Tina Memo No. 2016-006
Internal Report

When Bayesian Model Selection meets the Scientific Method.

N.A. Thacker.

Last updated
14 / 04 / 2016

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

When Bayesian Model Selection meets the Scientific Method

N.A.Thacker 14/4/16

Abstract

The scientific method includes the process of testing theories with data in order to reject incorrect ones. I was asked by a PhD student for my opinion of a recent paper which attempted to evaluate the adequacy of competing pharmacokinetic models for the analysis of MR data [2]. Based upon papers I have read over the last three decades, I believe that its description of Bayesian Model Selection could well be a representative example of the methodology from the literature. However, for reasons explained below, I do not believe this approach to be either quantitatively valid or capable of usefully addresing such an issue. As on several previous occasisons, a verbal explanation to the student simply could not do my numerous comments justice, so this time I decided to write them down.

The general approach taken is based upon, and consistent with, other documents found on our web pages (see for example [3]). The conclusions which follow may appear quite contentious, even inconvenient, but I don't expect anyone to simply take my word on this. It should be possible for any mathematically literate reader to independently confirm these criticisms with a little thought. I will modify this document subject to any constructive feedback.

Conventional Approach

For a set of mutually exclusive data generators m_i and associated parameters a_i we can define the joint probability of the data d and the model parameters using Bayes Theorem as

$$p(m_i, a_i|d, I) \;=\; \frac{p(m_i, a_i|I)p(d|m_i, a_i, I)}{p(d|I)}$$

where I is all prior information including the co-hort used to define set of example models (but excluding the current data d). In practical use $p(d|m_i, a_i, I)$ is taken to be the Likelihood for the data d given the model parameters a_i. This expression is sometimes optimised as a process called MAP estimation, and similar forms (a prior multiplying a Likelihood), motivated for model selection.

However, this "probability" is not directly suitable for this, or indeed for model selection, as it is really a density and depends upon our specific choice of parametric description (a_i). In order to compute the probability of the model independent of the parameter choice we must integrate over a meaningful interval of the parameters. In [2] the common argument is made to use all possible values of the parameter. Then

$$p(m_i|d, I) \;\propto\; p(m_i|I) \int p(a_i|m_i, I)p(d|m_i, a_i, I)\; da_i \qquad (1)$$

where we have discarded the normalisation and used the result

$$p(m_i, a_i|I) \;=\; p(m_i|I)p(a_i|m_i, I)$$

Where $p(m_i|I)$ is a scale factor which is assumed in advance (prior) and $p(a_i|m_i, I)$ is the distribution of model parameters sampled from data in the specified sample cohort.

This theory is very general, no specification is given for particular distributions, but what is known is the definitions of what we mean by p (Kolmogorov's axioms) and that terms in the expression must be a function of the parameters (and only those parameters) specified. We will also require for science that our probabilities reflect measurable distributions [3]. Where the method requires modification in a way that these requirements are not met, we will say below that it **contradicts** the theory. You can of course seek to extend the theory to incorporate additional terms or definitions, but can no longer expect to be able to appeal to the original "mathematically rigorous" derivation as justification for the modified approach.

Quantitative Application

We can imagine, and people often develop, all sorts of ways to construct the terms in equation (1) for real world problems. However, making an analysis scientifically useful requires a quantitative approach. Using equation (1) quantitatively requires us to attend to several issues. Firstly the Likelihood function needs to be appropriate to the **data measurement**. Secondly (and also third), the prior distributions $p(a|m_i, I)$ and the scalings $p(m_i|I)$ need

to be appropriate to the **set of data** being analysed. In practice all three of these requirements have technical (and philosophical) complications if this is to be done meaningfully.

On the first point, statistical texts often introduce Likelihood as any general function composed as a product of "theory to data" differences, which may be optimised in order to estimate parameters [1]. People will often assume that a least-squares approach is adequate or may go as far as weighting data points. However, for valid application of equation (1), $p(d|m_i, a_i, I)$ isn't simply any Likelihood, but the one which correctly describes the data, taking into account data correlations if necessary. This can be tested by confirming the assumed data density distributions such as the "pull" for the measurements and hypothesis tests for the parameters (see below). Bland-Altman plots also have a role here.

