BAYESIAN STATISTICS 9
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© Oxford University Press, 2010

What's the H in H-likelihood: A Holy Grail or An Achilles' Heel?

Xiao-Li Meng
Harvard University, USA
meng@stat.harvard.edu

Summary

H-likelihood refers to a likelihood function of both fixed parameters and random "unobservables," such as missing data and latent variables. The method then typically proceeds by maximizing over the unobservables via an adjusted profile H-likelihood, and carries out a Fisher-information-like calculation for (predictive) variance estimation. The claimed advantage is its avoidance of all "bad" elements of Bayesian prediction, namely the need for prior specification and posterior integration. This talk attempts to provide an in-depth look into one of the most intriguing mysteries of modern statistics: why have the proponents of the H-likelihood method (Lee and Nelder, 1996, 2001, 2005, 2009) been so convinced of its merits when almost everyone else considers it invalid as a general method? The findings are somewhat intriguing themselves. On the one hand, H-likelihood turns out to be Bartlizable under easily verifiable conditions on the marginal distribution of the unobservables, and such conditions point to a transformation of unobservables that makes it possible to interpret one predictive distribution of the unobservables from three perspectives: Bayesian, Fiducial and Frequentist. On the other hand, the hope for such a Holy Grail in general is diminished by the fact that the log H-likelihood surface cannot generally be summarized quadratically due to the lack of accumulation of information for unobservables, which seems to be the Achilles' Heel of the H-likelihood method.

Keywords and Phrases: Bartlizability; Bayesian inference; Fiducial inference; Latent variables; Missing Data; Predictive inference; Predictive likelihood; Predictive pivotal quantity; John Nelder.

Xiao-Li Meng is the Whipple V. N. Jones Professor of Statistics at Department of Statistics, Harvard University. The author sincerely thanks Professor Jose Bernardo for 1) inviting him to attend his first and the last Valencia meeting, 2) helping him to deal with the world's most indescribable visa application process, and 3) allowing him to be the last to submit this proceeding article. The author also thank Alan Agresti, Joseph Blitzstein, Tony Desmond, Thomas Louis, Piero Veronese, and Xianchao Xie for comments and proofreading, and National Science Foundation (US) for partial financial support.

1. WHAT IS H-LIKELIHOOD?

In order to answer the question in the article title, one first needs to answer the question in the section title above. Among several hundreds of attendees of Valencia 9, few raised their hands when I posed the latter question at the beginning of my talk. Of course I was not trying to identify who had read my talk abstract (same as this article's abstract), but rather to demonstrate that H-likelihood is not a well understood or even well-known notion.

There is a good reason for this lack of general recognition. Technically, the H-likelihood is very easy to define and understand. As my talk was based on my discussion (Meng, 2009a) of Lee and Nelder (2009), I shall follow the notation there (and only repeat its essence in this sequel). Let y denote our observation, v be any random "unobservable" such as missing data or latent variables that we want to or need to include in our model, and $f_\theta(y, v)$ be the joint probability distribution/density of $\{y, v\}$, where θ is the model parameter. The H-loglikelihoodthen is defined as

$$h(\theta, v; y) = \log f_\theta(y, v). \tag{1}$$

In other words, the only difference between an H-likelihood and an ordinary likelihood (see below) is that the former will include any "unobservables" also as a part of the argument of the likelihood function.

Most of us who have taken a basic course in parametric statistical inference may recall how our teachers emphasized that all "unobservables," other than the parameter θ, must be integrated out before forming a likelihood function (and before taking the log). That is, in contrast to (1), the ordinary log-likelihood function of θ is given by

$$\ell(\theta; y) \equiv \log f_\theta(y) = \log \int f_\theta(y, v)\mu(dv), \tag{2}$$

where the baseline measure μ depends on the problem at hand. Some of us surely had lost points on a homework or on an exam for accidently having used (1) instead of following the well accepted definition (2) (but see Bayarri, DeGroot and Kadane, 1988 and Berger, Liseo and Wolpert, 1999, for discussions on the lack of unique definition of a likelihood function in general).

I effectively did, on an exam at Fudan University in late 1970s when I took my first course in statistics. Had the "Bible" on the likelihood principle, Berger and Wolpert (1988), been available then and had my teacher read it, he might have given me extra points instead of deducting some. On page 21.2, Berger and Wolpert (1988) explicitly stress that the argument of a likelihood function should include "*all* unknown variables and parameters that are relevant to the statistical problem." (Emphasis is original.) They even went on to separate unobservable variables of interest from nuisance unobservable variables, just as we specify parameters of interest and nuisance parameters (see also Berger, Liseo and Wolpert, 1999).

Despite Berger and Wolpert's (1988) emphasis, few would be surprised to see homework or exam points continuously taken away if a student adopts (1) in place of (2). Indeed, I would not be reluctant to do the same to my students if they repeat what I did on my first statistics exam, unless they justify properly what they would *do* with (1). As I argued in Meng (2009a), there is nothing wrong with (1) as a definition. The central reason for its lack of general recognition is that there has not been well established (non-Bayesian) methods and theory, with appreciable generality, for making valid inference based on (1). Rather, there are an array of examples in the literature, some of which were reviewed in Lee and Nelder (2009) and its discussions (by Louis, Molenberghs *et. al.*, and myself), that demonstrate the kind of erroneous results from applying the methods established for (2) to (1). For example, maximizing over both θ and v often leads to an inconsistent estimator for θ and meaningless prediction for v.

For Bayesians, (1) is merely the log of the joint posterior of θ and v under constant prior on θ (up to a normalizing constant), and hence there is no need of a separate principle or justification. However, some researchers have been making persistent attempts to establish a framework for drawing valid inferences based on (1) by generalizing the standard likelihood methods such as MLE and profile likelihood that were designed for (2). Lee and Nelder (2009) appears to be the latest installment in this pursuit. The study I conducted in preparing Meng (2009a) convinced me that this pursuit would likely be an indefinite one, and this sequel provides additional reasons for my conviction. This sequel also makes links to several highly relevant articles in the literature, which unfortunately I was not aware of at the time of writing Meng (2009a).

2. OPTIMIZATION OR INTEGRATION?

At the heart of the matter is an age-old but critical question: how should we "marginalize" out nuisance quantities in our inference? I put "marginalize" in quotation marks because the term *marginalization* has different meanings for Bayesians and for likelihoodists.

For example, whereas mathematically a likelihood function is a (possibly unnormalized) posterior density or a probability function under the constant prior, a "marginal likelihood" is not necessarily a special case of "marginal posterior" in the same sense or in any sense. The meaning of the term "marginal posterior" has little ambiguity, but the term "marginal likelihood" has been used in the literature to mean very different quantities. When it is used as a synonym for *integrated likelihood* (Kalbfleisch and Sprott, 1970, 1974; Berger, Liseo and Wolpert, 1999), its use is consistent with the Bayesian meaning, namely, integrating out the nuisance parameter in a likelihood. This includes the naming of "marginal likelihood" as "evidence", that is, the density/probability of the data with all parameters integrated out, as in Bayes factor calculation (e.g., Kass and Raftery, 1995; Meng and Schilling, 2002). However, the term "marginal likelihood" has also been used in the sense of "partial likelihood" (Cox, 1975a), that is, when only a "marginal" part of data is used for forming the likelihood, such as in Kalbfleisch and Sprott's (1970) definition. In such cases, the marginalization is done on the data space, not on the parameter space. To add to the confusion, Lee and Nelder (2009) termed the ordinary likelihood (2) as a marginal likelihood of the H-likelihood (1); this naming is more consistent with the "partial likelihood" usage, as $f_\theta(y)$ is a marginal distribution/density of $f_\theta(y, v)$ on the data space (including the unobservables).

Regardless of its meaning, all marginalization processes mentioned in the preceding paragraph are carried out via *integration*, which is a probabilistic operation in the sense that the resulting function remains to be a (un-normalized) probability density if the parental function being marginalized is such on the joint space. In contrast, the well-known profile likelihood method and its many "adjusted" variations (e.g., Cox and Reid, 1987, 1993; Barndorff-Nielsen, 1994) achieve "marginalization" via maximization, which is not a probabilistic operation in general because the resulting profiled likelihood, adjusted or not, may not have any probabilistic interpretation, on either the parameter space or the data space. This of course is a well-known fact (e.g., Ghosh, 1988), and indeed the issue of integration verses maximization has been discussed at length in the literature (e.g., Berger, Liseo and Wolpert, 1999; Bjørnstad, 1999). It is therefore difficult to add anything really new. My intention here is to emphasize that the distinction between integration and maximization, or more generally optimization, is more blurred than what meets our eyes. The blurriness of course is not about the two operations mathematically, but rather about the underlying principles that lead to their adoption as tools for marginalizing out the nuisance quantities.

On the surface, the difference between integration and optimization is obvious, even from the inference point of view. On the one hand, optimization has this obvious "intuitive" appeal, which has seduced many investigators across all fields. What can possibly be better than "optimal?" For all those occasions where I needed to explain a statistical concept or method to a novice, explaining methods such as least-squares fitting or the maximum likelihood estimator turns out to be among the easiest. The person might not understand the concept of regression or likelihood at all, but whenever I said "Let us find the best-fitting line" or "Let us seek the most likely value", the frequency of head nodding just went up. What could possibly be more plausible than the parameter value that maximizes the probability/density of the data (in the absence of an informative prior)? But on the other hand, the very "most likely" appeal is a recipe for disasters in terms of overfitting – the lack of probabilistic propagation of uncertainties is the culprit. Indeed, the vast majority of the examples of the failure of MLE that I am aware of are examples where the signal/information in the data is not strong enough to overcome the overfitting (e.g., due to too many parameters); these include the well-known Neyman-Scott problem (Neyman and Scott, 1948) and the H-likelihood examples reviewed and discussed in Meng (2009a).

The danger of maximization was well emphasized by Berger, Liseo and Wolpert (1999), who argued effectively the safety of using integration. It is also safer psychologically because the very use of the term *integration* reminds us that more than one state needs to be explicitly taken into account, in contrast to *optimization* which puts all our stake on one state. Berger, Liseo and Wolpert (1999) also provided a list of reasons why an integrated likelihood can be viewed on its own, not as a special case of a posterior density. Nevertheless, its strong Bayesian flavor is hard to mask, especially with its explicit use of a "weight function" for the nuisance parameter, i.e., the conditional prior of the nuisance parameter given the parameter of interest, just as with the partial Bayes framework (Cox, 1975b, McCullagh, 1990, Meng, 1994, 2009b). Perhaps because of this strong association between using integration and Bayesian methods, those who do not wish to be associated with the Bayesian school have tried hard to avoid using integration for eliminating nuisance quantities. The recent literature on the H-likelihood, as represented by Lee and Nelder (1996, 2001, 2005, 2009) and Lee, Nelder and Pawitan (2006), highlights this effort.

