January 1978

To appear in Studies in Bayesian Statistics, in honor of Sir Harold Jeffreys
(Arnold Zellner, Editor)

MARGINALIZATION AND PRIOR PROBABILITIES*

Edwin T. Jaynes

Department of Physics

Washington University

St. Louis, Missouri 63130

---

*A preliminary account of this work was given at the 14th NBER-NSF Seminar on Bayesian Inference, Holmdel, N. J. (June 1977).

1. INTRODUCTION

A recent article of Dawid, Stone, and Zidek (1973) notes two Bayesian calculation methods--the first using an improper prior, the second avoiding it--that one feels intuitively ought to lead to the same result, but in general do not. This "marginalization paradox" has been interpreted widely as revealing a fundamental inconsistency in the common Bayesian practice of using improper priors to express prior ignorance.

We argue that, on the contrary, resolution of the paradox is very simple, the discrepancy arising not from any defect of improper priors, but from a rather subtle failure of the second method to take into account all the relevant information. This situation, far from revealing an inconsistency in Bayesian methods, shows that to violate them in seemingly harmless ways, can generate paradoxes; i.e., it is only by strict adherence to the Bayesian principles expounded by Jeffreys in 1939, that one can avoid inconsistencies in statistical reasoning.

The marginalization process is then turned to advantage by showing that it leads to a new means for defining what is meant by "uninformative" and for constructing noninformative priors, as the solution of an integral equation. This method draws only upon the universally accepted principles of probability theory, making no appeal to such additional desiderata as entropy, group invariance, or Fisher information. However, its range of applicability is still largely unexplored.

2. THE PARADOX

A conscientious Bayesian B_1 studies a problem with parameters θ which he partitions into two sets, $\theta = (\eta,\zeta)$, being interested only in inferences about ζ. Dawid, Stone, and Zidęk (1973; hereafter denoted DSZ) note that, in several examples where B_1 uses an improper prior for η, the data x may also be partitioned into two sets, $x = (y,z)$ in such a way that B_1's marginal posterior distribution for ζ "is a function of z only," while the sampling distribution of z depends only on ζ.

A lazy Bayesian B_2 then tries to derive the posterior distribution $p(\zeta|x) = p(\zeta|z)$ more easily by applying Bayes' theorem directly to the sampling distribution $p(z|\zeta)$; and finds that he cannot reproduce B_1's result whatever prior $\pi(\zeta)$ he assigns.

DSZ then point the accusing finger at B_1 thus: "B_2's intervention has revealed the paradoxical unBayesianity of B_1's posterior distribution for ζ." In the ensuing discussion there was near unanimity of all opinions expressed, holding that B_1 is the party at fault, his transgression lying in his use of an improper prior.

A group-theoretical analysis by DSZ showed that if the sampling distribution $p(dydz|\eta\zeta)$ has a certain group structure [invariant under the combined action of coupled homomorphic groups G, $\bar{G}$ which are exact and transitive on the spaces $S(y)$, $S(\eta)$], the paradox can be avoided by choosing the prior as the right-invariant measure on $S(\eta)$. This procedure has, indeed, been advocated by a long list of writers starting with Poincaré (1912). However, as soon as we pass beyond the case of location and scale parameters it is rather exceptional to find a problem with all that group structure; and the paradox persists even in problems that have no group structure at all. In general, therefore, no way emerged for avoiding the paradox.

Predictably, some have seized upon this as a new tool for the abrogation of Bayesian statistics in general. However unimportant the practical consequences may be, it is imperative for Bayesian theory that this puzzle be cleared up.

In the following we argue that (1) Resolution of the paradox is far too simple a problem to be in need of group-theoretical analysis. That it can be made to appear and disappear by different choices of the η-prior, shows immediately where the difficulty lies. (2) the real cause of the paradox is not B_1's use of improper priors, or indeed any transgression on B_1's part. On the contrary, it appears only when B_2 violates elementary Bayesian principles. B_2's transgression was concealed from view by concise notation. (3) Nevertheless, the prior $\pi(\eta)$ that "avoids the paradox" has a useful interpretation as being, in a certain sense, "completely uninformative." (4) Recognizing this, marginalization leads to a new means for constructing noninformative priors, via a set of simultaneous integral equations. This method is consistent with, but appears more general than, the group analysis.

3. THE RESOLUTION

We must be careful to note exactly what the first quoted statement of DSZ means. From the mathematics it is clear that to say the posterior distribution of ζ "is a function of z only" means that it depends <u>on the data</u> x only through the value of z. But of course, any posterior distribution depends not only on the data, but also on the prior information. As Jeffreys (1939) stressed, to avoid ambiguities the prior information (or hypotheses) on which our probabilities are conditional, ought to be stated explicitly to the right of the stroke in our probability symbols $p(A|B)$.

B_1's prior information includes the whole structure of the model, the qualitative fact of the existence of the components η and y; and the prior distribution of η. How, then, can one be sure that B_2 is justified in considering only the reduced problem in which (η,y) never appear at all? According to Bayesian principles, one may not disregard any part of either the data or the prior information, unless that part is shown to be irrelevant in the sense that it cancels out mathematically.

B_2's reduction appears, at first glance, to be reasonable; but so did a multitude of ad hoc procedures of non-Bayesian statistics, which were found eventually to contain defects. Surely, there is no room for personal opinions about this; the mathematical rules of probability theory are quite competent to tell us whether B_2's reduction is or is not justified.

As a constant reminder of the presence of prior information, we extend the notation of DSZ by introducing the symbols I_1, I_2 to stand for the totality of prior information used by B_1, B_2 respectively. The quoted first statement is then, more precisely,

$$p(\zeta|xI_1) = p(\zeta|zI_1) \quad . \tag{1}$$

Now the rules of probability theory tell us that

$$p(\zeta|xI_1) = \int d\eta\, p(\zeta|\eta x I_1) p(\eta|xI_1) \quad . \tag{2}$$

If, given I_1 and all the data x, additional knowledge of η would be irrelevant for inference about ζ; i.e., if

$$p(\zeta|\eta x I_1) = p(\zeta|xI_1) \tag{3}$$

then η integrates out of (2) trivially. But if (3) does not hold, then η is relevant, and the posterior distribution $p(\eta|xI_1)$ intrudes itself inevitably into the problem, bringing with it a dependence on the prior $p(\cdot|I_1)$.

In this case, we have to expect that the separation property (1) cannot hold for all η-priors. If (1) holds for some class C of priors, then while $p(\zeta|xI_1)$ is, in a sense, "a function of z only," it is a different function of ζ for different η-priors in C. But since B_2's posterior distribution $p(\zeta|zI_2)$ is independent of $\pi(\eta|I_1)$, it appears that we have at hand all the material needed to manufacture paradoxes. In other words, we suggest that this paradox has, fundamentally, nothing to do with improper priors; B_1 and B_2 obtain different results when, and only when, B_2 ignores relevant prior information (about η and/or the model), that B_1 is taking into account.

It remains to be shown that the mechanism just suggested is the one actually operative in the examples of DSZ. Since (3) is equivalent to

$$p(\eta,\zeta|xI) = p(\eta|xI)p(\zeta|xI) \tag{4}$$

we examine some of the DSZ examples for this factorization property.

Example 1. The model is described in DSZ. For present purposes we need note only that the raw data $x = \{x_1 \ldots x_n\}$ are partitioned into $y \equiv x_1$, and $z \equiv \{z_i = x_i/x_1,\ 1 \le i \le n\}$. The joint posterior distribution is

$$p(n\zeta|xI_1) \propto \pi(\zeta|I_1)c^{-\zeta}\, n^n \exp(-nyQ)\pi(n|I_1) \tag{5}$$

where

$$Q(\zeta,z) \equiv \sum_{1}^{\zeta} z_i + c \sum_{\zeta+1}^{n} z_i \tag{6}$$

is a function that is known from the data. Here B_1 has helped B_2's prospects as much as possible by assigning independent priors to η, ζ. Nevertheless, the likelihood function mixes them up and we find the conjectured lack of factorization (4). B_1's marginal posterior distribution is

$$p(\zeta|xI_1) \propto \pi(\zeta|I_1)c^{-\zeta}\int_0^\infty n^n\, e^{-nyQ}\, \pi(n|I_1)\, dn \tag{7}$$

from which we note several things:

(A) As predicted, the dependence on the prior $\pi(\eta|I_1)$ is manifest. Prior information about η is clearly relevant to inference about ζ; and B_2's reduction violates Bayesian principles by throwing it away.

(B) The dependence on y drops out on normalization, leading to (1), if and only if the η-prior is in the class $\{C;\ \pi(\eta|I_1) \propto \eta^k,\ -n-1<k<\infty\}$, which includes all those considered by DSZ.

(C) For any prior in class C there are no convergence problems. It is therefore difficult to see how use of an improper prior can in itself be grounds for reproach; all of B_1's conclusions can be approximated to any accuracy we please (e.g., one part in 10^{1000}) by use of a proper prior (as shown explicitly below, Eq. (24)).

(D) On the other hand, use of a proper prior, $\int\pi(\eta|I_1)d\eta = 1$, will take us out of the class C. But then the statistic y cannot be disentangled, and remains relevant; the separation property (1) is lost, and B_2 becomes superfluous.

(E) The proof of DSZ that use of proper priors avoids the paradox, rested on two assumptions: that B_1 uses a proper prior, and [DSZ, Eq. (1.20)] that the separation property still holds. But for this model, those assumptions are contradictory. DSZ supposed that, with proper priors, the paradox would disappear because B_1 and B_2 then agree. We now see that, at least in this example, the paradox disappears rather because the comparison disappears; B_2 can no longer play his game at all.

For efficient verbalization at this point, we need to coin a new term. A prior $\pi(\eta)$ that leads to the separation property (1) nullifies the effect of the data y for inference about ζ. Let us call such a prior nullifying (more precisely: y-nullifying within the context of a particular model).

What DSZ proved is then: If a proper prior is also nullifying, then it necessarily leaves B_1 and B_2 in agreement. However, except in the trivial case of complete independence: $p(dydz|n\zeta) = p(dy|n)p(dz|\zeta)$ one cannot assume ohne weiteres the existence of such a prior, as this example illustrates.

