Drawing Inference - Literally and by Individual Contribution: Recognizing and visualizing distinguishable components of evidence for parameters of interest.

K O'Rourke

March 26, 2011

Abstract

The end products of most modern statistical analyses are tests or intervals (confidence or credible) or more generally functions from which these tests and intervals can be obtained (posterior distributions or likelihood functions). Many different inputs contribute to these, whether from one or many studies, or sources. This paper will outline the mathematical basis for recognizing and visualizing distinguishable components of evidence. Distinguishing and displaying these various components may make their contributions more transparent and their combining or addition to the overall end product more open to critical assessment and review.

Graphical displays for these components will then be developed and implemented for examples involving single and multiple studies (meta-analysis). These are model rather than data visualizations - displaying the degree of support for various parameter values under an assumed model from distinguishable sources. Recall that in parametric models, different parameter values index different models.

The methods required to implement the approach are not particularly novel, mainly involving ways to obtain individual observation or unit of analysis likelihood and prior marginal contributions, but these may not be well known or easily implemented. This paper will primarily focus on obtaining intervals in parametric non-hierarchical models.

The most immediate application of such ideas has been in the field of meta-analysis where separate study likelihoods are distinct, of interest on their own and issues of replication or consistency are primary - if not paramount. Similar ideas and techniques though may be useful

in almost any area of statistics where component likelihoods can be defined per individual observation or unit of analysis. It is argued, that such a visualization is more a matter of due diligence than an attempt to identify and confirm outliers.

Contents

1 Introduction

Many, if not most current parametric applications in statistics are most faithfully represented by complex probability models resulting in both high dimensional likelihoods and priors. When using these parametric modelling approaches, one can and arguably should plot distinguishable or separable contributions to inferences regarding important parameters of interest. This is similar to plotting individual raw data or estimates and their standard errors when they come from separate sources or studies. Here though, we are plotting in the model or parameter space rather than the estimate or data space. Recall that in parametric models, different parameter values index different models. This will allow one to discern if the contributions are generally consistent or discrepant as well as assess how the individual contributions "add up" to the overall inference - given the prior and data models tentatively assumed. That is, allow one to discern the degree of replication between each distinguishable contribution as well as the impact of each contribution.

These separable components will usually consist of individual observation or other unit of analysis contributions and if working from a Bayesian perspective - always the prior. All contributions need to be distinctively shown together in single plots - but likely for only one or perhaps a few parameters of focus at a time. This a matter more of due diligence than choice and should not

be avoided on the basis that calibration maybe problematic. It is first and foremost exploratory not confirmatory, not primarily directed to detection and assessment of outliers, but rather to enhancing the opportunities to discover and possibly repair modeling deficiencies and failures.

The strategy then is to clearly show these various contributions of likelihoods and priors in terms of what the component contributions of each would suggest as inferences on their own and how these are consistent or conflicting and how they add or pool together for the overall inference. The objective is too further "model understanding" - that is, to be one of the tools for following the path from data and models to inferences - hopefully noticing possible modeling failures, opportunities to be less wrong and important limitations along the way. Put another way, the purpose is to avoid being "blind sided" by the more usual statistical methods and procedures that are mostly blind or at least not explicit about how individual observations combine to give the over all inference.

Likelihood provides a direct route to defining individual observation contributions to inferences regarding unknown parameters - since likelihood functions are validly defined for individual observations and their multiplication together completely captures all the information regarding unknown parameters (in the data, given the model)[17][6]. A review of likelihood and some of its perhaps less know properties is given in the appendix. Generalized inverses can be used to define individual observation contributions for linear statistical models[35], but these are less direct and perhaps less readily interpretable. More generally (and perhaps obviously), any scalar estimate, curve or surface may always be re-written as a sum of arbitrary scalars, curves or surfaces. However, a principled choice for such a decomposition such as likelihood, seems sensible. Other units of analysis, comprised of more than one observation, offer more options for defining their inference contributions - but we choose simply to stick with the multiple of the individual likelihoods of the individual observations which comprise the unit of analysis. That is, with L_i being an individual observation likelihood - $\prod_{i \in g} L_i$ for all observations i in a given group g defining a unit of analysis.

Starting with the simplest case of a scalar parameter θ, the prior and overall study likelihood contributions can been depicted as separate components (i.e. a factorization) that can then be contrasted and appropriately combined as highlighted with square brackets in

$$\pi(\theta|y) \propto [L(\theta; y)] * [\pi(\theta)].$$

Similarly, the likelihood contributions from the individual units of analysis in a given study can themselves be depicted as arising from those separate unit of analysis components that again can

be contrasted and combined[8].

$$L(\theta; y) = \prod_{i}^{m} [L_i(\theta; y_i)]$$

Putting these together we get

$$\pi(\theta|y) \propto [\prod_{i}^{m} [L_i(\theta; y_i)]] * [\pi(\theta)].$$

with square brackets highlighting the separable curve components. This can more easily be facilitated by working with log prior curves and log likelihood curves which add, the pooled log likelihood being the sum of the individual unit of analysis log likelihoods and the log posterior being the sum of log prior and pooled (summed) log likelihood

$$\log(\pi(\theta|y)) = [\sum_{i}^{m} [l_i(\theta; y_i)]] + [\log(\pi(\theta))].$$

In the plots, a choice of which levels to hide versus make explicit will be required - i.e. should the single observations curves that make up a unit of analysis curves, be themselves displayed or not. For instance, as will be seen in one of the examples to follow, with the assumption of a binomial model with a common parameter, displaying individual observations is completely uninformative. With hierarchical models that have qualitatively different types of levels, the choice of levels to display and whether to call them prior or likelihood becomes very challenging - for instance see page 395 of Gelman and Hill[18] or page 162 of O'Rourke[30]. Such challenges are being left for a subsequent paper. In this paper only non-hierarchical models will be addressed, but these can be a good first display for hierarchical models in that they can display the non-commonness of assumed to be common parameters.

Given high dimensional parameters are usually required in modern flexible parametric models and the pragmatic need to focus on one or at most a few parameters at a time - curves or surfaces for parameters of focus will need to be obtained by some sort of marginalization or dimensionality reduction. That is, for instance, the unit of analysis contributions integrated or maximized over all the other parameters - rather than viewed multivariately or in full dimensionality. Given a simulation from a Bayesian posterior, say from an MCMC run, this is trivial - just re-parameterize for the chosen focus of interest and then plot the observed distribution (of that parameter disregarding all other parameters).

More formally, in Bayes, one simply integrates out all other parameters. But here, for a given parameter of focus, we also want to plot a marginal log prior and combined marginal log likelihood that adds exactly (plus an arbitrary constant) to the same marginal log posterior as obtained from a full Bayesian analysis. That is, given full multivariate prior for the parameter of focus θ and other parameters λ, $\pi^B(\theta, \lambda)$can be factored into the

$$\pi^B(\theta, \lambda) = \pi^B(\theta)\pi^B(\lambda|\theta)$$

and the true marginal posterior for just θ is obtained by integrating out the λ -

$$\begin{aligned} \pi^B(\theta|y) &= \int L(\theta, \lambda; y)\pi^B(\theta)\pi^B(\lambda|\theta)d\lambda \\ & \int L(\theta, \lambda; y)\pi^B(\lambda|\theta)\pi^B(\theta)d\lambda \\ & \pi^B(\theta)\int L(\theta, \lambda; y)\pi^B(\lambda|\theta)d\lambda. \end{aligned}$$

Here it is simply being argued that the marginal $\log(\pi^B(\theta))$ and marginal $\log(\int L(\theta, \lambda; y)\pi^B(\lambda|\theta)d\lambda)$ and their sum (marginal log posterior) be clearly plotted. Fortunately, this is straightforward for most Bayesian models - the log integrated posterior for the parameter of focus will usually be proportional to the log integrated prior plus log integrated likelihood - subject to appropriate choice of implied marginal prior from the full prior[7]. The perhaps somewhat unusual focus here, in terms of a fully Bayesian analysis, will be on the explicit obtaining and careful assessment and display of marginal priors and marginal likelihoods - the multivariate likelihoods not being just as a black box way to get from a multivariate prior to a univariate marginal posterior - but as something critical to be extracted, then contrasted with and then added on the log scale to the log marginal prior to get the log marginal posterior. Additionally, it is argued that it should be broken down by unit of analysis as well. For displaying how the evidence adds up, log transformations arguably make sense, for other purposes the original (probability) scale may make more sense.

A frequency based marginalization is much more challenging and limited. A simple, but possibly very misleading marginalization, can be obtained by setting the value of all other parameters to their joint *mle* or some other estimate and then taking and treating them as known. This is sometimes referred to as estimated likelihood and is often used just for a subset of the parameters - a partial Bayes estimated likelihood. This partial Bayes estimated likelihood (some parameters taken

as unknown and given a prior and others taken as simply known to be equal to given estimates) is ubiquitous in statistics and often ends up in Bayesian analyses via simplifying assumptions such as in meta-analysis when the within study σ is taken as known and equal to the reported estimates. For some models this can be disastrous[31]. A less misleading marginalization would be to take the *mle* of the other parameters - as a function of the current value of the parameter of focus. This is sometimes referred to as the profile likelihood as it traces out the peak of a multidimensional likelihood surface in the direction of the parameter(s) of focus. That is, as the parameter of focus θ varies from $-\infty$ to $+\infty$, always stay on the highest point on the surface. The profile likelihood is trivial to split into individual observation likelihoods - simply evaluate the likelihood with a single observation along the profile path determined by all the data, the values on this path being taken as known nuisance parameters (see details later). It often gives an adequate approximation to integrated likelihoods when priors are not that informative.[7]

Modern higher order asymptotic methods suggest various modifications that can be made to this profile likelihood to make it less misleading, though not yet completely with success and especially not in a straightforward manner[24]. Alternatively, the integrated likelihood could be used simply by assuming a measure (or formally a prior) on which to evaluate the integral[7] (here a possibly partial Bayesian approach with just priors for parameters not currently of focus). All that is required is a convenient or formal prior for all the other parameters and an ability to evaluate the integral.