Similarly the distribution $p(a|m_i, I)$ is not just any sample of parameters obtained from some previous noisy fits. The distribution required by the theory is the noise-free ground truth[1]. Obtaining this may require something akin to deconvolution of the sampled parameter distribution with its estimation uncertainty. A further step is required if Bayesian model selection is to deal correctly with degree of freedom effects. More details on these issues will be given below.

Finally, the assumed prior distributions and scalings, if used to generate a total data density, must match the cohort under analysis. If this condition is not met then the computed probabilities have no real meaning for the data. In order to avoid this, then either the priors need to be perfect and the true generators of the data, or else they must be estimated appropriately for a finite (preferably the incoming) set of data.

This latter solution would require an algorithm similar to Expectation Maximisation. Indeed, for the correct set of theories, estimates of the required parameters and distributions can be made by weighting with the computed conditional probabilities. i.e.

$$p(m_i|I) \;\propto\; \sum_{d \in I} p(m_i|d, I)$$

However, as this involves the data this contradicts the definition of "prior" in the original theory. This perhaps explains why many researchers do not even consider using this approach.

Whilst these steps can be deduced logically from the original derivation and properties of the theory, any reader of this document may struggle to find publications which address them.

Applications in Science

Although Bayesian approaches are often motivated on the basis of being a rigorous theoretical approach, in practical applications of Bayesian model selection steps such as those described above are rarely taken. In addition, the problem generally tackled is that of selecting **one of several competing model descriptions** of a data set. Philosophically this is not the same as having data generated by **a set of mutually exclusive models** (which is the origin of equation (1)). Consequently, no quantitatively valid co-hort for $p(a|m_i, I)$ may even exist.

If we do succeed in building our model selection system, taking all of the steps necessary to ensure quantitative probability, then $p(m_i|d, I)$ may still not do what we require. It is not, as many might expect (based upon other approaches), a confirmation of the adequacy of the model to describe the data, i.e. a scientific selection of candidate theories. Under the data generator assumptions, the separate candidate models are in competition, so that two similar models must share the probability of explaining data. If one model is a more complex variant of another, then they may both describe some data subset equally well (when the added complexity is not needed). This will lead to a 50% probability split in model attribution, even though either model is quite clearly capable of describing all of this data. This situation is a challenge to the exclusive model assumption. Moreover, if all of our models are generally very poor but one is slightly less poor, then this model will have a high Bayesian probability. So a high probability is not a recommendation of fundamental suitability. Finally, the computed probabilities are dependent upon the set of models under comparison, and not a unique assessment of the adequacy of each. Our results will completely change if we add a new model into the analysis (i.e. they are in some respect arbitrary). We need to think carefully about these characteristics in order to avoid over-interpretation.

These things are not difficult to conclude, they can be deduced from simple inspection of equation (1). If it helps, you can think of Bayesian model selection as analogous to a political election. If there are more than two candidates, two similar candidates can split the vote, so that over a population of voters a worse candidate than

[1] In order to deduce this for yourself you can consider the simplified application of a Gaussian mixture model applied to grey-level pixel classification. The parameters a_i can be treated as mixture means, and the data d simply the grey level. In order to regenerate the Gaussian mixture classifier, the grey-level means must have a prior distribution which is a delta function as the remaining Likelihood term will already describe fully the pixel data distribution without a need for further convolution. Whereas a sample of mean estimates will of course vary according to their estimation errors.

either can get in. Also if the candidates are so disliked that only a small fraction of people vote, there will still be a winner. I leave it as an exercise for the reader to decide if it is any better in the situation where there are only two model candidates.