Example 2. We have parameters $\theta = (\mu_1, \mu_2 \sigma)$ and data $x = (u_1, u_2, s)$ with sampling density function

$$p(u_1 u_2 s|\mu_1 \mu_2 \sigma) = A(s^{\nu-1}/\sigma^{\nu+2})\exp(-Q) \tag{8}$$

where A is a normalizing constant and

$$Q \equiv \frac{1}{2\sigma^2}\left[(u_1 - \mu_1)^2 + (u_2 - \mu_2)^2 + \nu s^2\right] \quad . \tag{9}$$

However, we are interested in inference only about

$$\zeta \equiv \frac{\mu_1 - \mu_2}{\sigma\sqrt{2}} \tag{10}$$

and the sampling distribution of

$$z = \frac{u_1 - u_2}{s\sqrt{2}} \tag{11}$$

is found to depend only on ζ:

$$p(z|\mu_1 \mu_2 \sigma) = p(z|\zeta) = \sqrt{2\pi}\, A \int_0^\infty \omega^\nu\, e^{-R}\, d\omega \tag{12}$$

where

$$R(z,\zeta,\omega) \equiv \frac{1}{2}\left[\nu\omega^2 + (\omega z - \zeta)^2\right] \quad . \tag{13}$$

Making the additional change of variables

$$\mu = \frac{1}{2}(\mu_1 + \mu_2) \quad , \qquad u = \frac{1}{2}(u_1 + u_2) \tag{14}$$

the unwanted components are $\eta = (\mu,\sigma)$; $y = (u,s)$, and B_1's posterior distribution of ζ is

$$p(\zeta|xI_1) \propto \pi(\zeta|I_1) \int_0^\infty \omega^\nu \, d\omega \, e^{-R} \, f(u,\sigma) \tag{15}$$

where

$$f(u,\sigma) \equiv \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} d\mu \; \pi(\mu,\sigma) \; \exp\left[-\left(\frac{u-\mu}{\sigma}\right)^2\right] \tag{16}$$

and in the integration over $\omega = s/\sigma$, s is held constant while σ varies. Again, a certain class C of priors $\pi(\eta) = \pi(\mu,\sigma)$ is found to be nullifying, leading to the separation property (1). This includes the class

$$\left\{C': d\mu_1 d\mu_2 \, \sigma^{-k} \, d\sigma = \sqrt{2} \; d\zeta d\mu \; \sigma^{1-k} \, d\sigma\right\} \tag{17}$$

considered by DSZ, or indeed any prior $\pi(\mu,\sigma)$ independent of μ, for which (15) is independent of y:

$$p(\zeta|x\,I_1) \propto \pi(\zeta|I_1) \int_0^\infty \omega^{\nu-1} \, d\omega e^{-R} \, \pi(\sigma) \quad . \tag{18}$$

From this we confirm Eq. (1.3) of DSZ (with $k = 1$) and comparing with (12) it is seen that B_1 and B_2 will agree if $k = 2$. Once again it is clear from (15), (18) that in general their conclusions will differ because B_1 is taking into account relevant prior information about $\eta = (\mu,\sigma)$ that B_2 is ignoring.

Rather than continuing with a rather tedious, but still superficial, inspection of more examples, which would only reconfirm the mechanism already established, we can get a better understanding by returning to a second look at Example 1.

4. A REINTERPRETATION

We may take a more charitable view of B_2, if DSZ will grant a similar courtesy to B_1. In these examples, independently of all questions of priors, it is true that the marginal sampling distribution of z depends only on ζ. Suppose that we now regard B_2, not as a lazy fellow who "always arrives late on the scene of inference" and tries to simplify B_1's analysis; but merely, through no fault of his own, an uninformed fellow whose knowledge about the experiment consists only of the sampling distribution

$$p(z|\zeta I_1) = p(z|\zeta I_2) \tag{19}$$

and is unaware of the existence of the components (η,y). Then B_2 is following strict Bayesian principles, and

$$p(\zeta|zI_2) \propto \pi(\zeta|I_2)p(z|\zeta I_2) \tag{20}$$

will always represent the best inferences that can be made on the information he has--<u>whether</u> <u>or</u> <u>not</u> B_1's posterior distribution has the separation property $p(\zeta|yzI_1)$ independent of y, that initiated all this.

It then makes sense to compare B_2's results with B_1's in all cases, whether B_1's prior for η is proper or improper; and in all cases the comparison will reveal just how much difference B_1's extra information has made. For the most meaningful comparison, we suppose they have the same prior information about ζ:

$$\pi(\zeta|I_1) = \pi(\zeta|I_2) = \pi(\zeta) \quad . \tag{21}$$

Returning now to Example 1, from Eq. (1.2) of DSZ, B_2's conclusions are given by

$$p(\zeta|zI_2) \propto \pi(\zeta)c^{-\zeta}\,[Q(\zeta,z)]^{-n} \tag{22}$$

while B_1's are given by our Eq. (7). Let B_1 assign a proper η-prior of the conjugate form

$$\pi(\eta|I_1) \propto \eta^{k-1} e^{-t\eta} , \qquad t > 0 \tag{23}$$

which as $t \to 0$ goes into the family of improper priors used by DSZ. Then B_1's result is

$$p(\zeta|yzI_1) \propto \pi(\zeta)c^{-\eta}\left[\frac{y}{t + yQ(\zeta,z)}\right]^{n+k} \tag{24}$$

which, we note, goes smoothly and continuously into B_2's result (22) as $t \to 0$, $k \to 0$.

But no "paradoxical unBayesianity" or "impropriety" is apparent. Strictly speaking, the dependence on y drops out, leading to the separation property (1), only when $t = 0$, but for $t \ll yQ$ there is virtually no y-dependence, even though the prior is still proper. There is no discontinuous change; as t becomes smaller and the prior (23) becomes more nearly nullifying, the y-statistic just becomes less and less informative.

If then B_1, having noted that $t \ll yQ$, decides to simplify (24) by setting $t = 0$, this now appears, not as a paradox-creating transgression into the realm of improper priors, but rather as a perfectly harmless and reasonable approximation--indeed, an approximation far better justified than many that are accepted without question in non-Bayesian statistics.

If B_1 has very little prior information about η [i.e., if (t,k) are small], then there is virtually no difference between his conclusions and B_2's, whether his prior is proper or improper. If, on the other hand, (t,k) are large, then B_1 is in possession of additional, highly cogent, information relevant to inference about ζ; and it is only right and proper that his conclusions deviate from B_2's. Any statistical method that failed

to make use of this information although it was available to the user, would then be deserving of the epithet, "impropriety."

5. IMPROPER PRIORS — DISCUSSION

In view of the great emphasis on the issue of improper priors in DSZ and in the ensuing discussion--almost to the exclusion of all else--and subsequent attempts to use this as an argument against all Bayesian methods, some further exegesis defending the use of improper priors is needed.

A sequence $\{\pi_i\}$ of proper priors defines a corresponding sequence $\{P_i\}$ of posterior distributions. Often, even though the limit of $\{\pi_i\}$ is improper, the limit of $\{P_i\}$ is a proper, well-behaved, and analytically simple, distribution. The Bayesian will often take that limit for mathematical convenience, after it is clear--whether by specific calculation in the manner of (24) or through past experience with similar problems--that this will make no practical difference in the results.

Often, the experimental data are so much more informative than the prior information that to carry along all the details of any particular proper prior, although in principle the correct thing to do, would in practice only increase the amount of computation without yielding anything of value for the purposes at hand. Usually, it is so clear when we have this situation that there is no need to construct specific sequences of the type (23), (24); one proceeds immediately to the simpler limit.

In a similar way, a person using the Chi-squared test knows in advance, from common sense and the past experience of Statisticians in general, about how much data he needs, and how many categories, to give a test that is good enough for his purposes. Beyond that, further

refinements, although correct in principle, would only increase the amount of computation without useful return. In orthodox statistics, use of a little practical common sense in applying a method is not regarded as an inconsistency. Perhaps, when his methods are more widely understood, the Bayesian may hope to be granted an equal dispensation.

Now, as noted by several participants in the discussion following the DSZ paper and discussed at greater length in Jaynes (1976), it is just the Bayesian results based on noninformative improper priors that correspond closely--often exactly--with those obtained by orthodox methods. In these cases, it is difficult to see how one can reject the Bayesian use of an improper prior, without thereby rejecting with equal force the orthodox method which yields the same result.

On the other hand, in some cases the attempted passage to an improper prior may fail, by yielding a non-normalizable posterior distribution in the limit. This is symptomatic that the experiment is so uninformative that our prior information is, necessarily, still highly relevant to any inference that can be made; and in such a case we had jolly well <u>better</u> take that prior information explicitly into account by using the appropriate proper prior. In this case an orthodox method, by its nature incapable of taking prior information into account, is practically guaranteed to produce absurd or dangerously misleading results [for a specific example, see Jaynes (1976); reply to Kempthorne's comments].

The actual equations both in DSZ and in the present work, are not in any way changed by our reinterpretation of B_2's role; but the analysis is seen in a different and perhaps more constructive light. We are not merely exhibiting the folly of a defective Bayesian procedure--ignoring information or using improper priors. We are comparing two entirely

correct Bayesian procedures, making inferences about the same quantity ζ, at two different stages of knowledge. The parameter η is not merely an "unwanted" complication; it represents new information relevant to the desired inference about ζ.

The comparison is, in effect, a microcosm of an often-occurring real life phenomenon, the effect of advancing knowledge on a scientific inference.

For example, from the known rate ($z = 2 \times 10^{20}$ megawatts) of radiation of energy from the sun, estimate its future lifespan (ζ); how much longer can it continue pouring out energy at that rate? The datum (z) was known 120 years ago about as well as it is today, and on the basis of the laws of physics as then known, B_2 [better known as Lord Kelvin (1862)] estimated a future life of ζ = a few million years; an entirely valid conclusion from the information he had. But today we know of a new parameter (η = energy release from nuclear reactions) that has an important bearing on the question, and we have new data (y = abundance of various elements in the sun, and energy release of a number of nuclear reactions). As a result, B_1 [George Gamow (1945)] reestimated the future life of the sun to be vastly greater, about 10^{10} years.

