

Tina Memo No. 2000-003

Internal Report. This document is distributed as teaching material to our MSc and PhD students.

Tutorial: Statistics and Estimation in Algorithmic Vision.

N.A.Thacker.

Last updated
1 / 2 / 2000



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

Statistics and Estimation in Algorithmic Vision.

April 27, 2009
N.A.Thacker

1 Introduction

The purpose of this report is to introduce some basic concepts of probability and statistics necessary for computational algorithmic research. It is explained how evaluation of posterior (Bayes) probabilities is the optimal form of information for decision making processes and how under some characteristic circumstances these probabilities and related quantities can be estimated using probability and Maximum Likelihood theories. It is shown how knowledge of the standard mathematical forms allows the underlying assumptions, often not made explicit in many research papers, to be identified and evaluated.

2 Computer Vision Methodology

Computer vision is the process of extracting useful information from images in order to perform a specific task. This practical emphasis is often forgotten in some algorithmic research but is an important part of the definition. Clearly, algorithms which deliver information which is of no practical use will never be used. The first rule of algorithmic research is therefore to specify the information that we wish to obtain from the image. We can generally expect that once this information has been obtained it will be used as the basis for subsequent action resulting from a decision based on this data. The decision process may well require complex evaluation of several sources of data and for this reason the practical use of computer vision is closely related to artificial intelligence.

Given that we wish to determine information for a particular purpose we now need to know if there is an optimal way of presenting this data. Clearly a decision-making process based on delivered information will crucially require information regarding the expected success of a particular outcome given the data. There are two ways that the successful outcome can be affected, the first by a failure in action and the second an error in the data. As a consequence, an algorithm must not only deliver an estimate of the required data but also an estimate of data reliability. Anything less than this information will make subsequent decision formation unreliable and the algorithm could then never form part of a practical system. The most common form of such information is error covariance measures. Computation and manipulation of these quantities is thus fundamental to much computational research.

The most direct information regarding the successful outcome of a particular decision is the posterior (Bayes) probability. This is defined as the probability that a particular event will be true given a particular observation. Knowledge of Bayes probabilities of outcomes given a set of alternative states allows a direct assessment of attempting alternative actions. Probability theory is regarded as the only self-consistent computational framework for all data analysis and decision making. It is therefore not surprising that it forms the basis of all statistical analysis processes.

There are several ways of computing posterior probabilities and directly related quantities under limited circumstances. However, in many practical situations problems cannot be easily formulated to correspond exactly to a particular computation. Compromises have to be made, generally in assumptions about the statistical form of the processed data, and it is the adequacy of these compromises which will determine the success or failure of a particular algorithm. Clearly, therefore, understanding these assumptions and compromises is an important part of algorithmic development. We can conclude that algorithms which will work best on a particular application will be those which model most closely the underlying statistics of the measurement process and correctly propagate the effects of these through to the output of the algorithm. Algorithmic robustness goes hand in hand with getting this process correct and is something that can rarely be compromised in the interest of computational speed. One of the major criticisms of computer vision over the past few years has been due to a general lack algorithmic reliability. This has

largely been due to the neglect of the important role that statistics must play in algorithm development. Computer vision should strictly be regarded as a branch of applied statistics.

For reasons which will be discussed below, there are relatively few computational models for statistical comparison. Familiarity with the common forms of the equations often allows the statistical assumptions regarding the data to be deduced. Successful assumptions in one algorithm will generally be of use in other algorithms, often allowing rapid algorithmic development for new problems. Knowledge of the adequacy of these assumptions has a direct and beneficial effect on algorithmic evaluation, often regarded as a difficult problem in computer vision. Statistical methods can be adopted directly as the basis of algorithm evaluation and test [15, 13].

An algorithm which makes use of all available data in the correct manner must deliver an optimal result. This is not as uncommon occurrence in computer vision as may be assumed and many problems (camera calibration for example) do have optimal solutions [14]. If this can be established for an algorithm further evaluation of the algorithm on a large number of images (the other method of algorithmic evaluation) then becomes unnecessary, as the algorithm can only be bettered by one which takes account of more data. For this reason methods of optimal data combination are of fundamental importance.

It is for the purpose of increasing the understanding of these important aspects of algorithmic research that this document has been written. It covers the essentials of probability theory, maximum likelihood, covariance and error propagation. It should however be remembered that statistical validity is a necessary but not sufficient requirement for a successful algorithm. Good algorithm research methodology should also take into account practical aspects of computing cost, scalability (application of the algorithm to a larger problem domain) and practical implementation (software and hardware issues).