But are their methods truly maximization-based? Initially Lee and Nelder (1996) adopted the same MLE recipe for (1) as for (2), that is, maximizing over both θ and v in arriving at point estimation for both θ and v, the so-called MHLE (maximum H-likelihood estimate). After a number of discussants and authors—see Meng (2009a) for details—pointed out that MHLE often leads to inconsistent or even meaningless estimators (e.g., always taking the value ∞), Lee and Nelder (2001) adjusted their approach by adopting APHL (adjusted profile H-likelihood) for inference about the unobservable v, with the inference for θ restored to be based on the ordinary "marginal" likelihood as given in (2).

To see the essence of APHL, let us follow the notation of Lee and Nelder (2009), who adopted the notation of $\ell = \ell(\alpha, \psi)$ for a log likelihood, which can be either (2) or (1), where ψ is the quantity of interest and α is the nuisance quantity — here "quantity" can be either a fixed parameter or a random unobservable. Lee and Nelder (2009) then presented APHL as

$$p_\alpha(\ell; \psi) = \left[\ell - \frac{1}{2}\log\det\{D(\ell, \alpha)/2\pi\}\right]\Bigg|_{\alpha=\tilde{\alpha}}, \tag{3}$$

where $D(\ell, \alpha) = -\partial^2\ell/\partial\alpha^2$ and $\tilde{\alpha}$ solves $\partial\ell/\partial\alpha = 0$. Below we will replace $p_\alpha(\ell; \psi)$ with a mathematically more precise notation $p_{\ell,\alpha}(\psi; y)$, which makes it clear that APHL is a function of the quantity of interest ψ and data y only, but its *functional form* is determined by the choice of ℓ and α.

Although no integration is carried out in reaching (3), Lee and Nelder (2001) noted that "for random effects v the use of $p_v(\ell)$ is equivalent to integrating them out." (Lee and Nelder's (2001) $p_\alpha(\ell)$ is the same as Lee and Nelder's (2009) $p_\alpha(\ell; \psi)$.) This is because the right-hand side of (3) is the first-order Laplace approximation to $\log[\int \exp\{\ell(\alpha, \psi)\}d\alpha]$ (see Reid, 1996) – the irrelevant constant 2π (for defining likelihood) is a give-away.

Consequently, for Baysians, no additional principles or justifications are needed because the APHL is merely a convenient approximation to the log of marginal posterior of the quantity of interest. The writing of Lee and Nelder, in their series of articles cited above, makes it clear that their goal is to make inference about the unobservables without resorting to the Bayesian framework. At the same time, they emphasized that "We dislike the use of estimation methods without a probabilistic basis, because, for example, inferences for joint and conditional probabilities are not possible." (Lee and Nelder, 2009). This emphasis or desire, together with their APHL, makes it particularly difficult to decide whether we should classify APHL as a maximization method on its own or as an approximate integration method. The former classification carries no probabilistic justification. The latter does, but only when it is viewed as an approximate Bayesian method, a view that Lee and Nelder want to avoid. The central question thus lingers: is there a non-Bayesian but probabilistic-based principle for APHL?

3. BUT DOES IT REALLY MATTER?

From a practical point of view, some may question whether it really matters if APHL has its own principle or it somehow relies on the Bayesian principle. The following simple (but not toy) example illustrates that it does matter, precisely from a practical point of view.

Let $y = \{y_1, \ldots, y_n\}$ be i.i.d. observations from an exponential distribution with mean λ, and $u = y_{n+1}$, a future realization, is our unobservable. (We use the notation u instead of the generic v to follow the notation of Meng (2009a) for the same example; the following discussion supplements my investigation there, where v is reserved for the "right scale," as further discussed below.) Our task is to estimate λ as well as to predict y_{n+1}. This is an extremely simple model yet it has many applications (e.g., in reliability testing). It is illogical to expect its general applicability when a method cannot handle such a simple and common case. Yet whether APHL can handle this case depends on which principle one adopts.

Specifically, it is easy to see that the H-loglikelihood(1) in this case is given by

$$h(\lambda, u; y) = -(n+1)\log\lambda - \frac{n\bar{y}_n + u}{\lambda}. \tag{4}$$

Hence for any fixed u, it is maximized by

$$\lambda(u) = \frac{n\bar{y}_n + u}{n+1}. \tag{5}$$

Consequently, the APHL of (3) for u is, using our notation (and ignoring an irrelevant constant term),

$$p_{h,\lambda}(u; y) = -n\log\lambda(u). \tag{6}$$

This is a strictly monotone decreasing function of u. Hence, when it is treated as a "log-likelihood" and maximized, it would lead to $\hat{u}_{APHL} = 0$, regardless of the data.

In contrast, if we recognize that (6) is intended as an approximation to the log of the marginal (predictive) posterior of u under the constant prior for λ, then we can "recover" the posterior density as

$$p(u|y) \propto \exp\{p_{h,\lambda}(u; y)\} = \left[\frac{n+1}{n\bar{y}_n + u}\right]^n. \tag{7}$$

If we let $r = u/\bar{y}_n$, then (7) is equivalent to setting the posterior predictive density for r to be

$$p(r|y) = \frac{n-1}{n}\left(1 + \frac{r}{n}\right)^{-n}, \quad r \geq 0, \tag{8}$$

a Pareto distribution with order n. Clearly no one would/should use its mode for point estimation! The mean of r from (8) is $n/(n-2)$ when $n \geq 3$, and hence the posterior predictive mean for $u = y_{n+1}$ given y is $\hat{u}_1 = [n/(n-2)]\bar{y}_n$. This point prediction is not perfect because of the multiplier $n/(n-2)$ (an issue that will be discussed shortly), but it is certainly far more sensible than $\hat{u}_{APHL} = 0$! Note that this imperfection is not due to the Laplace approximation, which in fact is exact in terms of the functional form for u; the approximation is in the normalizing constant and hence it becomes immaterial after the re-normalization, as done in (8). This can be verified directly because integrating out λ in

$$p(\lambda, u|y) \propto \lambda^{-(n+1)}\exp\{-(n\bar{y}_n + u)/\lambda\}$$

will give the same function form as in (7). In other words, (8) is identical to the actual posterior predictive distribution of r given y, under the constant prior on λ.

Lee and Nelder (2009) showed that the problem of $\hat{u}_{APHL} = 0$ is avoided if one uses $v = \log u$ as the unobservable. In general, they emphasized that the choice of the scale for unobservables "in defining the H-likelihood is important to guarantee the meaningfulness of the mode estimation." This emphasis itself is an indication that it is the integration/Bayesian principle in guidance rather than the maximization/likelihood recipe in play because maximization is invariant to (one-to-one) transformations, whereas integration is not and hence a choice needs to be made. Lee and Nelder (2009) noted in particular that when normality holds approximately, their APHL method worked well. Whereas the normality assumption is obvious and sufficient, it is by no means necessary — one can find various examples where treating APHL as a regular log-likelihood and maximizing it will deliver acceptable results (at least in terms of point estimators), and yet the APHL curve is far from normal.

In fact, for our current example, with $v = \log(u)$, the APHL curve is (see Meng, 2009a)

$$p_{h,\lambda}(v; y) = -n \log(n\bar{y}_n + e^v) + v. \tag{9}$$

This clearly is far from being a quadratic function of v; indeed, for large v, it behaves like $-(n-1)v$. Nevertheless, it is maximized at $\hat{v} = \log(\bar{y}_n) + \log[n/(n-1)]$, which leads to the point estimate for u as $\hat{u}_2 = [n/(n-1)]\bar{y}_n$, almost identical to $\hat{u}_1 = [n/(n-2)]\bar{y}_n$, the posterior mean from (8). Note that on the v scale, if we directly maximize the H-loglikelihood(1), we would have arrived at the "perfect" estimator, $\hat{u}_3 = e^{\hat{v}_{MHLE}} = \bar{y}_n (= \hat{\lambda})$, as shown in Meng (2009a). The extra factor $n/(n-1)$ in $\hat{u}_2$ is due to the adjustment in the profile loglikelihood, because the original unadjusted profile loglikelihood is

$$p_{h,\lambda}(v; y) = -(n+1) \log(n\bar{y}_n + e^v) + v, \tag{10}$$

which is maximized at $\hat{v}_{MHLE} = \log(\bar{y}_n)$. This difference reflects the difference between joint MHLE for v from H-likelihood, which corresponds to (10), and the marginal MHLE for v from its marginal H-likelihood (with λ integrated out), which corresponds to (9).

Regardless of which MHLE we are after, it is clear that the v scale is far better than the original u scale. Mathematically speaking the reason for the scale to matter is the Jacobian factor needed to preserve probability mass via integration/transformation. A question of both theoretical and practical interest then is if there is any general theoretical result to guide the discovery of such scales. This led to the Bartlization results reported in Meng (2009a).

4. BARTLIZATION: AN HEROIC EFFORT?

As reviewed in Meng (2009a), a theoretical backbone for Fisher's ML paradigm is the Bartlett identities, especially the first two. That is, under mild regularity conditions, the (marginal) log-likelihood $\ell(\theta; y)$ of (2) satisfies

$$\mathrm{E}_\theta \left[\frac{\partial \ell(\theta; y)}{\partial \theta} \right] = 0, \quad \forall\, \theta \in \Theta, \tag{11}$$

and

$$\mathrm{E}_\theta \left[\frac{\partial^2 \ell(\theta; y)}{\partial \theta^2} \right] + \mathrm{E}_\theta \left[\left(\frac{\partial \ell(\theta; y)}{\partial \theta} \right) \left(\frac{\partial \ell(\theta; y)}{\partial \theta} \right)^\top \right] = 0, \quad \forall\, \theta \in \Theta, \tag{12}$$

where E_θ denotes the expectation under $f_\theta(y)$. The first Bartlett identity (11) ensures that the score function, $S(\theta; y) \equiv \partial\ell(\theta; y)/\partial\theta$, is *unbiased*, and the second identity (12) ensures it to be *information unbiased*, in the terminology of Godambe (1960) and Lindsay (1982). That is, from the estimating equation point of view, identity (11) is responsible for the consistency of MLE (more precisely of a root of the score equation). Identity (12) is the key in establishing that the score function is the optimal estimating equation, in the sense that

$$I^{-1}(\theta)\mathrm{Var}_\theta\left[S(\theta; y)\right] I^{-1}(\theta) \leq I_G^{-1}(\theta)\mathrm{Var}_\theta\left[G(\theta; y)\right] I_G^{-\top}(\theta), \quad \forall\ \theta \in \Theta, \tag{13}$$

for every $G \in \mathcal{G}$, the class of regular (unbiased) estimating equations as defined in Godambe (1960, 1976). Here, $I(\theta)$ is the usual expected Fisher information, and $I_G(\theta)$ is its generalization to unbiased estimating function G:

$$I_G(\theta) = \mathrm{E}_\theta\left[-\frac{\partial G(\theta; y)}{\partial\theta}\right], \tag{14}$$

where all expectations are with respect to $f_\theta(y)$ of (2). Note unlike $I(\theta)$, $I_G(\theta)$ is not in general guaranteed to be positive semi-definite, or even be symmetric (and of course it may not exist, just as Fisher information may not exist).

