The Use of Modern Statistical Techniques in Physics: Comments on Talks at the PHYSTAT Meeting 17/11/05.

N. A. Thacker.

Last updated
10 / 1 / 2006

TiNA

WWW.TINA-VISION.NET

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

I should start by saying that there is a lot of confusion within the area of statistics. This confusion can be traced back to a variety of underlying assumptions, definitions of probability and random selections for formulae which seem to have the correct properties and have been adopted for convenience. A lot of these approaches are then given names and calibrated via Monte-Carlo. The selection of specific techniques often includes some degree of arbitrariness. It is a subject which is generally presented as set of black box methods and people are never told how particular methods were derived. The reason probably being that if this were attempted the user community would be even more confused. There doesn't seem to be much of an audience for material which attempts to explain the processes involved. I expect most physicists would have misgivings about this. The following observations are therefore not intended to be a criticism of the meeting. In fact I can understand well how the material presented in these talks would naturally be accumulated by dipping into the standard statistical textbooks, even the good ones such as Kendall's [7].

Louis Lyons: Do's and Don't with Likelihoods.

This talk presented a number of problems with conventional use of Likelihood some of which were genuine qualitative features of Likelihood and others which were related to the approximation of the likelihood around the minimum as a Gaussian. All of the empirical examples given were valid and worth knowing if you are using Likelihood as a black box. Many of these features can be explained on the basis of the underlying assumptions, had the method been properly motivated from quantitative probability to begin with. The derivation of Extended Maximum Likelihood from the joint probability of a Poisson sample would have been useful here for this purpose [Appendix A]. Taking a fully quantitative frequentist approach, it is particularly important to make a distinction between probability (which makes definite statements regarding likely frequency of occurrence) and probability density (which does not). This issue arises again in the talk by the second speaker.

This talk also touched on issues of selection of appropriate models and the answer to the final question suggested the need for approaches which were sensitive to the shape of the data distribution. This issue is particularly interesting and has been addressed to a large extent by a number of the statistically experienced researchers in the subject of neural networks. You can actually reasonably define the best estimate of any parameters AND model to be that which gives the best ability to predict unseen data drawn from the same statistical distribution. Joint probability, from which likelihood is derived, is only an approximation to this process. First order corrections are generally accepted extensions to likelihood, Akaike Information Criteria (AIC) and Neural Information Criteria (NIC) [2] [Appendix B], so I don't think this is a contentious statement.

Roger Barlow: New Developments in Bayesian Priors.

This presentation started by introducing the idea of multiple definitions of probability, and the need for correct interpretation when summarising scientific results. The speaker pointed out, correctly in my view, that there is not much benefit to be gained in setting this difference up as a confrontation. However, the differences between interpretations of probability is a source of great confusion. The problem goes even deeper than just inconsistencies between strict Bayesian and Frequentist.

The standard definition of a frequentist probability is as the large sample limit. A second limit must be taken in order to produce a probability density. In fact there are several ways that we can take the sample limit, which depend upon how we define the measurement process, and follow from conventional statistics. For example we can take Poisson limits, Multi-nomial limits or a Bi-nomial limit. Which of these we take makes a difference if we are attempting tasks which compare probability densities. This has led to a multitude of probability similarity measures (and their use in algorithms) and no general understanding of when each is appropriate. I would say the method for determination of the information used in the construction of Jeffreys priors [3] (discussed below) has such problems. It uses something called the Kullback-Liebler Divergence when you could argue it should be using the Bhattacharrya measure [4].

The speaker went on to talk about use of Bayesian methods for summarising data and the difficulty of defining suitable prior distributions and the potential for multiple interpretations of data. He then

discussed the use of Jeffreys priors as a way of resolving this ambiguity but also stated that this approach was not generally accepted by statisticians.