On the other hand, although the integrated likelihood essentially solves the marginalization challenge, factoring it into individual or unit of analysis marginal likelihoods is problematic. The integrated likelihood required to obtain the marginal likelihood is essentially a sum of products of all individual likelihoods that share the same nuisance parameters (which are being integrated over). Since sums and products do not commute, it is not clear as to how to obtain the required factorization, i.e. how to obtain a $g_i(x_i;\theta)$ for each i such that

$$\int \prod_{i=1}^{n} f(x_i;\theta,\lambda)d\lambda^{\tilde{}} \prod_{i=1}^{n} \int g_i(x_i;\theta,\lambda)d\lambda.$$

However, an almost individual observation pseudo integrated log likelihood can be defined by differentially rescaling individual profile log likelihoods as a function of θ (which as well be shown below can be directly obtained from the combined profile log likelihood). This is accomplished by multiplying the individual profile likelihood by an $a(\theta)$ that is equal to the combined integrated

likelihood divided by combined profile likelihood

$$\frac{\int \prod_{i=1}^{n} f(x_i;\theta,\lambda)d\lambda}{\sup_{\lambda\in\Omega} \prod_{i=1}^{n} f(x_i;\theta,\lambda)}.$$

The individual observation profile log likelihoods adjusted in this manner add up exactly to the combined integrated log likelihood. So these will at least sum to the correct (combined) integrated marginal log likelihood. They are simply a deformation of likelihoods over the profile path so that they add exactly in this way. Alternatively, adjustments based on Laplace approximation methods can be applied to quadratic curves that approximate the almost individual profile likelihoods or arguably better to the almost individual profile likelihood (non-quadratic) curves themselves. This would provide some connection with asypmtotic theory and the almost individual observation pseudo integrated log likelihood can be factored and displayed as Laplace approximation plus the additional deformation $a(\theta)$ applied to the almost individual profile likelihood.

Unfortunately, dimensionality reduction, more so from the frequency perspective but also for obtaining almost individual pseudo integrated likelihoods, can be somewhat challenging and there are even certain situations where any marginal views will usually be misleading. These misleading situations are often referred to as Neyman-Scott situations involving incidental parameters[28][38] (i.e. the number of parameters remains approximately equal to the number of observations). Given these currently remain as open problems[9][3], it is not in general that one can always focus on one or two parameter(s) of interest at a time, without being mislead. In a Bayesian approach this can be remedied by utilizing informative priors[21], albeit often without strong justifications.

On the other hand, Neyman-Scott situations can be thought of arising from the lack of replication, in the sense that there is little (or no good) replication of the incidental parameters (i.e. not many observations for each parameter common to those observations). Perhaps the most widely known example is pair matched binary outcomes, with usual logistic regression maximizing out the incidental shared pair parameter to focus on the common odds ratio across pairs (i.e. using the profile marginalization). It is well known that this results in a seriously inconsistent *mle* for the odds ratio. Conditional logistic regression "fixes" this by conditioning out the shared pair (incidental) parameter and modified profile likelihood tries to "mimic" this. There are also Bayesian approaches that mimic this "fix"[36]. The inconsistency falls off rapidly - when matched pairs are increased to matched strata that provide better replication of the shared parameter (i.e. a few

observations for each common strata parameter). So a full assessment of replication should in principle help identify Neyman-Scott problems - which parameters have little to no replication? Again, another remedy would be to provide some replication by using informative priors[21]. For many if not most situations, where all parameters have some minimal replication (possibly just coming from an informative prior), the graphs can be illuminating without necessarily being misleading about the path from data given the model, to inferences.

Now, to get unit of analysis likelihoods that are simply curves or low dimensional surfaces rather than full likelihoods, the required marginalization over the other parameters, first entails a borrowing or pooling of information for the given individual unit of analysis from all other units of analysis – but only for the other (nuisance) parameters. That is, the individual unit of analysis inference contributions for the parameter of focus come only from each separate individual unit of analysis given the inference for all other parameters is first pooled. Hence the need for the qualification - almost individual observation likelihoods. (Perhaps this should be called partial combining but that risks being confused with the term partial pooling used in hierarchical modeling.) This "partial combining" would be easiest to grasp with estimated likelihood and a single parameter of focus - given the *mle* for all other parameters the marginalized log likelihood is now just a scalar function of the parameter of focus - a curve - defined for individual observation that adds up exactly to the marginalized total log likelihood evaluated at the *mle*. For the profile marginalization, the unit of analysis profile log likelihoods would add up to exactly the profile total log likelihood (the curve that traces the maximum of the high dimensional log likelihood surface, in the direction of the parameter of focus.)

Except under especially convenient distributional assumptions, any marginal view will lose some information. This loss of information is notoriously difficult to quantify[24] and is usually determined by showing that there is more information - in some sense - for the parameter of focus in the full dimensional likelihood than in the marginal likelihood. When this is case, differing marginalizations may lose differing amounts of information and it may not be clear which loss is least wrong or detrimental. The impact of this may be quantified to a certain degree by bootstrap sampling/cross validation of other unit of analysis given the individual unit of analysis and parameter of focus – i.e. how much does the pooling of contributions for other parameters depend on the particular sample values in hand (other than those in the unit of analysis)? This difficulty can entirely be avoided in a Bayesian approach, as long as the appropriate implied marginal prior is

used[7] and will not be discussed further in this paper.

Computational challenges vary depending on the data model specifications and the method for the marginalization over the nuisance parameters. In particular, the profile marginalization (i.e. taking the best values of the other parameters for each given value of the parameter of focus), has computational advantages and often approximates the more desirable marginalization methods such as integrating over the other parameters with respect to a fully specified prior distribution if the prior is sufficiently weak. Fortunately, MCMC sampling can often facilitate this - the marginal log prior just needs to be subtracted from the log posterior - given both can be adequately sampled from. More formally, in terms of differences in empirical distributions see Evans[12]. In particular, theorem 7 in this paper, suggests how general the approach will be. Additionally, for a theoretical justification for plotting integrated likelihoods as *the* likelihoods to plot, see Evans. [15]

The estimated marginalization, where the other parameters are simply set to estimates such as the *mle* and then treated as known constants, is especially simple but often inadequate. Estimated likelihood though, can often be a useful first step or platform in what is sometimes referred to as "scaffolding" - which refers to using in turn, less and less wrong modeling approaches, for the same problem. For instance one could start with the estimated likelihood, then obtain the profile likelihood and finally the integrated likelihood - even for a fully Bayesian approach where the implied best integrated likelihood is known. Plotting all of them, especially the incorrect ones, may often be illuminating, as will be demonstrated in some of the following examples.

2 Recognizing and visualizing distinguishable components of evidence

The use of multiple plots may allow all choices to be separately displayed. When the unknown parameter θ is simply a scalar, all these functions and their sums are all simply curves, and the plotting challenges are at their minimum. For more general, multivariate parameters θ, particular re-parameterizations and reduction or marginalization to one or at most two dimensions will be required to get manageable, interpretable static plots. (For a few parameters, dynamic plots may be more appropriate). However, once this has been achieved, the preceding factorization again applies (it does not depend on the dimension).

Dual plots can be constructed by plotting individual features that define the likelihood func-

tions and priors (e.g. the terms of a second or higher Taylor series expansion of the likelihood/prior/posterior function or sometimes equivalently the sufficient statistics and ancillaries) along with the uncertainties of these features. In certain cases, well known plots are obtained as the dual (for instance Forrest plots in meta-analysis using assumptions of Normally distributed study estimates of a common unknown parameter θ with known σ - i.e. here θ is simply a scalar and the curves are quadratic). When these dual plots omit important features of the likelihoods and priors they are likely to be less informative or even outright misleading - especially when the curves are not very quadratic[25]. The calibration of the frequency based uncertainties for the features plotted in the dual plot will likely be quite challenging when θ does not simply start out as a scalar parameter. This possible development or refinement will not be further addressed in this paper. From a Bayesian perspective the prior predictive generation of say 20 plot replications (replotting the 20 sets of generated data the same way) using informative enough priors would seem the immediate solution with further refinements depending on work on methods for refining posterior predictive model checking[5][4][11][15] and model prior conflict[13]. The "responsibility" to discern how the evidence is adding up, should not be spurned on the basis that suggestive patterns can not be fully or precisely calibrated. At the very least, basic prior predictive checking can be carried out using a decreasing range of slightly informative priors as a sensitivity analyses.

In the background section below, we demonstrate the basic ideas with a toy regression example originally used by Pena to demonstrate similar ideas for linear model single observation estimates using generalized inverses. Here the unit of analysis is simply the individual observation pair (x_i, y_i) and the estimated versus profile versus integrated marginalization methods are demonstrated. We then give a simple motivational example drawn from a recent cable TV news report about John Sides' scatterplot of unionization rates and budget deficits of US states posted online at

http://www.themonkeycage.org/2011/02/when_scatterplots_invade_cable.html. In commenting on this on his blog, Andrew Gelman suggested the newscasters' idea was that of "summarizing a bivariate pattern by comparing pairs of points" as the newscaster drew attention in turn to four individual states -

http://www.stat.columbia.edu/~cook/movabletype/archives/2011/02/on_summarizing.html. From that perspective it is a nice motivating example for plotting the almost individual log likelihoods of the various states.

Five more realistic examples will later be presented. The first, involves multiple assessments

(studies or batches) of a possibly common underlying proportion - where at least on hypothesis there is only a single parameter. Here there is no need to consider any marginalization at all, as the parameter is simply a scalar to start with and the only unit of analysis worth plotting is the study or batch. The individual observation log likelihoods are simply not informative being simply either $\log(p)$ or $\log(1-p)$. In this simple example, we will focus mostly on how to plot the log prior, pooled (summed) log likelihood, study log likelihoods and resulting log posterior curves. With a single parameter this is simply the (perhaps) well known Bayesian triplot on the log scale. The second example again involves a single parameter, but was posted by Radford Neal on his blog as an example of an inconsistent *mle*. The challenge here is to assess whether there truly is an inconsistency which would suggest any plot would only be misleading - i.e. unit of analysis log likelihoods would be apparently consistent but with their correct combination always becoming inconsistent. The third, is an example from regression where there is an interaction (i.e. a decidedly non-common parameter). Here the individual observation log likelihood contributions should be very different by the interacting feature. The fourth, is an example of a Meta-Analysis of two - two group RCTs with binary outcomes. Here the unit of analysis is the reported study summaries given in the published papers and we contrast a Bayesian versus frequency approach. The fifth and last example, is an especially challenging one from health economics where the unit of analyses are bivariate individual observations (outcome,cost) and the parameter of primary focus is a function of many other parameters. In this last example, although the observations or data are unavoidably bivariate, the chosen parameter of interest in the model space is simply univariate.

In conclusion, all parametric analyses can be recast in terms of separable multiple likelihood and prior contributions. These can often be effectively marginalized to facilitate the focussing on any given parameter of interest and conveniently plotted for the inspection of the consistency or discrepancy of evidence (replication) as well as to gain an appreciation of where the evidence comes from and how it all adds up to the overall inference for that parameter. That is display the path from the data and given model, to inferences for various parameters of focus. This arguably should be a responsibility rather than a choice.

We, however, are not suggesting that the best or final inferences should be adapted or restricted to that which can easily be plotted but simply that it is better that the "least wrong" (marginal) plots be plotted rather than none at all. (And here, perhaps also some "more wrong" but simpler and possibly instructive ones plotted as well.) Fortunately, for most if not all practical Bayesian

analyses, the sum of marginal log prior and marginal log likelihoods will correspond to the correct marginalization from the full posterior. In other situations where the plots likely will be misleading, for instance when there is little to no replication (e.g. not even from prior information) for some parameters that are being marginalized over, this needs to be noted along with (if possible) descriptions of the direction and amount likely.