As a side note, it is interesting that Godambe (1960) did not motivate the "sandwich" criterion in (13) from its obvious asymptotic justification, namely, its right-hand side is the asymptotic variance of any root of $G(\theta; y) = 0$ (under regularity conditions). Rather, it was motivated by the desire to have G as good an estimate of its mean, that is, zero, as possible (and hence smaller $\mathrm{Var}[G(\theta; y)]$), and to make G as sensitive as possible as a function of θ (and hence larger derivative, in magnitude, with respect to θ). Perhaps this was driven by the desire to provide a deeper insight via revealing individual ingredients of the "sandwich," and/or the desire to make a direct finite-sample generalization of the Cramér-Rao lower bound.

The optimality as formalized in (13) also holds more generally for conditional score functions; see Godambe and Thompson (1974), Godambe (1976), Lindsay (1980, 1982), and especially a comprehensive and very readable discussion paper by Desmond (1997; incidently, John Nelder was one of the discussants). Because a score function naturally possesses the Bartlett identities (11)-(12) and hence these identities effectively become necessary (but by no means sufficient) conditions for achieving optimality (13), efforts have been made throughout the literature to construct estimating functions that are both unbiased and information unbiased, such as with quasi-likelihood (e.g., McCullagh and Nelder, Chapter 9, 1989) and with profile likelihood (e.g., McCullagh and Tibshirani, 1990). As emphasized in McCullagh and Tibshirani (1990), identities (11) and (12) hold for regular likelihood (2) because it permits differentiation under integration (under mild regularity conditions):

$$\frac{\partial}{\partial\theta}\mathrm{E}_\theta[T(\theta; y)] = \mathrm{E}_\theta\left[\frac{\partial T(\theta; y)}{\partial\theta} + T(\theta; y)S^\top(\theta; y)\right], \tag{15}$$

where $T(\theta; y)$ is an arbitrary function but is differentiable with respect to θ (and all quantities in (15) are well defined).

Clearly taking $T(\theta; y) = 1$ leads to (11) and consequently taking $T(\theta; y) = S(\theta; y)$ yields (12). However, when we replace θ by $\phi = (\theta, v)$ as required by the

H-loglikelihood (1), (15) no longer makes sense because the unobservable v is a part of the integration variable and θ remains to be fixed, and hence the E_θ notation is unchanged in this replacement. Consequently, it is quite logical to suspect that Bartlett identities will not hold in general for H-likelihood, which would be an explanation why H-likelihood cannot be handled as a regular likelihood. It therefore was somewhat a surprise (at least to me) that it turns out that there exist almost trivially verifiable sufficient and necessary conditions on v such that the Bartlett identities hold for H-likelihood, as given in Theorem 1 and Theorem 2 of Meng (2009a). Perhaps the most surprising aspect of these results is that the required conditions only involve the *marginal* distribution of the unobservable v, and hence they can be checked (almost) irrespective of the observed-data loglikelihood (2).

In particular, as long as the density of $v, f_\theta(v)$, vanishes on the boundary of its support, the first Bartlett identity holds for the H-loglikelihooddefined in (1). Furthermore, if $f_\theta^{(1)}(v)$ *also* vanishes on the same boundary, where $f_\theta^{(k)}(v)$ denotes the k-th derivative with respect to v, then the second Bartlett identity holds as well. And these conditions are almost necessary (see Theorem 1 of Meng, 2009a, for the precise results, which are also illustrated in the following section). As verified in Meng (2009a), for our exponential example, the conditions are violated for the original scale $u = y_{n+1}$ because the exponential density $f_\theta(u) = \theta e^{-\theta u}$, $u \in R^+$ does not vanish at $u = 0$ as long as $\theta = \lambda^{-1} > 0$ (which always hold for $0 \leq \lambda < \infty$). However, once we transform it to $v = \log(u)$, $f_\theta(v) = \theta e^{v-\theta e^v}, v \in R$ vanishes on both $v = -\infty$ and $v = \infty$, as does its derivative, for any $\theta > 0$. The Bartlett identities (11)-(12) therefore hold for the H-likelihood when the unobservable is "parameterized" as $v = \log(y_{n+1})$.

The existence of such easily verifiable conditions for establishing the "right" transformation for the unobservables, a process that can be termed as Bartlization (Meng, 2009a), seems to lend some encouragement to the H-likelihood research (see Section 6). Part of the excitement is that in this example the v scale leads to "3-in-1." That is, under the common default prior, the constant prior on $\log(\theta)$ (not on θ), the posterior predictive distribution, the sampling pivotal predictive distribution (also can be viewed as a fiducial distribution), and the h-distribution (i.e., by exponentiating AHPL, as done in (7)) for $r = y_{n+1}/\bar{y}_n$ are all Pareto distribution of order $n + 1$, that is,

$$p(r|y) = \left(1 + \frac{r}{n}\right)^{-(n+1)}, \quad r \geq 0, \tag{16}$$

as shown in Section 7 of Meng (2009a).

Such "3-in-1", if it can be made to hold in general, of course would be a Holy Grail, as it unifies Bayesian, frequentist and fiducial perspectives. Unfortunately, but not surprisingly, this unification remains to be the legendary Holy Grail. Or as Professor Ed George, the discussant at my Valencia 9 presentation, put it, "H" stands for heroic effort, which is laudable but it also indicates potentially unsurmountable difficulties. Indeed, the unsurmountable difficulty of the MHLE methods is what I labeled, in Meng (2009a), as the lack of *accumulation of information* for unobservables by which I meant the following. Let $\hat{\phi}$ be the (joint) MHLE, and

$$S_h(\phi; y) = \frac{\partial h(\phi; y)}{\partial \phi} \quad \text{and} \quad I_h(\theta) = \mathrm{E}_\theta\left[-\frac{\partial S_h(\phi; y)}{\partial \phi}\right] \tag{17}$$

be the H-score and H-information (a.k.a. the expected Hessian information) respectively. Then the usual Taylor expansion of $S_h(\hat{\phi}; y) - S_h(\phi; y)$ leads to

$$\hat{\phi} - \phi = I_h^{-1}(\theta) S_h(\phi; y) + R. \tag{18}$$

If this is for the regular likelihood (2), then suitable regularity conditions would guarantee that the corresponding remainder term $R \to 0$ as the data size goes to infinity (or more generally as the Fisher information goes to infinity). This, however, cannot be made true for H-likelihood in general, regardless of whether the Bartlett identities hold or not. This is because no matter how much data we have, we cannot, for example, predict a future observation with certainty. The data only help us to learn as much as possible about our model. But even if we know our model perfectly, there is still uncertainty about a future realization, an uncertainty precisely our model intends to capture. Therefore, even if the first term on the right-hand side of (18) has mean zero and variance $I_h^{-1}(\theta)$, which is a direct consequence of the Bartlett identities, we still cannot use it to approximate the distribution of $\hat{\phi} - \phi$ because R may not be negligible; see Meng (2009a) for a detailed demonstration. The following extension of that demonstration illustrates further that Bartlization is by no means sufficient, and even within the Bartlized class of transformations, the choice of scale can still have significant impact even asymptotically, precisely because of the Achilles' Heel, that is, R fails to converge to zero.

5. HOW MUCH CAN THE BARTLIZATION PROCESS HELP?

For our exponential example, let us consider a general transformation of $u = y_{n+1}$ via $u = B(w)$, a function from $S_B = [a, b]$ to $[0, \infty)$, where a and/or b can be infinity. To simplify mathematics, we will assume $B(w)$ is monotone increasing and its kth order derivative $B^{(k)}(w)$ exists for at least $k \leq 3$. Under such a setting, the marginal density of w is given by

$$f_\lambda(w) = \frac{1}{\lambda} \exp\left\{ -\frac{B(w)}{\lambda} \right\} B^{(1)}(w). \tag{19}$$

Theorem 1 of Meng (2009a) implies that the first Bartlett identity holds for the corresponding H-loglikelihoodif and only if

$$f_\lambda(a) = f_\lambda(b), \text{ for all } \lambda > 0; \tag{20}$$

and given (20), the second Bartlett identity holds if and only if

$$f_\lambda^{(1)}(a) = f_\lambda^{(1)}(b), \text{ for all } \lambda > 0. \tag{21}$$

For a density support with infinite Lebesgue measure, the easiest way to make (20) and (21) hold is to make all quantities there zero, i.e., "vanish on the boundary." This leads to requiring

$$B^{(1)}(a) e^{-B(a)/\lambda} = B^{(1)}(b) e^{-B(b)/\lambda} = 0, \tag{22}$$

and

$$\left[B^{(2)}(a) - \frac{[B^{(1)}(a)]^2}{\lambda} \right] e^{-B(a)/\lambda} = \left[B^{(2)}(b) - \frac{[B^{(1)}(b)]^2}{\lambda} \right] e^{-B(b)/\lambda} = 0, \tag{23}$$

for all $\lambda \geq 0$. There are obviously infinitely many functions $B(w)$ that satisfy these two sets of conditions. For example, for any $m \in (2, \infty)$, $B_m(w) = w^m$, $w \in [0, \infty)$ satisfies both (22) and (23). This indicates that more conditions are needed to pinpoint the "optimal" transformation, unless all of them are equivalent, at least asymptotically. The derivation hereafter demonstrates the possibility for the former and the impossibility of the latter.