Abuse of Notation

As is generally the case, the above presentation makes no distinction between probability and density. For any scientific use we wish the result of equation(1) to be a probability P and not a probability density p. The integration over a_i in equation (1) ensures that the final expression, being a function of a state variable m_i, (i.e. not a continuous one such as a_i), can be interpreted as a probability. As always, the Likelihood function is proportional to a probability $P(d|a_i, m_i, I)$ defined over some unspecified but fixed interval over d. Whilst the sample distributions over the parameters can only be interpreted as densities.

As a consequence, if we wish to adapt equation (1) in order to specify an "uninformative prior" distribution, although we know that this would mean the probability of obtaining a parameter $P(a_i|m_i, I)$ should be uniform, we cannot achieve this by making the density $p(a_i|m_i, I)$ uniform.

In order to illustrate this, we can make use of the observation that the shape of a Likelihood function around the optima is related to parameter accuracy via the minimum variance bound. For example, imagine that the Likelihood function over a_i is Gaussian with width σ_a. Then the Likelihood function $P(d|a_i, m_i, I)$ can be expressed around the likelihood estimate a' by

$$P(d|a_i, m_i, I) \ \propto \ p(d|a_i, m_i, I) \ = \ p(d|a'_i, m_i, I) \ exp(-(a_i - a'_i)^2/2\sigma_a^2)$$

Applying equation (1), a uniform prior density $p(a'_i|m_i, I)$ would generate the result

$$p(m_i|d, I) \ \propto \ \sqrt{2\pi}\sigma_a \ p(m_i|I) \ p(a'_i|m_i, I) \ p(d|a'_i, m_i, I) \qquad (2)$$

We have no right to expect that any choice of a is a homoscedastic parameter, so if we insist that $p(a'_i|m_i, I)$ is uniform, all terms on the RHS are fixed numbers except for σ_a. The dependency on σ_a in an otherwise constant scaling of Likelihood is a problem. We can choose any equivalent parametric representation we like (a monotonic transformation ($b_i \ = \ f(a_i)$)), so this result is not invariant in the way we need for science. Therefore, for arbitrary $f()$, we need to divide the assumed uniform prior density by σ_b in order to gain self consistency (and so scientific integrity) for equation (1).

Equation (2) is not only true for a uniform prior but also true for a prior density which is linear over a range consistent with σ_a. Therefore, (following the frequently quoted mantra for pattern recognition), it will be a good approximation to (1) provided the distribution $p(a'_i|m_i, I)$ is much broader than the associated measurement error. This is equivalent to saying that $p(a_i|m_i, I)$ is unmodified by a convolution with the measurement process (see below).

Using the inverse parameter estimation error as the prior density may be interpreted as using a Jeffreys prior, but once again this contradicts the theory, as a prior (as specified by our notation) can be a function of any prior knowledge we like (I) but specifically **not** the incoming data. In addition this density is very likely to be "improper"[2], contradicting the very axioms the theory is based upon.

However, we can have our cake and eat it if we accept the following proposition, we make a distinction between probability and density such that

$$P(a_i|m_i, I) \ \propto \ p(a_i|m_i, I)\sigma_a \ \rightarrow p(a_i|m_i, I) \ \propto \ P(a_i|m_i, I)/\sigma_a$$

i.e. the probability of getting a_i should be defined over a range consistent with its estimation uncertainty (e.g. $P(a_i \pm \sigma_a|m_i, I)$). The inverse parameter estimation error takes the role of an interval and not a prior and consequently we avoid any contradiction of the axioms. Mathematically the uninformative prior is the Jeffreys prior, but would be better described as the Jeffreys "interval". The true uninformative prior is $P(a_i|m_i, I)$ (and not $p(a_i|m_i, I)$), and is indeed uniform as expected. If we do not make a distinction between probabilities and densities (as seems to be convention) then we must expect logical paradoxes.