Doubtless, an econometrician could give much more immediate examples; e.g., the effect of new knowledge (the role of oil prices) on prediction of economic activity from models in which, prior to 1973, oil price did not appear as a factor.

6. THE INTEGRAL EQUATIONS

Can we extract something of positive value from all this, leaving Bayesian theory with a net gain? As is now clear, there is no reason to be surprised when B_1 and B_2 disagree; that was only to be expected. What

is perhaps surprising, and calls for explanation, is rather that in some cases B_1's extra information was unavailing, and they *did* agree, after all. How is that possible?

Clearly, in all cases, B_2 was incorporating no prior information about η. If, nevertheless, they agree in one case, it seems natural to conclude that, in that case, B_1 must not have been incorporating any prior information either; at least, none that was relevant to ζ.

The prior $\pi(\eta)$ that leaves them in agreement should, then, have some close relation to the one describing "complete ignorance" of η, if such exists. Is it possible that marginalization is giving us a new, objective, and above all, *workable* criterion for defining precisely what is meant by "complete ignorance," and for telling us whether and when such priors do or do not exist?

But we must proceed cautiously. It is not clear how marginalization could tell us that a prior is "completely uninformative" without qualifications. But marginalization can and does provide an answer to the question whether, within the context of a given model, any proposed prior $\pi(\eta)$ is or is not "completely uninformative about ζ."

Two further cautions are necessary. A prior $\pi(\eta)$ that is held to be uninformative about η ought, one would suppose, to have the property that it is uninformative *a fortiori* about any other quantity ζ. Clearly, however, the converse need not hold. In any given model, a prior $\pi(\eta)$ might express very great knowledge about η; and still be ζ-uninformative because of the functional form of $p(yz|\eta\zeta)$. On the other hand, if one could prove that a given prior $\pi(\eta)$ is ζ-uninformative for all models in which η appears as a scale parameter, and unique for one, that would seem to be valid grounds for a stronger claim.

Finally, we note that in another respect the property of a prior $\pi(\eta)$ to leave B_1 in agreement with B_2 involves rather more than what one usually means by the term "uninformative." B_1's advantage over B_2 does not lie only in his prior knowledge of η; he has also the additional data y. If a prior $\pi(\eta)$ is to leave him in agreement with B_2, therefore, it is not enough for $\pi(\eta)$ to have the passive property of being ζ-uninformative (i.e., of not in itself providing any information relevant to ζ). It must perform also the active function of rendering the new data y irrelevant to ζ; that is what we have termed "nullifying."

The necessary and sufficient condition for a prior $\pi(\eta)$ to be nullifying independently of $\pi(\zeta)$ is that B_1's quasi-likelihood contains y and ζ in separate factors:

$$\int p(y,z|\eta,\zeta)\pi(\eta)d\eta = f(y,z)g(z,\zeta) \qquad (25)$$

for some functions f, g. The surprising discovery of DSZ was then that, while a proper prior that is nullifying is also necessarily uninformative [Proof: integrating y out of (25), it then reduces to B_2's likelihood $p(z|\zeta)$], an improper prior may be nullifying without being uninformative. In Example 1, the prior (23) is nullifying if $t = 0$; but it is not uninformative unless $k = 0$ also.

Let us seek the necessary and sufficient condition for agreement of B_1, B_2, subject to three assumptions. Firstly the property (19) without which we should have little reason to compare B_1 and B_2 at all. Secondly,

it would make little sense to ask whether an η-prior is uninformative about ζ if it contained ζ; so we assume that B_1 assigns independent priors:

$$\pi(\eta\zeta|I_1) = \pi(\eta|I_1)\pi(\zeta|I_1) \quad . \tag{26}$$

Thirdly, DSZ considered whether B_2 could, by any choice of his prior $\pi(\zeta|I_2)$, achieve agreement with B_1's posterior distribution. But for present purposes we cannot allow B_2 that much freedom; for if B_1 and B_2 had different priors for ζ, that would in itself lead to a difference in their conclusions, that really has nothing to do with B_1's prior knowledge about η, although agreement of the posterior distributions might be, fortuitously, restored by a particular η-prior. But in that case it would be very wrong to label such an η-prior as "uninformative about ζ." To avoid this, we must suppose rather that B_1 and B_2 start from the same state of prior knowledge about ζ as in (21):

$$\pi(\zeta|I_1) = \pi(\zeta|I_2) = \pi(\zeta)$$

and end up still in agreement as to the posterior distribution. And by "agreement", we do not mean that they agree for one particular sample or one particular prior. In order to justify saying that B_1's prior for η was completely ζ-uninformative and y-nullifying, they must remain in agreement for all data sets $x = (y,z)$, all sample sizes, and all priors $\pi(\zeta)$.

With these assumptions, B_1's posterior distribution is

$$p(\zeta|y,z,I_1) \propto \pi(\zeta)p(z|\zeta)\int d\eta\, \pi(\eta|I_1)p(y|z\eta\zeta I_1) \tag{27}$$

While B_2's is

$$p(\zeta|zI_2) \propto \pi(\zeta)p(z|\zeta) \quad . \tag{28}$$

Evidently, the necessary and sufficient condition for agreement is that

$$\int d\eta\ \pi(\eta|I_1)p(y|z\,\eta\zeta\,I_1) = p(y|z\,\zeta\,I_1) \tag{29}$$

shall be independent of ζ for all y,z. Denoting the parameter space and our partitioning of it into subspaces by $S_\theta = S_\zeta \otimes S_\eta$, we may write (29) as

$$\int_{S_\eta} p(y,z|\zeta,\eta)\pi(\eta)d\eta = \lambda(y,z)p(z|\zeta) \quad , \quad \zeta \text{ in } S_\zeta \tag{30}$$

This is a Fredholm integral equation in which the kernel is B_1's likelihood, $K(\zeta,\eta) = p(y,z|\zeta,\eta)$, the "driving force" is B_2's likelihood $p(z|\zeta)$, and $\lambda(y,z) \equiv p(y|zI_1)$ is an unknown function to be determined from (30).

For each possible data set $x = (y,z)$ we have an equation of the form (30); so if a single prior $\pi(\eta)$ is to suffice for all data sets it must satisfy not just one integral equation, but a large--in general infinite--class of simultaneous integral equations.

Now in other applications we are accustomed to find that a single Fredholm equation has already a unique solution. At first glance, therefore, it seems almost beyond belief that the system of equations (30) could fail to be grossly overdetermined; from which one would be forced to conclude, with the antiBayesian skeptics, that uninformative priors do not exist.

Clearly, the consistency of previous Bayesian thought, which presupposed the existence of uninformative priors, is being put here to a severe test. But it is also an eminently fair and "objective" test. The question whether, in a given model, the notion of an uninformative prior is contradictory, ambiguous, or well defined, is removed from the realm of philosophical debate, and reduced to the question whether a set of simultaneous integral equations is overdetermined, underdetermined, or well-posed.

7. AN EXAMPLE

We know at least that the system of equations (30) is not always overdetermined; for in several examples DSZ were able to recognize particular priors $\pi(\eta)$ which leave B_1 and B_2 in harmony for all samples. Each of the DSZ examples can tell us something about the mathematical structure of (30) and its correspondence with previous group invariance arguments.

Example 1. The sampling distribution is

$$p(yz|\eta\zeta) = \eta^n\, c^{n-\zeta}\, y^{n-1}\, \exp[-\eta y Q(z,\zeta)] \tag{31}$$

with $Q(z,\zeta)$ defined by (6). This gives the marginal sampling distribution

$$p(z|\zeta) = (n-1)!\; c^{n-\zeta}\, Q^{-n} \quad . \tag{32}$$

S_η is the positive real line, and the family of integral equations (30) becomes: for each possible sample (y,z),

$$\int_0^\infty \pi(\eta)\eta^n\, e^{-\eta y Q}\, d\eta = (n-1)!\; y\lambda(y,z)[yQ(z,\zeta)]^{-n}, \quad \zeta = 1,2,\ldots,n-1 \; . \tag{33}$$

Now choose any two values $\zeta \neq \zeta'$, and write $Q \equiv Q(z,\zeta)$, $Q' \equiv Q(z,\zeta')$. Equation (33) then requires that for all (y,z)

$$\int_0^\infty \pi(\eta)(\eta y Q)^n\, e^{-\eta y Q}\, d\eta = \int_0^\infty \pi(\eta)(\eta y Q')^n\, e^{-\eta y Q'}\, d\eta \tag{34}$$

or

$$\int_0^\infty \left[\pi(\eta) - \frac{Q}{Q'}\,\pi\left(\eta\,\frac{Q}{Q'}\right)\right]\eta^n\, e^{-\eta y Q}\, d\eta = 0, \quad 0 < y < \infty \; . \tag{35}$$

Since the Laplace transform is uniquely invertible, this requires that for all choices of $\{z,\zeta,\zeta'\}$ we must have, setting $a \equiv Q/Q'$,

$$\pi(\eta) = a\pi(a\eta) \; , \qquad 0 < \eta < \infty \tag{36}$$

But this is the same functional equation that was deduced earlier from the transformation group that expresses "complete ignorance" of a scale parameter. To complete the proof, note that from (6), if $\zeta' < \zeta$, Eq. (36) must hold in the continuous range $(1 \leq a \leq c)$, and so the only possibility is, to within a constant factor,

$$\pi(\eta) = \eta^{-1} \quad . \tag{37}$$

This argument shows that no $\pi(\eta)$ other than (37) can satisfy (33). Conversely, on substitution we see that (33) is indeed satisfied for all $\{y,z,\zeta\}$, with $y\lambda(y,z) = 1$. Again, we note several things:

(A) This argument made no use of the separation property (1). The solution (37) implies this as a necessary, if obvious, condition for agreement of B_1 and B_2.

(B) The problem turned out to be well-posed; there is one unique prior $\pi(\eta)$ that is "completely uninformative about ζ," and it is just the one that Jeffreys anticipated, on partly intuitive grounds, some forty years ago (as the prior representing "complete ignorance" of a parameter known to be positive). It follows also from the fact that η is a scale parameter, by some transformation group methods (for example, Jaynes, 1968; one of several quite different approaches all called "the transformation group method" or "the group invariance principle," although they utilize different groups which operate in different spaces, are chosen by different criteria, and yield different results. For further comments, see Appendix A).

