Tina Memo No. 2010-002
Presented as a seminar to the Department of Computational Science, University of Birmingham, 18/4/10

Seminar: A Quantitative Methodology for Design of Computer Vision Algorithms.

N. A. Thacker

Last updated
18 / 04 / 2010

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

Seminar: The Origins of a Quantitative Methodology for Design of Computer Vision Algorithms.

N.A.Thacker, ISBE, University of Manchester.

The seminar will start with basic properties of probability and quantitative use of conditional notation, including the restrictions on Bayesian methodologies when considered in a frequentist framework. I will then explain how the use of this approach can be used to understand Likelihood as a design principle, and the way that common errors in its use during algorithm design can be identified and avoided.

I will give answers to simple problems which are regularly claimed either impossible or a reason for introducing subjective probability in standard AI texts. I will illustrate the ideas with practical examples of more complicated computer vision algorithms intended for robotic, scientific and medical applications.

Motivation.

There are many methods relating to probability which are used regularly as the basis for algorithm design.

eg:

- Likelihood
$$\sum_i log[p(d_i|m)]$$
- Variational Method
$$Q^2/var(Q)$$
- Expectation Maximisation,
- Maximum A-Posteriori Optimisation (MAP)
$$p(d|m)p(m)$$

How are they related?

Do they guarantee optimal (quantitative) performance?

Definition of Probability.

Kolmogorov's axioms

- $P(E_i) >= 0$
- $P(E_i \; or \; E_j) \; = \; P(E_i) \; + \; P(E_j)$
- $\sum_{\forall i} P(E_i) \; = \; 1$

although you can use these axioms to derive several common results, for example;

$$P(E_i) \; = \; 1 \; - \; P(\tilde{E}_i)$$

they are devoid of any real meaning.

One way to overcome this is to introduce Frequentist probability;

$$n/N \rightarrow P$$

Conditional notation; $P(a|b)$ is probability of a given b.

Joint probability $P(a, b) \; = \; P(a|b)P(b)$

Independence $P(a, b) \; = \; P(a)P(b)$

This notation only describes correlation. (not causality)

If a causes b, and b causes a then the resulting conditional expressions are non-stationary and meaningless.

[TINA Memo 2007-008, Defining Probability for Science.]

Probability and Probability Density.

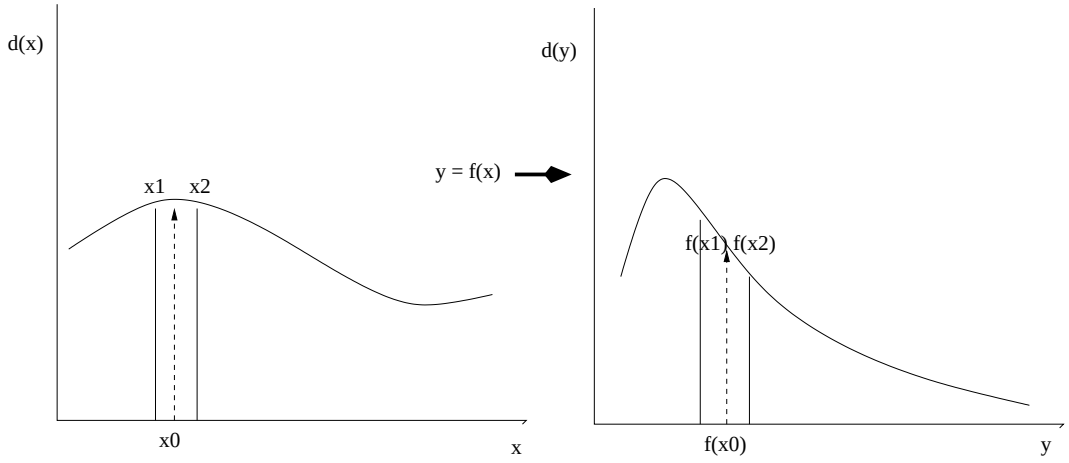
For discrete events, probability is a well defined concept.

For continuous variables we can define a probability density.

Probability densities p are related to probabilities P via an integration.

$$P(x_1 < x < x_2) \;=\; \int_{x_1}^{x_2} p(x)\; dx$$

Non-linear transformations change density behaviour ie: density maxima are not unique.



Given an interval, the transformation of parameters has no effect on the computed probabilities,

$$P(x_1 < x < x_2) \;=\; P(f(x_1) < f(x) < f_(x_2))$$

as the interval itself transforms to preserve the result.

For continuous variables, $P(x)$ is only ever a shorthand for

$$P(x_1 < x < x_2)$$

$P(theory|result)$ can never be considered a physical theory unless we insist that $P(theory)$ is constructed using an appropriate interval.

