderson, Model Selection and Multi-Modal Inference, Springer NY, p56, 1998.
- [19] N.A.Thacker, Defining Probability for Science, Tina Memo 2007-008, 2007.