Given a Bartlized transformation $B(w)$ (that satisfies the aforementioned monotonicity and differentiability assumptions), clearly the H-loglikelihoodis given by

$$h(\lambda, w; y) = -(n+1)\log\lambda - \frac{n\bar{y}_n + B(w)}{\lambda} + \log[B^{(1)}(w)]. \tag{24}$$

Consequently, the H-score equation becomes

$$\begin{aligned} \frac{\partial h}{\partial \lambda} &= -\frac{n+1}{\lambda} + \frac{n\bar{y}_n + B(w)}{\lambda^2} = 0, \\ \frac{\partial h}{\partial w} &= -\frac{B^{(1)}(w)}{\lambda} + \frac{B^{(2)}(w)}{B^{(1)}(w)} = 0. \end{aligned} \tag{25}$$

Whether the solution(s) of (25) correspond(s) to MHLE will depend on the nature of $B(w)$, but one thing is clear. That is, if we want the solution of (25) for λ to be the same as the MLE from the regular log-likelihood (2), that is $\hat{\lambda}_{MLE} = \bar{y}_n$, *for any data set,* then the following equation must hold for $B(w)$:

$$[B^{(1)}(w)]^2 = B(w)B^{(2)}(w), \quad \forall\, w \in S_B. \tag{26}$$

This now uniquely defines B up to an affine class, because (26) is equivalent to (noting $B(w) > 0$ for $w > a$) $[\log B(w)]^{(2)} = 0$ for all $w \in S_B$, which means $u \equiv B(w) = c_1 \exp\{c_2 W\}$, or equivalently

$$w = c_3 \log(u) + c_4, \quad \text{for any } c_3 \neq 0 \text{ and } c_4 \in R. \tag{27}$$

Given (27), the fact that u needs to vary from 0 to ∞ implies that S_B must be $(-\infty, \infty)$. Therefore, the log scale is the unique "optimal" Bartlization (up to an affine class) in the sense of retaining the MLE from the ordinary likelihood by MHLE. It would be quite interesting to investigate the existence and uniqueness of such optimal transformations more generally.

To illustrate that Bartlization process alone is not enough to determine even the asymptotic behavior of MHLE, let us concentrate on $B_m(w) = w^m$ when $m > 2$. In such cases, it is easy to derive from (25) that its (unique) solution is given by

$$\begin{aligned} &\hat{\lambda}_{m,n} = \frac{n}{n + m^{-1}}\bar{y}_n \\ &\hat{u}_{m,n} \equiv B(\hat{w}_{m,n}) = (1 - m^{-1})\hat{\lambda}_{m,n}. \end{aligned} \tag{28}$$

Hence, when $n \to \infty$, whereas $\hat{\lambda}_{m,n}$ is consistent for λ regardless of the value of m, $\hat{u}_{m,n}$ will converge to $u_m = (1 - m^{-1})\lambda$. No matter how one questions the meaning of "convergence" for unobservables, the fact that u_m depends on the choice of m, which clearly is an artifact of MHLE, is at least a discomfort. This result also shows the problem with taking $m = 1$ because it leads to $\hat{u}_{1,n} = 0$ regardless of the data, as we have seen before, as well as the advantage of taking $m = \infty$, which is equivalent to taking the optimal transformation $B(w) = e^w = \lim_{m\to\infty}(1 + w/m)^m$ (by changing $w = u^{1/m}$ to its affine equivalent $w = m(u^{1/m} - 1)$).

The need for $m > 2$ can also be seen from the Hessian calculation. Further differentiating (25) but with $B(w) = w^m$ yields

$$\begin{aligned}\frac{\partial^2 h}{\partial\lambda^2} &= \frac{n+1}{\lambda^2} - 2\frac{n\bar{y}_n + w^m}{\lambda^3};\\ \frac{\partial^2 h}{\partial\lambda\partial w} &= \frac{mw^{m-1}}{\lambda^2};\\ \frac{\partial^2 h}{\partial w^2} &= -\frac{m(m-1)w^{m-2}}{\lambda} - (m-1)w^{-2}.\end{aligned} \tag{29}$$

Noting that $w = u^{1/m}$ and hence

$$\mathrm{E}_\lambda(w^k) = \mathrm{E}_\lambda(u^{k/m}) = \lambda^{k/m}\Gamma(1 + k/m) \tag{30}$$

for any k such that $1 + k/m > 0$. Consequently, the expected Hessian matrix is given by (where $\phi = (\lambda, w)$)

$$I_h(\lambda) = E_\lambda\left[-\frac{\partial^2 h}{\partial\phi^2}\right] = \begin{pmatrix} \frac{n+1}{\lambda^2} & -\frac{m\Gamma(2-1/m)}{\lambda^{1+1/m}} \\ -\frac{m\Gamma(2-1/m)}{\lambda^{1+1/m}} & \frac{(m-1)^2\Gamma(1-2/m)}{\lambda^{2/m}} \end{pmatrix}. \tag{31}$$

Hence, $I_h(\lambda)$ exists *and* is positive definite if and only if $m > 2$. (Note that $I_h(\lambda)$ does exist when $m = 1$, but it is not non-negative definite because its second diagonal element is zero.) Therefore in this case the condition needed for Bartlization is actually the same for ensuring $I_h(\lambda) > 0$; how generally this phenomenon holds is worth some investigations.

Without getting into the further details of Taylor expansion (18) and the non-convergence of its remaining term R, we already have seen enough issues with the choice of the scale for the unobservable even in this simplest non-trivial case. One therefore has to wonder about the difficulties in pushing the H-likelihood methods with reasonable generalities via the maximization route. Even if it is not impossible, it does require heroic effort to make significant progresses, with unclear impact in terms of both theory and practice. However, judging from my email exchanges with Professor Nelder, it appears that he (and his co-authors) had a bigger picture in mind in pushing the H-likelihood research. The next section, a tribute to John Nelder, documents my reasoning for this speculation.

6. INFERENCE FUSION: AN UNREALIZED (UN-REALIZABLE?) DREAM OF JOHN NELDER?

“At last! Someone who takes our work seriously!” This was the opening line of an email of June 29, 2009, from Professor John Nelder, with whom I never had any exchange, in person or in writing, prior to that correspondence. Apparently, Nelder was pleased to see the Bartlization results reported in Meng (2009a), which was submitted to the editor on June 24, 2009. He wrote, in the next sentence,

> "I have wanted general results for a long time, BUT we use the method only for a particular model class (double hierarchical GLMs), as explained in our book. I am going to send you a copy of the book; please send your full address. .... The H in H-likelihood originally stood for 'hierarchical' because we were thinking of hierarchical classifications, extending the normal case as first put forward by Henderson the cattle breeder in the 50s, but later withdrawn by him. We later found that we could apply the method to cross-classifications, (where it works especially well) but the 'h' stuck. (I am very bad at finding catchy name for these things). It is unreasonable of me to ask you to amend your contribution, but I do wish you would continue our work using the model class in the book. I look forward to hearing from you."

This description spells out the origin of the term "H-likelihood" (I might add that "h" for is "hierarchical" and "H" for "Henderson"!) and Nelder's wish that this line of work be continued. In subsequent emails, Nelder expressed strong interests in comparing and connecting the h-distribution with Fisher's fiducial distribution. In particular, he speculated that the results in Barnard (1995) may help for this purpose:

> "You may know ..., that Barnard 'solved' the fiducial problem. i.e., gave conditions under which a probability can be associated with a parameter. Fiducial inference was not Fisher's 'great mistake', but he overestimated the scope of its use. However, it may be that other distributions may be close to satisfying Barnard's conditions. I believe that Fisher's transform of the correlation coefficient may be one of this class. ..." (July 2, 2009)

His interest in comparing and connecting the two was even more vividly described subsequently:

> "I reread the Barnard paper, which I think is a masterpiece. Have you had time to look at it? I had a half-baked idea that perhaps fiducial distributions, when they exist, form a way of scaling some appropriate likelihood without the use of prior distributions. When you have had time to read our book, I very much hope that we could write a paper combining your insights with our formulation. Does this sound like a good idea to You?" (July 10, 2009)

I must confess that I neither had read Barnard's article then nor was I ready to accept Professor Nelder's invitation to work with him. I was of course very flattered by his invitation, and in other circumstances I would have jumped into such a precious opportunity of working with one of the most preeminent figures in statistics. But I was already completely overwhelmed by my teaching, research and administrative commitments, and that the project Nelder had in mind is not something that could be completed over weekends, considering the attempts made by many great minds, including Fisher, throughout history.

Nelder obviously sensed my reluctance; in one of his emails, he wrote "If I appear to be pressurising you, it is because (1) I am naturally impatient and (2) because I am an old man (85 in Oct.)." [Nelder's sense of urgency was also reflected in his conversation with Senn (2003), where he made an analogy between partial likelihood and H-likelihood: "Partial likelihood was a new kind of quantity for which Cox didn't give a full justification (Cox, 1972) but was later shown by other people to have the right sort of properties. I don't know why at the moment we have this resistance, but I hope to get over it before I die."] Nevertheless, he continuously encouraged me to join him to pursue his ultimate dream, the fusion of schools of inference. In almost every email he sent me subsequently, this dream was revisited:

> "We have finally finished our rejoinder, and think we have made some progress towards integrating the three modes of inference, Fisherian, frequentist, and Bayesian. ... I find it quite exciting and hope we may be able to make a synthesis." (July 25, 2009)
>
> " ... I specially want to know if you think we have at least started a fusion of the three schools of inference. " (July 30, 2009)
>
> " ... Youngjo has finally finished the rejoinder for our paper and will send you a copy. He has made a real effort to join the three schools of inference, but there is much to be done. I do think it is a worthwhile effort to make statistics whole." (August 28, 2009)
>
> " ... It would be marvelous to find a common framework for the schools of inference." (October 13, 2009)

From all these writings, it became increasingly clearer to me that Nelder's (and possibly his co-authors') ultimate interest is not in avoiding specifying a prior per se, but rather in unifying different schools of inference. This is a very laudable goal, a dream that many of us share, although our beliefs in its realizability may differ greatly.

My reply to Nelder (on November 6, 2009) clearly reflected that we had different expectations:

> "As for my general impression of the rejoinder, my reading so far has not generated new insights, as the central message seems to be the same one emphasized in your original article and echoed in my discussion, that is, the choice of scale is critical, and there is a possibility that there is one scale that can render the same result for all three approaches. If this can be established more generally, yes of course it is exciting. And I agree that it is unlikely this is possible in completely generality. Indeed, my current thinking is that the existence of such scale perhaps should be taken as a characterization of a family of models. Once we can get that characterization, I believe it might provide new insights into the similarities and differences of the three schools of thoughts beyond what we already know. I of course fully recognize that the lack of new insight is likely due to my haphazard reading. I really wish I would have more time to devote to this topic, as I have been very intrigued by it. Unfortunately I seem to manage to overwhelm myself with too many "yeses" ..."