Notice that the previous problem is not apparent when obtaining the prior densities from a cohort, as transformation of the parameters is compensated directly in equation (1) by changes in the sample densities[3]. Only with the above analysis does it become possible to start to pull apart the mathematics and see a relationship between

[2]It does not have unit normalisation over a_i, and fails to satisfy Kolmogorov's axioms, which were needed to derive the theory.

[3]In the same way, we also do not have to worry about the parameters a_i forming a linearly independent orthonormal basis when computing the multi-dimensional integral.

sample parameter densities (e.g. histograms over a_i) and the terms people call "priors". Several facts then begin to emerge regarding the interaction between the measurement process and these distributions.

As a consequence of this, the only definition for a_i which allows us to obtain a meaningful prior probability by sampling typical parameter values is when a_i is homoscedastic. Under these circumstances the sampled distribution ($p'(a_i|m_i, I)$) will be a noisy convolution of a 'true' parameter distribution ($p(a_i|m_i, I)$) with the fixed measurement perturbation process (as already described above). As equation (1) effectively applies a convolution around a' it is unnecessary and we can use equation (2) instead with $p'(a_i|m_i, I)$[4]. It is therefore legitimate to consider the use of a cohort further with the homoscedastic constraint, this will be done below[5].

Degree of Freedom Effects

In scientific applications we are interested in dealing appropriately with degree of freedom effects, i.e. not just picking the most complex model, as would happen if we use an unmodified Likelihood. Instead we find we need to penalise the probability of the more complex model by a factor of e for each additional parameter. Alternatively this can be seen as a finite ($\sqrt{2}\sigma_a$) change in the default value of the new variable. This makes sense, as the new parameters must allow the Likelihood function to be improved by at least this factor before we would choose the more complex model, and this is consistent with conventional hypothesis tests. The conventional Bayesian assertion is that Likelihood bias is due to it being a point estimator and integrating over the parameters (equation(1)) deals with the problem. We challenge this assertion with the following two examples.

Case 1

For uninformative prior probabilities (P), homoscedastic variables will have a uniform sample density (p). So we can investigate this special case independently of any possible objection to Jeffreys priors and make all prior information equal for all models (any σ terms, being independent of a_i now simply get absorbed into our definition of $p(m_i|I)$, which are all made equal). Then, as equation (2) shows, using Bayesian model selection reduces to selection using nothing more than the original Likelihoods $p(d|a'_i, m_i, I)$. It must therefore have the usual Likelihood bias.

Case 2

Consider the effects of adding a uninformative parameter to an existing model based upon a cohort in the Bayesian approach. Trying to consider uninformative parameters in a Bayesian framework is difficult as it rapidly throws up infinities and we must approach the problem by taking a limit (here a Gaussian prior of finite width). The addition of a new parameter can be said to penalise the prior density by a factor of $\approx \sigma_a/\sigma_p$, the ratio of the measurement accuracy to the (much broader) assumed prior width respectively. In this case as we let $\sigma_p \to \infty$ the probability of the more complex model ($p(m_i|d, I)$) goes to zero for any finite change in the Likelihood following addition of the parameter, i.e. the limiting case description of an uninformative parameter is not consistent with known DOF effects. It appears that basing a prior on a sample is a completely different thing to embodying our prior knowledge of uncertainty (as previously).

Crucially, however we do it, neither of these two approaches allows us to derive the accepted quantitative DOF correction to Likelihood (i.e. Akaike). So the DOF problem is due to the Likelihood function itself and not just an issue with using point estimators. Indeed any derivation of a DOF correction will never mention Bayes Theorem.

We know that Likelihood can not be used directly for model selection due to degree of freedom effects and requires a bias correction. However, none of the prior terms in equation (1) can be **justifiably** modified[6] to accommodate this (a DOF bias is a function of d)[7]. As for use of Jeffreys priors, this contradicts the original theory. Using separate multiplicative correction terms (e.g. Akaike) must be seen as additional to deconvolution of distributions obtained from a cohort (as mentioned above). A correction is simply something needed to make the Likelihood estimate "honest", and it is naive to refer to this as a prior simply because it can be written as a multiplicative term.