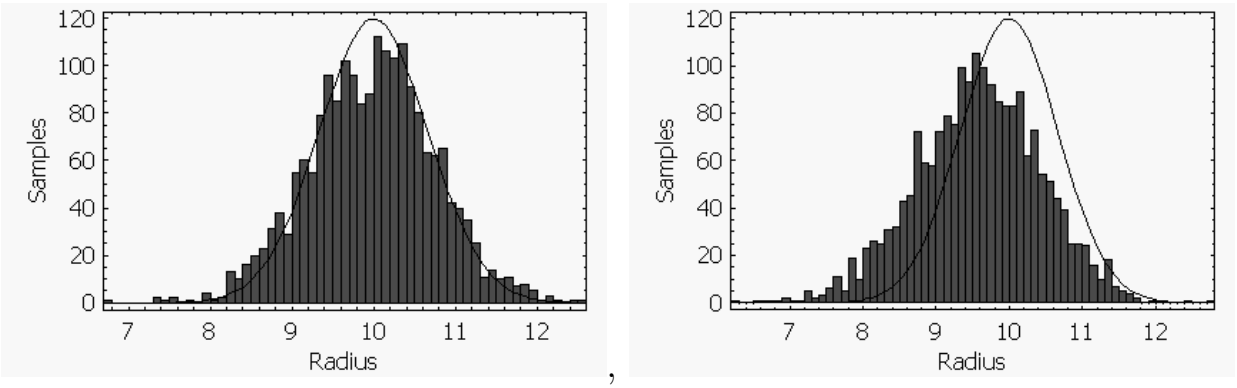
Example: Circle Size Estimation

Imagine that we are doing an experiment and wish to estimate the mean size of a set of circles.

We will use histograms of data distribution in order to estimate probability densities and use these to construct a Likelihood.

Due to the quadratic relationship between the parameters, the likelihoods computed from the empirical distributions of radius and area *will be different*, that is,

$$\frac{p(r_i|\theta)}{2\pi r_i} = p(a_i|\theta) \Rightarrow \sum_i \ln p(r_i|\theta) \neq \sum_i \ln p(a_i|\theta).$$



Size estimates using area have been converted back to radius for comparison.

Data is shown along with the generator distribution.

Estimated parameter covariances are also affected (try it!).

Quantitative algorithms need to be based upon quantitative probability!

note: the above result occurs because $p(a_i|\theta)$ changes as a function of θ, estimating $p(a_i|\theta)$ from samples will give us a residual distribution of fixed shape. This is a good reason for using Bland-Altman plots. Plots here courtesy of Paul Bromiley

The Probability of Getting a Measurement.

We need a way of relating the density distribution to a meaningful probability.

Choose the interval to be proportional to measurement accuracy σ, ie; the probability of getting data like that we have **measured**.

$$P(x_0 - \kappa\sigma_x < x_0 < x + \kappa\sigma_x) \;=\; \int_{x_0-\kappa\sigma_x}^{x_0+\kappa\sigma_x} p(x)\; dx$$

As $\kappa \to 0 \qquad P(x) \propto p(x)\sigma_x$

This value gives a consistent interval which can be estimated from data, regardless of how we define the measurement system.

Equivalently we can choose the space $y \;=\; f(x)$ in which σ_y is constant ie: f is an equal variance transform.

It has been noted previously that use of equal variance spaces is mathematically equivalent to using a Jeffreys' Prior (questionable). [Kendal and Ord]

[TINA Memo 2004-005, The Equal Variance Domain]

Defining Uninformative Priors.

Define a MAP based optimisation function data Likelihood $\times$ prior

$$P(model|data) \propto P(data|model)P(model)$$

Suppose that the world circumstances change, and we find we can have an additional source of knowledge $P'(model)$. If it is independent we should be able to write this as

$$P(model|data) \propto P(data|model)P(model)P'(model)$$

For consistency if both sources of data are uninformative, both approaches must give identical results

$$P(data|model)P(model) \propto P(data|model)P(model)P'(model)$$

Unless we want two different definitions of uninformative

$$P(model) = P'(model)$$

So that $P(model)^2 = P(model) \rightarrow P(model) = 0$, *const* or.. only a uniform prior has no effect on parameter estimation!

We can use uninformative priors to define ranges of allowed model parameters.

The notation $P(model)$ implies a term conditional on nothing.

A hypothesis will often constrain our priors such that they are either true $P(model) = 1$ or false $P(model) = 0$

Beyond this, uninformative priors cannot have structure.

Quantitative Bayes Priors.

Imagine that we wish to formulate a MAP estimate using a non-uninformative $P(model)$.

$$P(model|data) \propto P(data|model)P(model)$$

Where do we get this from? [Popper]

In a quantitative framework $P(model)$ must be the probability of the specified model being the generator of data in a real world sample

$$P(model) \rightarrow P(model|sample) = constant$$

eg: medical image segmentation

If the model parameters A are continuous variables then this is an analogous problem to previously and strictly

$$P(model) \rightarrow P(A - \sigma_A < A < A + \sigma_A|sample) = P(A|sample)$$

To get this from sampled density distributions $p(A|sample)$ we must put

$$P(A|sample) \propto p(A|sample)\sigma_A$$

Assuming $P(data|model) \rightarrow P(data|A, sample) = P(data|A)$, any attempt to use sampled distributions of parameters to define a 'prior' results in a MAP formulation equivalent to

$$P(data|A)P(A|sample) = P(data, A|sample)$$

which is just the joint Likelihood of seeing the data and the parameters but not yet $P(model|data)$ or $P(A|sample, data)$

Effects of Priors on Covariances.

Often we wish to combine estimates from multiple experiments, what are the consequences when using MAP?