3 Basic Definitions of Notation

In this tutorial we will follow the following definitions of probability:

- $P(A)$ probability of event A .
- $P(\tilde{A}) = 1 - P(A)$ probability of non-event A .
- $P(A, B)$ probability of simultaneous events A and B .
- $P(A + B)$ joint probability of events A or B .
- $P(A|B)$ probability of event A given event B .
- $P(A|B, C)$ probability of event A given events B and C .
- $P(A|B; C)$ probability of event A given events B or C .
- $P(A = B)$ probability of equivalence of events A and B .
- $p(x)$ probability density over real variable x (note lower case).
- $P(x_1 < x < x_2)$ probability that data x lies between x_1 and x_2 (= $\int_{x_1}^{x_2} p(x) dx$).

An event may be defined in any way we wish. An event is always a binary measure but continuous distributions, such as a probability density distribution, can be modelled by defining measures within discrete ranges which may then become infinitesimal. Care should be taken when attempting to interpret probability equations as the notation is only a shorthand for the underlying model and does not contain all the information necessary to understand the range of allowable manipulations for a particular set of data. For example, in the sections which follow several probabilistic comparison measures will be derived. Even though each of these contain the same probabilities of events ($P_1(x), P_2(x)$) these measures should only be used for data which has the statistical properties for which they have been defined.

4 Bayes Theorem

The basic foundation of probability theory follows from the following intuitive definition of conditional probability.

$$P(A, B) = P(A|B)P(B)$$

In this definition events A and B are simultaneous and have no (explicit) temporal order we can write

$$P(A, B) = P(B, A) = P(B|A)P(A)$$

This leads us to a common form of Bayes Theory, the equation:

$$P(B|A) = P(A|B)P(B)/P(A)$$

which allows us to compute the probability of the interpretation of an event in terms of observations and prior knowledge.

5 Probability Recoding

If we consider a set of mutually exclusive events y we can write

$$P(A|y_0; y_1; y_2; \dots y_N) = \sum_i^N P(A|y_i)P(y_i)$$

Provided the set of alternatives y define all possibilities then this defines $P(A)$ the total probability of event A occurring under all circumstances. If events A and B both map uniquely to the domain in y then we can extend this equation to

$$P(A|B) = \sum_i^N P(A|y_i)P(y_i|B)$$

Which in the limit of y_i becoming a continuous variable S is

$$P(A|B) = \int ds P(A|S)P(S|B)$$

Under practical situations $P(S|B)$ is computable and $P(A|S)$ is obtainable from example data. This is the defining equation for probability recoding (integrating over a marginal variable) and finds common usage in mixtures methods for pattern classification [10] and is suitable for parallel neural network formulation [12].

6 Maximum Likelihood

Starting with Bayes theorem we can extend the joint probability equation to three and more events

$$P(A, B, C) = P(A|B, C)P(B, C)$$

$$P(A, B, C) = P(A|B, C)P(B|C)P(C)$$

For n events with probabilities computed assuming a particular interpretation of the data Y (for example a model)

$$P(X_0, X_1, X_2, \dots X_n|Y)P(Y) = P(X_0|X_1, X_2, \dots X_n, Y)P(X_1|X_2, \dots X_n, Y) \dots P(X_n|Y)P(Y)$$

Maximum Likelihood statistics involves the identification of the event Y which maximises such a probability. In the absence of any other information the prior probability $P(Y)$ is assumed to be constant for all Y . For large numbers of variables this is an impractical method for probability estimation. Even if the events were simple binary variables there are clearly an exponential number of possible values for even the first term in $P(XY)$ requiring a prohibitive amount of data storage. In the case where each observed event is independent of all others we can write.

$$P(X_n) = P(X_0)P(X_1)P(X_2)...P(X_n)$$

Clearly this is a more practical definition of joint probability but the requirement of independence is quite a severe restriction. However, in some cases data can be analysed to remove these correlations, in particular the use of an appropriate data model (such as in least squares fitting) and processes for data orthogonalisation (including principle component analysis). For these reasons all common forms of maximum likelihood definitions assume data independence.

Probability independence is such an important concept it is worth defining carefully. If knowledge of the probability of one variable A allows us to gain knowledge about another event B then these variables are **not** independent. Put in a way which is easily visualised, if the distribution of $P(B|A)$ over all possible values of B is constant for all A then the two variables are independent. Assumptions of independence of data can be tested graphically by plotting $P(A)$ against $P(B)$ or A against B if the variables are directly monotonically related to their respective probabilities.