(C) But the prior (37) is now derived in a way that is completely independent of anybody's intuition or any additional desiderata such as entropy, group invariance, or Fisher Information. Given the sampling distribution (31), the result (37) follows by straightforward mathematical steps. [Indeed, on sufficiently fine analysis, it will be seen that the only elements of probability theory used in the transition from (31) to (37) are the product rule $p(AB|C) = p(A|BC)p(B|C)$, and the sum rule $p(A|B) + p(\bar{A}|B) = 1$].

(D) This, however, recalls the oft-quoted remark of Lindley (1971): "Why should one's knowledge, or ignorance, of a quantity depend on the experiment being used to determine it?" The answer, in our view, is that the prior distribution should, of course, be based on <u>all</u> the prior information available. But the role a parameter plays in a sampling distribution is always a part of that information. Indeed, that is the irreducible minimum information without which a problem of inference cannot be formulated. Often, in pedagogical examples, it is the <u>only</u> prior information at hand, because (as in all the DSZ examples) the person formulating the problem simply neglects to provide any more. In this case--and only in this case--the prior distribution is, necessarily, determined (not necessarily uniquely) by the sampling distribution. But this is just the case we are solving by (37).

(E) In a real problem, a parameter will be, in general, "a physically meaningful quantity about which we know something." But for the mechanics of incorporating that something into our informative prior there are, to the best of the writer's knowledge, only two known principles: Bayes' theorem and maximum entropy; and both of these still require an ignorance pre-prior like (37) as their starting-point (Jaynes, 1968).

Therefore, for any problem of inference we see no way to avoid the notion of "complete ignorance," any more than we could avoid the concept of zero in arithmetic. Nor should we wish to avoid it; for clearly, to ask, "What is our state of knowledge after receiving information I?" cannot have any definite answer until we specify: What was our state of knowledge before receiving I? And this holds with equal force whether we choose to classify I as part of the data, or part of the prior information (see, however, Appendix A for some further comments).

(F) In this example, η was a scale parameter, the sampling distribution (31) having the functional form $p(y,z|\zeta,\eta) = y^{-1}\, g(z,\zeta;y\eta)$. For any sampling distribution of this form [or equally well, $y^{-1}\, g(z,\zeta;y/\eta)$] one readily verifies that the Jeffreys prior $\pi(\eta) \sim \eta^{-1}$ satisfies (30), and $y\lambda(y,z)$ is then a constant. Whether this solution is unique depends, of course, on how z,ζ enter into the function g.

8. THE ONE-DIMENSIONAL CASE

With the insight gained from the DSZ Example 1, we are able to give a more general discussion of the case where y, η are one-dimensional. We started cautiously, asking only for a prior $\pi(\eta)$ that is uninformative about ζ within the context of a given model. We now see that for a scale parameter η, the Jeffreys prior is ζ-uninformative for all models, and unique for one. But this is already enough to establish it as the only prior for a scale parameter that is "completely uninformative" without qualifications.

Since the location and scale parameter cases are equivalent by the transformation $\mu = \log \sigma$, it follows that the uniform prior $d\mu$ is similarly general and unique for a location parameter (but in this case the result is so intuitive that it had never been doubted anyway).

The analysis may be generalized in the following way [suggested to the writer by a remark of W. D. Fisher]. Consider any sampling distribution with the functional form

$$p(yz|\eta\zeta) = g[z,\zeta;h(y,\eta)]\,\frac{\partial h}{\partial y} \tag{38}$$

for which the property $p(z|\eta\zeta) = p(z|\zeta)$ underlying marginalization theory follows at once. The integral equation (30) for an uninformative prior becomes

$$\int g(z\zeta;h)\,\frac{\partial h}{\partial y}\,\pi(\eta)d\eta = \lambda(y,z)\int g(z\zeta;h)dh \quad . \tag{39}$$

If this is to hold without further assumptions about the functional form of $g(z\zeta;h)$, it is necessary that $\lambda(y,z) = \lambda(y)$ be independent of z, and that

$$\frac{\partial h}{\partial y}\,\pi(\eta) = \lambda(y)\,\frac{\partial h}{\partial \eta} \quad . \tag{40}$$

But then, making the change of variables $(y,\eta) \to (\bar{y},\bar{\eta})$ where

$$\bar{y} \equiv \exp\int\lambda(y)dy$$

$$\bar{\eta} \equiv \exp\int\pi(\eta)d\eta \tag{41}$$

Eq. (40) reduces h to a function of $(\bar{y}\bar{\eta})$:

$$h(y,\eta) = \bar{h}(\bar{y}\bar{\eta}) \tag{42}$$

and (38) assumes the functional form $p(\bar{y},z|\bar{\eta},\zeta) = \bar{y}^{-1}\,\bar{g}(z,\zeta;\bar{y}\bar{\eta})$ of remark (F) above, where $\bar{g}(z\zeta;\alpha) \equiv g(z,\zeta;\bar{h}(\alpha))$. Thus the class of functions $h(y,\eta)$ for which $\pi(\eta)$ and $\lambda(y)$ can be constructed as in (40) takes us back, to within the change of variables (41), to the scale parameter case.

For example, if

$$h(y,\eta) = \tanh\sqrt{y^n + \eta^m - a} \quad , \qquad 0 < \eta < \infty$$

we have at once from (40) that the uninformative prior is

$$\pi(\eta) = \eta^{m-1} \quad .$$

Likewise, if

$$h(y,\eta) = f(y)[\tan\alpha\eta]^{3/2} \quad , \qquad 0 < \alpha\eta < \frac{\pi}{2}$$

the uninformative prior is

$$\pi(\eta) = \csc(2\alpha\eta) \quad .$$

and if

$$h(y,\eta) = \log\left[\frac{(\eta y - 1)(\eta + y)}{\eta y}\right]$$

the uninformative prior is

$$\pi(\eta) = 1 + \eta^{-2}$$

Now, although the result (40) is rather special in the class of all problems with one-dimensional (y,η), it is easily seen to exhaust the possibilities of the DSZ group analysis for that class. They took the sampling distribution as [DSZ, Eq. (2.6)] $p(dydz|\eta\zeta) = f(yz|\eta\zeta)\mu_G(dy)dz$, where μ_G is "a fixed general measure element" and defined the group structure by [DSZ, Eq. (2.7)]:

$$f(y,z|\eta,\zeta) = f(gy,z|\bar{g}\eta,\zeta) \tag{43}$$

where g, $\bar{g}$ are corresponding elements of the groups G, $\bar{G}$ mentioned in Sec. 2 above. Evidently, if (y,η) are one-dimensional, (43) says only that we have the functional form [compare (38)]:

$$f(yz|\eta\zeta) = g[z,\zeta;h(y,\eta)] \tag{44}$$

the "combined action of the groups" signifying a kind of hydrodynamic flow in the (y,η) plane, whose streamlines are the contours $h(y,\eta) = \text{const}$. But just as our Eq. (40) cannot be satisfied for all functional forms of $h(y,\eta)$, so the group structure (43) restricts the form of $h(y,\eta)$ in (44).

The form of that restriction can be anticipated at once by the following argument. A continuous exact group of mappings of the real line onto itself is necessarily a one-parameter group [for in $y' = gy$ with fixed y, each group element g is represented by one and only one value of y'; thus y' parameterizes the group]. But a one-parameter continuous group is isomorphic with the group of simple translations $(x' = x + a)$. We infer that the group structure (43) must restrict us to problems that are equivalent, to within a change of variables, to the location/scale parameter case. Indeed, on following through the analysis (Hamermesh, 1962) we find that the condition imposed

on (44) by the group structure is just our Eq. (40); i.e., the functions denoted $h(y,\eta)$ in (38), (44) are identical.

The condition found here is the same as that given by Lindley (1958) for agreement of a Bayesian posterior with a fiducial distribution; such relations were noted also by Fraser (1961) and Villegas (1971).

Now we arrive at the really interesting question: What happens in the one-dimensional case if we try to go beyond the class of problems just discussed? Do we continue to find uninformative priors from (30) beyond those obtainable by group analysis; or do we come up against that threatening overdetermination? This opens up a wide class of new mathematical problems, interesting in their own right and of obvious importance for the future of Bayesian statistical theory. At the time of writing (January 1978) progress on them is far from complete, consisting mostly of isolated results.

The following example, due to C. L. Mallows, shows that further solutions do exist beyond those resulting from the group structure (43); and that the apparent overdetermination is not always real.

Let y, z be non-negative integers, and

$$p(yz|\eta\zeta) \propto \frac{\zeta^z \eta^y (1-\eta)^{z-y}}{y!(z-y)!} \quad , \quad \begin{matrix} 0 \le \zeta,\eta < \infty \\ 0 \le y \le z \end{matrix} \quad . \tag{45}$$

Then the marginal sampling distributions are Poisson:

$$p(z|\zeta\eta) = p(z|\zeta) = e^{-\zeta} \frac{\zeta^z}{z!} \tag{46}$$

independent of η, as required by marginalization theory and

$$p(y|\zeta,\eta) = e^{-\zeta\eta} \frac{(\eta\zeta)^y}{y!} \tag{47}$$

depending on both parameters; thus seemingly leading to a nontrivial marginalization problem. This example lacks the group structure (43), since y is discrete, η continuous. But we now find the peculiarity that

$$p(y|z,\zeta,\eta) \tag{48}$$

is independent of ζ, and as a consequence the integral equations (30) are satisfied trivially; all priors $\pi(\eta)$ are nullifying and uninformative about ζ.