3 Application Background

When a probability model is being specified for an application, there are some aspects that perhaps deserve more attention and emphasis than is usually the case in order to more clearly delineate the path from data and given models to inferences. At the heart of any statistical analysis is replication - or the repeated observation of a phenomenon. Each replication is considered a unit of analysis - and it is these very unit of analysis contributions that we wish to understand and clearly display. For a replication to be a true replication, there must not be complete dependence and for a replication to be strong there must be as much independence as is possible. In fact, often a unit of analysis is taken as that unit that gives complete independence under reasonable assumptions. Examples are individual observations in an individually randomized study, clusters in a cluster randomized study and the reported study estimates in a meta-analysis of separately conducted studies. The prior being independent of the data, if not also often "prior" in time to the data, may or may not be replicated by the data[14]. Defining units of analysis that are independent simplifies their individual display as well as combination (as is shown in the appendix - they multiply and hence add on the log scale).

Once they are defined, it will often be worthwhile to discern how parameters vary - or even must vary - with the units of analysis. Now, the parameters in differing units of analysis may be common, common in distribution (implied by the units of analysis being exchangeable in what the parameter reflects) or arbitrarily different (e.g. incidental parameters that only involve that unit of analysis as for instance in Neyman-Scott problems where the unit of analysis involves only two or a few observations). We provide a very simple example here to demonstrate the essential ideas (further examples will be given later).

Consider a univariate regression with one independent variable - $y = B0 + B1X1 + error$ - under the usual assumptions of independent identically distributed random error distributed as Normal$(0, \sigma)$. There are just three parameters here and the model implies that each observation

is drawn from Normal($B0 + B1X1, \sigma$), each and every time with the same values of $(B0, B1, \sigma)$. If we choose to focus on $B1$ we need to somehow marginalize out $(B0, \sigma)$ to get a $L(B1)$ - a scalar function just involving $B1$. Additionally, with a Bayesian approach, one needs to marginalize the prior over $(B0, \sigma)$ to get a prior just involving $B1$. In this simple example, we first use a classical approach (no explicit prior) and choose to undertake the estimated and then profile marginalization - i.e. the maximization over $(B0, \sigma)$ for each value of $B1$ (the profile likelihood for $B1$). We then redo it with an explicit prior using integrated likelihood by specifying a semi-conjugate slightly informative prior distributions to help highlight the differences (for convenience, the same prior for both $B0$ and $B1$). We use simple example data contrived by Pena, with x values $(1 : 10, 17, 17, 17)$ and y values $(1 : 10, 25, 25, 25)$ - the three last observations are meant to be tricky outliers in an other wise perfectly correlated example. The graph of resulting individual observation log likelihoods is shown below with the three non-replicating log likelihoods being fairly obvious (some random shifting of the arbitrary height of the curves could be used to make the three distinct - they are actually identical functions). It is important to keep in mind that this is a contrived example and guard against any intuitions about the observations being generated randomly - they simply were not.

In this example, one could incorrectly assume that σ is known and equal to its *mle* (i.e. use the estimated likelihood marginalization) as an initial start. Interestingly, although this is "more wrong" - σ is not known - the plot perhaps more clearly identifies the non-replication. In a real sense, the less correct (more wrong) assumptions providing a less possibly misleading (wrong) display. A likely good strategy perhaps almost always being; first get the joint *mle*, then for all other parameters except the current parameter of focus - set those to their *mle* and then plot the marginal estimated log likelihoods to discern the non-replicated ones (outliers?). A more carefully investigate using profile and or integrated log likelihoods would then follow - but now with the "suspects" clearly in one's sights. If calibration of the plots was needed to confirm an outlier was indeed an outlier, predictive checking of the integrated likelihoods using an informative but not too informative prior would be the current route recommended here. We start first omitting an intercept parameter in Figures 1-3, displaying the estimated, profile and integrated respectively.

The Bayesian approach is presented in the integrated likelihood plots. Note the addition of the log prior curve with maximum at 0.Next the intercept parameter is added: note the somewhat less clear picture (Figures 4-6, estimated, profile and integrated respectively). This likely would

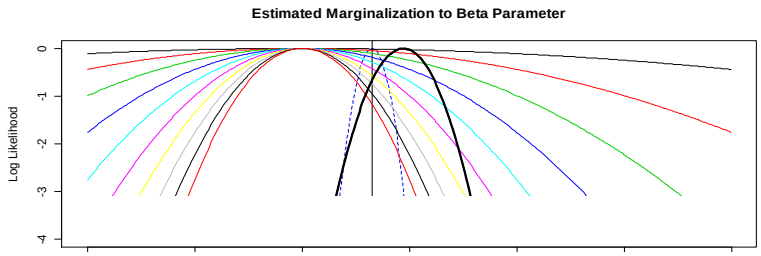

Figure 1: Figure 1

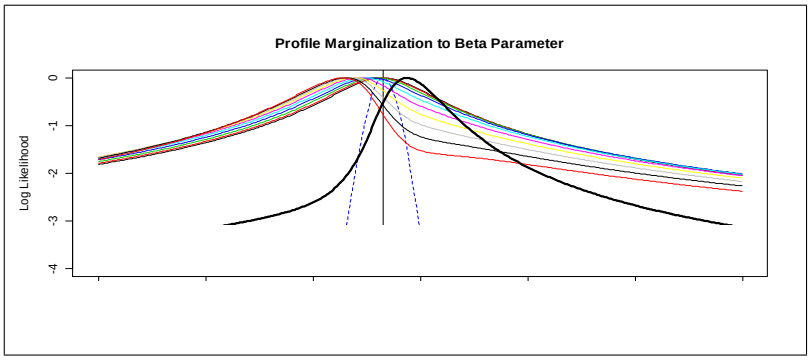

Figure 2: Figure 2

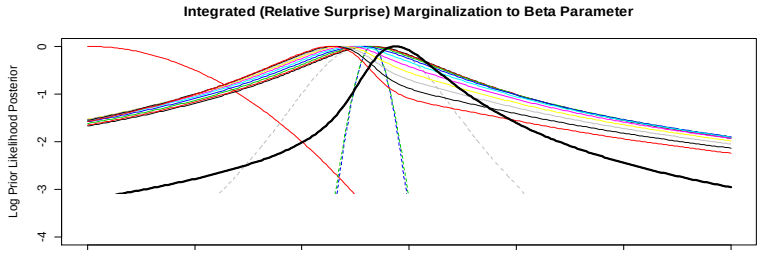

Figure 3: Figure 3

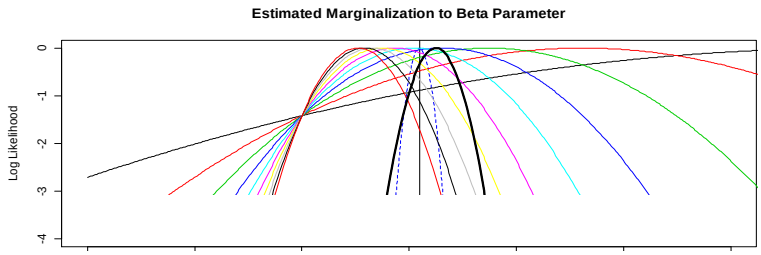

Figure 4: Figure 4

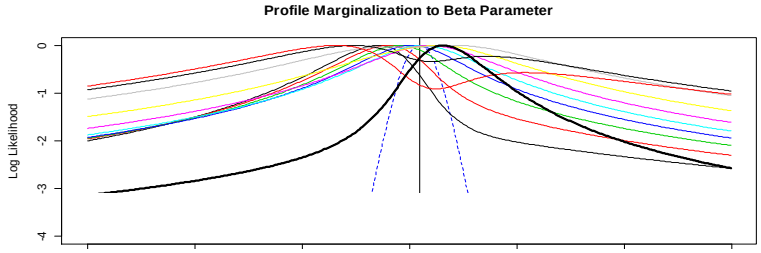

Figure 5: Figure 5

be true most often, that as more nuisance parameters are added to the model the pictures will become less clear. Note: for the integrated likelihood the same prior for the intercept was used as for Beta (which again is shown in the plot with maximum at 0).

We now provide a motivational example, again based on a simple univariate regression. The example as was explained earlier, was drawn from a recent cable TV news report about John Sides' scatterplot of unionization rates and budget deficits of US states posted online at