7 Dealing with Binary Evidence

If we make the assumption that the event X_i is binary with probability $P(X_i)$ then we can construct the probability of observing a particular binary vector X as

$$P(X) = \prod_i P(X_i)^{X_i} P(\tilde{X}_i)^{\tilde{X}_i}$$

or

$$P(X) = \prod_i (P(X_i)^{X_i} (1 - P(X_i))^{(1-X_i)})$$

The log likelihood function is therefore

$$\log(P) = \sum_i X_i \log(P(X_i)) + (1 - X_i) \log(1 - P(X_i))$$

This quantity can be minimised or directly evaluated in order to form a statistical decision regarding the likely generator of X . This is therefore a useful equation for methods of statistical pattern recognition.

If we now average many binary measures of X into the vector O we can compute the mean probability of observing the distribution O generated from $P(X)$ as

$$\langle \log(P) \rangle = \sum_i O(X_i) \log P(X_i) + (1 - O(X_i)) \log((1 - P(X_i)))$$

It should be noted that this is not the log probability that O is the same distribution as P as it is asymmetric under interchange of O and P . To form this probability we would also have to test for P being drawn from the distribution O . The resulting form of this comparison metric is often referred to as the log entropy measure as the mathematical form (and statistical derivation) is analogous to some parts of statistical mechanics in physics.

8 Dealing with Data Distributions

Often when working with measured data we need to interpret frequency distributions of continuous variables, for example in the form of frequency histograms. In order to do this we must know the

statistical behaviour of these measured quantities. The generation process for a histogram bin quantity (making an entry at random according to a fixed probability) is described by the Poisson distribution. The probability of observing a particular number of entries h_i for an expected probability of p_i is given by

$$P(h_i) = \exp(-p_i) \frac{p_i^{h_i}}{h_i!}$$

For large expected numbers of entries this distribution approximates a Gaussian with $\sigma = \sqrt{h_i}$. The limit of a frequency distribution for an infinite number of samples and bins of infinitesimal width defines a probability density distribution. These two facts allow us to see that the standard χ^2 statistic is appropriate for comparing two frequency distributions h_i and j_i for large measures.

$$\chi^2 = \sum_i (h_i - j_i)^2 / (h_i + j_i)$$

This equation has the restriction that it is not defined in the region where $h_i + j_i = 0$. We can overcome this problem by transforming the data to a domain where the errors are uniform by taking square roots. This removes the denominator and leads to the common form of probability comparison metric known as the Matusita distance measure L_M .

$$L_M = \sum_i (\sqrt{P_1(X_i)} - \sqrt{P_2(X_i)})^2$$

This can be rewritten in a second form

$$= 2 - 2 \sum_i \sqrt{P_1(X_i)P_2(X_i)}$$

Where the second term defines the Bhattacharyya distance metric L_B .

$$L_B = \sum_i \sqrt{P_1(X_i)} \sqrt{P_2(X_i)}$$

which can also be written for continuous variables for application to probability densities ($p(x)$).

$$L_B = \int_{-\infty}^{\infty} \sqrt{p_1(x)} \sqrt{p_2(x)} dx$$

This second metric may be more familiar to those with a physics background as the projection operator in Quantum Mechanics, used to compare the probability density distributions for two wave functions $\Phi_1(x)$ and $\Phi_2(x)$.

$$\int dx \Phi_1^*(x) \Phi_2(x)$$

where $\Phi^* \Phi = P(x)$ is the probability of observing a particle at x . Under these circumstances this distance measure is directly interpretable as a probability that a particle drawn from the distribution $P_1(x)$ could have been consistent with a particle drawn from $P_2(x)$. Our event probabilities have now become particles in a state space but the behaviour of the statistical distributions must still have exactly the same properties.

9 Dealing with Functions

As we have mentioned previously, one way of de-correlating probabilities is to use a model. Take for example a set of data described by the function $f(a, Y_i)$ where a defines the set of free parameters defining f and Y_i is the generating data set. If we now define the variation of the observed measurements X_i about the generating function with some random error we can see that the probability $P(X_0|X_1X_2...X_N a Y_0)$

will be equivalent to $P(X_0|aY_0)$ as the model and generation point completely define all but the random error.