This was in response to the email he sent to me on the same day, which continually displayed his enthusiasm for "fusing":

> "I ought not to be bothering you, but I would like to know what you thought of our rejoinder to the discussion in our Stat. Sci. paper. The

> possibility of fusing methods of inference I find very exciting; I am sure it will need some restriction on the model class, but this is not surprising to me. ... Do let me know if you get to London. I have officially retired from Imperial College, but we could still meet there." (November 6, 2009)

Sadly, I never had and will never have a chance to meet Professor Nelder (and to enjoy his legendary singing and piano playing; see Senn, 2003). He passed away on August 7, 2010. I received the news right in the midst of preparing this sequel. I felt a profound loss, more so than loss of a friend. I was given the opportunity to meet and work with him, and I was even warned with his candid "I am an old man", yet all I can tell my students and grand students now is my deep regret. I will never know how disappointed he must be upon receiving my "lack of new insight" response above, for that was the last time I heard from him. But I hope he had forgiven my reluctance and would have permitted me to share with the world one more time of his never diminishing enthusiasm for our beloved subject – his devotion to statistics was infectious and well known (e.g., Senn, 2003; Payne, 2010; Payne and Senn, 2010). Regardless of whether we share Nelder's enthusiasm for H-likelihood methods, just as whether we share Fisher's conviction to fiducial arguments, Nelder's contribution and commitment to statistics is a tremendous inspiration for generations to come.

Indeed, if I am lucky enough to live to 85 and still have half as much energy as Nelder had, I promised myself that I would push future "Xiao-Li Meng"'s as hard as he did to me. I literally would never have written Meng (2009a) or this sequel if not for his strong belief in what he had been pursuing.

Thank you, John.

REFERENCES

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis.* Cambridge: University Press

Barnard, G. A. (1995). Pivotal models and the fiducial argument. *Internat. Statist. Rev.* **63**, 309–323.

Bayarri, M. J., DeGroot, M. H., and Kadane, J. B. (1988). What is the likelihood function? *Statistical Decision Theory and Related Topics IV* (S.S. Gupta and J. O. Berger, eds.). New York: Springer

Barndorff-Nielsen, O. E. (1994). Adjusted versions of profile likelihood and directed likelihood, and extended likelihood. *J. Roy. Statist. Soc. B* **56** 125–140.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* New York: Springer

Berger, J. O., Liseo, B. and Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters (with discussion). *Statist. Science* **14** 1–28.

Berger, J. O. and Robert, L. W. (1988). *The Likelihood Principle.* IMS Lecture Notes **6**, Hayward, California: IMS

Bjørnstad, J. F. (1999). Comment on "Integrated likelihood methods for eliminating nuisance parameters" by Berger, Liseo and Wolpert. *Statist. Science* **14**, 23–25.

Bock R. D. and Aitkin M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, **46**, 443–459.

Butler, R. W. (1986) Predictive likelihood inference with applications. *J. Roy. Statist. Soc. B* **48**, 1–38 (with discussion).

Cox, D. R. (1972). Regression models and life-tables . *J. Roy. Statist. Soc. B* **34**, 187–220 (with discussion).

Cox, D. R. (1975a). Partial likelihood. *Biometrika* **62**, 269–276.

Cox, D. R. (1975b) A note on partially Bayes inference and the linear model. *Biometrika* **62**, 399–418.

Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. B* **49**, 1-39 (with discussion).

Cox, D. R. and Reid, N. (1993) A note on the calculation of adjusted profile likelihood *J. Roy. Statist. Soc. B* **55**, 467–471.

Desmond, A. F. (1997). Optimal estimating functions, quasi-likelihood and statistical modelling. *J. Statist. Planning and Inference.* **60**, 77–121 (with discussion).

Ghosh, J. K., ed. (1988). *Statistical Information and Likelihood. A Collection of Critical Essays by D. Basu.* New York: Springer

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208–1211.

Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63**, 277–284.

Godambe, V. P. and Thompson, M. E. (1974). Estimating equations in the presence of a nuisance parameter. *Ann. Statist.* **2**, 568–571.

Hinkley, D. V. (1979). Predictive likelihood. *Ann. Statist.* **7**, 718–728 (corrig. **8**, 694).

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *Ann. Statist.* **21**, 1359–1378.

Junker, B. W. and Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement* **24**, 65–81.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773–795.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *J. Roy. Statist. Soc. B* **32** 175–208 (with discussion).

Kalbfleisch, J. D. and Sprott, D. A. (1974). Marginal and conditional likelihood. *Sankhyā A* **35** 311–328.

Lauritzen, S. L. (1974). Sufficiency, prediction, and extreme models. *Scandinavian J. Statist.* **1**, 128–134.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalised linear models (with discussion). *J. Roy. Statist. Soc. B* **58**, 619–678.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* **88**, 987-1006.

Lee, Y. and Nelder, J. A. (2005). Conditional and marginal models: another view. *Statist. Science* **19**, 219–238 (with discussion).

Lee, Y., and Nelder, J. A. (2009). Likelihood inference for models with unobservables: another view. *Statist. Science* **24**, 255–269 (with discussion).

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood.* London: Chapman and Hall.

Lindsay, B. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Phil. Trans. R. Soc A.* **296**, 639–665.

Lindsay, B. (1982). Conditional score functions: some optimality results. *Biometrika* **69**, 503–512.

McCullagh, P. (1990). A note on partially Bayes inference for generalized linear models. *Tech. Rep.*, The University of Chicago, USA.

McCullagh, P. and Nelder, J. A.(1989)- *Generalized Linear Models* (2rd Ed.) London: Chapman and Hall

McCullagh, P. and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *J. Roy. Statist. Soc. B* **52**, 325–344.

Meng, X.-L.(1994). Posterior predictive *p*-values. *Ann. Statist.* **22**, 1142–1160.

Meng, X.-L. (2009a). Decoding the H-likelihood. *Statist. Science* **24**, 280–293.

Meng, X.-L. (2009b). Automated bias-variance trade-off: Intuitive inadmissibility or inadmissible intuition? *Frontiers of Statistical Decision Making and Bayesian Analysis: In honor of James O. Berger* (M.-H. Chen, D. K. Dey, P. Mueller, D. Sun, and K. Ye, eds.). New York: Springer, 95–112.

Meng, X.-L. and Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *J. Amer. Statist. Assoc.* **91**, 1254–1267.

Meng, X.-L. and Zaslavsky, A. (2002). Single observation unbiased priors. *Ann. Statist.* **30**, 1345–375.

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

Payne, R. (2010) John Ashworth Nelder. VSN International.
http://www.vsni.co.uk/featured/john-nelder/.

Payne, R. and Senn, S. (2010). John Nelder obituary: Statistician whose work was influential in a range of sciences. *Guardian*, Thursday 23 September 2010.
http/www.guardian.co.uk/technology/2010/sep/23/john-nelder-obituary

Rasch, G. (1960/1980). *Probabilistic Models for some Intelligence and Attainment Tests.* (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980) with foreword and afterword by B. D. Wright. Chicago: The University of Chicago Press.

Reid, N. (1996). Likelihood and Bayesian approximation methods. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press 351-368 (with discussion).

Senn, S. (2003). A conversation with John Nelder. *Statist. Science* **18**, 118–131.

Sijtsma, K. and Junker, B. W. (2006). Item response theory: past performance, present developments, and future expectations. *Behaviormetrika*, **33**, 75–102.

DISCUSSION

EDWARD I. GEORGE (*University of Pennsylvania, USA*)

H Stands for Hopeful. In his attempt to find merit in the H-likelihood approach, Professor Xiao-Li Meng has provided some deep insights into what is needed at the very least if H-likelihood methods were to work. Ironically, his success only underscores the ultimate limitations of the H-likelihood approach.

H-likelihood, proposed by Lee and Nelder in a series of papers as a potential tool for likelihood inference, is given by $h(\theta, v; y) = \log f_\theta(y, v)$ where y is the observed data, θ is the unknown parameter and v is an unobserved random variable such as a latent variable, missing variable or future realization from f_θ. At first glance, it is very tempting to treat θ and v alike, to think of h as the joint likelihood of θ and v. After all, they are both unknown entries in the likelihood, and if we can find enough similarities in their roles, we should be able to at least formally use the same methods to make inference about their values. But the more and more one tries to find similarities, the more one actually finds differences. This is the ironic conclusion of Professor Meng's deep insights into this problem.

The fundamental reason why likelihood methods for v based on h should not be expected to work is that as a function of v given y, $f_\theta(y, v)$ is simply not a likelihood. Ultimately, a likelihood is a reversal of a conditional probability distribution. More precisely, $p_\theta(y)$ is a likelihood as a function of θ given y if and only if $p_\theta(y)$ is a conditional probability distribution as a function of y given θ. It is this feature that

Edward I. George is the Universal Furniture Professor of Statistics at the University of Pennsylvania.

allows likelihood to fit in with the consistent probability calculus that, for example, gives rise to coherent Bayesian inference. Clearly, $f_\theta(y, v)$ is not a likelihood as a function of v given y because it is not a conditional probability distribution of y given v.

H Stands for Heroic. In spite of this lack of appropriate motivation for H-likelihood, Xiao-Li perseveres and in an effort to reveal its hidden potential, heroically investigates the extent to which the fundamental properties of a likelihood analysis carry over for H-likelihood. Focusing on the basic Bartlett identities, he notes that the usual formulation of expected score and expected information under (θ, v) does not really make sense for H-likelihood. Ultimately, this problem can be seen as stemming from the fact noted above, that as a function of v given y, $f_\theta(y, v)$ is simply not a likelihood.

Undeterred by this observation, in Meng (2009a), Xiao-Li brilliantly observes that hidden inside the H-likelihood is a bonafide likelihood, namely $f_\theta(y \mid v)$, the likelihood of v given y corresponding to the conditional distribution of y given v. Indeed, the H-likelihood can be decomposed as $h(\theta, v; y) = \log f_\theta(y \mid v) + \log f_\theta(v)$, the second term being the marginal distribution of v. The key then to making the Bartlett identities hold for H-likelihood is to require conditions which would make this marginal disappear under the Bartlett expectations. These conditions, given in Theorem 1 and 2 of Meng (2009a), can at least in some cases be met by using an appropriate transformation of v, a process which Xiao-Li has colorfully termed Bartlization. Further investigation in Section 5 reveals that the determination of the optimal transformation, if it exists, can be subtle and difficult. So there you have it, in some cases, under a suitable transformation of v which can be difficult to find, the H-likelihood will satisfy the Bartlett identities. This contrasts sharply with the general appeal of likelihood methods which are typically at least straightforward. Ironically, Xiao-Li insightful discoveries seem to underscore the limitations, rather than the potential, of H-likelihood as a useful practical tool.