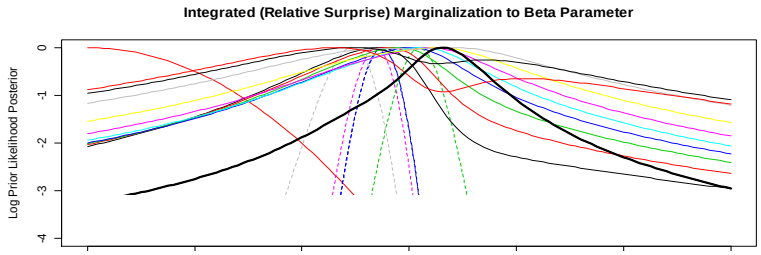

Figure 6: Figure 6

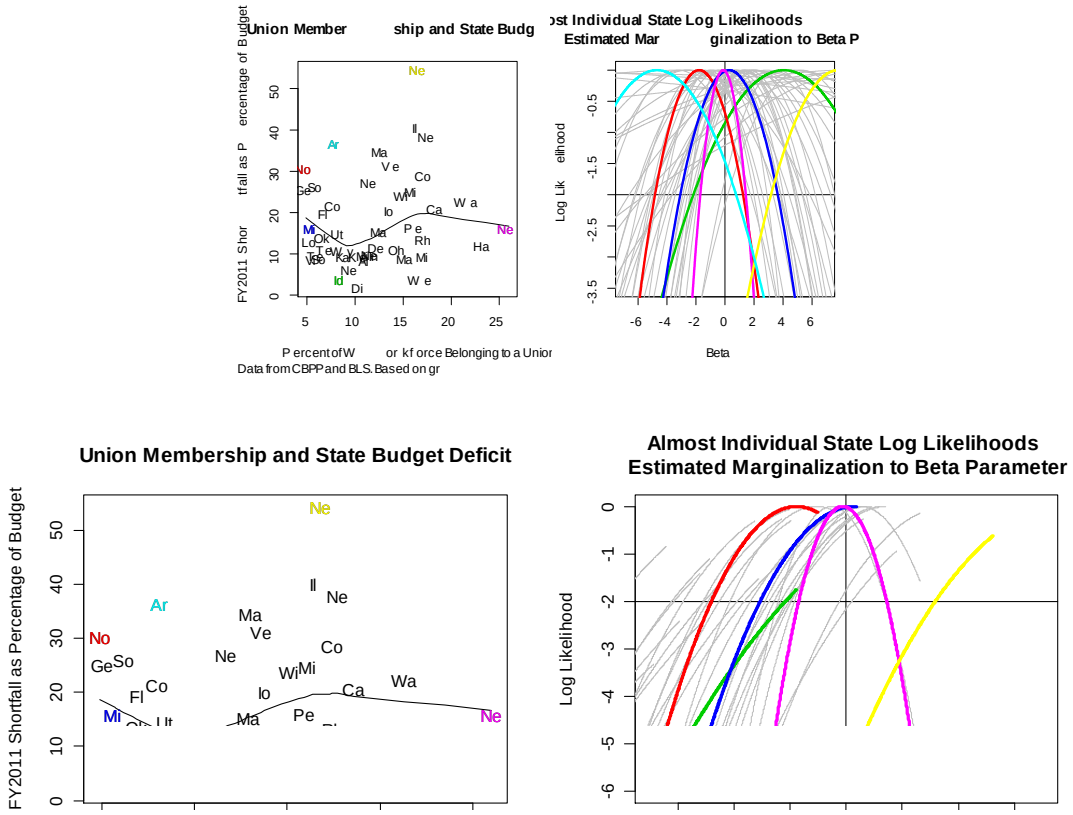

Figure 7: Figure 7

http://www.themonkeycage.org/2011/02/when_scatterplots_invade_cable.html. In commenting on this in his blog, Andrew Gelman suggested the newscaster's idea was that of "summarizing a bivariate pattern by comparing pairs of points" - http://www.stat.columbia.edu/~cook/movabletype/archives/2011/02/on_summarizing.html. From that perspective, it is a nice motivating example for plotting the almost individual log likelihoods for the various states. The only modifications of the programs used for the above example that are required for this example, are the change of the data and the use of less informative prior for the nuisance intercept parameter (here shrinking the intercept parameter towards 0 is a really bad idea). For simplicity and clarity we only show the marginal estimated log likelihoods in Figure 7. The plot on the left is a recreation of John Sides' scatter plot that the newscaster was discussing and the plot on the right displays the almost individual state log likelihoods for the slope parameter (Beta) from a simple linear regression model. (A non-linear model fit curve shown on the scatter

plot, but here we just evaluate a simple linear regression slope). John Sides' scatter plot was presented by the newscaster as a way to evaluate politically charged arguments that public sector unions were or are largely responsible for the current growing state budget deficits in the US. In the regression model used here, the dependent variable (y) is the FY2011 Shortfall as Percentage of Budget and the independent variable (x) is the Percent of Workforce Belonging to a Union. An intercept was included in the regression model. The newscaster first pointed to the state of Mississippi as an example of a state which did not provide evidence of positive relationship as was being claimed between the percent of workforce belonging to a union and the FY2011 Shortfall as Percentage of Budget. This state is shown in blue in both plots. As the newscaster seemed to be claiming that this state offered little evidence of an increasing relationship, this is confirmed by that state's log likelihood (given a simple linear regression model). The newscaster next pointed to the state of New York (shown in purple). Here there is an even stronger suggestion in the log likelihood (given the curvature) that the relationship is very weak, if not almost zero. The newscaster next brought in background knowledge about the percent of the public servant workforce belonging to a union and pointed to North Carolina (red) and Idaho (green) as being very low and low respectively (the same x axis though was maintained). North Carolina presumedly being given as a state that actually suggests a decreasing relationship and similarly so for Idaho. This is observed to be the case for North Carolina, but for Idaho there is actually a weak suggestion of an increasing relationship in the log likelihood. In fact, the individual states most suggestive of increasing versus decreasing relationships (again under this simple univariate linear regression model) in the log likelihoods are Nevada (yellow) and Arizona (light blue), respectively. That is, in terms of an approximate confidence interval (i.e. the log likelihood at minus two from its maximum) excluding positive versus negative slope parameters. Hopefully this example has shown that at least some people do wish to discern what individual observations (states) suggest (provide evidence for) about particular unknowns of interest and "how it all adds up" (democratically so to speak). Being a politically charged news example, one might go as far to say - "the people have the right to know".

4 Examples

The first example is fairly simple one, at least on the basis of the hypothesis of a number of separate studies each of which is estimating the same single common parameter value of an assumed binomial

model. The data was obtained from a published meta-analysis that provided the raw data in a table[33] and although the assumption of a single common parameter is almost surely wrong here - it is a convenient example to start with. The log likelihood is very straightforward

$$x \log(p) + (n - x) \log(1 - p)$$

and even for very large x and or n presents little computational challenges. However, how to best plot these may not be obvious. So that the individual unit of analysis curves will add to the combined curve, log likelihoods will be plotted.

Given that only relative values of the likelihoods are to used for inference, only changes in the height of each unit of analysis log likelihood or combined log likelihood will matter. Therefore arbitrary constants (even different ones) can be added to each individual log likelihood and the combined log likelihood (calculated from the sum of these) - with no change for any possible inference if correctly "read" off the plot. The same is true for log prior and log posteriors - if the graph is only meant to correctly portray their proportionality - with the probabilities being perhaps given in a separate graph or table. We therefor choose to set the value of each of m unit of analysis log likelihood to $2/m$ at the combined *mle*. This results in the maximum of the pooled (summed) log likelihood being 2 at the combined *mle*. Additionally, the points at which the log likelihood curves intersect the zero horizontal line provide approximate 95% confidence intervals (pooled or individual depending on the curve). The maximum of the log prior and log posterior is also arbitrarily set to 2 at the parameter value at which its respective maximum occurs. For this example, a Beta(.5,.5) prior was used and the resulting posterior is so similar to the combined likelihood they cannot be distinguished in the plot (the prior is red, likelihood blue and the posterior purple and dashed). Similarly, very approximate 95% credible intervals can equivalently be read off the plot from the posterior curve.

This is a design choice to facilitate the display of approximate intervals (both confidence and credible intervals). A different choice to facilitate the display tests might be to set the value of each of m unit of analysis log likelihoods to 0 at the null parameter value to be tested against the *mle*. Then one could discern how they add/subtract to get the pooled (summed) maximum at the *mle* [personal communication, Steve Goodman].

In the plot shown below, the x-axis for the parameter value was simply plotted on the original scale. A logit transformation (re-parameterization) would likely make the log likelihoods more

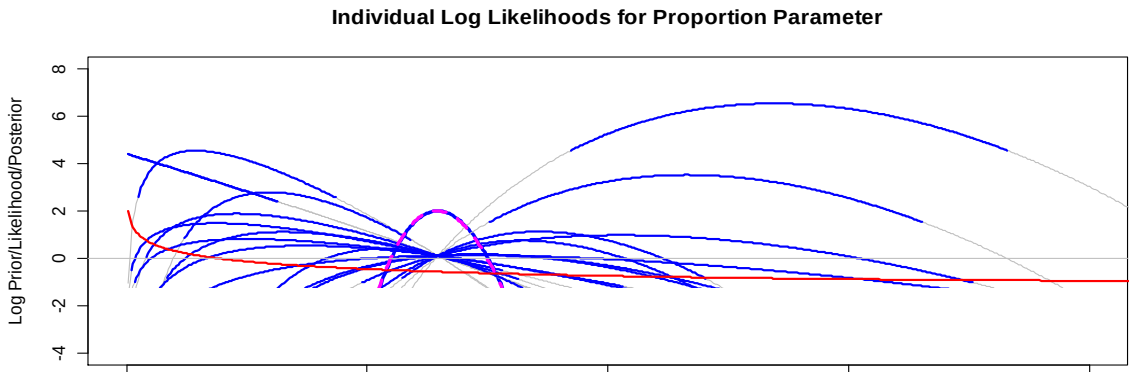

Figure 8: Figure 8

quadratic - which might be important if a dual plot was to be constructed. Here though, for the observations with 0 events observed out of n trials, the log likelihood is simply $n\log(1-p)$ and one would need to be careful when summarizing its features. Note (in Figure 8) that the log likelihoods were plotted as bolded thick lines when within 2 units of their (individual) maximums and grey thin lines elsewhere. The pooled log likelihood was plotted as a (tentative) dashed line.

The next example is of an apparently inconsistent *mle* with a single parameter that was posted on Radford's Neal Blog - see material at

http://radfordneal.wordpress.com/2008/08/09/inconsistent-maximum-likelihood-estimation-an-ordinary-example/#comments.

The probability specification is i.i.d. observations, drawn from the distribution -

$$f(t) = (1/2)N(0,1) + (1/2)N(t, exp(-1/t\hat{}2)\hat{}2),$$

where t is a positive real parameter. These likelihoods involve a single parameter - so the plotting of individual log likelihoods and their sum is very straight forward. A few plots clearly shows that with increasing sample sizes there is very likely to be a positive observation near 0 that has a very spiked log likelihood. Because of this, at that point, it dominates the sum of all the other log likelihoods. Hence clearly displaying the apparent inconsistency of the *mle*. With an increasing numbers of samples - it gets more and more likely that such a dominant one will occur (hence the inconsistency).

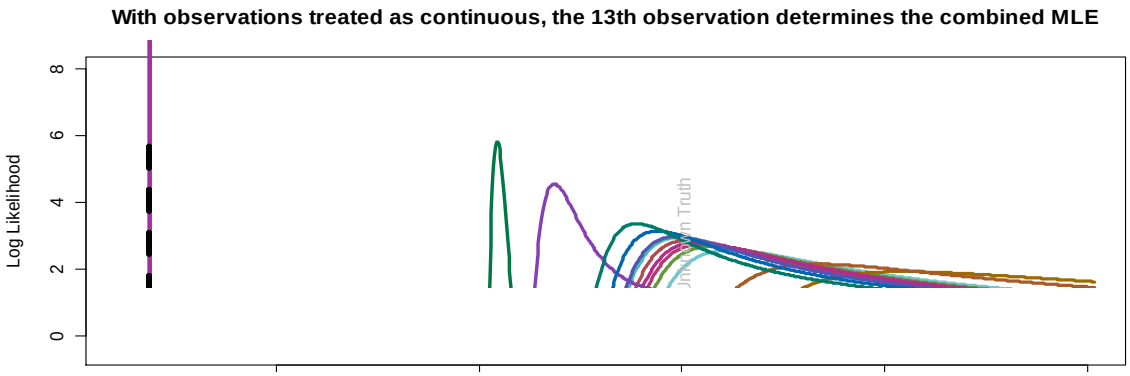

Figure 9: Figure 9

This then would seem to be an example where a plot could only mislead - most, if not all but one single observation log likelihood suggest something other than the *mle* that is guaranteed to be wrong - the *mle* will almost always be very close to 0. This happens even with a correct specification and no nuisance parameters unless the sample size was small and we were lucky (see Figure 9). That is, until one recalls that the likelihood is the probability of what was observed and the observations are never observed to infinite precision.

When the likelihood is rewritten to correct for this, i.e. an integral over some small interval that represents the accuracy of recording the observations, the inconsistency disappears (see Figure 10). Specifically on the same blog see - Radford Neal | August 26, 2008 at 2:47 pm ("if one took account of observations being actually discrete rather than truly continuous and replaced the density with the integral from obs – e to obs + e the inconsistency would go away – if e was big enough?" Yes). The inconsistencies that arise in Neyman-Scott problems are not so easy to overcome in other problems such as the Neyman-Scott problems discussed earlier and remain serious open problems.