Choosing Gaussian random errors with a standard deviation of σ_i gives

$$P(X_i) = \int_{X_i-\delta}^{X_i+\delta} A_i \exp\left(\frac{-(x - f(a, Y_i))^2}{2\sigma_i^2}\right) dx \approx B_i \exp\left(\frac{-(X_i - f(a, Y_i))^2}{2\sigma_i^2}\right)$$

where A_i and B_i are normalisation constants, and δ is the interval in which the measurement is located. We can now construct the joint probability function

$$P(X) = \prod_i B_i \exp\left(\frac{-(X_i - f(a, Y_i))^2}{2\sigma_i^2}\right)$$

which leads to the χ^2 definition of log likelihood

$$\log(L) = \frac{-1}{2} \sum_i \frac{(X_i - f(y_i))^2}{\sigma_i^2} + \text{const}$$

This expression can be maximised as a function of the parameters a and this process is generally called a least squares fit. Whenever you encounter least squares there is therefore a built in assumption of independence and Gaussian distribution. In practical situations the validity of these assumptions should be checked by plotting the distribution of $X_i - f$ to make sure that it is Gaussian.

The choice of a least squares error metric gives many advantages in terms of computational simplicity and later we will see that it is also used extensively for definitions of error covariance and optimal combination of data. However, the distribution of random variation on the observed data X is something that generally we have no initial control over and could well be arbitrary. This may initially be seen as an overwhelming problem but in most circumstances it is possible to make distributions handleable (Gaussian) by transformation $g(X_i)$ and $g(f(a, y_i))$, where g is chosen so that the initial distribution of X_i maps to a Gaussian distribution in g^1 .

One good example of this is in the location of a known object in 3D data derived from a stereo vision system. In the coordinate system where the viewing direction corresponds to the z axis, x and y measures have errors determined by image plane measurement. However, the depth z_i for a given point is given by

$$z_i = fI / (X_{li} - X_{ri})$$

where I is the interocular separation, f is the focal length and X_{li} and X_{ri} are image plane measurements. Attempts to perform a least squares fit directly in (x, y, z) space results in instability due to the non-Gaussian nature of the z_i distribution. However, transformation to $(x, y, 1/\sqrt{2}z)$ yields Gaussian distributions and good results.

Once the least squares cost function has been obtained there are several way of minimising this quantity to obtain the maximum likelihood parameters. These include numerical minimisation methods, such as matrix inverse methods, Gauss/Newton and Conjugate gradient which work well with well behaved data and can be found in standard texts. There are also more robust methods which cope reasonably well with local minima and discontinuities which include simplex minimisation simulated annealing and genetic algorithms. Robust methods are generally computationally more expensive than their analytic counterparts. There is also the incremental time solution to the least squares problem, the Kalman Filter. This achieves a least squares solution to a multi-parameter fit by incrementally updating the optimal solution for new data. The method is related to optimal data combination, which will be discussed later. There is a further report dedicated to the description of this statistical method [7].

Under many circumstances, even after taking care to obtain Gaussian variation on the fitted quantities, there is still one final problem which needs to be addressed. This is the problem of fliers or outliers.

¹It is slightly more complicated than this if we wish to maintain a formal link between probability and Likelihood. For a more in-depth discussion see the Equal Variance document on this web site.

Fliers are the name given to the data generated by any real system which do not conform to the assumed statistical distribution. These are generally caused by complete failure of the data measurement system and generated well away from the expected mean of the distribution. If ignored they can completely dominate the fitting process giving meaningless results. For example, measurement of the distance to an object pre-supposes that we have selected the correct object. The correct way to deal with these measures is to modify the expected probability distribution to include the long tails from fliers, this leads to the branch of numerical methods known as robust statistics. The simplest way to do this which allows us to continue to use standard methods for covariance estimation and optimal data combination, which assume Gaussian distribution, is to limit the contribution to the χ^2 distribution from any data point to some maximum value $n^2\sigma^2$. This makes the assumption that the statistical distribution is constant for any gearing point greater than $n\sigma$ from the expected position. Unfortunately this process precludes the use of standard least squares solution methods and solution must generally be iterative as the gearing point will vary for each data point during parameter estimation. This process is efficiently executed by the probabilistic Hough transform for small numbers of parameters.

10 Parameter Estimation

Often describing the cost function for an algorithm is not the only part of the algorithmic solution. We also need a way of searching a space of possible models in order to find one that optimises this function. If we take the view that the parameters (or variables) in our system model correspond to unique, perhaps even physical, meaningful values then the algorithmic goal becomes that of parameter estimation (along with the associated covariances so that we can make practical use of the result).