The principle of Fisher information would appear to be a good one, so we have to ask ourselves why the use of Jeffreys priors is not more widely accepted. I believe that this is because Jeffreys priors are generally used to fix the problem of estimation bias, (a frequentist problem not a Bayesian one). In addition you could have applied the principle of Fisher information not to the parameters space but to the measured data, during the formulation of the likelihood. In fact this is exactly the way to maintain the quantitative link between probability and probability density. It is possible to show that this eliminates parameter estimation bias [Appendix C]. This additional modification to likelihood may also explain to the problem of "Punzi Bias" which was mentioned briefly at this meeting. Statisticians have several common data transformation methods, mapping Binomial and Poisson distributions to better approximate Gaussian distributions of fixed variance, which can be used to achieve this [6]. Jeffreys priors appear to approximate this correction, but by doing so in the parameter space this correction is sensitive to the specific sample being analysed. Thus application of the Jeffreys priors to different datasets from an equivalent experiment generates different corrections. This analysis is backed up by the multitude of papers published describing various prescriptions for prior selection in areas such as speech recognition, computer vision and neural networks, all contradictory and all supported by empirical evidence. Dependency on the data sample means that a Jeffreys prior is not a prior at all. In summary, Frequentists will not accept the method because they see no need for it, and Bayesians do not accept it because they see its not a true prior. In fact as the existence of the method and its apparent empirical success really does not help the cause for either the frequentist or Bayesian camp, the method and the issues it raises are simply ignored. No academic statistician builds a career from publishing Likelihood anymore.

Glen Cowen: Nuisance Parameters and Systematic Uncertainties.

A good talk and accurate in every respect. Methodologically what was described is the accepted way to apply Bayesian analysis in the situation where you would like to improve the stability of estimates using prior information. I remain to be convinced that integrating over the unknowns removes the dependency on the assumed prior distribution, as implied in the talk.

I want here to mainly repeat what I said in the discussion. The use of Bayesian methods for the analysis of data will at least require that the prior distributions are specified in any publication. There are associated with this two problems, one I mentioned, that publications can only use consistent priors if results are to be related, thus restricting the application of these techniques to aspects of the data which are common to multiple experimental designs. The other thing I forgot to say is that use of consistent priors will produce analysis results which cannot be combined easily. It requires first the removal of the effect of the prior in order to work back to the objective evidence available in the data (ie: the Likelihood term), if double counting of the prior is to be avoided. A simple example is;

$$P(D_1, D_2|C) = P(D_1|C)P(D_2|C) \quad but \quad P(C|D_1)P(C|D_2) \propto P(D_1|C)P(D_2|C)P(C)^2$$

where is $P(D|C)$ is the conditional probability of the data D given the model assumption C.

Bill Murray: Machine Learning

This talk presented a very brief overview of various methods including Neural Nets, Support Vector Machines, Optimal Observeables, and Boosted Trees. The speaker made a very important observation which is correct in every respect. If the problem you are trying to solve is that of separating multi-dimensional data into groups (eg: signal and background) then you can order the locations populated within the space according to the ratio of signal to background. Separation of the data is optimal when done on the basis of this single (local) ratio variable. The speaker was slightly inaccurate however to refer to this process as use of Likelihood, as strictly this is a frequentist application of Bayes Theorem. This makes a big difference, as the importance of maintaining an understanding of the difference between probability and probability density becomes less important when we start taking ratios of densities. The

problems of Likelihood identified by the first speaker therefore do not apply to pattern recognition tasks. The subsequent descriptions of algorithms was good and I thought fair, with the important issue of dimensionality discussed appropriately.

The discussion following the meeting made one additional point, following from the observation that all of these methods are aiming to achieve the same goal; the identification of the boundary between samples defined as an iso-contour on the probability density ratio. All algorithms can be expected to have the same limiting performance, the inherent separability of the data at this boundary. (I have to say this is fine provided you can measure the background distribution, otherwise techniques based upon confidence intervals following the methods suggested by Neyman [5] are more appropriate). We can therefore have no a-priori expectation that any one technique will be any better in general than any other for arbitrary problems. Any technique with the ability to map this boundary with a complexity limited by a particular number of degrees of freedom (parameters) is likely to be just as effective as any other with the same level of complexity.

The difficulty comes with selection of parameters within the method, which ideally you would like to come from a theoretical understanding of the statistical problem. This generally involves the introduction of the concept of generalisation, which is related to the problem of model and parameter estimation mentioned in the discussion of the first talk (above). Though many neural network papers have addressed this issue I would have to say that methods based upon support vector machines and Boosting have not. They have instead focused on the problem of rapid extraction of decision boundaries from large multi-dimensional data. Successful application of the methods then becomes an issue of the expertise of the researcher applying the techniques and ease of use of the software.

Fundamentally, for a subject like particle physics, you must bear in mind how you will publish the results from applications of complex techniques. You need something simple but with clear statistical rigour and minimal difficulties with dimensional scaling. I would suggest you start by looking at Gaussian (and Gaussian mixture) models trained using Expectation Maximisation [2]. I believe you could make this work for you in a particle physics context.

Liliana Teodorescu : Gene Expression programming

This talk was different in nature to the others in that it concentrated not on a statistical understanding of the problem, but more a method to find an optimal solution once the cost function needed has been specified. I would have selected a cost function which optimised the statistical significance of the scientific question rather than just the signal.