Alas, Xiao-Li goes on to show us that even if Bartlization can be obtained, H-likelihood will still not enjoy all the appealing asymptotic properties that are usually associated with likelihood inference. The basic problem is that information about v does not accumulate with more data so that uncertainty about v will not be eliminated as the number of observations on y goes to infinity. For example, as Xiao-Li points out, in the fundamental likelihood approximation of $\hat{\phi} - \phi$ by $I_h^{-1}(\theta)S(\phi; y)$ in (18), the error of approximation fails to go to zero as the sample size goes to infinity. So even with Bartlization, the inferential benefit of H-likelihood is limited.

H Stands for Hiding the Motivation. From the Bayesian point of view, the appropriate adjustment of $f_\theta(y, v)$ for inference about θ is the marginal distribution obtained by margining out v with respect to a distribution. Similarly for inference about v, one would use the marginal obtained by margining out θ. This is in fact exactly the sensible motivation for the APHL (adjusted profile H-likelihood) given by (3), which can easily be seen as a first order Laplace approximation to the marginal obtained by integrating out with respect to a uniform distribution. So, the recommendation by Lee and Nelder (2001) to use maximized APHL rather than raw H-likelihood for estimation is at least reasonable. Unfortunately, Lee and Nelder in their seeming obsession to avoid crediting the Bayesian paradigm, promote APHL as a likelihood method, setting Xiao-Li off on his valiant investigation of the extent to which this might be justified. To my mind, here is a place where Occam's Razor can help us choose the best motivation for APHL. Compared to a likelihood motivation, I choose the Bayesian motivation because it is vastly simpler and transparent.

Let me conclude by congratulating Xiao-Li for a fascinating investigation that provides tremendous insight into the inner workings of likelihood methods. As further food for thought, I would be interested in Xiao-Li's answers to the following questions: (i) Does the success of the transformation $v = \log y_{n+1}$ in your exponential example in Section 3 fundamentally have to do with transformation to a location family for which an implicit uniform prior is working? (ii) In particular, what role does invariance play in these methods? (iii) Can decision theory approaches shed further light on these methods? (iv) Why are Lee and Nelder so invested in avoiding a prior on θ?

ANTHONY F. DESMOND (*University of Guelph, Ontario, Canada*) and
CHANGCHUN XIE (*McMaster University, Ontario, Canada*)

Introduction. One of us (Desmond) had the great pleasure of attending Professor Meng's presentation at Valencia 9. We greatly appreciated the clarity and wit with which Professor Meng presented his paper. It prompted us to read the original paper of Lee and Nelder (2009) of which Professor Meng was a discussant. Having experimented with the use of H-likelihood in our own work in biostatistics (Xie *et al.*, 2008), we appreciate the opportunity to comment on this stimulating presentation, which raises interesting and deep foundational issues about the nature of likelihood and predictive inference. In our discussion, we would like to ask some questions, motivated to some extent by recollections of the oral presentation at Valencia 9, and also by our subsequent reading of Meng (2009a).

Issue of terminology. One issue that is important, and is often raised, is the issue of terminology. For example should we talk about 'estimation' or 'prediction' of unobservables. When it comes to inference for random effects or latent variables we prefer the term 'estimation', as do Lee et al (2006); see also Robinson (1991). On the other hand Lee and Nelder (2009) use the term unobservable for both random effects and future observations. Meng (2009a) appears to agree, stating that " 'unobservables' is semantically more appropriate". We wonder about this and feel that this is not merely a semantic issue. In our view, there is a fundamental logical difference between 'unobservable' random effects and future observations. The latter are at least potentially observable. The same could be said of missing data. This leads us to ask whether, perhaps, the concept of H-likelihood might be more appropriate for one, but not the other? Related to this is the phenomenon well described by Professor Meng that information does not accumulate for unobservables such as random effects. The situation seems logically somewhat different for future observations, in that past data surely increases information for prediction of future observations. For example, standard textbook prediction intervals for future observations based on samples from normal distributions for both homogeneous and regression situations get more precise (narrower) as n or $(X'X)^{-1}$ increases. Finally, unknown parameters are themselves unobservable, but Lee and Nelder clearly wish to distinguish them from, say, random effects.

Fiducial prediction. H is for history! We were most intrigued by Professor Meng's discussion in (6.3) on fiducial ideas and predictive probability as this led us to revisit some of the writings of R. A. Fisher. There is at least formally a strong connection between Professor Meng's pivotal predictive distribution (7.14) of Meng (2009a) and a thought provoking section in Fisher (1956), Chapter V, entitled 'Fiducial prediction'. Fisher is here concerned with a situation in which one observes a random sample N_1 of exponentially distributed inter-emission times of a radioactive source with rate θ. From the sufficient statistic, the sum of the inter-emission times $X_1 = \sum_{i=1}^{N_1} y_{i1}$, $y_{i1} \sim \exp(\theta)$. Fisher wishes to derive a fiducial distribution of the sum of N_2 future times $X_2 = \sum_{i=1}^{N_2} y_{i2}$, $y_{i2} \sim \exp(\theta)$. He has previously, in chapter 3, used this example to illustrate the fiducial argument for the unknown parameter θ based only on the observed sample. Fisher considers the ratio of X_2 to X_1, which is a predictive pivot (although this is not Fisher's terminology) for X_2 distributed independently of θ and obtains, what he refers to as the 'distribution of X_2 given X_1' given by his expression (70). With $N_1 = n$, $N_2 = 1$ and, converting to Professor Meng's notation, this becomes

$$f(y_{n+1}|\bar{y}_n) = \frac{n(n\bar{y}_n)^n}{(n\bar{y}_n + y_{n+1})^{n+1}}, \qquad 0 < y_{n+1} < \infty.$$

Transforming to $r = y_{n+1}/\bar{y}_n$, Fisher's (70) leads to

$$f(r|\bar{y}_n) = (1 + \frac{r}{n})^{-(n+1)}, \qquad 0 < r < \infty,$$

which is the same as (7.14) of Meng (2009a). Fisher continues stating that: "Without discussing the possible values of the parameter θ, therefore, the exact probability of the total time recorded in a second series of trials lying within any assigned limits is thus calculable on the basis of the total time observed in the first series." Meng (2009a) notes, in Section 7.6, that (7.14) is obtainable with an improper "noninformative prior" on $log(\lambda)$ (Note Meng's $1/\lambda$ is Fisher's θ) and finds it "somewhat intriguing that this un-realizable posterior distribution via random λ is easily realizable via the pivotal predictive distribution." We have another instance here of Fisher's fiducial argument resulting in formally similar results to Bayesian inferences with "noninformative" priors. In the famous words of Savage, Fisher appears "to make the Bayesian omelette without breaking the Bayesian eggs", or to quote Meng, enjoy "the Bayesian fruits without paying the B-club fee." Fisher (1956), however, on page 118 makes a strong claim that his fiducial predictions are empirically verifiable and states: "Probability statements about the hypothetical parameters are, however, generally simpler in form and once their equivalence is understood to predictions in the form of probability statements about future observations, they are not seen to incur any logical vagueness by reasons of the subjects of them being relatively unobservable." On another historical note, Meng states that Nelder and Lee emphasize Pearson's (1920) point that Fisher's likelihood is not useful for predicting future observations. Fisher (1956, Chapter 5, Section 7) does in fact develop a type of predictive likelihood for future binomial observations (precisely Pearson's problem).

PIERO VERONESE (*Bocconi University, Milano, Italy*)

Professor Meng raises a very interesting issue concerning the relationship among pivotal predictive distribution, posterior predictive distribution and h-distribution.

In Section 7 of Meng (2009a), the author considers the general points previously discussed with reference to the exponential distribution in detail. In particular, in Section 7.5 of Meng (2009a), he emphasizes how it is important "moving from the original scale of y_{n+1} to the $v = \log(y_{n+1})$ scale" in order to obtain a *predictive* *pivotal quantity* and in Section 7.6 of Meng (2009a) he adds "the scale of the parameter also plays a role, especially for the adjusted profile h-likelihood ... (making) in the current example ... the adjustment ... immaterial". Furthermore he compares the pivotal predictive distribution, the posterior predictive distribution (under a *non-informative prior*) and the h-distribution and concludes that there exists an intimate connection "a truly 3-in-1!"

This final result is not completely surprising and part of the explanation can be found by extending a result due to Lindley (1958), and considered also by Consonni and Veronese (1993), who explains the relationship between a fiducial distribution and a posterior distribution. More precisely, Lindley shows that a fiducial distribution for a real parameter θ is, under some regularity conditions on the model, a posterior distribution if and only if:

i) the distribution function (d.f.) of the sufficient statistics U_n given θ, where n denotes the sample size, can be written as

$$F(u_n|\theta) = G_n(t(u_n) - \eta(\theta)), \quad n = 1, 2, \dots, \tag{32}$$

for some (known) d.f. G_n, which we assume defined on $\mathbb{R}$, and monotone function t and η,

ii) a constant prior on the parameter $\eta(\theta)$ is assumed.

It's interesting to note that equation (32) establishes automatically the *correct scale* of both variables and parameters, advocated by the Author. Thus, from now on, we will work with $T_n = t(U_n)$ and $\eta = \eta(\theta)$. It is immediate to verify that the density of T_n given η is given by

$$f(t_n|\eta) = g_n(t_n - \eta), \tag{33}$$

where g_n is the density corresponding to the d.f. G_n.

Now suppose that condition (33) holds, and let $y = (y_1, \dots, y_n)$ denote the sample. It follows that the likelihood of η is proportional to $f(t_n|\eta)$ and consequently the maximum likelihood estimate (M.L.E.) of η is given by $\hat{\eta} = t_n - C_n$ where $C_n = \text{argmax}_x\, g_n(x)$.

The *likelihood* of ν, using the plug-in technique as far as η is concerned, is proportional to $g_1(\nu - \hat{\eta})$ and consequently the M.L.E. of ν is given by $\hat{\nu} = \hat{\eta} + C_1 = t_n - C_n + C_1$, where $C_1 = \text{argmax}_x\, g_1(x)$.

Since the distribution of $\hat{\nu}$ can be derived from that of T_n, we can compute the distribution of $w = \hat{\nu} - \nu$, given η, which is

$$f(w|\eta) = \int f_\nu(\nu|\eta) f_{\hat{\nu}}(\nu + w|\eta) d\nu = \int g_1(\nu - \eta) g_n(\nu + w + C_n - C_1 - \eta) d\nu. \tag{34}$$

Making the change of variable $z = \nu - \eta$, it follows that the result of the integration does not depend on η. This shows that, under condition (33), the distribution of $\hat{\nu} - \nu$ is a real fiducial distribution.