There are many parameter estimation techniques, the commonest being those that seek to optimise a well behaved function $E(a)$ subject to change in the model parameters a . A common approach here assumes that the functional form of $E(a)$ is approximately quadratic such that

$$E(a) = E(O) + \sum_i \partial E / \partial a_i + \sum_{ij} \partial^2 E / \partial a_i \partial a_j x_i x_j + \dots$$

This equation can be seen to be directly analogous to that used later in the discussion of covariance. For simplicity we will associate these terms with the expression

$$E(a) = c - b.a + 1/2a.A.a + \dots$$

if the quadratic approximation is accurate then clearly by differentiation

$$\nabla E = A.a - b$$

which has a minimum when

$$A.a_m = b$$

thus

$$A.a = \nabla E + b$$

therefore

$$a_m - a = A^{-1}\nabla E$$

This solution suggests an iterative update scheme for estimates of a close to a quadratic minimum and this idea forms the basis of ‘quasi-Newton’ methods which provide iterative methods for approximating A^{-1} . The popular Kalman filter is effectively an iterative form of such an algorithm which provides a mechanism for optimal combination of new data into the current estimate of the parameters.

These methods, and the associated ‘conjugate gradient’ scheme require accurate derivative information which may not always be available. Under these circumstances methods which require only function evaluations such as the ‘downhill simplex’ method may be used. This method can be recommended as a

starting point for getting results quickly from a particular minimisation problem. This method and the entire issue of function minimisation is covered well in [9].

Sometimes the minimisation function is not only badly behaved but so too is the parameter space a . Under these circumstances the whole concept of functional optimisation needs to be reconsidered. At this point algorithmic research switches from an evaluation of the best optimisation functions to designing new optimisation methods. Such approaches are best typified by algorithms such as simulated annealing [5] and genetic algorithms [6]. The main difference between these approaches is that the former operates on what are effectively single point trials of the optimisation function while the latter operates with a set of interacting solutions. These interactions provide a very useful means for controlling the range of locations searched by the algorithm which is lacking from simulated annealing.

These algorithms can cope not only with multiple, discontinuous functions and parameter spaces but also non-stationary (even stochastic) evaluation functions. These properties are not however, obtained freely, the first obvious cost is the computation requirement. The success of simulated annealing is heavily influence by the choice of a problem specific annealing schedule. Genetic algorithms require that the parameters be cast in an appropriate representation (generally binary), suitable for the processes of cross-over and mutation which will drive the search efficiently through variants of the parameters in its search for the optimum. For these reasons these techniques are often looked upon as methods for solving specific long standing problems rather than automatic algorithmic solution to a class of problem.

The main point to be made on the process of function minimisation is that the choice of minimisation method (if adequate for the task) will not alter the performance of an overall algorithm. The main effect on algorithmic performance in the final evaluation lies with adequate definition of the cost function itself, which must have its roots in probability theory. If a set of data are insufficient to determine a set of parameters with sufficient accuracy for the task in hand then changing the method of optimisation will not improve the situation.

11 Covariance Estimation

The concept of error covariance is very important in statistics as it allows us to model linear correlations between parameters. For locally linear fit functions f we can approximate the variation in a χ^2 metric about the minimum value as a quadratic. We will examine the two dimensional case first, for example:

$$z = a + bx + cy + dxy + ex^2 + fy^2$$

This can be written as

$$\chi^2 = \chi_0^2 + \Delta X^T C_x^{-1} \Delta X$$

where C_x^{-1} is defined as the inverse covariance matrix

$$C_x^{-1} = \begin{vmatrix} u & v \\ w & s \end{vmatrix}$$

Comparing with the above quadratic equation we get

$$\chi^2 = \chi_0^2 + \Delta X^2 u + \Delta Y \Delta X w + \Delta X \Delta Y v + \Delta Y^2$$

where

$$a = \chi_0^2, b = 0, c = 0, d = w + v, e = u, f = s$$

Notice that the b and c coefficients are zero as required if the χ^2 is at the minimum. In the general case we need a method for determining the covariance matrix for model fits with an arbitrary number of parameters. Starting from the χ^2 definition using the same notation as previously.

$$\chi^2 = \frac{1}{2} \sum_i^N \frac{(X_i - f(y_i, a))^2}{\sigma_i^2}$$

We can compute the first and second order derivatives as follows:

$$\frac{\partial \chi^2}{\partial a_n} = \sum_i^N \frac{(X_i - f(y_i, a))}{\sigma_i^2} \frac{\partial f}{\partial a_n}$$

$$\frac{\partial^2 \chi^2}{\partial a_n \partial a_m} = \sum_i^N \frac{1}{\sigma_i^2} \left(\frac{\partial f}{\partial a_n} \frac{\partial f}{\partial a_m} - (X_i - f(y_i, a)) \frac{\partial^2 f}{\partial a_n \partial a_m} \right)$$