[4] It is this form of Bayes Theorem (without the integration, but also without the σ_a!) which is encountered in many practical applications.

[5] I should emphasise that this approach is only valid for constant σ_a, in the more general case it would be more difficult to determine the true parameter distribution and this would later need to be convolved according to the incoming data.

[6] Researcher often do include bias correction terms and justify them in terms of Bayesian priors. Correction terms are often justified using Bayes Theorem when Likelihood can not be made to work.

[7] The correction needed for degree of freedom effects is not simply the number of parameters we write down in the model and known a-priori, but the number of linearly independent parameters (rank of the parameter error covariance), as calculated from the data.

Old Fashioned Statistics

In order to emphasise the difference between Bayesian Model selection and other methods it is worth giving an example.

For data generated with Gaussian random variables we can define a χ^2 function. This value can be interpreted on the basis of the number of degrees of freedom in order to construct a hypothesis probability which tests that the theory accounts for the data up to the limits of the measurements (and takes into account Likelihood bias). This hypothesis probability should be uniform, by definition, for the correct theory applied to the data. Any low probability result is otherwise a candidate failure of the theory. Counting these failures gives us an assessment of the ability of the model to describe data independently of any other competing theories. Such an approach does not require knowledge of prior distributions over the parameters in order to make an assessment, indeed we could claim that for most theories we wish to test (and as illustrated in [2]), there is no predicted parameter distribution to compare to. If we do happen to also have a prediction for the distribution of parameters this is something which can be assessed separately.

Any model fit of data can be used as the basis for a hypothesis test, supporting an equivalent test of the theory to a χ^2. Of course, as for Bayesian model selection above, the procedure will always require us to take care and ensure that Likelihood functions match practical data generation processes.

Conclusions

Whilst [2] is typical of Bayesian approaches to model selection, **mathematical rigour** is something different to **appropriate use of a theory**. Bayesian model selection used in this context, far from being rigorous, is instead misleading. Crucially, the actual question specified in the paper title, "Are Complex DCE-MRI Models Supported by Clinical Data?" can not be directly answered via a Bayesian approach. **There is a confusion here between assessing the probability of how often a model can be said to have generated data (Bayes Theorem), and the probability that one model can be taken as generally true (the actual science question).** Clues to this can be seen in some of the less useful properties of this approach. Competition between similar models ensures that the probabilities computed using Bayesian model selection do not provide a unique assessment of model viability. Bayesian selection also does not embody the conventional statistical notion of goodness of fit. These observations remain true even if we abandon quantitation and adopt the (unscientific) subjective definition of probability. Non-the-less this paper might well be counted as one of the many successful applications of Bayesian analysis by its proponents, if the criteria are; it was applied, it was published.

The different underlying assumptions and approximations for the pharmacokinetic models tested are not necessarily valid as the basis for descriptions of tissue contrast concentration. We cannot therefore consider them as valid data generators. Even if we accepted that the approach taken was informative, quantitative use of probability is needed for scientific analysis, but making Bayesian model selection quantitative is difficult. The key problem can be summarised in the standard criticism, "Where do the priors come from?". This isn't just an issue of having any procedure, but having a procedure which is consistent with the theory we intend to use (equation (1)) and not contradicting the specified definitions of mathematical terms.

Integrating over the parameters when performing Bayesian model selection, and using the Likelihood as the probability of the data given the model, does not deal with DOF problems in the same way as conventional approaches. Attempting to justify additional terms as a prior to do this (as often done), is not consistent with the theoretical notation.

In [2], no attempt is made to empirically justify the Likelihood used. Had this been done then the obvious flaw of fitting a model to contrast quantities rather than the raw image data might have been discovered. Using Bayes Theorem (effectively taking ratios) will mask any problems with poor construction but it would be optimistic in the extreme to believe it makes them go away. Also, prior probabilities are estimated using an entire data-set (in effect $p(a_i|I)$, note the missing m_i), there is no ground truth to establish the actual origins of the data and so determine $p(m_i|I)$ or $p(a_i|m_i, I)$. Finally, graphs are shown of the distribution of model probabilities and assessed, in an absolute way, to determine the adequacy of each model. These issues contradict the way we deduce these probabilities should be constructed and constitute over-interpretation.