That the opposite behavior can also occur, is shown by Example 5 of DSZ. The concluding message of their Sec. 1 was that all is well as long as B_1 uses proper priors. Later, they consider the model:

$$p(yz|\eta\zeta) \propto \int_0^\infty t^{2n-1} \exp\left\{-\tfrac{1}{2}[t^2 + n(zt-\zeta)^2 + n(yt-\eta))^2]\right\}dt \tag{49}$$

and note that, if $y = 0$, the posterior distributions of B_1 and B_2 are

$$p(\zeta|zI_1) \propto \pi(\zeta)\int_0^\infty t^{2n-1} \exp\left\{-\tfrac{1}{2}[t^2 + n(zt-\zeta)^2]\right\}dt \quad , \tag{50}$$

$$p(\zeta|zI_2) \propto \pi(\zeta)\int_0^\infty t^{2n-2} \exp\left\{-\tfrac{1}{2}[t^2 + n(zt-\zeta)^2]\right\}dt \quad . \tag{51}$$

But if $z > 0$, the ratio of the integrals is (by a Schwarz inequality) a monotonic increasing function of ζ; and so B_1 and B_2 cannot agree unless they assign a singular prior $\pi(\zeta) = \delta(\zeta-\zeta_0)$, in which case their posterior distribution is independent of the data.

DSZ (Appendix 2) term this situation, "The Inevitable Paradox of Example 5." It is, perhaps, even more inevitable and more paradoxical than they intended; for it is clear from (49) that this situation arises for all priors $\pi(\eta)$, proper or improper! What, then, are we to make of their proof in Sec. 1, that this discrepancy "could not have arisen if B_1 had employed proper prior distributions"?

Passing over this query, the DSZ Example 5 is particularly instructive, just because at first glance the trouble appears so acute. The only nullifying prior is the uniform one $\pi(\eta) = \text{const.}$; and it leads us back to (50) for all y. Surely, we have now run up against that overdetermination;

it is simply a mathematical fact that there is no prior $\pi(\eta)$ that can leave B_1 and B_2 in agreement for all data sets (y,z) and all priors $\pi(\zeta)$.

Yet we would argue that there is still no real paradox here. This situation should not be disconcerting to anyone who has noted, in other Bayesian problems, that the effective sample number n often drops by one unit when we integrate out an unwanted parameter; or who, in using the Chi-squared test, has reduced the number of degrees of freedom by one unit to take account of a parameter estimated from the data.

In fact, the explanation was noted in our Sec. 3 above; the mere qualitative fact of the existence of the components (η,y); i.e., the knowledge that other parameters are present in our model beyond those of interest--already constitutes prior information relevant to B_1's inferences, that B_2 is ignoring. For further discussion, see Sec. 11 below.

These examples demonstrate that two opposite extremes of behavior are possible; presumably, many or all of the conceivable intermediate cases are also possible. It is evident that a great deal more insight into the content of the integral equations (30) will be needed before any overall understanding of marginalization and its implications for Bayesian theory can be reached. In the writer's judgment, the remaining space available here is best used, not in communicating a mass of further isolated results like the above (which the reader can easily invent for himself), but by giving a preliminary survey of a more general attack on the structure of those integral equations, not restricted to the one-dimensional case. But before turning to that, we note some further pertinent clues from the DSZ examples with higher dimensionality.

9. HIGHER DIMENSIONALITY

It appears from the foregoing that the case of a single location or scale parameter--or one that can be reduced to this by a change of variables--is disposed of once and for all; the only remaining function of the integral equations (30) is to determine whether, in a given model, the result is unique. Mathematically, this is the question whether the kernel of the integral equation is complete.

If the parameter η is two-valued, comprising both a location and scale parameter; i.e., if $\eta = (\mu,\sigma)$ and the corresponding data y can be separated into two components $y = (u,s)$ such that the sampling distribution has the form

$$p(zus|\zeta\xi\sigma) = s^{-2}\, g\left(z,\zeta;\frac{u-\mu}{\sigma};\frac{s}{\sigma}\right) \tag{52}$$

then we can verify that the element of prior probability

$$\pi(\mu,\sigma)d\mu d\sigma = \frac{d\mu d\sigma}{\sigma} \tag{53}$$

will satisfy (30) with $s\lambda(s,u,z)$ a constant. Clearly, then, whether or not (53) is uniquely determined by (30), no disagreement of B_1 and B_2 can arise from its use. Yet DSZ produce apparent counter-examples, in which a prior of the form (53) does lead to disagreement! The DSZ paradoxes must, then, have been in part illusory. In the following examples we will see just how this has come about.

Example 2. Here we appear to be in the aforementioned difficulty, for DSZ note that the "paradox" (i.e., disagreement of B_1 and B_2) does not disappear for the "widely recommended prior" $d\mu_1\, d\mu_2\, d\sigma/\sigma$; but it does for $d\mu_1\, d\mu_2\, d\sigma/\sigma^2$ for which "no recommendations appear to exist." Of course, in a problem with two parameters the prior $d\mu\, d\sigma/\sigma$ is indeed widely--and as we have just seen, justifiably--recommended; but that is a very different problem. In Appendix A we discuss the present problem from the standpoint of the transformation group method recommended by the writer (Jaynes, 1968) and show that either result may be obtained depending on further details of the "real-life" situation which are not conveyed by the mere words "location parameter" or "scale parameter".

Our sampling distribution is, in the notation of Eqs. (9)-(18),

$$p(usz|\mu\sigma\zeta) = A\,\frac{s^\nu}{\sigma^{\nu+2}}\exp\left[-\left(\frac{u-\mu}{\sigma}\right)^2 - R(z,\zeta,s/\sigma)\right] \quad . . \qquad (54)$$

Note particularly that from (14), $\mu = \frac{1}{2}(\mu_1+\mu_2)$. Since (54) has the form (52), the prior $d\mu\, d\sigma/\sigma$ must be a solution of (30). To see this directly, and to see whether the result is unique, we can write (30), using (12), in the form

$$\int_0^\infty \omega^\nu\, e^{-R}[f(u,\sigma) - s\lambda(u,s,z)]d\omega = 0 \qquad \begin{matrix} 0 < s < \infty \\ -\infty < u,z,\zeta < \infty \end{matrix} \qquad (55)$$

where $\lambda(usz) = p(us|zI_1)$, $f(u,\sigma)$ is given by (16), and in the integration over $\omega = s/\sigma$, s is held constant.

But (55) is an integral equation with complete kernel, since e^{-R} is the generating function for a complete set of (Hermite) functions:

$$e^{-R} = e^{-\frac{1}{2}\omega^2(\nu+z^2)}\sum_{n=0}^{\infty} H_n\left(\frac{\omega z}{\sqrt{2}}\right)\frac{(\zeta/\sqrt{2})^n}{n!} \quad . \qquad (56)$$

Substituting this into (55), it is apparent that each term of the summand must vanish separately. A function orthogonal to a complete set must vanish almost everywhere, and so the necessary and sufficient condition (NASC) for an uninformative prior reduces to

$$f(u,\sigma) = s\lambda(u,s,z), \qquad \begin{matrix} 0 < s,\sigma < \infty \\ -\infty < u,z < \infty \end{matrix} \tag{57}$$

from which we infer that $f(u,\sigma)$ is independent of σ, and the undetermined function

$$g(u) \equiv s\lambda(u,s,z) \tag{58}$$

is independent of s, z. Using (16), the NASC is then

$$\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} d\mu \; e^{-\left(\frac{\mu-u}{\sigma}\right)^2} \pi(\mu,\sigma) = g(u) , \qquad \begin{matrix} 0 < \sigma < \infty \\ -\infty < u < \infty \end{matrix} \tag{59}$$

Evidently, for any $\pi(\mu,\sigma)$ that could be taken seriously as a prior, the integral (59) converges so well that $g(u)$ must be an entire function. But then appealing again to completeness and generating function relations of the form (56), the most general function satisfying (59) is

$$\pi(\mu,\sigma) = \sum_{n=0}^{\infty} a_n \, \sigma^{n-1} \, H_n\left(\frac{\mu}{\sigma}\right) \tag{60}$$

where a_n are arbitrary constants, and H_n are the Hermite polynomials. We then find $g(u) = \Sigma a_n (2u)^n$. Conversely, on substituting (60) back into (15), we find that B_1's posterior distribution reduces to B_2's, all the arbitrary constants a_n cancelling out upon normalization.

Of course, not all functions of the form (60) satisfy the further requirement $\pi(\mu,\sigma) \geq 0$; but (60) includes many non-negative priors. For example, if $\omega(q)$ is any non-negative function with moments of all orders, the choice

$$a_n = \frac{1}{n!}\int_{-\infty}^{\infty} \omega(q) q^n \, dq \tag{61}$$

leads to a non-negative prior

$$\pi(\mu,\sigma) = \sigma^{-1} \int dq \omega(q) \exp(2q\mu - q^2\sigma^2) \tag{62}$$

for which B_1 and B_2 will be in agreement. As a special case, if $\omega(q)$ goes into a delta-function $\delta(q-t)$ we get $a_n = t^n/n!$ which in turn yields the anticipated Jeffreys prior $d\mu\, d\sigma/\sigma$ in the special case $t = 0$.

Also in this example, then, our early fears for the poverty of overdetermination disappear in an embarrassment of riches; from a mathematical standpoint, the problem is grossly underdetermined. Nevertheless, out of the many different solutions of (59) the Jeffreys prior $\pi(\mu,\sigma) \sim \sigma^{-1}$ still appears to hold a favored position. Out of the class of solutions (62) it is the only one that does not become exponentially large as $|\mu| \to \infty$. We conjecture that some further restriction on the allowable behavior at infinity [for example that $\pi(\mu,\sigma)$ shall be at most $O(|\mu|^N)$ for some $N < \infty$] may lead, after all, to the Jeffreys prior as the unique solution.

Example 3. We have n independent observations of a bivariate (x_1,x_2) with model structure

$$x_1 = \sigma_1 e_1 \quad , \qquad x_2 = \gamma x_1 + \sigma_2 e_2 \tag{63}$$

where e_1, e_2 are independent and $N(0,1)$. We require inference about the correlation coefficient $\zeta = \gamma\sigma_1/(\gamma^2\sigma^2 + \sigma_2^2)^{1/2}$. The prior that avoids the paradox should then express complete ignorance about those components of

the parameter space that can be varied with ζ held constant. DSZ note that the "recommended" prior

$$d\gamma \frac{d\mu_1}{\sigma_1} \frac{d\mu_2}{\sigma_2} \tag{64}$$

yields a posterior distribution identical with Fraser's structural distribution; but that the marginalization paradox is still present.