Imagine two Likelihood experiments computing

$$L_1 \;=\; P(data_1|A) \quad and \quad L_2 \;=\; P(data_2|A)$$

For independent data

$$L_1(A) \times L_2(A) \;=\; P(data_1, data_2|A)$$

which can be approximated by use of parameter covariances. However, for MAP we have

$$M_1(A) \;=\; P(data_1|A)P(A|model) \;\; and \;\; M_2(A) \;=\; P(data_2|A)P(A|model)$$

If we try to simple combine these by multiplication

$$M_1 \times M_2 \;=\; P(data_1|A)P(data_2|A)P(A|model)^2$$

Direct combination of MAP estimates using covariances will double count the prior, unless the prior is uninformative.

In general we cannot combine the data from MAP estimations without first undoing the effect of the priors.

This might be impossible if the prior is buried in the algorithm.

Conclusion; design modules using Likelihood with uninformative priors and don't use priors until the end of an analysis.

Poisson Statistics.

For a Poisson distribution, the probability of observing n samples when the generator of the distribution has mean μ is

$$P(n) \;=\; \frac{exp(-\mu)\mu^n}{n!}$$

When we observe 0 samples of a particular event, we know that this does not mean that the event will never occur.

There are more ways of generating $n \;=\; 0$ than just $\mu \;=\; 0$.

eg: "Principle of Maximum Ignorance"; $\mu \;=\; n \;+\; 1$

Including the interval term described above, the **probability** of μ generating n can be written as

$$P(\mu|n) \;\propto\; exp(-\mu)\mu^n$$

$$=\; (interval) \;\times\; (density) \;\times\; (uninformative\ prior)$$

where $(interval) \propto \sqrt{\mu}$ and $uninformative\ prior \;=\; const$ so that

$$(density) \propto exp(-\mu)\mu^n/\sqrt{\mu}$$

the expected mean frequency μ' is given by

$$\mu' \;=\; n \;+\; 1/2$$

When we observe n events the most likely Poison generating model has a mean of $n \;+\; 1/2$... NOT n or $n+1$!
(see Tina memo 2009-008)

Binomial statistics.

Analogous with the previous case, we know that a sample of $n = N$ from N only puts an upper limit on the most likely ratio, it does not guarantee that we will continue to see data with 100 % probability.

Binomial distribution

$$P(n|\mu, N) \;=\; \frac{N!}{n!(N-n)!}\mu^n\;(1-\mu)^{N-n}$$

$$\sigma \;\propto\; (\mu \;-\; \mu^2)^{1/2}/\sqrt{N}$$

Therefore, for fixed N

$$density \;\propto\; \mu^n(1 \;-\; \mu)^{N-n}/(\mu \;-\; \mu^2)^{1/2}$$

ie:

$$\mu' \;=\; \frac{n+1/2}{N+1}$$

A new definition of frequentist probability

$$(n+1/2)/(N+1) \rightarrow P$$

[Von Mises] A probability can never be exactly 0 or 1.

[Laplace] "Rule of Succession" $(n{+}1)/(N{+}2)$, computed as the expectation of μ from the conventional Likelihood.

.

Example: Reference Class problem.

In the end, even a strict frequentist position involves subjective analysis... The **reference class problem** *illustrates the intrusion of subjectivity. Suppose that a frequentist doctor wants to know the chances that a patient has a particular disease. The doctor wants to consider other patients who are similar in important ways - age, symptoms, perhaps sex - and see what proportion of them had the disease. But if the doctor considered everything that is known about the patient - weight to the nearest gram, hair colour, mother's maiden name, etc. - the result would be that there are no other patients who are exactly the same and thus no reference class from which to collect experimental data. This has been a vexing problem in the philosophy of science.*

[Russell and Norvig, Artificial Intelligence]

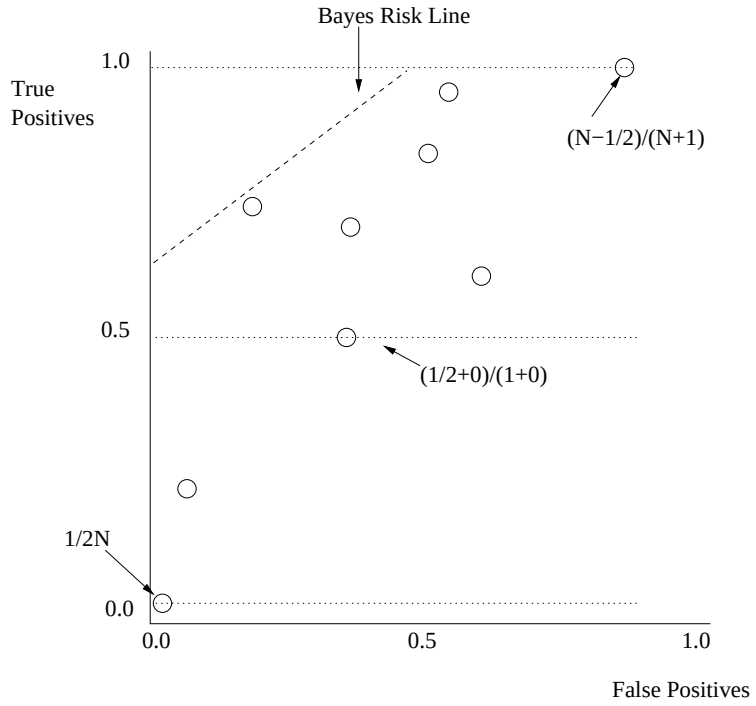
This problem needs us to appreciate several issues;

- How to deal with small numbers (and 0 in particular).
- There are multiple solutions, which arise according to how we want to consider classification errors.

