Tina Memo No. 2016-004
Internal Report

"Adopting the Conventional Mis-use of Probability Notation..."

N.A. Thacker and S.Notley.

Last updated
27 / 03 / 2016

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

Abstract

Some are well aware that when deriving expressions in probability the notation generally adopted is not sufficient to fully describe what we are doing. On several occasions I have even seen researchers introduce a derivation with the phrase "Adopting the Conventional Mis-use of Probability Notation...", or similar.

This document is intended as a tutorial for students and RA's, on the use of conditional notation for the derivation of cost functions for use in algorithm design. We attempt to provide a more logical approach to use of notation, whilst showing the consequences of conventional usage. As in other documents on these web pages we make a clear distinction between probabilities and densities (P and p). We use this notation to examine the theoretical origins of various probability based optimisation functions and the problems associated with domains of applicability.

"Adopting the Conventional Mis-use of Probability Notation..."

Probability theory is defined for discrete variables and not continuous ones, which are described by densities. For scientific purposes results should be invariant to redefinition of parameters and data. However, whilst probabilities behave as required, parameter densities do not. We should therefore use probabilities and not densities as the basis for algorithmic cost functions. In order to show this for the simplest possible cases we will derive expressions for a two dimensional system, for data x and parameter y. However, these expressions will generalise simply to multi-dimensional systems.

In general we need to work with continuous variables, and our knowledge of distributions is in the form of densities p and not probabilities P. We can choose to adopt the usual convention for probability for values which lie within $\pm\Delta$ of our parameters[1].

$$P(x \pm \Delta_x | y \pm \Delta_y) \;\equiv\; P(x|y) \propto p(x|y)$$

but given probability is defined only for events and not continuous variables, this constitutes an abuse of notation, and hides some potential problems with valid application of probability theory. As will be explained below, using a notation which makes no distinction at all between densities and probabilities encourages such errors.

Background

Using the above notation, we find in general statistical text books that

$$P(x \pm \Delta_x) \;=\; \int_{x-\Delta_x}^{x+\Delta_x} p(x)\; dx$$

We are also taught that the distinction between probabilities and densities is important, but once we have been told this, the issue is rarely mentioned again.

In practice, data analysis requires us to work with conditional notation. For conditional probabilities the situation is slightly more complicated. We only know that the observed variables must lie in the observed ranges according to the joint density $p(x, y)$. Then we must write

$$P(x \pm \Delta_x | y \pm \Delta_y) \;=\; \frac{\int_{x-\Delta_x}^{x+\Delta_x} \int_{y-\Delta_y}^{y+\Delta_y} p(x,y)\; dx\; dy}{\int_{-\infty}^{\infty} \int_{y-\Delta_y}^{y+\Delta_y} p(x,y)\; dx\; dy}$$

$$= \frac{\int_{x-\Delta_x}^{x+\Delta_x} \int_{y-\Delta_y}^{y+\Delta_y} p(x,y)\; dx\; dy}{\int_{y-\Delta_y}^{y+\Delta_y} p(y)\; dy} \qquad (1)$$

Under some circumstances simplifications are possible. If $p(x, y)$ is approximately constant or linear over x and y then

$$P(x \pm \Delta_x | y \pm \Delta_y) \;\approx\; \frac{4\; p(x,y)\Delta_x\Delta_y}{2\; p(y)\Delta_y}$$

and the Δ_y terms cancel. When taking the limit of these intervals $\Delta \to \delta$, such approximations become exact.

$$P(x \pm \delta_x | y \pm \delta_y) \;=\; 2\; \delta_x\; \frac{p(x,y)}{p(y)} \;=\; 2\; \delta_x\; p(x|y) \qquad (2)$$

[1] Equivalently we could use either $x \pm \Delta_x$ or $x \in (x - \Delta_x, x + \Delta_x)$, we choose the former mainly for convenience, although this does restrict the theory which follows to compact distributions over x and y.

However, as we will see below, taking a limit does not allow us to entirely forget the intervals. "Adopting the conventional misuse of probability notation", leads to expressions such as

$$P(x|y) \propto p(x|y) \qquad (3)$$

the difference between discrete events and continuous variables is ignored and we work with densities as though they are probabilities. We can see above that this is inaccurate (if a density function highly nonlinear as a function of y over the implied interval), and wrong (if the interval over data is allowed to vary), when our theoretical expressions are not written to include the δ_x terms.

Likelihood Functions

In the context of use of a cost function, equation (2) is directly equivalent to use of Likelihood, here $p(x|y)$. We can see that a link to probability is maintained provided that interval terms (Δ_x) are constant. This can often be safely assumed if the data is fixed, as is the case in the majority of Likelihood based analyses. In addition, the interval terms relating to y play no role, and likelihood is therefore invariant to monotonic parameter transformation ($f(y)$). i.e.

$$p(x|y) = p(x|f(y))$$

It makes no difference how conditional information is represented. This is perfectly logical.