The second term in this equation is expected to be negligible compared to the first and with an expected value of zero if the model is a good fit. Thus the cross derivatives can be approximated to a good accuracy by

$$= \sum_i^N \frac{1}{\sigma_i^2} \left(\frac{\partial f}{\partial a_n} \frac{\partial f}{\partial a_m} \right)$$

The following quantities are often defined.

$$\beta_n = \frac{1}{2} \frac{\partial \chi^2}{\partial a_n}$$

$$\alpha_{nm} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial a_n \partial a_m}$$

As these derivatives must correspond to the first coefficients in a polynomial (Taylor) expansion of the χ^2 function then,

$$C = \alpha^{-1}$$

And the expected change in χ^2 for a small change in model parameters can be written as

$$\Delta \chi^2 = \Delta a^T \alpha \Delta a$$

12 Optimal Combination

Given two estimates of a set of parameters a_1 and a_2 and their covariances we can combine the two sets of data as follows

$$a_T = \alpha_T^{-1} (\alpha_1 a_1 + \alpha_2 a_2)$$

with

$$\alpha_T^{-1} = \alpha_1^{-1} + \alpha_2^{-1}$$

This method combines the data in the least squares sense, that is the approximation to the χ^2 stored in the covariance matrices has been combined directly to give the minimum of the quadratic form. The method can be rewritten slightly giving

$$a_T = a_1 + \alpha_T^{-1} \alpha_2 \Delta a$$

where $\Delta a = a_2 - a_1$. In this form the method is directly comparable to the information filter form of the Kalman filter.

13 Error Propagation

In order to use a piece of information $f(X)$ derived from a set of measures X we must have information regarding its likely variation. If X has been obtained using a measurement system then we must be able to quantify measurement accuracy. Then we require a method for propagating likely errors on measurement through to $f(X)$. Assuming knowledge of error covariance this can be done as follows

$$\Delta f(X) = \nabla f^T C_X \nabla f$$

The method simply uses the derivative of the function f as a linear approximation to that function. This is sufficient provided the expected variation in parameters ΔX is small compared to the range of linearity of the function. As an example we can take the Poisson distribution itself which for large numbers is expected to have a standard deviation of \sqrt{N} where N is the mean of the distribution. We will call a sample random variable from this distribution s . If we now construct a new measure given by

$$t = \sqrt{s}$$

then we can show, using a simplified form of error propagation for one parameter, that the expected variance on t is given by

$$\begin{aligned} \Delta t &= \frac{\partial t}{\partial s} \Delta s \\ &= \frac{-1}{2\sqrt{s}} \sqrt{s} \\ &= \frac{-1}{2} \end{aligned}$$

Thus the distribution of the square-root of a random variable drawn from a Poisson distribution with large mean will be constant. This result will be used to generate the Matusita probability distribution comparison metric defined later.

When the problem not permit algebraic manipulation in this form due to significant non-linear behaviour in the range of $\Delta f(X)$ or functional discontinuities then numerical approaches may be appropriate. These techniques are often referred to as Monte-Carlo approaches because they make use of random number generation techniques to generate sample distributions.

14 Data Orthogonalisation

As we have already said, most probabilistic comparison metrics make an assumption regarding independence of each component of data X_i . Under any practical circumstance the data delivered by a system may not have this property. It is then that we may need to preprocess the data to remove correlations. We can define the correlation matrix

$$R = \sum_j (X_j - X_m) \otimes (X_j - X_m)$$

where X_j is an individual measurement vector from a data set and X_m is the mean vector for that set. This equation is directly comparable to that for generation of the covariance matrix with the derivatives replaced by a finite difference and there are several mathematical similarities. Provided that the data has been generated by one process (assumed to be characterised by the mean X_m) then we can extract from this matrix estimates of linear correlation between vector components. The best form for this information is as the orientation of principle axes of variance of the original distribution of data in parameter space. It can be shown that these orthogonal axes correspond to the eigenvectors V_k of the matrix R . Solution of the eigenvector equation

$$RV_k = \lambda_k V_k$$

can be done analytically for small matrices but is a complex problem for large matrices. However, there are standard matrix manipulation techniques which make this computationally tractable. The method known as Singular Value Decomposition (SVD)[9] approximates a matrix by a set of orthogonal vectors W_l and singular values w_l .