The next example, is a univariate regression with two independent variables - $y = B0 + B1X1 + B2X2 + error$ - under the usual assumptions that the random error is i.i.d. $N(0, \sigma)$. The unit of the analysis is almost always taken as the individual observations given the i.i.d. assumption implying no serial correlation. There a four common parameters here - $B0, B1, B2$ and σ. Every unit of analysis is assumed to have been generated from a probability model that has these same four parameters at exactly the same value. The expected value of y for each unit of analysis is linearly modulated by the values of $X1$ and $X2$ via exactly the same parameter values. Now if

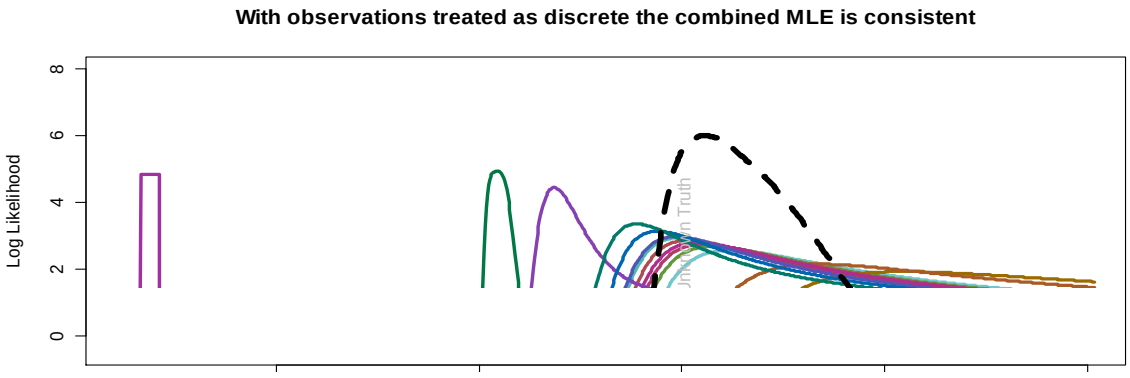

Figure 10: Figure 10

say $X2$ is gender and we add an interaction $B3X1X2$, now there are five parameters - a common intercept α_m and common slope β_m amongst the males, a different common intercept α_f and common slope β_f amongst females and a common σ amongst every one. (A rather strange, but common assumption.) If we choose to do the regressions separately by gender, we will get exactly the same estimates, except for the estimate of σ which will differ depending on how the residuals errors split between the gender groups. We can also see this difference when plotting the individual log likelihoods for β_m - the male slope parameter - from the female subjects versus the male subjects. If we assume σ is known and equal to its estimate (e.g. its mle) versus unknown - the information separates. This is obvious in Figure 11, in the flat likelihoods for the individual female log likelihoods for β_m (females provide no information at all for the males slope).

On the other hand, with σ taken as unknown and hence being estimated as in the profile marginalization, the individual female log likelihoods for β_m are curved but symmetric about the mle for β_m (i.e. females do provide some information for the males' slope, as is clear in Figure 12). This information arises from the σ parameter being unknown but making some observations more likely than others - for both males and females.

We now present example of a meta-analysis of two studies - each involving randomization to two groups for the assessment of the same treatment on the same binary outcomes. Here we do a fully Bayesian analysis but also plot a classical approach using profile marginalization. It is a common assumption in meta-analysis to allow the Pcs to be arbitrarily different in each trail but

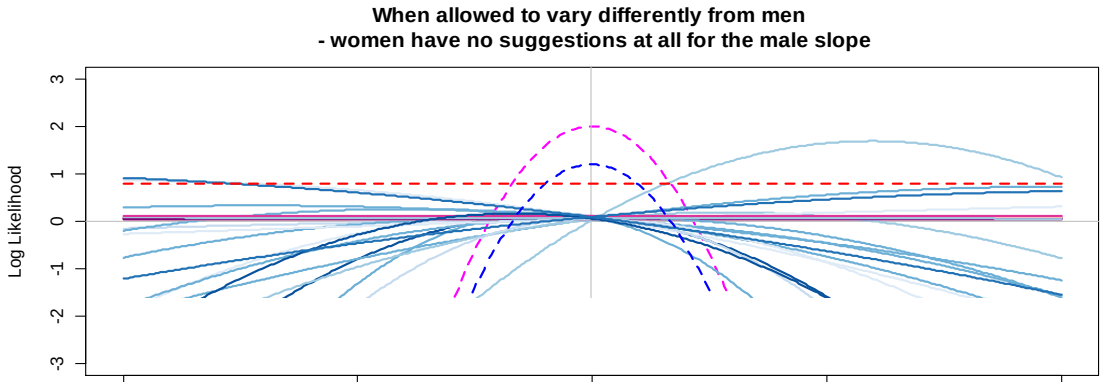

Figure 11: Figure 11

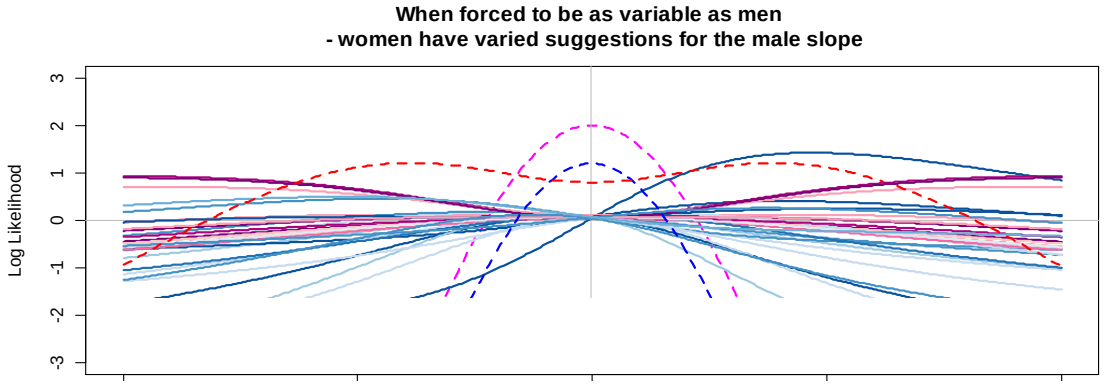

Figure 12: Figure 12

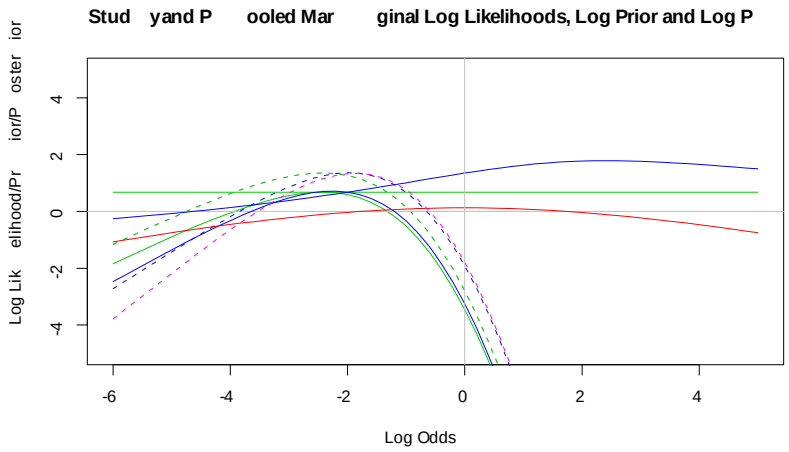

Figure 13: Figure 13 Integrated in blue, prior red, posterior purple and profile green. With 0 events in both groups, the profile likelihood is a straight line - no information about the odds.

then argue for an essentially similar relative treatment effect. This relative treatment effect is often, but not always, taken to be the log odds ratio. This results in two independent, two dimensional likelihoods based on two parameters, one parameter that is common to both and another that arbitrarily differs. The Bayesian analysis requires a joint prior for Pc and Pt. The pooled or meta-analysis likelihood is simply the multiple of these two likelihoods and this multiple now has three parameters, $Pc1$, $Pc2$ and OR. Here it would be feasible to attempt a three dimensional plot of some sort (one dimension for each parameter) but instead we illustrate our strategy to use a marginalization for both the individual and pooled likelihoods and prior to get one dimensional functions for the common parameter OR of focus to plot (see Figure 13). The profile reduction simply maximizes out the non-common parameter(s) and except for occasional numerical problems can always be conveniently obtained. In the Bayesian reduction, it can challenging to obtain the integrated likelihoods and here was facilitated by adopting the assumptions and approach suggested by Wolpert[39] and the use of default numerical integration provided in R. More generally, MCMC sampling will be required.

In Figure 14, we just display the study with zero outcome successes separately to note the difference between prior and data model information about the log odds. The prior suggests that positive and negative log odds ratio of the same magnitude are equally supported. The profile log likelihood suggests the same. But the posterior via the integrated log likelihood suggests positive log odds ratios are better supported than negative ones. For an account of this, using Simpson's

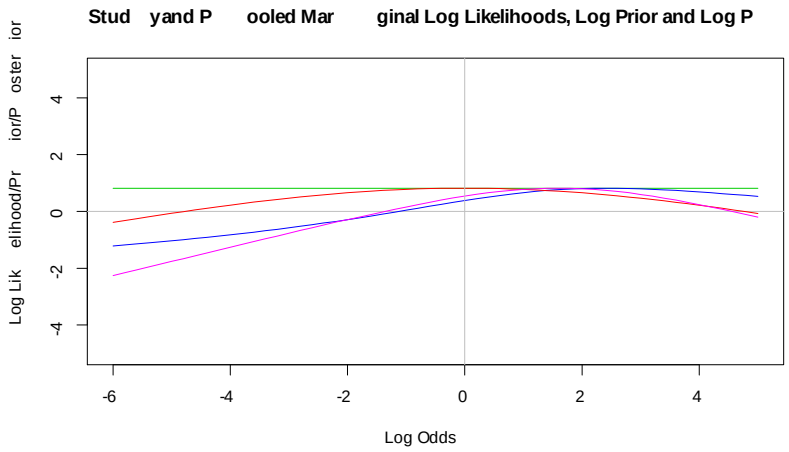

Figure 14: Figure 14 Integrated in blue, prior red, posterior purple and profile green. With 0 events in both groups, the profile likelihood is a straight line - no information about the odds.

paradox, see Greenland[20]. The important point here, is just to (always) be aware that this is happening. If the prior assumptions were not carefully thought of, especially with regards to impacts on marginal inferences, some modification may be sensible. Peter Thall has brought much attention to this point for adaptive (very small sample size) trials via the concept of effective sample size[26].

The final example is from health economics. Here it is critical to consider both effects and costs of treatments jointly in order to appropriately assess costs versus benefits of adopting new treatments in clinical practice. To do this, both effects and costs are modelled jointly and this quickly leads to complicated data model specifications[29] that are hard to motivate, justify and especially assess their robustness. Given these challenges in specifying appropriate joint data models - we here just use an especially convenient joint data model - a joint Normal distribution where the relationship between effects and costs is simply linear. Here it is also obvious how the bivariate distribution is factored into two univariate ones, (one marginal and one conditional) and this factorization greatly facilitates the programing.