However, from the standpoint of the writer's transformation group method, the difficulty with (64) is obvious. For the model equation (63) is written in such a way that the parameter γ is not decoupled [i.e., it gets entangled in the change of scale transformations which express the fact that σ_1, σ_2 are scale parameters]; and so of course, it cannot be assigned an independent prior. If we rewrite (63) as

$$x_1 = \sigma_1 e_1 \quad , \qquad x_2 = \sigma_2(e_2 + \tau e_1) \tag{65}$$

then $\tau \equiv \gamma\sigma_1/\sigma_2$ is decoupled, an arbitrary change of scale $\sigma_1' = a_1 \sigma_1$, $\sigma_2' = a_2 \sigma_2$ inducing no change in τ. But since the parameter of interest ζ is a function of τ [if $\tau = \tan \alpha$, then $\zeta = \sin \alpha$], the prior assigned to τ should not have anything to do with the paradox. Re-examining the equations of DSZ we find, as expected, that use of the prior

$$f(\tau)d\tau \frac{d\sigma_1}{\sigma_2} \frac{d\sigma_2}{\sigma_2} \tag{66}$$

avoids the paradox, where $f(\tau)$ is an arbitrary function.

The same result can be reasoned out without introducing τ. For in the model equation (63), γ appears not as a location parameter, but as a scale parameter [note that the product or ratio of two scale parameters is still a scale parameter]. Complete ignorance of all three parameters should then be represented by a product of three Jeffreys priors. But again the prior assigned to the quantity of interest ζ should not matter;

so we should be able to insert an arbitrary function $g(\zeta)$ without disrupting the agreement of B_1 and B_2. Indeed, one can verify that the prior

$$g(\zeta)\,\frac{d\gamma}{\gamma}\,\frac{d\mu_1}{\sigma_2}\,\frac{d\mu_2}{\sigma_2} \tag{67}$$

leads to agreement, and is equivalent to (66). The prior which DSZ noted as "avoiding the paradox" is a special case of (67), corresponding to $g(\zeta) = \zeta/(1-\zeta^2)^{1/2}$.

The joint likelihood function, which forms the kernel of the integral equations, can be read off from Eq. (1.8) of DSZ. However, to avoid cumbersome expressions we introduce the notation

$$\beta \equiv \log \frac{\sigma_1}{\sigma_2}\sqrt{1-\zeta^2} \tag{68}$$

$$b \equiv \log\sqrt{\frac{S_{11}}{S_{22}}} \tag{69}$$

$$\omega \equiv \frac{\sqrt{S_{11}\,S_{22}}}{\sigma_1\,\sigma_2} \tag{70}$$

where S_{11}, S_{22} are the sample moments as defined by DSZ. The joint likelihood is then

$$L(\zeta,\sigma_1,\sigma_2) = \omega^n\, e^{-\omega T} \tag{71}$$

where

$$T(z,b;\zeta,\beta) \equiv \frac{\cosh(\beta - b) - z\zeta}{\sqrt{1-\zeta^2}} \tag{72}$$

and $z \equiv S_{12}/\sqrt{S_{11}\,S_{22}}$ is the sample correlation coefficient, whose sampling distribution depends only on ζ [DSZ, Eq. (1.10)]. The "unwanted" components of the parameter and sample spaces may then be taken as $\eta = (\sigma_1,\sigma_2)$; $y = (S_{11},S_{22})$; and the NASC that a prior $\pi(\sigma_1,\sigma_2)$ shall be completely

uninformative about ζ is

$$\int_0^\infty d\omega \int_{-\infty}^{\infty} d\beta[\pi(\sigma_1,\sigma_2) - \omega\lambda]\omega^{n-2}\, e^{-\omega T} = 0 \quad , \quad \left\{ \begin{array}{l} -1 < z,\zeta < 1 \\ 0 < S_{11},S_{22} < \infty \\ n = 2,3,\ldots \end{array} \right\} \qquad (73)$$

where $\lambda(z,S_{11},S_{22})$ is an undetermined function, $\{S_{11},S_{22},\zeta\}$ are held constant in the integrations over ω, β.

Evidently, if the kernels $\omega^{n-2}\exp(-\omega T)$ are complete on the domain of integration we shall be led to the Jeffreys prior $(d\sigma_1/\sigma_1)(d\sigma_2/\sigma_2)$ as the unique solution. We conjecture that this is the case; however we have not succeeded in finding a fully rigorous proof of this, or a counter-example. Therefore, in view of the writer's astonishment at discovering the non-uniqueness of (59) after long believing it unique but being unable to prove it, we leave this an open question which others may perhaps answer.

Example 4a. At this point in the DSZ narrative, the sense of paradox increases sharply; for they produce two versions of a problem that appear not only paradoxical, but unavoidably inconsistent with each other. We obtain the sample $\{x_{11},x_{12},\ldots x_{1n}\}$ from $N(\mu_1,\sigma)$ and $\{x_{21},x_{22},\ldots x_{2n}\}$ from $N(\mu_2,\sigma)$. In version (1) B_1 is interested in inference about $\zeta \equiv \{\zeta = \mu_1/\sigma, \zeta_2 = \mu_2/\sigma\}$. The "unwanted" component is $\eta = \sigma$, and the corresponding data separation is in part $\{z_i = (ns)^{-1}\ \Sigma_j\ x_{ij},\ i = 1,2\}$; y was not specified. Using the class of priors

$$\sigma^p d\mu_1 d\mu_2 d\sigma \qquad (74)$$

DSZ show that B_1 and B_2 cannot agree unless $p = -3$.

But then in version (2) B_2 "asserts his interest in ζ_1 alone." Now ζ_2 becomes part of the "unwanted" parameters: $\eta = \{\sigma, \mu_2/\sigma\}$, and the paradox is resurrected; B_1 and B_2 cannot agree unless $p = -2$. Not only do the two priors contradict the Jeffreys rule; they contradict each other!

In view of our earlier demonstrations, something must be very amiss; yet we can find no error in the DSZ calculations. So the resolution must be--and is--very much simpler. We have here a case of paradox by optical illusion.

B_1 and B_2 are free to partition the parameter space $S_\theta = S_\zeta \otimes S_\eta$ in any way they please; but having chosen any such partition, the mathematical problem is; what prior $\pi(\eta)$ in the space S_η, i.e., with ζ held constant, is uninformative about ζ? The trouble was simply that, after choosing a partition, DSZ continued to write their prior (74) in terms of the old variables (μ_1, μ_2, σ), thereby failing to make the condition ($\zeta = \text{const.}$) visible. Had DSZ transformed their priors to the new variables $\pi(\zeta)\pi(\eta)$ they would have found, in all these examples, that the "paradox" disappears just for the priors $\pi(\eta)$ recommended by Jeffreys. Far from suggesting any inconsistency in Bayesian principles, marginalization thus demonstrates again the power and basic soundess of the notions introduced into this field by Jeffreys some forty years ago.

10. SINGULAR SOLUTIONS: KNOWLEDGE IS IGNORANCE

In the DSZ example 3, the correlation coefficient was considered the quantity of interest, $\rho = \zeta$, and we found that the prior $\pi(\eta)d\eta = (d\sigma_1/\sigma_1)(d\sigma_2/\sigma_1)$ was completely uninformative about ζ. Can we reverse our viewpoint, and find an uninformative prior $\pi(\rho)d\rho$ for the correlation coefficient? Most people, facing the problem of expressing ignorance of ρ, have chosen the form

$\pi(\rho) \sim (1 - \rho^2)^{-k}$ more or less instinctively; but complete agreement on the value of k still eludes us.

We would like to make the choice: $\{\zeta = (\sigma_1, \sigma_2),\ \eta = \rho\}$, and our method then requires that we find a separation of data $x = (y,z)$ for which $p(z|\zeta\eta) = p(z|\zeta)$. Perhaps this is possible, but our first guess: $\{y = (S_{11}, S_{22});\ z = r = S_{12}/(S_{11}\ S_{22})^{1/2}\}$ does not work. The joint sampling distribution of (S_{11}, S_{22}) still depends on ρ, containing as a factor the modified Bessel function

$$I_{\frac{n}{2}-1}\left[\frac{\rho}{\sigma_1\sigma_2}\sqrt{\frac{S_{11}\ S_{22}}{1-\rho^2}}\ \right] \quad .$$

However, the sampling distribution of S_{11} depends only on σ_1; so let us marginalize using the choice: $\{\zeta = \sigma_1,\ \eta = (\sigma_2,\rho),\ z = S_{11},\ y = (S_{22},r)\}$. In view of what we found above, we shall perhaps be willing to take from the start $\pi(\eta) = \pi(\sigma_2,\rho) = \sigma_2^{-1}\,\pi(\rho)$. From Eq. (1.8) of DSZ, we are led to the integral equation

$$\int_{-1}^{1}\pi(\rho)d\rho\int_0^\infty \frac{d\sigma_2}{\sigma_2^{n+1}}\exp\left\{-\frac{1}{2}\left[\frac{S_{11}}{(1-\rho^2)\sigma_1^2} - \frac{2\rho S_{12}}{(1-\rho^2)^{1/2}\sigma_1\sigma_2} + \frac{S_{22}}{\sigma_2^2}\right]\right\}$$

$$= \lambda\exp\left(-\frac{S_{11}}{2\sigma_1^2}\right) \qquad (75)$$

where λ must be independent of σ_1. For the class of samples: $\{S_{22} = 2,\ S_{12} = 0\}$, (75) collapses to

$$\left(\frac{n-2}{2}\right)!\int_{-1}^{1}\exp\left\{-\frac{\rho^2 S_{11}}{2(1-\rho^2)\sigma_1^2}\right\}\pi(\rho)d\rho = \lambda(S_{11}) \quad . \qquad (76)$$

But this cannot be independent of σ_1 unless $\pi(\rho)$ is singular:

$$\pi(\rho) = \delta(\rho) \qquad (77)$$

i.e., the prior information must be that $\rho = 0$ with certainty! Conversely, the prior (77) satisfies (75) for all samples.