Bayes Risk, the consequences of making misclassifications.

Reference Class Solution.

The ROC solution (candidate samples shown as circles for $P(c' = c|subset, c)$ vs $P(c' = c|subset, \tilde{c})$



The Bayes risk calculation quantifies the consequences of particular forms of error (ie: total deaths).

Locations of equal risk lie along a line.

Use of the above theory prevents us from making errors for $n = 0$, $N = 0$ and $n = N$.

A subjective approach prevents us from assessing the quantitative outcome of any particular decision! (Not acceptable in clinical practice.)

[Tina Memo 2009-008, Avoiding Zero and Infinity in Sample Based Algorithms]

Derivation of Likelihood.

The probability of observing a given number of samples n_i within each discrete region within a region with non-zero probability is;

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

The log probability for the model in terms of the data is $n_i \;=\; 0$,$n_i \;=\; 1$,$n_i \;=\; 2$ ...;

$$log(P(\lambda)) \;=\; \sum_i^{N_0} log\ P(0, \lambda(X_i)) + \sum_i^{N_1} log\ P(1, \lambda(X_i)) + \sum_i^{N_2} log\ P(2, \lambda(X_i)) + ...$$

the first being N_0 empty cells, the second being N_1 cells containing one sample and N_2 two samples, and so on..

Equally we can write this as;

$$\begin{aligned} log(P(\lambda)) \;=\; & \sum_i^{N_0+N_1+N_2+...} log\ P(0, \lambda(X_i)) \;+ \\ & \sum_i^{N_1} log\ P(1, \lambda(X_i)) - log\ P(0, \lambda(X_i)) \\ & + \sum_i^{N_2} log\ P(2, \lambda(X_i)) - log\ P(0, \lambda(X_i)) \;+\; ... \\ = \sum_i^{N_1+N_2+...} & n_i log(\lambda(X_i)) \;- \sum_i^{N_0+N_1+N_2+...} [\lambda(X_i)] \;+\; k_1 \qquad (1) \end{aligned}$$

where k_1 is a constant.

[Fermi] "Extended Maximum Likelihood"

Statistical Consistency.

Generalisation of (1) to continuous variables (x) is now straight forward, but requires us to define the way in which probabilities for discrete values relate to those for continuous variables

$$\lambda(X) \rightarrow \lambda(x)$$

and so probability densities from a parametric model $p(x|A)$.

$$\lambda(x) \ \propto P(x|A) \ \propto \sigma(x,A)p(x|A)$$

The expectation of the probability of the data given the model, for continuous valued observations x follows from (1),

$$< logL > \ = \ \int p(x) log(\sigma(x,A)p(x|A))\ dx \ - \ \beta \int p(x|A)\ dx \ + \ k_2 \qquad (2)$$

where β is a constant .

We can then ask what distribution $p(x|A)$ must take in order to maximise Q.

Differentiating (2) with respect to a specific $p(x|A)$ (at a single value of x) and setting to zero we get

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

In order for the technique to regenerate the data distribution with the available model parameters, we must have $\sigma(x,A) \ = \ constant$

Statisticians would call this a homoscedastic space [Kendal and Ord].

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

$$L = -\frac{1}{N}\sum_{n}^{N} log(\sigma(x_n, A)p(x_n|A)) \qquad (3)$$

subject to the constraints;

- Fixed density integral

$$\int p(x|A)\ dx$$

will generate consistent estimates of parameters from a Likelihood which is quantitatively related to the probability of generating the data.

The method suggested by Fisher, and found in text books for integral normalised probability densities is the following

$$L = -\frac{1}{N}\sum_{n}^{N} log(p(x_n|A)) \qquad (4)$$

s.t.

$$\int p(x|A)\ dx = 1$$

This isn't quite the same!

As we have seen it leads to multiple interpretations.

.

Example: χ^2 not Likelihood.

If we apply (4) to Gaussian samples

$$p(x_i|A) \;=\; \frac{1}{\sqrt{2\pi}\sigma} exp(-(x_i - A)^2/2\sigma^2)$$

then for N measured data x_i

$$L \;=\; \sum_i^N (x_i - A)^2/2\sigma^2 \;+\; log[\sqrt{2\pi}\sigma]$$

Note: the second term is not constant, and varies as a function of σ.

However, if we apply (3)

$$L \;=\; \sum_i^N (x_i - A)^2/2\sigma^2 \;+\; log[\sqrt{2\pi}]$$

where the second term is now constant, and the first term is a χ^2 with N degrees of freedom.

Expressions derived from (3) are quantitatively related to probability, those from (4) are not.

J. Berkson, Minimum Chi-Square, not Maximum Likelihood! The Annals of Statistics, Vol. 8, 3, 457-487, 1980.

[TINA Memo 2005-008, Beyond Likelihood]

Example: Localising Objects.

For a model and set of camera parameters θ we can write the probability of detecting an edge at location (x, y) with orientation ψ as;

$$P(x, y, \psi|\theta) = P(x, y|\theta)P(\psi|x, y, \theta)$$

Computing $P(x, y|\theta)$:

Define a feature localisation Likelihood based upon the probability of detecting an edge, ie: The edge strength g must be above threshold and greater than at least n of its neighbours h.