I don't want to say much more about this talk, except the difficulties of setting up the representation of a problem to get any advantage from using genetic programming is widely overlooked. In particular the model definition inherent in the gene sequence must benefit from the linear nature of its construction. This observation is at odds with the idea of downloading software from the internet and just running it. As a first attempt the work was a good illustration of the kind of problem that can be tackled. However, I was not convinced that sufficient effort had been put into use of the approach to achieve much more than random search. The author herself did mention the need to look at the implementation in more detail.

Final Comments

I have found it very difficult to accumulate a self consistent set of derivations for applied methods and am unaware of any standard text which attempts to explain these issues. In recent years I have therefore taken to attempting my own derivations, some of them I have included here in order to try to make specific points and more are available from *www.tina-vision.net*. It is understandable that people who's main focus is another scientific discipline would have difficulty in determining what is relevant to their area. I don't expect anyone reading this document to simply to take my word for what I say here. Particularly as for the sake of brevity I have had to make some quite sweeping statements. However, physicists, probably more than any other area, have the knowledge, principles, mathematical skills and numerical tools needed to penetrate this and develop a real understanding of what is needed and why.

Appendix A: Deriving Likelihood from Probability

The aim is to define the probability of observing a set of sample data $\mathbf{x}_i$ in a multi-dimensional (continuum) space based upon a specific probability density function $\lambda(\mathbf{x}_i)$. To do this we must define a finite set of sample cells $(\mathbf{X}_i)$ and compute the probability of observing data within each and take the continuum limit. The correct statistical model for this is the Poisson distribution, so that the probability of observing a given number of samples within each discrete region within the non-zero portion of the probability density is;

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

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

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

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

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

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

$$=\; \sum_i^{N_1+N_2+...} n_i log(\lambda(\mathbf{X}_i)) \;-\; \sum_i^{N_0+N_1+N_2+...} \lambda(\mathbf{X}_i) \;+\; k_1 \qquad (1)$$

where k_1 is a constant. This measure (first described by Fermi) is referred to as Extended Maximum Likelihood (EML) [1]. It is extended in the sense that, rather than simply computing the likelihood for a fixed density distribution it allows parameters describing the density distribution to also be determined. This expression can therefore be taken as the probability of obtaining a particular sample of data from a set of finite states. This sampling process can be thought of as the basis for the assessment of measurements and samples which must be compared to any theoretical model. We would therefore expect this to be the basis of standard fitting routines (such as regression using least squares and its relatives) and also density estimation (such as Expectation Maximisation). We can see that the second term (which corresponds to the probability of a zero entry in every location) is simply the integrated probability density which we can make constant as part of the Likelihood process. Therefore, only the first term is data dependant.

Generalisation of (1) to continuous variables (x) is now straight forward, but does require us to define the way in which frequencies such as $\lambda(\mathbf{X})$ relate to probability densities. What we require is the probability of observing a measurement like the one we observe located at x. For this we need a consistent way of defining an interval around this value with which to compute a probability $(P(x|A))$. Such a process necessarily requires us to define what we mean by "like". For a measurement there is only one correct way to do this and it requires knowledge of the statistical location of the measurement $\sigma(x)$ (that is the measurement or repeatability error not necessarily just the width of the probability density distribution). This can also be considered as the definition of the "information" in the Fisher sense.

Appendix B: Bias on Likelihood Statistics

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

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

The limit of the bias is defined directly as ;

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

where $p(x_i)$ is the true probability density from the correct model and $<X>$ denotes the expectation operation. We can expand this about the true solution θ_0 as;

$$q = <2\sum_{i=1}^{N}[log(p(x_i, \theta_0)) + (\theta - \theta_0)\partial log(p(x_i, \theta_0))/\partial\theta$$

$$+\frac{1}{2}(\theta - \theta_0)^T H(x_i, \theta_0)(\theta - \theta_0) + h.o.t] > - <2\sum_{i=1}^{N} log(p(x_i))>$$

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

$$q' = <2\sum_{i=1}^{N} log(p(x_i, \theta_0)) - 2\sum_{i=1}^{N} log(p(x_i))>$$

$$+ <\sum_{i=1}^{N}(\theta - \theta_0)^T H(x_i, \theta_0)(\theta - \theta_0)>$$

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

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

which can be reduced further to

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

This result is now directly interpretable, as for the correct model the Kullback-Liebler distance is expected to be zero [1]. The remaining term contains the information matrix for the data, more frequently used to approximate the inverse covariance of the parameters and the covariance on the parameters. For a well determined system we would expect the trace of the product of these matricies to be the rank of the parameter covariance. This is simply the number of model parameters k and leads to the standard form of the AIC measure

$$AIC = \chi^2 + k \quad (3)$$

For badly determined parameters the information matrix may not be full rank and the trace will be the number of linearly independent parameters determined in the model with this data set.