However, if we try to formulate a Likelihood function which contains a transformation of the data $g(x)$, and this includes something as simple as a scaling or a more general monotonic transform (such as an Anscombe transform, used to obtain data with equal variance) then the interval terms for data need to be transformed in accordance with the change of limits in equation (1).

$$P(g(x) \pm \delta_g|y \pm \delta_y) \propto p(g(x)|y)\ \delta_{g(x)}$$

This can be more conveniently expressed as

$$p(g(x)|y)\ \frac{\partial g}{\partial x}\ \Delta_x \qquad (4)$$

In this form Δ_x can once again be assumed fixed and ignored after taking the limit. If we do not do this then successive evaluations of the cost function during optimisation will not support meaningful comparison over changes in g.

Bayesian MAP estimation

As a consequence of the above, expressions such as

$$P(x, y) \propto p(x|y)p(y) \qquad (5)$$

where no distinction is made between densities and probabilities, cannot be taken as generally correct, or even meaningful. For equation (5), the nearest form of optimisation function is found when using Bayesian approaches, i.e. a Likelihood term with an additional "prior". Here we can observe that

$$\begin{aligned} P(x \pm \Delta_x, y \pm \Delta_y) &= P(x \pm \Delta_x|y \pm \Delta_y)P(y \pm \Delta_y) \\ &\approx 4\ p(x|y)p(y)\Delta_x\Delta_y \end{aligned}$$

as $\Delta \to 0$

$$P(x \pm \delta_x, y \pm \delta_y) = 4\ p(x|y)p(y)\delta_x\delta_y$$

but note also as $\Delta \to 0$

$$P(x \pm \delta_x, y \pm \delta_y) = 4\ p(x, y)\delta_x\delta_y$$

so that

$$p(x, y) = p(x|y)p(y)$$

as has been derived by others from measure theory and used for the derivation of equation (2) above. On the face of it, it looks like we have simply exchanged the probabilities in the original theory with densities. We now have two valid alternative expressions which take the form of a Likelihood multiplied by a prior term, either a prior probability or a prior density. You may conclude from this that we no longer need to make a distinction,

but it really does make a difference which is used. For example, uninformative prior probabilities are uniform (flat), whilst uniformative prior densities are constructed from a uniform probability using the Jeffreys prior (see below). It is possible to determine the consequences of choosing either by considering the effects of parameter transformation, and the former generates a scaled probability as output, whilst the latter generates a density.

Which you need for your own analysis will depend upon what you are trying to do. A density is suitable for a MCMC style Bayesian analysis of the distribution over parameters. **However, this does not mean that $p(x, y)$ is now a good basis for a cost function.** In this case a cost function which excludes the interval terms (i.e. $p(x|y)p(y)$) does not maintain a link to probability for the parameters. As a consequence the location of the maximum density will depend upon the specific choice of parameter definition (e.g. $f(y)$), making any point estimate made this way arbitrary. Probabilities, not densities, are needed for the construction of point estimators.

Changing the Prior Won't Always Fix Things

In a quantitative (frequentist) framework, the original behaviour can be regained by putting in appropriate terms to correctly describe the change in interval in equation (1). As

$$p(x|y)\ p(y)\ \delta_y\ =\ p(x|f(y))\ p(f(y))\ \delta_{f(y)}$$

In a "strong" Bayesian framework (which eschews the frequentist axiom), interval limits cannot be discussed meaningfully (they involve measurable quantities) but the non-invariance of expressions is recognised. Such problems are then solved instead by the use of Jeffreys "priors" $p'(y)$, as a modification to the prior density. In effect priors are defined such that

$$p(x|y)\ p'(y) = p(x|f(y))\ p'(f(y))$$

This solution may be fine for simple extensions of Likelihood, but will not solve the problems of transforming the data itself, as outlined above. This can only be understood in terms of interval change for x, as priors (as implied by the notation $p(y)$) can strictly only be a function of y.

Even if a MAP estimate is something you want, it is not advisable to try to construct one using a prior density (for example obtained from data samples) without integrating over the uncertainty in the parameters first. Interestingly, this amounts to dividing (rather than multipling) by the Jeffreys prior, which is not the way people normally expect to see them used[2].

Equal Variance Domains

Indeed, the only way that densities can be freely interchanged with probabilities is to specify a unique domain for all data x and parameters y and forbid any alternative representations of information. The domain which regenerates Likelihood solutions for parameter estimates would be the equal variance space (following appropriate Anscombe transforms). Only then can we claim in general that the joint probability of x and y can be generally written as;

$$P(x \pm \delta_x, y \pm \delta_y) \propto p(x|y)p(y)$$

This can equally be defined as using homoscedastic variables and is also the domain which eliminates skewing of Likelihood functions (and so conventional definitions of parameter bias). Once again, outside of our own documents, we know of no general discussion of the need for equal variance transforms in order to use formulae expressed using "the conventional misuse of notation" as the basis for cost functions.