Generalise the Binomial sampling formula for noisy data to obtain a soft rank.

$$R(g > h \in N) \;=\; \frac{1/2 \;+\; \Sigma_i^N \, erf(\frac{g-h_i}{\sqrt{2}\sigma_I})}{N+1}$$

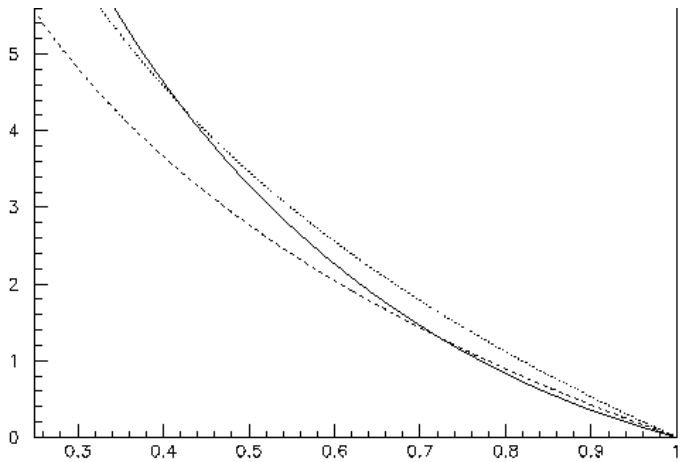
Taking a bootstrap approach, the probability that it is greater than 6 or more neighbours (P_e) is then

$$P_e \;=\; P_{(6/8)} \;+\; P_{(7/8)} \;+\; P_{(8/8)} \;=\; 28R^6 \;-\; 48R^7 \;+\; 21R^8$$

the corresponding hypothesis test H_e is given by the normalised integral

$$H_e \;=\; 12R^7 \;-\; 18R^8 \;+\; 7R^9$$

This can be approximated by a polynomial $P'_e \;\approx\; R^M$ estimated in real data.



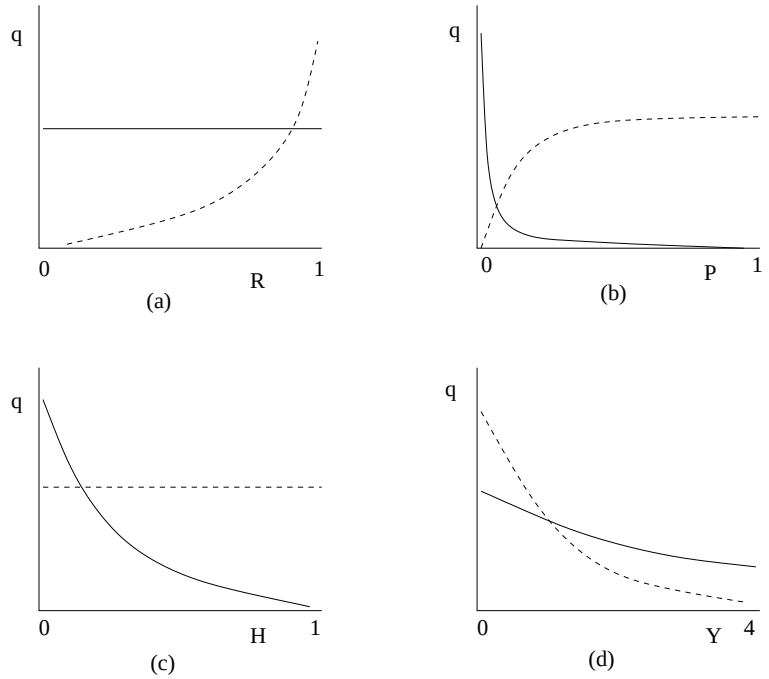
General transformation of variables $y = f(x)$ results in a transformation of density distribution $q_x \rightarrow q_y$ according to

$$q_y = q_x / \frac{\partial y}{\partial x}$$

so that (for example) the histogram of all data in an image (uniform in R) for variable $P = R^M$ is given by

$$q_P \propto 1/\frac{\partial P}{\partial R} = 1/R^{M-1} = P^{(1-M)/M}$$

Histogram	**All pixels**	**Edges**
R	constant	$R^{M'}$
$P = R^{M'}$	$P^{(1-M')/M'}$	$P^{1/M'}$
$H = R^{M'+1}$	$H^{-M'/(M'+1)}$	constant
$Y = -log(H)$	$exp(-Y/(M'+1))$	$exp(-Y)$



$$P(x, y|\theta) = P_t R^{M'}$$

Computing $P(\psi|x,y,\theta)$:

Compute the feature orientation likelihood based upon error propagation (variational method).

$$-ln\ p(\psi|x,y,\theta) = \frac{(\phi(x,y) - \psi(x,y,\theta))^2}{2\ var(\psi)}$$

$$+\ \frac{1}{2}\ log(2\pi\ var(\psi))$$

$$\psi = \arctan(\frac{dI/dx}{dI/dy}) \ = \ \arctan(\frac{u}{v})$$

Given equal errors δ on the derivatives u and v

$$\delta_\psi^2 = (\frac{v\delta}{v^2+u^2})^2 + (\frac{u\delta}{v^2+u^2})^2$$

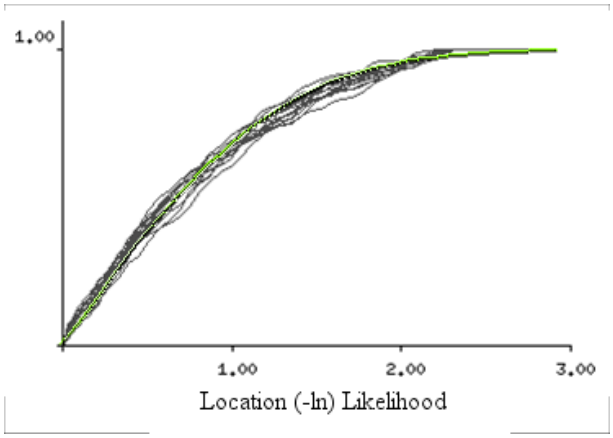
Since the variance is the square of the error, and taking 'r' to be the edge magnitude