Consider now the predictive posterior density of ν given y, under the constant prior on η, $\pi(\eta)$. We have

$$f^B(\nu|y) \quad = \quad \int f(\nu|\eta)\pi(\eta|y) d\eta. \tag{35}$$

Because $\hat{\nu} = T_n - C_n + C - 1$ is a linear transformation of the sufficient statistics T_n, also $\hat{\nu}$ will be sufficient for η and thus $\pi(\eta|y) = \pi(\eta|\hat{\nu})$. Consequently

$$\begin{aligned} f^B(\nu|y) = f^B(\nu|\hat{\nu}) &= \frac{\int g_1(\nu-\eta) g_n(\hat{\nu} + C_n - C_1 - \eta) d\eta}{\int g_n(\hat{\nu} + C_n - C_1 - \eta) d\eta} \\ &= \int g_1(\nu-\eta) g_n(\hat{\nu} + C_n - C_1 - \eta) d\eta. \end{aligned} \quad (36)$$

Recalling that $\hat{\nu} = \nu + w$ and that g_n is defined on $\mathbb{R}$, it follows that $f^B(\nu|y)$ coincides with the fiducial distribution (34).

In the example of the exponential distribution it is easy to see that condition (32) holds, with sufficient statistic $U_n = \sum Y_i$, $T_n = t(U_n) = \log(\sum Y_i)$ and $\eta = \log(\lambda)$ with the function $g_n(x) = 1/\Gamma(n) \exp(nx - e^x)$. It follows that $\nu = t(y_{n+1}) = \log(y_{n+1})$ and thus we have the scaled transformations suggested by the Author. Furthermore it is easy to check that $C_n = log(n)$ and $C_1 = \log(1) = 0$. Thus $\hat{\eta} = T_n - C_n = \log(\sum Y_i) - \log(n) = \log(\overline{Y})$ and $\hat{\nu} = \hat{\eta} + C_1 = \log(\overline{Y})$, as expected.

Conditions (32) realizes 2-in-1, but it must be stressed that it is a strong conditions. For example, inside the exponential family it holds only for distributions that can be reinterpreted as normal or exponential.

It would be interesting to investigate the role of condition (32) from an asymptotical point of view and, in this case, relate it also to the third, and more crucial element of the paper, the H-likelihood.

REPLY TO THE DISCUSSION

H for Heartfelt Thanks!

In my now 20 years of professional career, I had over half a dozen opportunities to prepare a discussion article with a rejoinder. I do not recall having had a more enjoyable time than the current one. All three discussants have been superb, offering constructive insights and real food for thoughts. My heartfelt thanks therefore go to all of them: to Professor George for being a fabulous “podium-mate” at Valencia 9 and for the witty discussion, both in oral presentation and in writing, and to Professors Desmond, Xie and Veronese for deep and historical insights – I learned a great deal by studying the discussions. Thanks also to Desmond and Veronese for correcting much of my Chinglish!

I, of course, want to thank Professor Jose Bernardo again for inviting me and for insisting that I prepare a written article in addition to my presentation at Valencia 9, which was mostly based on my discussion of Lee and Nelder (2009). For that piece (Meng, 2009a), I am grateful to Professor David Madigan, the Executive Editor of *Statistical Science*, who is responsible for starting my journey to the land of H-likelihood and for publishing my journey diary in its entirety.

Response to George

Professor Ed George is known for his great clarity and abundant humor in delivering speeches, technical or otherwise, something I also strive to mimic. It was therefore a true professional joy to have Ed, a great friend, share the Valencia 9 podium. In between all the laughter, however, are his four insightful and critical questions (i)-(iv), which are reproduced at the end of his written discussion.

As for (i), my investigation so far supports a “yes” answer, especially in view of Lindley’s (1958) and Veronese’s results discussed in the previous section. I, however, need to emphasize that the investigation so far is in a rather restrictive setting of “unobservables”, namely univariate future observations. When dealing with more general unobservables, especially in high dimensions, things could be much more complicated or unexpected. Similarly for (ii) — pivotality is a form of invariance, and indeed invariance has played a critical role in the literature of *predictive likelihood* (e.g., Lauritzen, 1974; Hinkley, 1979; Butler, 1986). In particular, the sampling pivotal predictive distribution I discussed in Meng (2009a) is closely related to both the *marginal predictive likelihood* based on an ancillary quantity $a(y, u)$ and the *conditional predictive likelihood*, which is constructed by conditioning on a sufficient quantity $s(y, u)$ (Butler, 1986). For our exponential model, we can choose

$$a(y, u) = \log(y_{n+1}/\bar{y}_n), \quad s(y, u) = \sum\nolimits_{i=1}^{n+1} y_i$$

(recall $u = y_{n+1}$). But until a more general investigation is conducted, especially in multiple dimensions, I had better resist the temptation of drawing too many conclusions from the investigation so far – I perhaps already have milked the “exponential cow” too much!

The answer to question (iii) perhaps can be a safe “yes”, since it is almost always useful to consider the decision theoretic angle, even if it is just to confirm what we already know. Indeed, it may even shed some light on (iv), the answer to which seems to lie in understanding Lee and Nelder’s “utility” consideration in their quest for avoiding specifying a prior. The Section 6 of my article indicated their desire to infuse different schools of inference, and avoiding prior specifications seems to be an integrated part of that effort.

I have some additional remarks inspired by Professor George’s written discussion. First, George is absolutely correct that the theoretical results I obtained demonstrate the limitations of H-likelihood more than its applicability. As mentioned in Meng (2009a), I ended up devoting five weekends to H-likelihood because I was intrigued by Lee and Nelder’s perseverance despite the fact that nearly all the published feedback they had received was on the negative side. Like George, the Bayesian interpretation of APHL was obvious to me. But I told myself to keep as open minded as possible – after all, it is healthy especially in foundational research to push arguments as hard as one can, even to play as the devil’s advocate. I was indeed a bit surprised by how easy it is for an H-likelihood to satisfy the Bartlett identities, relative to what I initially expected. However that “easiness” also reminded me of a hidden message, namely Bartlett identities are *minimal requirements.* Without them we can almost be sure that the corresponding “likelihood” will not deliver sensible results if we use it as if it were a real one (e.g., making inference based on its “Fisher information”). But H-likelihood provides a vivid demonstration that Bartlett identities alone do not guarantee correct inference. This was not a message that I had come across prior to my study of H-likelihood, though surely I hope this negative implication is not the only tangible benefit of my “heroic effort”! $\ddot\smile$

Second, George attributed the failure of H-likelihood largely to the fact that it is not a genuine likelihood. Whereas a genuine likelihood obviously does not suffer the kind of problems H-likelihood does, by now there are plenty of artificial likelihoods in the literature that generally do not satisfy George’s description that “a likelihood is a reversal of a conditional probability distribution.” These include partial likelihood, empirical likelihood, dual likelihood, quasi-likelihood, composite likelihood, etc. Unlike H-likelihood, these "likelihoods" are much better received in the literature even by some Bayesians (I now can claim to be one of those, having finally been inducted to the Valencia Hall of Fame), because they lead to useful methods that generally cannot be recast from the Bayesian perspective. Indeed, how to conduct Bayesian inference with artificial likelihoods is still an underdeveloped area (see Lazar, 2003).

Third, I noticed that George was careful in using the term "conditional probability distribution" instead of "conditional probability density." The difference is not semantic, because defining a likelihood via a *density* is a trickier business than we routinely tell our students. A good example is to explain to students why the likelihood function is unbounded when our model is a mixture of $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, and when the parameter $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \alpha)$ is unrestricted, other than the obvious constraints such as the mixing proportion α is between 0 and 1 (and the order of the mixture is known). A student may wonder why MLE does not exist regardless of the sample size, given clearly we can estimate θ consistently, and that the method of moments dates back to Pearson (1894). The non-existence of MLE actually carries a hidden message: there is a problem in defining likelihood using the mixture of normal densities with respect to the Lebesgue measure. The family of models admits a mixture of a continuous component, say, $N(\mu_1, \sigma_1^2)$ and a singleton, $\delta_{\{y=\mu_2\}} \equiv N(\mu_2, 0)$. This mixture forms a sub-class of non-degenerated models, yet it does not have a density with respect to the Lebesgue measure.

Finally, there is no known principle to explain why the *reversal* in "a likelihood is a reversal of a conditional probability distribution" is the right thing to do. The reversal was of course a huge success for Fisher's likelihood formulation, building the foundation of much of statistical science as we have today. However, when Fisher invoked a similar "reversal" operation to $f(y|\theta)$ as a *distribution* instead of an objective function, he ended up with his "biggest blunder", as viewed by many to this date, namely, the fiducial distribution. Another byproduct of my H-likelihood journey is the detour to the fiducial land, but the more I understand it (I hope!) the more I feel Fisher's agony, or at least as I imagined. How could the reversal operation work so beautifully for interval inference but so frustratingly for distributional inference? What could be the hidden message here?

Response to Desmond and Xie

Regarding Professors Desmond and Xie's question about terminology, I fully agree that "what's in the name?" is often more than a trivia question. I also agree that "unobservable" random effects and future observations (and other form of unobserved but potentially observable quantities) do have some logical differences. The reason that I agreed that " 'unobservables' is semantically more appropriate" than "missing data" is because of what my thesis advisor Donald Rubin once told me about Sir David Cox's objection to the phrase "missing data." The word *data* is the plural of *datum*, which is a Latin word meaning *something given.* Therefore, semantically, "missing data," is a self-contradictory phrase, meaning "something given but is not given." The use of "unobservables" as an all-encompassing term at least avoids this contradiction. But it does have its own problems, one of which is that it might leave the impression that the quantities being described are unobservable under any circumstance. This may be true for some constructed latent variables, such as a person's true ability in item response theory or latent trait models (e.g., Rasch, 1960/1980; Bock and Aitkin, 1981; Meng and Schilling, 1996; Sijtsma and Junker, 2006), but not so for other "unobservables" such as a future observation. Just because something is not observed for a problem at hand does not automatically imply that it can never be observed.

From Desmond and Xie's wording, I gather they were wondering if the difference between "potentially observable" and "never observable" has something to do with whether or not Lee and Nelder's H-likelihood methods are applicable. I had a similar suspicion, but then I realized that the matter is rather complicated. For example, whereas latent variables typically are "unobservable" in the real sense of the phrase, aspects of them can produce observable manifestations that can be tested against the observed data; see Junker (1993) and Junker and Sijtsma (2000). Although these manifested signals tend to be weak, they nevertheless pose theoretical difficulties in our quest for separating random-effect like latent constructs from potentially observable "unobservables" for the purpose of identifying when Lee and Nelder's H-likelihood methods may provide acceptable results.