The full specification with choice of common and arbitrarily different parameters is

$$Placebo(\theta; e, c) = N(c; \mu_{pc}, \sigma_{pc})N(e|c; \mu_{pe} + \beta_p(c - \mu_{pc}), \sigma_{pe|c})$$

for the placebo group, with e, c being the effects and costs respectively and the subscripts p, pe, pc

referring to the placebo, placebo effects and placebo costs (i.e. β_p is the slope coefficient of effects on costs in the placebo group). The treatment group's specification is

$$Treatment(\theta; e, c) = N(c; \mu_{tc}, \sigma_{tc})N(e|c; \mu_{te} + \beta_t(c - \mu_{tc}), \sigma_{te|c}).$$

The common parameters here being restricted to being within the treatments groups (i.e. only replication within the groups is being specified.) Many alternative specifications are possible, one reasonable with replication over the groups would specify the same slope in both treatment groups by dropping the subscript on β. Often the parameter of interest is a function of the underlying parameters as in the Net Monetary Benefit parameter

$$NMB = K(\mu_{te} - \mu_{pe}) - (\mu_{tc} - \mu_{pc})$$

This represents the excess of valued treatment effect over treatment cost. We "insert" this parameter of focus into the specification by the re-parameterization

$$\mu_{te} \rightarrow \frac{NMB}{K} + \mu_{pe} + \frac{\mu_{tc}}{K} - \frac{\mu_{pc}}{K}$$

and then directly obtain the marginalized unit of analysis log likelihoods for the NMB parameter that is now explicitly in the specification.

The units of analysis here are simply the individual bivariate observations by patient. The example in the following plot (Figure 15) is based on simulated data drawn from this bivariate Normal specification and with one of the observations purposely changed to be an outlier (i.e. it was made not to replicate) in Figure 16. The profile marginalization was used to obtain the unit of analysis log likelihoods - here being the individual observations. The non-replicating observation appears as the unusual red curve in the plot. Although the probability model is bivariate with numerous parameters, it was straightforward to display replication, by individual observations, for a function of the parameters.

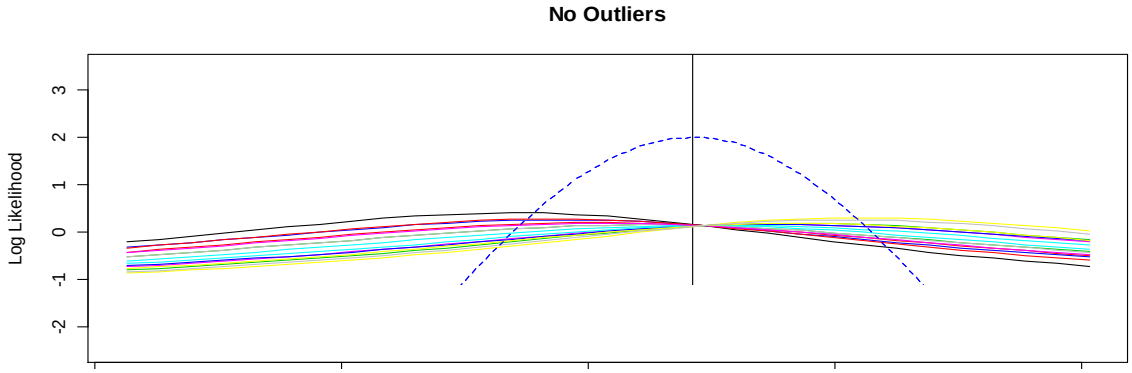

Figure 15: Figure 15

Figure 16: Figure 16

5 Discussion

Plotting separable contributions to inferences regarding important parameters of interest can be challenging but illuminating. Just how illuminating needs to be better informed by more and wider experience. On the other hand, it has been argued that this is a matter of due diligence - especially with the use of complicated models (and complicated is in the eye of the beholder). The only barrier would seem to be computational feasibility - can the integrated likelihood of interest be approximated well enough? In some cases though, even the joint likelihood itself is not readily available. Well known examples occur with missing or incompletely observed data. Of particular concern in meta-analysis is the obtaining of likelihoods from reported summaries rather than raw individual data. Methods to over come this difficulty will be addressed in O'Rourke (to be submitted).

It is much more common to plot individual raw data summaries or estimates and their standard errors when they come from separate sources or studies. Here though, we plotted in the model or parameter space rather than an estimate or data space. This plotting in the model space may be fairly novel and might require some adjustment as to what should be looked at in applications. It may allow one to discern if the contributions are generally consistent or discrepant as well as assess how the individual contributions "add up" to the overall inference - given the prior and data models tentatively assumed. It will display this and exactly in fully Bayesian applications. That is - show how inference under the assumed model actually "did add up". Should that be an option or a usual requirement in applying statistics? Perhaps it should be an obligation in statistical applications to explicitly discern the degree of replication between each contribution as well as the impact of each contribution in the overall inference - for each and every set of models entertained?

As simply another graphical method, especially in regard to the assessment of the commonality of parameters, the approach could be viewed as an extension of the method of locating several outliers in multiple-regression, using elemental sets[23] to single observations as opposed to sets of observations of size p. Elemental sets of size p are the smallest set of observations that will allow consistent estimates of p parameters. All possible such sets need to be formed and investigated. It is also possible to just write the slope estimates as a single double sum of difference ratios[19]

$$\beta = \frac{\sum_{i,j} \frac{y_i - y_j}{x_i - x_j}(x_i - x_j)^2}{\sum_{i,j}(x_i - x_j)^2}.$$

It is also distinct from the graphical method for assessing the fit of a logistic regression model by Pardoe[34] or more generally methods that use posterior predictive values as theses are based on Box's global approach to assessing joint model fit (prior and data models). The method here separates the prior model versus data model, displaying any possible conflict between these two, and then focuses directly on data model fit. This is done separately by units of analysis and addresses just the commonality of parameters, rather than a mixture of commonality of parameters and distribution shape. Whether parameters appear common or not depends heavily on distribution shape - what appears as common with heavy tailed distributions will appear non common with a light tailed distribution. If the distributional assumptions are fixed and one focuses on likelihoods under those assumptions (recall the likelihood is minimal sufficient) one is focusing on commonness. Hence, it supports the emphasis on a strategy of considering parameters for individual units of analysis, one at a time, and discerning which components are common, common in distribution or arbitrarily different across individual observations.

If calibration is required, prior predictive simulations are likely though to be the most direct route to calibrate these plots. That is, for patterns that are discerned in one of the these plots, how often would similar patterns appear in prior predictive simulations. If the apparent discrepancy or lack of replication could easily be due to chance, perhaps it should be ignored.

Distributional specifications could be investigated by varying the prior and or data model assumptions and discerning how this changes the apparent commonness in the plots. A dramatic example perhaps being common versus common in distribution data model assumptions (i.e. a hierarchical model)- any apparent lack of consistency between the unit of analysis log likelihoods under a common parameter assumption usually disappears with the alternative assumption the parameter of focus being common in distribution instead of just plain common. The topic of hierarchical models will be addressed in a later paper. Another example would be with regard to apparent outliers. These can change dramatically with differing data model assumptions. Nelder apparently suggests this as a way of avoiding outliers all together[27] - choosing a data model distributional specifications in which there appear to be no outliers. This seems a bit too convenient. On the other hand, David Cox warns that unpleasant features in the model space (e.g. multi-modality) may simply be warnings that the model being assumed - is just way too wrong[9].

References

[1] Barnard, G., and Copas, J. Likelihood inference for location, scale and shape. *Journal of Statistical Planning and Inference 108* (2002), 71–83.

[2] Barndorff-Nielsen, O. Likelihood theory. In *Statistical Theory and Modelling*, D. V. Hinkley, N. Reid, and E. J. Snell, Eds. Chapman and Hall, London, 1990.

[3] Barndorff-Nielsen, O., and Cox, D. R. *Inference and asymptotics*. Chapman and Hall, London, 1994.

[4] Bayarri, M., and Castellanos, M. Bayesian checking of the second levels of hierarchical models. *Statistical science 22*, 3 (2007), 322–343.

[5] Bayarri, M., Castellanos, M., and Morales, J. MCMC methods to approximate conditional predictive distributions. *Computational Statistics & Data Analysis 51*, 2 (2006), 621–640.

[6] Berger, J., and Wolpert, R. The likelihood principle. IMS.

[7] Berger, J. O., Liseo, B., and Wolpert, R. L. Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science 14* (1999), 1–28.

[8] Cox, D. R. Some remarks on likelihood factorization. *IMS Lecture Note Series 136* (2000), 165–172.

[9] Cox, D. R. *Principles of statistical inference*. Cambridge University Press, Cambridge, 2006.

[10] Cox, D. R., and Hinkley, D. V. *Theoretical statistics*. Chapman & Hall Ltd, 1974.

[11] Evans, M. Comment-Bayesian Checking of the Second Levels of Hierarchical Models. *Statistical Science 22*, 3 (2007), 344–348.

[12] Evans, M., Guttman, I., and Swartz, T. Optimally and computations for relative surprise inferences. *Canadian Journal of Statistics 34*, 1 (2006), 113–129.

[13] Evans, M., and Moshonov, H. Checking for prior-data conflict. Tech. rep., University of Toronto, 2005. University of Toronto Technical report No 0413.

[14] Evans, M., and Moshonov, H. Checking for prior-data conflict. *Bayesian analysis 1*, 4 (2006), 893–914.

[15] Evans, M., and Shakhatreh, M. Optimal properties of some Bayesian inferences. *Electronic Journal of Statistics 2* (2008), 1268–1280.

[16] Fisher, R. *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh, 1959.

[17] Fraser, D. A. S. *Probability and Statistics: Theory and application.* Duxbury Press, North Scituate, 1976.

[18] Gelman, A., Hill, J., and Corporation, E. *Data analysis using regression and multilevel/hierarchical models*, vol. 625. Cambridge University Press Cambridge, 2007.

[19] Gelman, A., and Park, D. Splitting a predictor at the upper quarter or third and the lower quarter or third. *The American Statistician 63*, 1 (2009), 1–8.

[20] Greenland, S. SimpsonŠs Paradox From Adding Constants in Contingency Tables as an Example of Bayesian Noncollapsibility. *The American Statistician 64*, 4 (2010), 340–344.

[21] Gustafson, P., Gelfand, A., Sahu, S., Johnson, W., Hanson, T., and Joseph, L. On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statistical Science 20*, 2 (2005), 111–140.

[22] Gustafson, P., and Greenland, S. Interval estimation for messy observational data. *Statistical Science 24*, 3 (2009), 328–342.

[23] Hawkins, D., Bradu, D., and Kass, G. Location of several outliers in multiple-regression data using elemental sets. *Technometrics 26*, 3 (1984), 189–233.