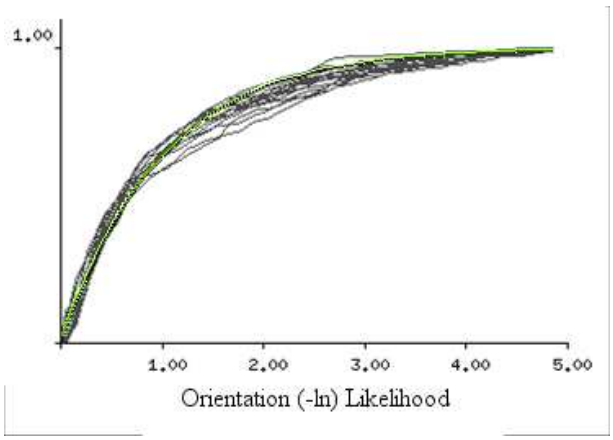
$$var(\psi) \approx \frac{\delta^2}{r^2}$$

i.e. the error on the local edge orientation is inversely proportional to the edge strength r .

$$log[P(\psi|x,y,\theta)] \ \propto \ r^2\ (\phi(x,y) - \psi(x,y,\theta))^2$$

Confirm Models using the cumulative distribution functions.

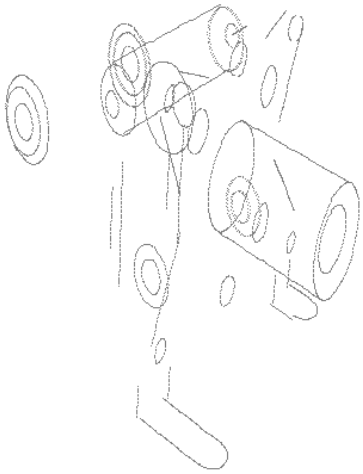


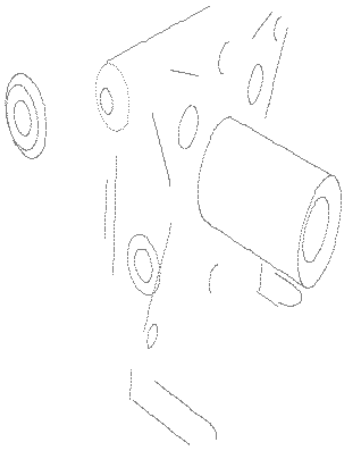


The orientation term destabilises alignment if the propagated error is allowed to vary during optimisation!

This confirms our assertion that quantitative use of Likelihood precludes changing the density function during optimisation.

[TINA Memo 2007-011, A Methodology for Constructing View-Dependent Wireframe Models]





[TINA Memo 2006-007, Quantitative Verification of Projected Views Using a Power Law Model of Feature Detection]

Recommendations.

Make a distinction between probability and density.

Any theoretical justification for an algorithm must not change under non-linear transformation.

Use “equal variance” or “interval term”

Keep the Likelihood distribution fixed.

Try to avoid using anything other than uninformative priors for quantitative data analysis.

see www.tina-vision.net for details.

P.A. Bromiley and N.A. Thacker, The Effects of an Arcsin Square Root Transform on a Binomial Distributed Quantity., Tina Memo 2002-007.

N.A.Thacker, P.Bromiley, The Equal Variance Domain: Issues Surrounding the use of Probability Densities for Algorithm Construction., Tina Memo, 2004-005.

N.A.Thacker, Beyond Likelihood., Tina Memo, 2005-008.

S. Coupe and N.A. Thacker, Quantitative Verification of Projected Views Using a Power Law Model of Feature Detection. , Tina Memo, 2006-007.

N.A.Thacker, Defining Probability for Science., Tina Memo, 2007-008.

N.A.Thacker, Avoiding Zero and Infinity in Sample Based Algorithms., Tina Memo 2009-008.

Acknowledgements:

Paul Bromiley, Simon Coupe, Visvanathan Ramesh (Siemens Corporate Research, USA)

Notes:

Fishers own specifications for Likelihood as a tool for scientific analysis include invariance under parameter transformation, but not invariance under data transformation. The Likelihood definition is suggested as a form which will satisfy his requirements. The mathematical form chosen (not derived) fails to do this but is still described the same way to this day in text books.

Popper says that once you have taken the subjective definition of probability no further manipulation is possible, it certainly seems impossible to derive Bayes theorem without being able to write $P(A|B)P(B) = P(B|A)P(B)$ (you cannot meaningfully equate non-quantitative expressions)

Even before Fisher, Laplace had derived the principle of succession but missed out the interval term, in accordance with what we now call Likelihood. Others have noted that this approach is self contradictory.

Fermi derives Extended Likelihood, again missing an interval term. Otherwise this is the nearest thing we have to a derivation.