Desmond and Xie are also correct that as we collect more data our information about future observations should also increase. My point is that there is a limit to this accumulation, same as Professor George's emphasis in his presentation, that is, no matter how much past data we have, at the best we can only pin down our model perfectly, but not any future observations. This is in contrast to the usual inference of the model parameter, where the increase of the data size will eventually accumulate the information to infinity, that is, reduce the uncertainty in our estimator to zero (at least in theory). Retrospectively, perhaps I should have adopted the term "non-vanishing of uncertainty" instead of "lack of accumulation of information".

I am also literally flattered by Professor Desmond and Xie's identification that my "3-in-1" distribution (16) is a special case of Fisher's (1956) "fiducial prediction" distribution, his (70). There is no other pioneer's work I'd like to reproduce (unknowingly) more than Fisher's! On the other hand, it is not hard to do so either because R. A. Fisher had done so much that I yet need to find a major modern advance that I would be willing to bet my annual salary on it that it absolutely cannot be traced back to Fisher's work in some way. I could only invent excuses for not having read Fisher (1956) (e.g., it was published before I was born). If there is any silver lining in my ignorance of Fisher's work, it is that Desmond and Xie's identification boosted my self-confidence for having accidently wondered about the type of philosophical issues that seemed to be on Fisher's mind when he wrote the statement on page 118 of Fisher (1956), as quoted by Desmond and Xie.

Fisher's statement also solved a minor puzzle I had initially, that is, why it is necessary to invoke the label "fiducial prediction" when there is a perfectly clear sampling interpretation of (16) on the joint space. My initial thinking, along the line as documented in Section 6.3 of Meng (2009a), was that the term "fiducial" was used to turn the probability statement on the joint space of future and current observations into a conditional statement of the future observation given the current ones. But Fisher's statement seems to emphasize more the use of such distributions for inferring "hypothetical parameters" once the probability statements about them can be made—or rather, *understood*—to that of predictions. Fisher is not known for invoking unnecessary arguments, but he did have the tendency of making statements without crisply spelling out their meanings. I surmise that this *equivalence transformation* from an *estimation* problem (for hypothetical parameters) to a *prediction* problem (for a future observation), albeit not having a clearly explained meaning, is nevertheless Fisher's best attempt of bringing an empirically verifiable statement (on the aforementioned joint space)—and hence avoiding "logical vagueness"—into an inferential statement about the "relatively unobservable" hypothetical parameter, and without resorting to Bayesian philosophy.

Incidently, this in a way also answers Desmond and Xie's question about as whether we should talk about "estimation" or "prediction" of unobservables. Fisher's statement suggests that both terms are relevant because it is the interplay between them that permits the *equivalence transformation*. Although it is unclear how this transformation can be done in general, the "somewhat intriguing" phenomenon I was wondering about, as noted by Desmond and Xie, does seem to have a close connection with this transformation. But of course I had better read Fisher this time before trying to figure out what the connection is!

Response to Veronese

Professor Veronese quoted a result of Lindley (1958), which I had not read either so I can only invoke the same invented excuse. But I cannot even invent any excuse for not knowing Consonni and Veronese (1993), for I actually studied it at the time of cooking SOUP (Meng and Zaslavsky, 2002). Although that cooking was for a different dish, namely identifying a prior such that the corresponding posterior mean of a parameter is an unbiased estimator of the same parameter, in hindsight, the key ingredient is the same. Both are about determining prior densities such that the resulting posterior densities have certain pre-specified characteristics; in the current content it is about when a posterior distribution coincides with Fisher's (1956) fiducial distribution. Now I feel really ashamed for writing about fiducial arguments without reading Fisher (1956), but it is nice to be reminded once again that unexpected returns on research investment only take positive sign, unlike the stock market!

Lindley's (1958) and Veronese's results demonstrate further the impossibility of having "3-in-1" in general, even when Fisher's fiducial distribution exists. Lindley's (1958) results show that even without "unobservables", within the exponential families, Fisher's fiducial distribution can be viewed as a Bayesian posterior distribution only if the underlying problem can be transformed into a normal distribution or a Gamma distribution (which includes the exponential distribution). Although Lindley's setting is restrictive (e.g., his requirement that the distribution admits univariate sufficient statistics for any sample size), Veronese's derivation for its generalization to unobservables suggested that such restrictions are perhaps inevitable in order to maintain mathematical tractability or theoretical interpretability.

Indeed, in my attempt to extend Professor Veronese's result to include the h-distribution, I came to appreciate why he concludes his discussion, where all results are based on finite-sample exact calculations, with a call for its investigation only from an asympototic point of view. Specifically, following Lindley (1958), Veronese started with a model $f(Y_1, \ldots, Y_n|\theta)$ such that there exists a univariate sufficient statistics $T_n = S_n(Y_1, \ldots, Y_n)$, where n is arbitrary, such that its CDF belongs to a location family with parameter $\eta = \eta(\theta)$ (which does not depend on n):

$$F_n(t|\theta) \equiv \Pr(T_n \leq t) = G_n(t - \eta), \quad n = 1, 2, \ldots. \tag{37}$$

Since n here is arbitrary, this setting also implies that for a future (independent) realization y_{n+1}, the transformation given by $v = S_1(Y_{n+1})$ has the CDF $G_1(v - \eta)$.

Hence, by the usual sufficiency reduction argument, the H-likelihood for (η, v) is

$$H(\eta, v|y_1, \dots, y_n) = g_n(t_n - \eta)g_1(v - \eta), \tag{38}$$

where g_n is the density function of G_n (for arbitrary n) and t_n is the observed value of T_n, that is, $t_n = S_n(y_1, \dots, y_n)$, where y_i is the observed value of $Y_i (i = 1, \dots, n)$. Note that by "usual sufficiency reduction argument" we mean that $f(y_1, \dots, y_n|\theta)$ can be replaced by $f(t_n|\eta) = g_n(t_n - \eta)$ in arriving at (38). It would be a mistake to conclude, however, that we can also replace $f(y_1, \dots, y_n, Y_{n+1}|\theta)$ with $g_{n+1}(T_{n+1} - \eta)$, where $T_{n+1} = S_{n+1}(y_1, \dots, y_n, Y_{n+1})$, which would imply that the H-likelihood is given by $g_{n+1}(S_{n+1}(y_1, \dots, y_n, S_1^{-1}(v)) - \eta)$, assuming the function $S_1(\cdot)$ is invertible. Its discrepancy with (38) is because that in invoking sufficiency for θ via T_{n+1}, we have ignored a factor that depends on the unobservable v, which is not legitimate when v itself is a part of the likelihood argument.

Given (38), it is quite obvious that the MHLE for (η, v) is any $(\hat{\eta}, \hat{v})$ such that

$$t_n - \hat{\eta} = C_n \quad \text{and} \quad \hat{v} - \hat{\eta} = C_1, \tag{39}$$

where C_m is any global maximizer of $g_m (m = 1, n)$. This yields the result Veronese reported. However, in order to derive the APHL of (3), we need to maximize (38) with respect to η for *any given* v, which does not permit any closed-form expression in general. Therefore, we do not have a useful expression for the corresponding profile H-likelihood, even if we ignore the adjustment part.

Intriguingly, Veronese defines the *likelihood* for v as the "plug-in likelihood", that is, with η in (38) replaced by its MLE (which is also MHLE because of the factorization in (38)), leading to the simple expression $g_1(v - \hat{\eta})$. Whereas this simplicity is of considerable appeal, it is well-known that "plug-in" methods generally lead to "misleadingly precise" (e.g., Aitchison and Dunsmore, 1975; Butler, 1986) inference statements because they ignore the uncertainty in the plug-in estimator. Of course, Veronese did not treat his "plug-in likelihood" as the H-likelihood for v, nor did he use it for inference. Rather, he showed that the sampling distribution of $w = \hat{v} - v$, as a random variable on the joint space of $\hat{v}$ (which is determined by $f(Y_1, \dots, Y_n|\theta)$ only) and of v (which is independent of $\hat{v}$) is identical to the posterior predictive distribution of v under the constant prior on η, achieving "2-in-1". Clearly there is little chance for 3-in-1 even under this restrictive setting because we do not even have a workable profile H-likelihood expression for v under (38). Nevertheless, as usual, it is easier to expect that asymptotically different schools of inferences tend to produce similar results. For the current setting, since APHL of (3) is simply the Laplace approximation to the Bayesian integration, we obviously can expect a 3-in-1 asymptotically, as long as the errors in the Laplace approximation become negligible as $n \to \infty$.

Indeed, even for the "plug-in" predictive distribution $g_1(v - \hat{\eta})$ the same asymptotics kicks in when the posterior for η, $g_n(t_n - \eta)$, becomes increasingly concentrated around $g_n(t_n - \hat{\eta}) = g_n(C_n)$, the maximal value, the usual asymptotic phenomenon. As a trivial demonstration, for our exponential example, the "plug in" predictive distribution for v is $g_1(v - \hat{\eta})$, where $\hat{\eta} = \log(\bar{y}_n)$ and $g_1(x) = \exp(x - e^x)$. Consequently, the corresponding distribution for $r = y_{n+1}/\bar{y}_n = e^{v - \hat{\eta}}$ (with $\hat{\eta}$ considered as fixed) will simply be the exponential distribution $f(r) = e^{-r}$. This clearly is the limit of the "3-in-1" distribution in (16) as $n \to \infty$. It is intriguing to note that the finite-sample difference between them is analogous to that between a t density and a normal density. Mathematically the difference essentially is between $(1+x/n)^n$ and e^x, and statistically the difference is between whether or not we take into account the uncertainty in the "plug-in" estimator (for scale parameter in both cases) in forming our predictive/influence distributions.

ADDITIONAL REFERENCES

Consonni, G. and Veronese P. (1993). Unbiased Bayes estimates and improper priors, *Ann. Inst. Statist. Math.* **45**, 303–315.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference.* Edinburgh: Oliver and Boyd.

Lazar, N.A. (2003) Bayesian empirical likelihood. *Biometrika* **90**, 319-326.

Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *J. Roy. Statist. Soc. B* **20**, 102–107.

Pearson, K. (1894) Contributions to the mathematical theory of evolution. *Philos. Trans. Roy. Soc. London*, **185**, 71-110.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statist. Science* **6**, 15–51 (with discussion).

Xie, C., Singh, R. S., Desmond, A. F., Lu, X. and Ormsby, E. (2008). Hierarchical quasi-likelihood approach to bioavailability and bioequivalence analysis. *Communications in Statistics-Theory and Methods* **37**, 1641–1658.