$$R = \sum_l \frac{1}{w_l^2} W_l \otimes W_l$$

If we multiply both sides of the equation by one of these vectors W_k

$$RW_k = \sum_l \frac{1}{w_l^2} W_l \otimes W_l \cdot W_k$$

we see that the singular vectors satisfy the eigenvector equation with

$$\lambda_k = \frac{1}{w_k^2}$$

Thus SVD determines the axes of maximal variation within the data. A limited approximation to the full matrix R^*

$$R^* = \sum_l^{l_{max}} \frac{1}{w_l^2} W_l \otimes W_l$$

gives an optimal approximation to the matrix R in the least squares sense $(R - R^*)^2$, allowing the selection of a reduced number of orthogonal descriptor variables. This process is sometimes referred to as principle component analysis and is useful for limiting the effects of numerical stability and singularity during the process of matrix inversion. Knowledge of the equivalence of eigenvectors and principle axes justifies our initial assertion that the eigenvectors align with the axes of maximum variance. These methods can also be applied to standard covariance matrices in order to identify the principle axes of variation.

These and other techniques including the Karhunen Loeve transform and optimal subset identification are common in the field of statistical pattern recognition [16]. However, on pattern data these methods require linear correlation at least locally. If this correlation is non-linear alternative techniques, such as neural networks, have to be used.

15 Neural Networks

When the area of neural networks re-emerged as a popular topic in the mid 80's much was claimed about the expected capabilities regarding flexibility, suitability for system identification and robustness. Most of these claims were subsequently shown to be optimistic. However, one problem that neural networks are relatively good at is non-linear orthogonalisation. A neural network when trained on an appropriate form of data with the correct algorithm will approximate Bayes probabilities as outputs.

The mathematics describing this process is given in [8] but a more intuitive argument is as follows. Each input vector pattern X defines a unique point in input space. Associated with each data point is the ideal required output, for example a binary output classification. As the number of samples grows large the number of examples of data in the region of each point also grows large. If training with a least squares error function the target output for each point in pattern space will be the mean of local values. For a binary coding problem the mean value is the Bayes probability.

Thus when using a least squares training function and training with binary class examples in the limit of an infinite amount of data and complete freedom in the network to map any function the network will approximate Bayes probabilities as outputs.

In practice however, there will not be an infinite data set and there will be no proof that a particular, a-priori, choice of network architecture will map the required problem. There are also other problems associated with neural networks and the research still has far to go. However, the methods are of some merit for problems of a small scale.

In particular, no mention has yet been made of optimal combination of probabilities. Given $P(A|B)$ and $P(A|C)$ can we compute $P(A|BC)$? We can clearly solve this problem provided these probabilities are independent by simple multiplication. If however the measures are correlated there is no standard statistical method for this process. This is unfortunate as we would expect a modular (AI) decision system to need to solve this task. Standard neural network architectures trained [4] in the standard way will however approximate $P(A|P(A|B)P(A|C))$ for the reasons described above. This suggests the eventual possibility of hierarchal reasoning systems.

16 Inverse Statistic Identification

If we examine standard image processing techniques it is possible to deduce the statistical assumptions made about the data. For example if we take the block matching algorithm commonly used for image compression and object tracking we see that the most stable of the popular forms uses the Mean Absolute Difference (MAD) between pixels in a block. If this algorithm is to conform to a standard Maximum Likelihood definition then the use of a sum implies that we are dealing with log probabilities. The conclusion is that the MAD algorithm assumes independent double sided exponential distributions for the grey level data around the matching image patch values. This distribution has quite long tails and so we would expect it to be quite robust to fliers (ie incorrect data from occluded regions). If we wanted to check the adequacy of this distribution assumption we could generate a true distribution from a set of example data. Clearly, if we wanted to improve the algorithm we may even decide to use the observed distribution directly rather than rely on the ad-hoc choice of double sided exponential. Clearly we would need to know the expected improvement from this approach before we could make a judgement regarding the trade off against computational simplicity.

Another algorithm which has found widespread application and has a reputation for robustness is the Hough transform. In this algorithm an array representing the allowable range of possible model parameters is incremented at all valid parameter locations for each piece of data. There are two possible statistical interpretations of this algorithm. The first is an accumulation of probabilities towards a hypothesis by addition and the other is a Maximum Likelihood log probability. In this second interpretation there is a built in assumption about a uniform background distribution across all possible parameter values, this is because an exponentiated zero is one. Thus once again there is an account taken of fliers in the data and we would expect this algorithm to be robust. Also, if we wished, we could improve the performance of the Hough transform by entering the log probabilities properly as has been suggested in [11].