Jeffreys observes that starting from Bayes Theorem the mathematical form required to achieve the characteristics that Fisher originally wanted for Likelihood can be achieved using Jeffreys Priors. However, these terms are neither consistent with the definition of a prior, nor the axioms of probability. This amounts to a proof by absurdum of the failure of subjectivity in scientific use. Jeffreys Priors have never been universally accepted as 'the' uninformative priors.

Berkson notes that conventional Likelihood is not quantitative, χ^2 seems more fundamental, but he does not develop an explanation. The interval/equal variance interpretation provides one.

Kendal and Ord note that equal variance Likelihoods are equivalent to using a Jeffreys prior, but go no further.

Starting from a subjective definition, Jaynes uses priors to 'fix' specific problems which arise with Frequentist methods. At no point does he make a distinction between probability and probability density, thus dooming his frequentist calculations to failure. At the same time he claims that the Bayesian approaches solve the problems for quantitative use (note contradiction). However, people like these ideas, it increases the scope for publication.

Russell and Norvig use similar arguments, based upon an inability to understand how to get Frequentist methods to work, to advocate general use of subjectivity in AI. Again people like this, though presumably they do want to quantitatively "minimise the number of times" that their robot will walk of the edge of a cliff, rather than assessing whether such a move might be "a rather bad idea".

Expectation Maximisation is derived from conventional Likelihood, but in a peculiar twist of fate, the proof of convergence demands that the Likelihood distributions remain fixed during the Maximisation step, thereby obtaining consistency with quantitative use of probability almost by accident.

.

12 Reasons why people use subjective probability.

Mathematical validity.

BUT it can be argued that the convergence issue is a mathematical irrelevance which arises as a consequence of mathematical sequences not being truly random. It also appears that we can't derive Bayes theory for subjective axioms ($p(a|b)p(b) = p(b|a)p(a)$ is an assertion). Once we have asserted $p(a|b)p(b) = p(b|a)p(a)$ and $p(c|b)p(b) = p(b|c)p(c)$ does $p(c|a)p(a) = p(a|c)p(c)$? Not unless the prior terms are non-arbitrary.

Origin of Likelihood.

BUT although many believe that Bayes theorem is the origin of likelihood, it can be derived without using Bayes as the joint probability of the parameters which would be most likely to generate the data. Logical arguments suggest that uninformative priors should be uniform, so that Likelihood $P(params|data)$ and MAP ($P(params|data)$ will be identical, but for logical constraints on parameter ranges. The arguments for this seem to be valid for both quantitative and subjective probability. Conventional (subjective) use, which allows the use of real values in conditional expressions, is not quantitatively valid use of notation and consequently varies under non-linear transformation of the parameter space.

$P(params|data) \neq P(data|params)P(params|cohort)$

Improved Algorithm performance.

BUT adding extra degrees of freedom to a flawed methodology can always be used to fine tune performance (legitimised hacking). and quantitative algorithmic tests are frequentist!

Algorithmic stability.

BUT the need to avoid numerical problems is completely different to the issue of defining the correct cost functions. We could have used a regulariser (turn off at solution as previously).

Local Minima

BUT the information in the data is only summarised with use of uninformative priors. Otherwise, use of priors will bias the results. However, it could be that the main benefit here is to constrain the optimisation process so that it converges more reliably to the required minimum. Therefore we can use them in the same way as Lagrange multipliers and turn them off at the minimum, achieving the same constraint without needing a subjective interpretation.

Use of prior knowledge.

BUT subjective Bayesian approaches are not the only ways to incorporate prior knowledge. Combination of likelihoods from two sets of data $P(data1|params)P(data2|params)$ also does the same job. Uninformative priors can be used to encode logically allowable ranges. We can also use sample distributions over parameter values to construct a nested likelihood. The only thing which seems to be left out of this list is the use of arbitrary (guessed) priors to embody vague constraints. Aside from the obvious criticism that this is unscientific, the question is; are these necessarily subjective, or just guesses at frequentist distributions?

Solving ill posed problems.

BUT if we are 'honest' derived data has no additional information when summarised as an estimate with its error covariance (parameters constrained only by the prior strictly should have infinite error). Theoretical arguments, that the use of priors is not valid, can be constructed from the observation that estimates are biased and cannot be combined without first removing the effect of the prior. It might be argued that in the absence of additional quantitative data all we are achieving here is avoiding numerically unstable regions of the parameter space (eg. NaN and Inf) see below. If this is the case, then as with the local minim, we should turn the priors off at the minimum or set them to be so weak at to have no noticeable effect. This process therefore makes no contribution to the available information and does not require legitimisation from subjective probability as it is equally justifiable in a quantitative approach.

Model selection.

BUT logical consideration of the model selection task seems to point to the aim being the identification of the model which has best predictive capabilities (ie. optimal generalisation). However, this requires terms which are a function of the data, not just the parameters. Consequently, subjective priors can only achieve model selection as a first order correction, calibrated to equivalent statistical samples. The quantitative optimisation of generalisation is possible, and requires knowledge of the measurement errors in order to set the information limit

available for necessary parameters.