On further reflection, we see that this result does, after all, make sense. B_2 is (in our reinterpretation) given data $\{x_1 \ldots x_n\}$ whose sampling distribution depends only on σ_1; and uses it to make the standard Bayesian inference about σ_1. B_1 has in addition the other data components $\{y_1 \ldots y_n\}$. But if B_1 also knows that $\rho = 0$, then these additional data cannot help him to estimate σ_1; uncorrelated normal distributions are independent. B_1 and B_2 will then agree, not because B_1 is totally ignorant of ρ, but for the opposite reason that his perfect knowledge of ρ makes his extra data irrelevant.

To recognize this puts a new dimension into the marginalization game. A prior $\pi(\eta)$ that is uninformative about ζ does not necessarily express ignorance about η; it depends on the structure of the model. If B_1 did not know ρ, his extra data $\{y_1 \ldots y_n\}$ would always be relevant and he would always revise B_2's conclusions about σ_1; there is no ignorance prior $\pi(\rho) \sim (1 - \rho^2)^{-k}$ that can avoid this. But if B_1 had far greater knowledge, he might throw away the new data and accept B_2's conclusions after all!

This is not paradoxical, but is a natural and necessary part of consistent plausible reasoning. We can see this phenomenon in generality already in Eq. (2). If for some particular value $\eta = \eta_0$ we should have

$$\frac{\partial}{\partial y}\, p(\zeta|\eta_0, y, z, I_1) = 0 \tag{78}$$

then the singular prior $\pi(\eta|I_1) = \delta(\eta - \eta_0)$ will bring about agreement of B_1 and B_2 by making the data y irrelevant. Whenever the property (78) exists, the integral equation will have singular solutions representing ignorance of ζ due to perfect knowledge of η.

Now, at long last, we have enough clues in hand to commence a general attack on the integral equations.

11. STRUCTURE OF THE INTEGRAL EQUATIONS

For any fixed data set $x = (y,z)$, (30) is an integral equation which we write, for suggestiveness, in the form

$$\int_{S_\eta} K(\zeta,\eta)\pi(\eta)d\eta = \lambda f(\zeta) \quad , \qquad \zeta \in S_\zeta \qquad . \tag{79}$$

Already at this stage, it is possible to have "ζ-overdetermination". The set of all functions on S_ζ forms a Hilbert space H_ζ. As η ranges over S_η the functions $K(\zeta,\eta)$ span a certain subspace H'_ζ of H_ζ. If $f(\zeta)$ does not lie in H'_ζ, there can be no solution of (79). In these cases, the mere qualitative fact of the existence of the components (η,y)--irrespective of their numerical values--already constitutes prior information relevant to B_1's inferences [because introducing them restricts the space of B_1's possible posterior distributions to $\pi(\zeta)H'_\zeta$]. We saw an example in (49). In this case, the shrinkage of H_ζ cannot be restored by any prior on S_η and the integral equations (79) ask an ill-advised question. In the following we consider only the case where the problem is free from ζ-overdetermination.

If we think of $\pi(\eta)$ as a vector in a Hilbert space H of functions on S_η, then for each ζ, Eq. (79) specifies the inner product of $\pi(\eta)$ with the function $K(\zeta,\eta)$. If as ζ varies over S_ζ these functions span the full space of H then the kernel $K(\zeta,\eta)$ may be said to be "complete," and the function $\pi(\eta)$ is defined "uniquely"; i.e., almost everywhere.

On the other hand, if the functions $\{K(\zeta,\eta):\zeta \in S_\zeta\}$ are not complete on S_η, they span some subspace $H_0 \in H$, and (79) determines only the projection

$\pi_0(\eta)$ of $\pi(\eta)$ onto H_0; i.e., $\pi(\eta) = \pi_0(\eta) + \pi_1(\eta)$ where $\pi_1(\eta)$ is orthogonal to $\pi_0(\eta)$ but is otherwise undetermined. Since the coefficient λ is at this stage arbitrary, $\pi_0(\eta)$ is determined to within a multiplicative constant.

But all this has referred to only one particular data set x. For every different data set we can have a different kernel

$$K_x(\zeta,\eta) = p(yz|\eta,\zeta) \tag{80}$$

a different "driving force"

$$f_x(\zeta) = p(z|\zeta) = \int dy\ p(yz|\eta\zeta) \tag{81}$$

and a different coefficient λ_x. The equations

$$\int_{S_\eta} K_x(\zeta,\eta)\pi(\eta)d\eta = \lambda_x\ f_x(\zeta) \quad , \quad \zeta \in S_\zeta \tag{82}$$

will, for two different data sets x, x', determine the projections $\pi_x(\eta)$, $\pi_{x'}(\eta)$ of $\pi(\eta)$ onto two different subspaces H_x, $H_{x'}$ of H. If H_x, $H_{x'}$ are disjoint, the two integral equations (82) determine no relation between these solutions; i.e., the arbitrary constants C_x in $[\pi_x(\eta), \lambda_x]$ and $C_{x'}$ in $[\pi_{x'}(\eta), \lambda_{x'}]$ may be specified independently. But if H_x and $H_{x'}$ are disjoint (i.e., they have a common linear manifold M), then there are several possibilities:

Case I. If M has dimensionality greater than one, the two integral equations for x and x' may determine different (i.e., linearly independent) projections of $\pi(\eta)$ onto M. If these are both non-zero, then formally we can still escape overdetermination by setting $\lambda_x = \lambda_{x'} = 0$; and then hoping that some other data point x" will allow $\lambda_{x"} \neq 0$. If one of the projections (say of $\pi_{x'}$) vanishes, then we need set only $\lambda_x = 0$. But in either case there will

be an embarrassing situation; since λ_x has the meaning: $\lambda(y,z) = p(y|z\,I_1)$, we are escaping overdetermination only by assigning a prior $\pi(\eta)$ which says that the trouble-making data set $x = (y,z)$ is impossible!

One would be very reluctant to accept such a prior as "uninformative"; indeed, it would seem to be a rather obvious minimum requirement of any prior deserving of that name, that it should not exclude in advance any data set permitted by the sampling distribution $p(y,z|\eta\xi)$. A fully acceptable solution ought to lead to $\lambda > 0$ over the entire sample space. Case I thus represents a kind of moral--even if not formal mathematical--overdetermination. If it should occur for many pairs of data points, we could have also mathematical overdetermination, the only solution of (82) being $\lambda_x \equiv 0$, $\pi(\eta) \equiv 0$.

Case II. The two integral equations for x, x' agree that $\pi(\eta)$ is orthogonal to M. Then the situation is basically as if the subspaces H_x, $H_{x'}$ were disjoint; i.e., no connection is established, and λ_x, $\lambda_{x'}$ may still be specified independently. As far as x, x' are concerned, the "unused" manifold M could be removed from the Hilbert space with no essential change in the problem (of course, if some other data point x" should determine a non-zero projection onto M, we are back to Case I).

Case III. The two integral equations agree in assigning nonzero projections of $\pi(\eta)$ onto M, that are the same within a multiplicative constant. Then the existence of a single function $\pi(\eta)$ demands that these multiplicative constants be equal. A connection is thus established, so that, given λ_x, $\lambda_{x'}$ is determined. This can happen whatever the dimensionality of M. If we can find a third data set x" for which $\lambda_{x'}$ and $\lambda_{x''}$ have common manifold, then $\lambda_{x''}$ is in turn determined by λ_x.

In this way, by a sequence of points $\{x \to x' \to x'' \to \ldots\}$ with overlapping manifolds the constants λ_x, originally arbitrary and independent at each point of the sample space, become tied together by the requirement of a single solution $\pi(\eta)$, into a function $\lambda(x)$ defined at many points. The existence or nonexistence of unique and "morally acceptable" noninformative priors then depends on whether by this process a single function $\lambda(x) > 0$ can be set up over the entire sample space.

Let us call any sequence $\{x_1, x_2, x_3, \ldots\}$ such that H_{x_i} and $H_{x_{i+1}}$ overlap, a continuation path P. Then for any two points x, x' that can be connected by a continuation path, the ratio $[\lambda(x')/\lambda(x)]$ is determined by the integral equations (82). The process is somewhat analogous to analytic continuation [but very different topologically; i.e., a sequence $H_x, H_{x'}, \ldots$ of overlapping manifolds does not in general correspond to a continuous path in the sample space].

Case III is thus in turn comprised of three possibilities:

Case IIIa. "Nonintegrability." Two points x, x' can be connected by two different continuation paths P_1, P_2; but they yield different ratios $[\lambda(x')/\lambda(x)]_1 \neq [\lambda(x')/\lambda(x)]_2$. Then there is no single-valued nonvanishing function $\lambda(x)$ and as in Case I the problem is morally--and, depending on how much of the sample space is infected with this disease, perhaps also mathematically--overdetermined. The avoidance of this case is analogous to a condition of integrability [but again, very different topologically!].

Case IIIb. "Intransitivity." The sample space S_x can be decomposed into two subspaces $S_x^{(1)}$, $S_x^{(2)}$ in such a way that there is no continuation path from any point in $S_x^{(1)}$ to any point in $S_x^{(2)}$. Then no connection is established between $\lambda^{(1)}(x)$ and $\lambda^{(2)}(x)$; i.e., they can be assigned independent arbitrary

multiplicative constants. The problem is then underdetermined, and more than one "noninformative prior" exists. If there are K disconnected subspaces $\{S_x^{(1)} \ldots S_x^{(K)}\}$, then the prior $\pi(\eta)$ determined by (30) will contain K arbitrary constants.

Case IIIc. "Integrable Transitivity." Any two points x, x' in the sample space can be connected by a continuation path, and if more than one such path exists, all paths assign the same ratio $[\lambda(x')/\lambda(x)]$. Then a single-valued function $\lambda(x) > 0$ exists over the entire sample space, and the equations (82) define one unique noninformative prior $\pi(\eta)$, to within a normalization constant.

Previous Bayesian thought (including the writer's) which simply took for granted the existence of unique noninformative priors, has thus in effect assumed that we always have Case IIIc. But looked at in this new way, it seems astonishing that such a thing could be true. If for any two points x, x' in our sample space we should have Case I or Case IIIa, then it is all over with our search for a "morally acceptable" noninformative prior. Yet we have the counter-examples of DSZ where such solutions do exist. What, in the structure of the problem, prevents these cases from occurring?