The AIC approach therefore gives a bias corrected estimate of the probabilty of the data given the model, such that we obtain a value which is close to what we would have observed at the true value of the parameters. This analysis does not alter our idea of the best estimate of the parameters given the data, but it does change our idea of the conformity of the model for a particular number of degrees of freedom. Notice therefore that an unbiased estimate of the test statistic at the parameters actually estimated (θ and not their true value θ_0) for an equivalent second sample of data, requires an additional correction of k as suggested by the NIC approach (Figure 1).

---

[1] We are not claiming here that this measure is the correct way of performing this comparison, only that KL is zero for identical PDFs.
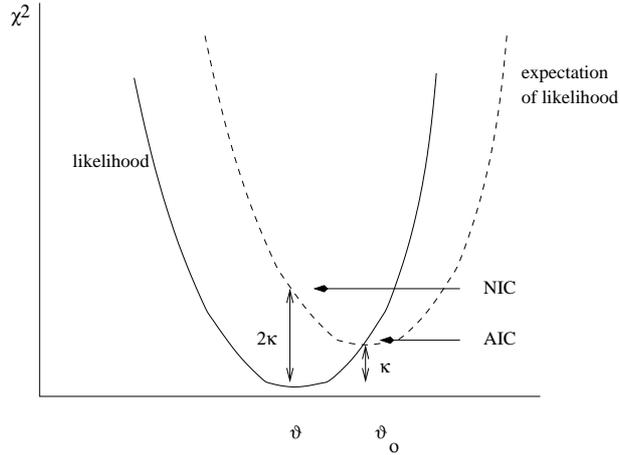
Figure 1: The relationship between AIC (defined on the expectiation of the likelihood around the true solution θ_0) , NIC (defined at the likelihood estimate of the parameters θ) and the original likelihood function.

Appendix C: Bias on Likelihood Parameter Estimates

In order to estimate the bias for the case of infinite data we can use the true probability distribution of the data $p(x)$ and compute the expectation of the Likelihood (excluding the $\sigma(x, A)$ term).

$$< L > \; = \; - \int p(x) \; log(p(x|A)) \; dx / \int p(x) \; dx \qquad (4)$$

For fixed integrated probability density.

We now need to assess this definition for the possibility of bias on the parameters A. The expectation of the probability of the data given the model, for continuous valued observations x follows from Apendix A and is written as

$$Q \; = \; \int p(x) log(\sigma(x, A) p(x|A)) \; dx \; - \; \beta \int \sigma(x, A) p(x|A) \; dx \; + \; k_2 \qquad (5)$$

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

$$\frac{p(x)}{p(x|A)} \; = \; \beta \sigma^2(x, A)$$

In order for the technique to be unbiased, the observed distribution $p(x)$ must match the one assumed in the likelihood model $p(x|A)$, so that all parameters of the data generation process are regenerated by the estimation. This result tells us that Q is maximised for this condition when $(\sigma(x, A) \; = \; constant)$. This is the domain in which σ (or alternativey the Fisher information) is constant and therefore independent of x.

References

[1] R.J. Barlow. Statistics: A Guide to the use of Statistical Methods in the Physical Sciences. John Wiley and Sons, U.K., 1989.

[2] C.M.Bishop, Neural Networks for Pattern Recognition, pp. 66 ff. Clarendon Press, Oxford, 1995.

[3] H. Jeffreys, Theory of Probability. Oxford Univ. Press, 1939.

[4] Devijver, P. R. and Kittler, J., "Pattern Recognition: A Statistical Approach", Prentice Hall, 1982.

[5] J. NEYMAN, *X-Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability*, Phil. Trans. Royal Soc. London, **A236**, pp. 333-380, 1937.

[6] Snedecor, G. W. and Cochran, W. G., Statistical Methods (8th Edition) Ames (IA), Iowa State University Press, 1989.

[7] A. STUART, K. ORD AND S. ARNOLD *Kendall's Advanced Theory of Statistics* Vol. 2A, Classical Inference and the Linear Model, Sixth Edition, Arnold Publishers, 1999.