Show Me the Errors

Equivalently, we can define the required intervals from the outset as proportional to our measurement accuracies, rather than hoping to ignore them (note, the Jeffreys prior can also be interpreted in terms of the minimum variance bound on the parameters). The derivative terms suggested in equation (4) above, for monotonic transformations of compact intervals over x, is better understood in general as the application of error propagation applied to measurement uncertainties (σ).

$$p(g(x)|y)\ \frac{\partial g}{\partial x}\ \Delta_x\ \propto\ p(g(x)|y)\ \frac{\partial g}{\partial x}\ \sigma_x$$

[2] As a consistency check, if the sample distribution actually matched the Jeffreys prior, the two terms would cancel so regenerating maximum Likelihood as the MAP estimator. Though from a frequentist point of view we are simply using the integration interval and this does not imply the physical existance of an uninformative prior density distribution.

i.e.

$$\sigma_x \propto \Delta_x \qquad and \qquad \sigma_{g(x)} \approx \frac{\partial g}{\partial x}\,\sigma_x$$

This more general strategy for dealing with changing definitions of variables ensures that the intervals scale in the required way for any transformation of variables, and can be considered as a definition of a fixed amount of the associated "information".

For the case of Gaussian random variables, making the interval scale with the expected error has the effect of cancelling out the normalisation term normally seen in Likelihood. i.e.

$$p(x|y) = \frac{1}{2\pi\sigma_x} exp(-(x-y)^2/2\sigma^2)$$

whilst

$$p(x|y)\sigma_x \propto exp(-(x-y)^2/2\sigma^2) \qquad (6)$$

Taking the logarithm of (6) generates a Chi-Square (χ^2) statistic, which is invariant to rescaling of x, making it suitable as an absolute goodness of fit. Other examples include the "Variational Method", which assess how well data conforms to an assumed model on the basis of propagated (approximately) Gaussian errors.

Examples

Several cost functions are suggested for quantitative analysis of data, and observed to work under different circumstances. Three are specifically worthy of mention;

- Likelihood: where normalisation terms are included as part of a density definition (equation 3).
- Chi-squared (χ^2): where (c.w. Likelihood) nomalisation terms are omitted (e.g. equation 6).
- The variational method: where the square of a contraint function is scaled by the propagated variance on the constraint.

Below are some simple estimation tasks which illustrate applicability.

Variance of a Sample

The variance of a sample of Gaussian random variables with known (exact) mean can be estimated using Likelihood, but a value of infinity will be obtained if we minimise the χ^2. The famous bias on the variance (which leads to the $N-1$ correction), is due to uncertainty on the mean.

Fitting Scaled Data

If we have two measured curves with Gaussian noise (or equivalently image regions of pixel values) and wish to fit them assuming using a model which includes data scaling, then the variational method (using error propagation to estimate the uncertainty), will estimate the correct scaling whilst Likelihood will not.

Fitting Bland-Altman Plots

A Bland Altman plot can be fitted to an error model using Likelihood and not χ^2. However, if we first apply a monotonic transform to the data and try to adjust it, to obtain a homoscedastic density model, Likelihood will fail to identify the correct transformation and a cost function based upon equation (4) is necessary.

Summary

In each of the above examples (including the variance estimate with error on the mean), using an interval based interpretation of the required cost function (which takes account of uncertainties on data and parameters) regenerates the working approach. Specifically, Likelihood only works if the data is not modified during cost-function construction.

Such errors are likely to occur when trying to construct a cost function to fit data density distributions and we choose to non-linearly transform the input data (in order to better fit a Gaussian for example). This warning applies to popular approaches such as Kernal methods and Probabilistic PCA, for example, which would need to optimise suitably modified cost functions in order to work correctly. To date we know of no published account of this issue.

There is one simple test which we can always apply in order to ensure that our work is not invalidated by these problems. Any suggestion for a cost function must be consistent for all equivalent definitions of data and model. This can be tested by assessing the effects of monotonic transformations. If our results depend entirely on the way we choose to do things then this is unscientific.

A more involved approach requires that we confirm the distributions generated by our theory at each stage in analysis. In particular we can check that distributions over estimated parameters behave as theory predicts. Any theoretical approach which does not make quantitative predictions of estimation uncertainty should accordingly be treated with scepticism. We leave it as an exercise for the reader to consider how optimisation functions such as "Mutual Information" rise to this challenge.

In order to head off the usual response to such comments, we say this; empirical evidence that our methods "seem to work" (in a finite number of cases), is not enough to validate a methodology proposed for general use (induction). If it were then we might expect either Likelihood or chi-squared statistics to work for every problem having seen them work for one. Whilst a single logical argument which can be shown to contradict the theory can be used to entirely discredit it (falsification).

As a consequence, we deduce that optimising

$$p(x, y) \;=\; p(x|y)p(y)$$

is in general invalid for use as a scientific tool. So too is the optimisation of

$$p(g(x)|y)$$

when varying data transformations $g(x)$ or scalings are allowed. Our inability to appreciate this issue is probably hindered by conventional abuse of notation, which encourages us to believe that probabilities and densities can simply be interchanged.