[24] Jorgensen, B. The rules of conditional inference: Is there a universal definition of nonformation? *Statistical Methods and Applications 3*, 3 (1994), 355–384.

[25] Meng, X. Decoding the h-likelihood. *Statistical Science 24*, 3 (2009), 280–293.

[26] Morita, S., Thall, P., and Meuller, P. Determining the effective sample size of a parametric prior. *Biometrics 64*, 2 (2008), 595–602.

[27] Nelder, J. A. There are no outliers in the stack-loss data. *Student 3*, 3 (2000), 211–216.

[28] Neyman, J., and Scott, E. L. Consistent estimates based on partially consistent observations. *Econometrica 16* (1948), 1–32.

[29] O'Hagan, A., and Stevens, J. Bayesian methods for design and analysis of cost-effectiveness trials in the evaluation of health care technologies. *Statistical Methods in Medical Research 11*, 6 (2002), 469.

[30] O'Rourke, K. Meta-analysis: Conceptual issues of addressing apparent failure of individual study replication or "inexplicable" heterogeneity. In *Empirical Bayes and likelihood inference* (2001), pp. 161–183.

[31] O'Rourke, K. *The Combining of Information : Investigating and Synthesizing What is Possibly Common in Clinical Observations or Studies Via Likelihood.* PhD thesis, University of Oxford, 2007.

[32] Pace, L., and Salvan, A. *Principles of statistical inference:from a Neo-Fisherian perspective*. World Scientific, London, 1997.

[33] Pal, T., Permuth-Wey, J., Kumar, A., and Sellers, T. Systematic review and meta-analysis of ovarian cancers: estimation of microsatellite-high frequency and characterization of mismatch repair deficient tumor histology. *Clinical Cancer Research 14*, 21 (2008), 6847.

[34] Pardoe, I., and Cook, R. A graphical method for assessing the fit of a logistic regression model. *The American Statistician 56*, 4 (2002), 263–272.

[35] Pena, D. Combining information in statistical modeling. *The American Statistician 51*, 5 (1997), 326–332.

[36] Rice, K. Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications. *Journal of the American Statistical Association 99*, 466 (2004), 510–522.

[37] Rubin, D. Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics* (1984), 1151–1172.

[38] Stigler, S. The epic story of maximum likelihood. *Statistical Science 22*, 4 (2007), 598–620.

[39] WOLPERT, R., AND MENGERSEN, K. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Statistical Science 19*, 3 (2004), 450–471.

6 Appendix: Review of likelihood and some of its properties

The likelihood is the probability of the observations for various values of the parameters and will be denoted as $c(y)\Pr(y;\theta)$, where Pr is the tentatively assumed probability model that generated the observations, the observations y are taken as fixed and the "parameter" θ is varied. The generic positive constant function $c(y)$ emphasizes that only relative values are of interest in applications and we will only be using relative values of the likelihoods in the plots. A perhaps preferable choice of $c(y)$ would be $1/\Pr(y;\hat{\theta})$ to make $L(\theta;y) = \Pr(y;\theta)/\Pr(y;\hat{\theta})$ simple - stripping off any function that just depends on y . A more formal definition of likelihood is that it is simply a mathematical function

$$L(\theta;y) = c(y)\Pr(y;\theta).$$

The formal definition as a mathematical function though, may blur that the likelihood is the probability of re-observing what was actually observed. In particular, one should not condition on something that was not actually observed such as a continuous outcome, but instead some appropriate interval containing that outcome (i.e. see page 52 of Cox & Hinkley[10][1]). We later give an example where this distinction was important.

In the event there is more than one unit of analysis, say y_1 and y_2, a combined inference is sought and as the likelihoods involved are probabilities, they multiply. Recall that

$$\Pr(y_1, y_2; \theta_1, \theta_2) = \Pr(y_1;\theta_1) * \Pr(y_2|y_1;\theta_1,\theta_2) = \Pr(y_2;\theta_2) * \Pr(y_1|y_2;\theta_1,\theta_2)$$

and simply

$$\Pr(y_1;\theta_1) * \Pr(y_2;\theta_2)$$

for independent units of analysis. Re-emphasized as likelihoods - as a function of θ for fixed y

$$L(\theta_1, \theta_2; y_1, y_2) = L(\theta_1; y_1) * L(\theta_1, \theta_2; y_2|y_1) = L(\theta_{_2}; y_2) * L(\theta_1, \theta_2; y_1|y_2)$$

and

$$L(\theta_1; y_1) * L(\theta_2; y_2)$$

or independent units of analysis.

If some parameter in the probability models used to represent each of the units of analysis is common, then there is a combination for that parameter - simply by that multiplication. In this way the combined likelihood concentrates (or increases in information) for common parameters. For parameters that are different or arbitrary (i.e. incidental parameters) there is an expansion of dimension in the combined likelihood without an increase in information[38]. Common parameters are clearly identified by the repeated appearance of the same parameter in the likelihoods that are multiplied together - common simply means repeated in two or more individual observation likelihoods. For a discussion of the factorization of likelihoods, i.e. the inverse of combination, see Cox[8].

These may or may not be obvious and may even be actually hidden in some parameterizations. Because of this, a common parameter reparameterization may be helpful to discern this. In notation: $\omega = \omega(\theta_1, \theta_2) = (\gamma, \chi)$, where $\gamma = \gamma(\theta_1, \theta_2)$ (the something that is possibly common) and $\chi_i = \chi(\theta_1, \theta_2)_i$ (the something that arbitrarily differs) and conversely, $\theta_1 = \theta_1(\gamma, \chi_i)$ and $\theta_2 = \theta_2(\gamma, \chi_i)$. The multiplication for independent observations then gives

$$\begin{aligned} L(\theta_1, \theta_2; y_1, y_2) &= L(\gamma(\theta_1, \theta_2), \chi(\theta_1, \theta_2)_{(1,2)}; y_1, y_2) \\ &= L(\gamma(\theta_1, \theta_2), \chi(\theta_1, \theta_2)_1; y_1) * L(\gamma(\theta_1, \theta_2), \chi(\theta_1, \theta_2)_2; y_2) \end{aligned}$$

(one concentration $\gamma(\theta_1, \theta_2)$ and one expansion $\chi(\theta_1, \theta_2)_i$) versus

$$\begin{aligned} L(\theta_1, \theta_2; y_1, y_2) &= L(\chi(\theta_1, \theta_2)_{(1,2)}; y_1, y_2) \\ &= L(\chi(\theta_1, \theta_2)_1; y_1) * L(\chi(\theta_1, \theta_2)_2; y_2) \end{aligned}$$

with no common $\gamma(\theta_1, \theta_2)$ - just expansion. This is slightly more complicated for dependent observations where even with no common parameters some concentration arises from the dependency

itself

$$\begin{aligned} L(\theta_1, \theta_2; y_1, y_2) &= L(\chi(\theta_1, \theta_2)_{(1,2)}; y_1, y_2) \\ &= L(\chi(\theta_1, \theta_2)_1; y_1) * L(\chi(\theta_1, \theta_2)_{(1,2)}; y_2 | y_1) \end{aligned}$$

(with $\chi(\theta_1, \theta_2)_1$ in both likelihoods) and hence some concentration. Getting commonness is key to getting likelihood concentration and the good statistical properties associated with that.

In between parameters that are common and parameters that are arbitrarily different, there are parameters that differ by units of analysis but are drawn (or more formally are implied by exchangeability to be equivalent to being drawn) from distributions that have common parameters (common in distribution). In such cases, one would likely wish to denote such parameters as random variables χ_i^* drawn from $\Pr(\chi_i^*; \Theta)$. Here for there to be commonness, it is the Θ that must have components that repeat in the likelihoods multiplied together. This will be further clarified and more fully discussed later.

By focussing on unit of analysis likelihoods, it is made clear which parameters are common, arbitrarily different or common in distribution - by unit of analysis. For example in

$$L(\theta_1, \theta_2; y_1, y_2) = L((\mu, \sigma_1); y_1) * L((\mu, \sigma_2); y_2)$$

μ is common and the σ_i are arbitrary - $\gamma(\theta_1, \theta_2) = \mu$ and $\chi(\theta_1, \theta_2)_i = \sigma_i$. When there are common in distribution parameters, unobserved random parameters differ by study but are related by being drawn from the same "common" distribution. For example in

$$L(\theta_1, \theta_2; y_1, y_2) = L((\mu_1 \tilde{} N(\mu|\Theta), \sigma_1); y_1) * L((\mu_2 \tilde{} N(\mu|\Theta), \sigma_2); y_2)$$

Θ is common, μ_i random (implied by exchangeability of the means of y_1 and y_2) and the σ_i are arbitrary - $\gamma(\theta_1, \theta_2) = \Theta$, with μ_i unobserved and hence integrated out[7][25] and $\chi(\theta_1, \theta_2)_i = \sigma_i$.

Given likelihoods combine by multiplication, this combination is likely best displayed as addition on the log scale. An early proponent of this was Fisher [16]. On page 75, he states

> "It is usually convenient to tabulate its [the likelihoods] logarithm, since for independent bodies of data such as might be obtained by different investigators, the "combination of observations" requires only that the log-likelihoods be added."

We have adopted this suggestion for plotting log likelihoods and also log priors and log posteriors - all on the log scale.

Furthermore on page 165 Fisher states the need to use the likelihood function itself rather than approximations

> "...it is the Likelihood function that must supply all the material for estimation, and that the ancillary statistics obtained by differentiating this function are inadequate only because they do not specify the function fully."

providing an early warning against the uncritical use of dual plots. For more modern warnings, see Meng[25].

Apart from convenience of always being able to obtain unit of analysis likelihoods that simply combine by multiplication to give the overall likelihoods, why is it important to plot likelihoods? Perhaps because many, if not most techniques in applied statistics are based on or are closely related to the likelihood[3]? Essentially, under an assumed model, the likelihood captures all "useful inputs (from the data)" for statistical procedures. Why the likelihood is so useful in developing statistical techniques - in fact so useful as to eliminate the need for any other aspect of the observations - has been the subject of a long literature. From a mainly non-Bayesian perspective, Barndorff-Nielsen[2] states

> "Likelihood is the single most important concept of statistics." and further states it is mainly just the relative likelihood - "We are almost always interested only in relative values of the likelihood at different values of θ."

Pace and Salvan[32] further suggest sufficiency as a basis to restrict one's attention to relative likelihood values as they are sufficient. More formally, the distribution of any statistic conditioned on the observed relative likelihood is independent of the parameters (in the assumed model) - and hence they can provide no further "information". This would argue that it is only important that any plots maintain and highlight "relative" features of the likelihoods - height (y axis) is purely arbitrary.