Finally, we can examine the Pairwise Geometric Histogram representation for object recognition [3]. The use of a dot product metric for comparison of histograms implies a Bhattacharyya distance metric which is appropriate for comparing probability density distributions. This is consistent with the definition of these histograms.

17 Conclusions

This report has explained the importance of statistical approaches to algorithmic development and performance. I started by defining some general concepts of probability theory and explained the important role of the independence assumption. I have then discussed the properties of binary measures, model fitting and data distributions and the related analysis procedures. I have described what I consider to be some of the essential methods necessary for research and have explained how these methods are related. This relationship is summarised in Figure 1. Also discussed was their likely limitations and practical solutions including data transformation, orthogonalisation and robust statistics. The report then goes on to explain how successful algorithms can be reverse engineered to identify the statistical assumptions so that we can get a better understanding of algorithmic performance.

It should be apparent to the reader that all of these methods have been developed from a broad range of research areas including physics, pattern recognition, artificial intelligence and statistics. The limit of this list is only because of the bias from my own particular background. Any subject which involves the construction of statistical models may generate a useful result. Results in speech recognition in particular are likely to have a direct impact on computer vision research. Research in the field of algorithm development cannot be done in isolation and there are many methods available in other subjects which may find a useful application in your research.

I conclude by saying that in general any successful algorithm will have some statistical properties which are well suited to the analysed data. A knowledge of data behaviour and the appropriate statistical techniques is therefore crucial for algorithm interpretation, analysis and development. I would encourage this approach to algorithmic evaluation wherever possible as a means of improving the quality of research and the robustness of resulting algorithms.

References

- [1] D.Booth, N.A.Thacker, M.K.Pidock and J.E.W.Mayhew. 'Combining the Opinions of Several Early Vision Modules Using a Multi-Layered Perceptron.' *Int.Journal of Neural Networks*,2,2/3/4,June-December,75-79,1991.
- [2] A.C.Evans, N.A.Thacker and J.E.W.Mayhew, A Practical View Based 3D Object Recognition System, IEE conference on neural networks, Brighton, 1993a;
- [3] A.C.Evans, N.A.Thacker and J.E.W.Mayhew, The Use of Geometric Histograms for Model-Based Object Recognition, *Proc. 4th BMVC93*, Guildford, 21-23 pp429-438 Sept. 1993b
- [4] D.S.Broomhead, D.Lowe, Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks. RSRE Memorandum 4148 March 1988.
- [5] D.B.Fogel, An Introduction to Simulated Evolutionary Optimisation. *Neural Networks*, 5, 1, pp3-15, 1994.
- [6] D.E.Goldberg, *Genetic Algorithms (In Search of Optimisation and Machine Learning)*, Addison Wesley pub. Reading, Jan 1989.
- [7] A.J.Lacey and N.A.Thacker, The Kalman Filter: A Tutorial, ESG Report 93/7, 1993.
- [8] M.D.Richard and R.P.Lippmann,'Neural Network Classifiers Estimate Bayesian a posteriori Probabilities.' *Neural Computation*,3,461-483,1991.
- [9] W.H.Press B.P.Flannery S.A.Teukolsky W.T.Vetterling, *Numerical Recipes in C*, Cambridge University Press 1988.
- [10] P.Pudil,J.Novavicova and J.Kittler, Automatic Machine Learning of Decision Rules for Classification Problems in Image Analysis, *proc. BMVC 93*, vol 1, pp 15-25, 1993.
- [11] R.S.Stephens, A Probabilistic Approach to the Hough Transform, *Proc. 5th. Alvey Vision Conference*, pp 55-60, 1989.
- [12] N.A.Thacker and J.E.W.Mayhew, 'Designing a Network for Context Sensitive Pattern Classification.' *Neural Networks* 3,3, 291-299, 1990.
- [13] N.A.Thacker, P.Courtney. 'Statistical Analysis of a Stereo Matching Algorithm.' *Proc. British Machine Vision Conference.* ,Leeds, 316-326, 1992.
- [14] N.A.Thacker and J.E.W.Mayhew, Optimal Combination of Stereo Camera Calibration from Arbitrary Stereo Images, *Image and vision computing*, pp.27-32 vol 9 no 1 Feb., 1991.

- [15] R.Lane, N.A.Thacker and L.Seed, 'Stretch Correlation as a Real-Time Alternative to Feature Based Stereo Matching Algorithms.' Submitted to Image and Vision Computing 1993.
- [16] S.Watanabe, Patern Recognition : Human and Mechanical, John Wiley and Sons, 1985.