Quantitative methods don't work

BUT the use of probability notation with free interchange of probability and probability density contradicts the basic definition of probability notation. This should only be applied to sets of (non-ordered) logical events. Misuse of notation will lead to inconsistent and contradictory results. These should not be taken as a failure of frequentist probability but instead as a failure of understanding and methodology. Appropriate use of notation, including defining probabilities as the integral of a pdf over a meaningful interval, should avoid any such problems, eg: binomial counting, Likelihood. If quantitative probability really doesn't work someone will have to tell the all the physicists that their experiments only give the results their theories predict by accident. Also, we should stop confirming the success of subjective methods via quantitative (frequentist) testing.

We don't know the correct distributions and measurement errors.

BUT there is a difference between true subjectivity and guessing at frequentist terms. There will always be some level of uncertainty regarding probabilities which model sample distributions. Contrary to the pattern recognition mantra, there are many ways to obtain the noise distribution, we can sample the distributions in data, and measurement characteristics can be built in during system design or understood via repeatability studies/error propagation/Monte Carlo/factor analysis. As we know that knowledge of the error distribution is fundamentally required in order to obtain a unique solution we are left with a simple choice; Sample them or don't claim the method is unique/optimal/valid.

Subjectivity is inevitable eg: Reference Class, rain tomorrow.

BUT there is a quantitative solution for the reference class and quantitative prediction can be made based upon knowledge (monte-carlo). Subjectivity generally arises when we introduce Bayes Risk, but we can still be quantitative regarding the outcome of any decision policy.

The way we think.

BUT learning and evolution are frequentist tasks.

Number 13:

If we agree to eliminate ways of testing theories it is easier to get work accepted by reviewers and published.

There will be more of them, it will be popular... popularity is the new truth!

Does any of this make a difference?: The Bayesian Fudge and other Convenience Foods.

By definition, any approach to algorithm design which not only allows but encourages people to be subjective can never deliver a unique theory for data analysis. Bayesian priors are left to the whim of the researcher to select, according to whatever problem they are most interested in. Any problem encountered with an algorithm (numerical stability, inappropriate Likelihood formulation, inconvenient mathematical formulae, model selection, insufficient data to solve the chosen problem) can be 'fixed' in restricted situations with careful choice of prior. You might therefore describe the approach as "effective" rather than "fundamental".

However, a strictly quantitative approach should deliver a self- consistent theory and unique solutions to well defined data analysis tasks. We would expect the results of such research to be applicable to other algorithms and datasets. The conventional Bayesian approach which makes no distinction between between probabilities and their densities (eg: Jayne) can deliver a multitude of different but equally "valid" alternatives within the theory. As quantitative agreement between assumed distributions and data cannot be used to assess validity (and is actively avoided on subjective principle) these can only be compared through algorithmic shoot outs rather than on the basis of deeper design principles or theory. Though this is good source of publications, as the results do not transfer to other work, it is not effective use of intellectual effort. This scatter gun approach to algorithm design also demands the highest standards in experimental testing (eg: elimination of experimenter bias). In my view many publications are currently uninformative or worse misleading. Now, do you think this makes a difference?

.

The Effects of Non-Linear Transformations on Expressions Constructed Using Probability Densities.

Try out the test of transformation invariance on your favourite probability expression.

For $P(x) = p(x)\delta x$

applying the non-linear transformation

$x \rightarrow y = f(x)$

s.t. $\partial y/\partial x$ is positive.

Then $P(y) = p(y)\delta y = P(x)$

1. Likelihood (as defined by Fisher)

$$l_x = -\sum_i log[p(x_i)]$$

$$l_y = -\sum_i log[p(y_i)] = -\sum_i log[p(x_i)\partial x/\partial y] \neq l_x$$

The standard definition is non-unique and consequently statistically inconsistent (biased).

alternatively the joint probability of observing the measured data ($\delta x \propto \sigma_x$)

$$L_x = -\sum_i log[P(x_i)] = -\sum_i log[P(y_i)] = L_y$$

This is a consistent estimator.

2. Entropy

$$S_x = -\sum_i P(x_i)log[P(x_i)] = -\sum_i P(y_i)log[P(y_i)] = S_y$$

this is the unique concept of entropy as used in physics and analogous to Shannon Entropy.

but alternatively the expectation of the Likelihood

$$s_x = -\int p(x)log(p(x))dx$$

$$s_y = -\int p(y)log(p(y))\ dx = -\int p(x)log[p(x)\partial x/\partial y]\ dx \neq s_x$$

This is the concept of entropy as applied in many pattern recognition tasks which, in the absence of a mechanism to define a specific domain, is arbitrary and therefore unscientific (c.w. mutual information).

3. KL Divergence

$$KL = \int p_1(x)log[p_2(x)/p_1(x)]\ dx = \int p_1(y)log[p_2(y)/p_1(y)]\ dy$$

The extra term in the logarithm cancels the problematic term present for entropy. This is a divergence not a measure, ie: its only property is that it can identify identical distributions. It is not the correct way to compare PDFs.

4. Bhattacharyya

$$B_x = \int \sqrt{p_1(x)}\sqrt{p_2(x)}\ dx$$

$$B_y = \int \sqrt{p_1(y)}\sqrt{p_2(y)}\ dy = \frac{\int \sqrt{p_1(x)}\sqrt{p_2(x)}\ dy}{\partial y/\partial x} = B_x$$

This is a measure, as the square-root transforms the space to one of equal variance so that the dot-product like construction is monotonically related to a Euclidean distance (Matusita), c.w. overlap integrals in quantum mechanics. This is the correct way to compare PDFs.