Recalling the definition of the likelihood as the probability of re-observing what was actually observed - a more intuitive explanation of the primacy of the likelihood in frequency statistics can be given here. Under any data model specification there is only a finite set of possible observations (since without lack of generality continuous observations are excluded) and their probabilities are

fully determined by the data model for each parameter point (value). If for every parameter point in the data model, these probabilities are different, then in principle the is a one to one function from these probabilities of possible observations to the parameter points. If not, the data model is not identifiable and no amount of data would ever be able to discern which of the various parameters points gave rise to the observed data - no matter how much data is observed. Recall that the likelihood records the various probabilities of the actual observations observed for each parameter point. So in principle, with an unending amount of data collection, this would allow the parameter point actually sampled from to be identified (technically the *mle* is said to be consistent). Importantly, this could be accomplished with simply the relative versus absolute probabilities of the possible observations. Now, the holy Grail of frequency statistics, is perhaps the construction of a set of parameter points, that for a known and set percentage times under repeated sampling, catches the actual parameter point sampled from.

This is known to be impossible unless "set percentage" is relaxed to either less than a set percentage or roughly within (+ or -) a set percentage. For simple data models, especially those involving a scalar parameter, getting such a set of parameter points can be quite straight forward. From a Bayesian perspective, the relative likelihood is all that is needed to get the posterior and how often intervals from the posterior would catch true parameter points may seem irrelevant. Many though would argue that it should not be too bad at this.[37][15][22]

It is unargued though, that the joint model (data model plus prior model) is paramount in any Bayesian analysis. So here, both the data model and prior model needs to be plotted. It would be a mistake to just plot the posterior (joint model conditioned on the observed data) because, as we will see, the prior and likelihood (apparently separable components) may highly conflict with each other. How the marginal prior is modified to obtain the marginal posterior may not be at all obvious - see the meta-analysis example below. This would indicate that the joint model is "too wrong" to be taken seriously as a model for the research in hand. Perhaps more importantly for applied Bayesian analysis, the joint model may never have been meant as or more or less faithful model for the research in hand - where the priors need to represent salient subjective probabilities - nothing less nor nothing more. In such a case, the conditioning of that joint model on the observed data does not provide salient posterior probabilities - but only formal probabilities that are not related directly for uncertainties for the research in hand. Here plots may be even more necessary to help discern the role the priors played in the credible intervals they were involved in determining.

Perhaps arguably more important - the multivariate likelihood should not simply be considered as a nuisance black box for obtaining marginal posteriors for a parameter of focus from a multivariate prior.

Given the plots pragmatically need to be only one or a few dimensions, only marginal priors, marginal likelihoods and marginal posteriors will be plotted. Fortunately, for a fully Bayesian approach, unless one insists on using a marginal prior other than the one implied by the full dimensional prior, the marginal posterior that results from the displayed marginal prior added to the marginal likelihoods does correspond exactly (up to proportionality) to what would be obtained from the full posterior later marginalized to the chosen parameter of focus. In a full Bayesian analysis, there will be a full multivariate prior which will provide a specific integrated likelihood for each parameter of focus, but we need to verify that the true marginal posterior that results from the full multivariate prior times full multivariate likelihood will be proportional to this specific marginal likelihood times marginal prior. The full multivariate prior can be factored into the parameter of focus θ and other parameters λ

$$\pi^B(\theta, \lambda) = \pi^B(\theta)\pi^B(\lambda|\theta)$$

and where the superscript B indicates this is the intended prior for the Bayesian analysis. The question becomes when would the true marginal posterior for just θ which integrates out the λ -

$$\pi^B(\theta|y) = \int L(\theta, \lambda; y)\pi^B(\theta)\pi^B(\lambda|\theta)d\lambda$$

- be proportional to the marginal likelihood for just θ times marginal prior for just θ, i.e. would $\pi(\theta|y) \propto \pi^B(\theta)L(\theta; y)$?

Now recall that the full posterior for (θ, λ) is

$$\pi^B(\theta, \lambda|y) \propto L(\theta, \lambda; y) * \pi^B(\theta)\pi^B(\lambda|\theta)$$

and again the marginal posterior for θ is

$$\pi^B(\theta|y) \propto \int L(\theta, \lambda; y) * \pi^B(\theta)\pi^B(\lambda|\theta)d\lambda.$$

Any possible integrated likelihood can be written as

$$L(\theta; y) = \int L(\theta, \lambda; y)\pi(\lambda|\theta)d\lambda$$

where the lack of the subscript B indicates a possibly different conditional prior or measure used and note that

$$\begin{aligned} \pi^B(\theta)L(\theta; y) &= \pi^B(\theta) \int L(\theta, \lambda; y)\pi(\lambda|\theta)d\lambda \\ &= \int L(\theta, \lambda; y)\pi(\lambda|\theta)\pi^B(\theta)d\lambda \end{aligned}$$

and if $\pi^B(\lambda|\theta)$ is obtained from $\pi^B(\theta, \lambda)$ and is used

$$\begin{aligned} L(\theta; y)\pi^B(\theta) &= \int L(\theta, \lambda; y)\pi^B(\lambda|\theta)\pi^B(\theta)d\lambda \\ &= \int L(\theta, \lambda; y)\pi^B(\theta, \lambda)d\lambda \end{aligned}$$

So then, if the specific integrated likelihood "that should be used" is used

$$\pi^B(\theta)L^B(\theta; y) = \pi^B(\theta) \int L(\theta, \lambda; y)\pi^B(\lambda|\theta)d\lambda$$

the true posterior is recovered

$$\pi^B(\theta) \int L(\theta, \lambda; y)\pi^B(\lambda|\theta)d\lambda \propto \pi^B(\theta|y).$$

The assumption of a full dimensional prior provides the appropriate conditional prior for the parameter of focus to integrate over the other parameters in both the prior and integrated likelihoods to achieve a complete and fully accurate separability by the units of analysis. The full Bayesian analysis can be marginally displayed for any parameter of focus without any loss of any information for the parameter of focus. The marginal likelihoods obtained by integration also have theoretical support via work Relative Surprise Inference by Good and Evans[15]. Note the marginal likelihoods obtained by integration represents the amount by which the marginal prior is changed to obtain the marginal posterior. This provides support for the shape of credible intervals to match the integrated likelihood as well as Bayes factors for testing. Hence, at least for the

combined marginal integrated likelihood, there is direct theoretical motivation for plotting these.

We now look at how almost individual marginal likelihoods are obtained. In general, the full likelihood from m independent units of analyses with only common and arbitrary parameters is given as

$$\prod_{i}^{m} L(\theta, \lambda_i; y_i)$$

with θ representing the common parameter and λ_i representing the arbitrary parameters. Again, a general and computationally convenient route to focus on a given parameter by marginalizing out the other parameters by maximization, is the profile likelihood. The profile likelihood for a given θ is

$$L(\theta; y_{ij}) = \sup_{\lambda_i \in \Omega} \prod_{i}^{m} \prod_{j \in i}^{n_i} L(\lambda_i; y_{ij}, \theta) \quad j \text{ within } i \text{ given the same } \lambda_i$$

and as long as the λ_i are variation independent components (i.e. $\lambda_i \in \Omega_i$ and $\Omega_1 \times \Omega_2 \times ... \times \Omega_n = \Omega$), the profile likelihoods can be obtained by unit of analysis since

$$\sup_{\lambda_i \in \Omega} \prod_{i}^{m} \prod_{j}^{n} L(\lambda_i; y_{ij}, \theta) \; = \prod_{i}^{m} \sup_{\lambda_i \in \Omega} \prod_{j}^{n} L(\lambda_i; y_{ij}, \theta).$$

The analysis can be fully updated for the addition of a new unit of analysis by multiplication by of the above by

$$\sup_{\lambda_{i+1} \in \Omega} \prod_{j}^{n} L(\lambda_{i+1}; y_{(i+1)j}, \theta).$$

This, assumes that there is no modification of the assumed probability specifications in the later study. But as outlined above, a full and completely accurate marginal representation is available from

$$\pi^B(\theta; y) = \pi^B(\theta) \int L(\theta, \lambda; y) \pi(\lambda|\theta) d\lambda$$

and although the profile marginal likelihood is easily split into almost individual profile likelihoods, it may not provide an accurate enough approximation.

$$\pi^B(\theta; y) \sim \pi^B(\theta) \sup_{\lambda \in \Omega} \prod_{j}^{n} L(\lambda; y_j, \theta)$$

To get a better sense of this as an approximation we can look to Laplace approximation methods.

$$\pi^B(\theta; y) = \pi^B(\theta) \sup_{\lambda \in \Omega} [\frac{\pi^B(\lambda; \theta)}{\sqrt{I(\theta, \lambda)}} \prod_j^n L(\lambda; y_j, \theta)]$$

versus the proposed approximation given later.

First, we consider integrated likelihoods and how they factorize. The integrated likelihood value for a given $\theta^{'}$ becomes factorized as

$$L(\theta; y_{ij}) = \int \cdots \int \prod_i^m \prod_j^n L(\lambda_i; y_{ij}, \theta) d\lambda_i$$

and as long as the λ_i are variation independent components (i.e. $\lambda_i \in \Omega_i$ and $\Omega_1 \times \Omega_2 \times ... \times \Omega_n = \Omega$), the integrated likelihoods can be obtained separately since

$$\int \cdots \int \prod_i^m \prod_j^n L(\lambda_i; y_{ij}, \theta) d\lambda_i = \prod_i^m \int \prod_j^n L(\lambda_i; y_{ij}, \theta) d\lambda_i.$$

We now give the proposed almost individual marginalized pseudo integrated likelihood

$$\begin{aligned} \pi^B(\theta; y_j) &\equiv \pi^B(\theta) \frac{\int L(\theta, \lambda; y) \pi(\lambda|\theta) d\lambda}{L(\hat{\lambda}_\theta; y_j, \theta)} L(\hat{\lambda}_\theta; y_j, \theta) \\ \text{where } \hat{\lambda}_\theta &= \{\lambda : \sup_{\lambda \in \Omega} \prod_i^n L(\lambda; y_i, \theta)\} \end{aligned}$$

The sole motivation being that this pseudo likelihood adds on the log scale exactly to the combined integrated log likelihood.

The convenient factorizations pointed out above, do not continue to hold when the full likelihood from m independent units of analyses has additional common in distribution (random) parameters. The full likelihood for a given unit of analysis i is given then as

$$L^i(\theta, \lambda_i, \Theta, ; y_{(1,..i,..m)j}) = \int \prod_j^n L(\theta, \alpha_i \tilde{} g(\alpha|\Theta), \lambda_i; y_{(1,..i,..m)j}) dg(\alpha|\Theta)$$

where the $y_{1,..i,..m}$ emphasizes that data from all other units of analysis are required, now with θ and Θ representing the common parameters and λ_i representing the arbitrary parameters. Here, clearly there will be not factorization by unit of analysis as all the data will needed from each and every unit of analysis likelihood and the subsequent profiling or integrating of the remaining λ_i

and Θ must be done jointly over all the units of analysis. The analysis can not be fully updated for the addition of a new unit of analysis just simply by multiplication of the marginalized likelihood for the old units of analysis by the marginalized likelihood for the new unit. But again, once the marginal likelihood is updated using all the data, it can be displayed as a multiple of the revised previous unit of analysis likelihoods and the new one.