Conventional quantitative statistical methods, based upon goodness of fit and hypothesis tests with appropriate DOF corrections, along with someone who knows how to use them, are what is generally needed in science. In the paper considered here, if more care had been taken with Likelihood construction, these methods could have been used to directly assess the ability of the models to describe data in specific tissue regions. The philosophical issues associated with quantitative use of probability in science have been discussed in other documents on our web site [3].

References

[1] A. Stuart, K.Ord and S.Arnold,. Kendall's Advanced Theories of Statistics. Volume, 2A, Classical Inference and the Linear Model, Oxford University Press, 1999.

[2] C Duan, J F. Kallehauge, L Bretthorst, K Tanderup, J. J.H. Ackerman and J. R. Garbow., Are Complex DCE-MRI Models Supported by Clinical Data?, Magnetic Resonance in Medicine, 2016.

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

End Notes

Some may ask what the consequences of the ideas in this document are for their own work. Here is a short list of conclusions which accord with the above analysis and may be more directly relevant. I have put them in logical order. Most of them should be obvious (assuming you have already accepted the ones above), others will require a careful reading and some thought. For fun I have marked them out of ten, 1/10 being obvious and 10/10 being complicated[8]. Though these points may in cases be contentious, I would expect most scientists to accept those with low scores relatively quickly. If you are a student, I would seriously recommend that you at least take those below 5/10 seriously while you think about the others.

- It is illogical to expect a subjective definition of probability to maximise a quantitative (therefore frequentist) performance metric. Scientific use of probability for data analysis therefore requires that we make sample densities relate to real world samples. (1/10)
- Maximum Likelihood is not an absolute goodness of fit (i.e. like a chi-square). (1/10)
- Maximum Likelihood works for parameter estimation but associated chi-squares require degree of freedom (DOF) corrections in order to compare goodness of fit between alternative models. (2/10)
- MAP estimation, where we multiply by a prior but do not integrate over parameters is logically inconsistent and therefore unscientific. (2/10)
- Bayes probability estimates based upon integration over parameters are justifiable and consistent provided we have a valid co-hort from which to sample the prior density. (2/10)
- If we can not justify a co-hort then the prior density is subjective or (more critically) arbitrary and therefore open to being criticised as unscientific. (1/10)
- MAP estimation, as a concept, cannot be used to justify conventional Likelihood DOF corrections, as it is inconsistent with the theory (even though it looks like a reason to multiply with something). (7/10)
- MAP estimation, where we assume a uniform uniformative prior (or more likely a Gaussian with very large variance), is dependent upon the parameter definitions and therefore arbitrary and (again) unscientific. (6/10)
- The logical uninformative "prior" is the Jeffreys prior. Equivalently you can choose to use homoscedastic parameters. (7/10)
- A Jeffreys prior is neither a prior nor a probablility (improper) and invalidates Kolmogorov axioms (alone) by absurdum. (1/10, you need to understand the previous two points first)
- Jeffreys priors are mathematically equivalent to using a frequentist interval, based upon measurement errror, with which to relate probabilies to densities(c.w. minimum variance bound). (8/10)
- Use of Bayes theorem when the "priors" are determined from, and consistent with, data (and therefore not priors but sample quantities) is valid and self consistent. (3/10)
- Use of Bayes theorem for model selection must follow the same restrictions we conclude are required for MAP in order to be self consistent and therefore scientific. (1/10)
- Bayesian model selection is an inverse attribution to a **set of valid generators** and not the same as scientific model selection (i.e. choosing **the valid one**). (3/10)

[8] I did consider the idea of marking them in terms of inconvenience, but decided the reader would probably be doing that already.