For enlightenment let us turn back, still another time, to our faithful DSZ Example 1, which has never yet failed to give us an interesting and useful answer to any question we have put to it.

12. EXAMPLE 1 — A FOURTH LOOK

We have seen already, in Eq. (37), that the integral equations determine the Jeffreys prior $\pi(\eta) = \eta^{-1}$ uniquely; now we want to examine in minute detail the mechanism by which this is accomplished. Introducing

the Laplace transform of $\eta^n \pi(\eta)$:

$$F(a) \equiv \int_0^\infty \eta^n \pi(\eta) e^{-a\eta} \, d\eta \tag{83}$$

the set of integral equations (33) becomes

$$F[yQ(z,\zeta)] = \frac{(n-1)!\, y\lambda(y,z)}{[yQ(z,\zeta)]^n} \quad , \quad \zeta = 1,2,\ldots,(n-1) \; . \tag{84}$$

For a fixed data set $x = (y,z)$ this determines the value of $F(a)$ to within a multiplicative factor, at $(n-1)$ discrete points $a_\zeta = yQ(z,\zeta)$. The set of points $\{a_1(y,z) \ldots a_{n-1}(y,z)\}$ will be called the spectrum of x. We can suppose without loss of generality that $c > 1$. From (6), the a_i are nonincreasing: $a_1 \geq a_2 \geq \ldots \geq a_{n-1}$, and all lie within a factor c of each other; i.e.,

$$1 \leq a_1/a_{n-1} \leq c \qquad . \tag{85}$$

The subspace H_x is the one spanned by the set of functions

$$\phi_i(\eta) = \eta^n e^{-a_i \eta} \quad , \qquad i = 1,2,\ldots(n-1) \tag{86}$$

linearly independent if the a_i are all distinct.

Clearly, for $n > 3$ we can in general find another data set $x' = (y',z')$ such that

$$a_1(y,z) = a_2(y',z') \tag{87a}$$

$$a_2(y,z) = a_3(y',z') \tag{87b}$$

and the subspaces H_x, $H_{x'}$ then have a two-dimensional linear manifold M in common, consisting of all functions of the form

$$\phi(\eta) = C_1 \phi_1(\eta) + C_2 \phi_2(\eta) \tag{88}$$

with arbitrary coefficients C_1, C_2.

Therefore unless both data sets x, x' determine the same projection of $\pi(\eta)$ onto this manifold:

$$\frac{F(a_1)}{F(a_2)} = \frac{F(a_2')}{F(a_3')} \tag{89}$$

we shall have Case I, and the problem will be morally overdetermined. Now from the data set x we have

$$\frac{F(a_1)}{F(a_2)} = \left[\frac{Q(z,\zeta_1)}{Q(z,\zeta_2)}\right]^n \tag{90}$$

and from x' = (y'z')

$$\frac{F(a_2')}{F(a_3')} = \left[\frac{Q(z',\zeta_2')}{Q(z',\zeta_3')}\right]^n . \tag{91}$$

But (87) is, more explicitly,

$$yQ(z,\zeta_1) = y'Q(z',\zeta_2') \tag{92a}$$

$$yQ(z,\zeta_2) = y'Q(z',\zeta_3') \tag{92b}$$

from which we see that (90) and (91) are indeed equal. We have escaped overdetermination only because of the connection (87) required to produce a common linear manifold.

Likewise, we could have a three-dimensional common manifold by adding to (87) the condition

$$a_3(y,z) = a_4(y',z') \tag{93}$$

which is generally possible if n > 4. But again the three conditions (87a), (87b), (93) are just sufficient to bring about equality of the three ratios $F(a_i)/F(a_j)$; and so on. We continue to have Case IIIc.

We see now how different this problem is from the usual theory of integral equations with complete kernel. It is just the very great incompleteness of our kernel $K(\zeta,\eta)$ that, so to speak, creates room for agreement so that all the integral equations (82) can be satisfied simultaneously. Because of this incompleteness the subspaces H_x are so small, and scattered about so widely in the full Hilbert space H like stars in the sky, that it requires a very special relation between x, x' to bring about any overlapping manifold at all.

But it still seems magical that the relation required to produce overlapping should also be just the one that brings about agreement in the projections. So we still have not found the real key to understanding how Case I is avoided.

Since there is a unique solution (37), a single-valued function $\lambda_x > 0$ must have been determined over the sample space S_x. Evidently, then, our integral equations must be transitive on S_x; i.e., there must exist a continuation path connecting any two points x, x'. What are these paths? Are there more than one for given x, x'? If so, how was nonintegrability (Case IIIa) avoided?

We suppose the spectra $\{a_1,a_2,\ldots,a_{n-1}\}$, $\{a'_1,a'_2,\ldots,a'_{n-1}\}$ of x, x' to have no point in common (otherwise H_x, $H_{x'}$ have already a common manifold M and there is no need for a continuation path). If any point a_i is within a factor c of some point a'_j we define a new data point $x'' = (y'',z'')$ by

$$y'' = \frac{ca'_j - a_i}{c - 1} \quad , \qquad z''_2 = \frac{a_i - a'_j}{ca'_j - a_i} \tag{94}$$

and $z''_3 = z''_4 = \ldots = z''_n = 0$. Then the first two points of the spectrum of x" are, from (6),

$$a''_1 = a_i \quad , \qquad a''_2 = a'_j \tag{95}$$

and $x \to x'' \to x'$ is a continuation path. If the spectra of x, x' are more widely separated (i.e., if $c^{k-1} < a_{n-1}/a_1'' < c^k$), then [because of the restriction (85) on the spectrum of any one point x"] it will require a continuation path with at least k intermediate points to connect them; but this can always be done by repetition of the above process. The reason for the transitivity is thus clear.

Now, how does this determine the function $\lambda(y,z)$? Writing the family of integral equations (33) as

$$G(a) \equiv \int_0^\infty d\eta\, \pi(\eta)(\eta y Q)^n\, e^{-\eta y Q} = (n-1)!\, y\lambda(y,z) \tag{96}$$

for any given data point x, the necessary and sufficient condition that $\pi(\eta)$ be uninformative about ζ was that the integral in (96) take on equal values at (n-1) discrete values of ζ, or $G(a_1) = G(a_2) = \ldots = G(a_{n-1})$. Introducing new data points x', x", ... connected by continuation paths, this equality is extended to further values a',a",.... Now G(a) is a continuous function. As we continue to all points of the sample space S_x, if the set of spectral points a' where $G(a') = G(a_1)$ becomes everywhere dense on $0 < a' < \infty$, then G(a) = const. is the only possibility. Equation (96) then reads:

$$\int_0^\infty d\eta\, \pi(\eta)\eta^n\, e^{-\eta a} = (\text{const})\times a^{-n}\,, \quad 0 < a < \infty \tag{97}$$

and on inverting the Laplace transform we have again the unique solution (41). On setting $\pi(\eta) = \eta^{-1}$, (97) reduces to

$$\lambda(y,z) = y^{-1} > 0 \quad . \tag{98}$$

Our questions have now been answered. Uniqueness of the solution requires that the set of spectral points a' be everywhere dense on the positive real line; and nonintegrability was avoided because extension of λ_x along any continuation path connecting two points took the eminently satisfactory

form that a function of (x,ζ) was a constant. So, by study of Example 1 we see how all the conditions can be met, leading to the Case IIIc most pleasing to a Bayesian.

The structure thus revealed will, of course, generalize readily to other problems. But our story has already grown too long, and the next Chapter must be told elsewhere.

13. CONCLUSION

While the full implications of marginalization for Bayesian statistical theory are still far from explored, the analysis given here represents at least the necessary beginnings. However, in research of this type, more than half the game usually lies in the slow process of recognizing the existence of an important solvable problem, and learning how to reduce vague conceptual questions to definite, clearly formulated mathematical ones. After that, further progress to the limit of our mathematical capabilities generally comes rapidly.

Viewed in this way, one is encouraged to think that the slow initial stages are now over, and we may hope to see major advances in the determination of prior probabilities by logical analysis, in the near future.

The integral equations introduced here may or may not prove to be more widely useful, in practice, than previous desiderata for uninformative priors. At present, they seem to have at least the advantage of being general and noncontroversial; i.e., they express only the universally accepted principles of probability theory, making no use of intuitive ideas (symmetry, entropy, indifference, group invariance, "letting the data speak for themselves", etc.) which appeal differently to different people. Of course, with full understanding, those integral equations may

in time be seen as stepping-stones to a still more appropriate and useful method, as yet unimagined.

The most encouraging sign of all is simply that, at last, the first prerequisite for progress in Bayesian theory is now an accomplished fact. The blind alleys have been tracked to their ends, and after decades of neglect and worse--even from some who professed to be Bayesians--the program started by Jeffreys is recognized as the true road to progress. Mathematical problems that might have been solved by Wald or Fisher in 1940 are, at last, being taken seriously and actually worked on.

At present, the crucial problem before us is: What is the necessary and sufficient condition on the functional form of $p(y,z|n,\varsigma)$ for the integral equations (30) to possess nontrivial and "morally acceptable" solutions? Our analysis in Sec. 11 above does not yet answer this; only the future will tell how close it has come to that goal.

14. APPENDIX A — COMMENTS ON GROUP ANALYSIS

The explicit mathematical use of group invariance as a criterion for assigning probability distributions goes back to Poincaré (1912), although of course the intuitive recognition of symmetry in gambling devices was present from the very beginnings (Cardano and Pascal). It appears to the writer that, in the final analysis, all applications of probability theory are based necessarily on such considerations, however much those motivations have been disavowed.

Since the term "group analysis" has several different meanings, we try here to indicate how they are related to each other and to the general problem of inference.

In the group structure (43) considered by DSZ the sampling distribution is invariant under two groups G, $\bar{G}$ operating simultaneously on the sample space and the parameter space. The status of that approach can be seen as follows: (A) Whatever group structure of this kind a problem may possess, is determined by the functional form of the sampling distribution. (B) Ergo, whatever results may be deduced from that group structure, must also be deducible directly from the functional form. (C) Since the group analysis cannot be more general than a "functional form" analysis--and is easily seen to be less general--the question of method reduces to whether, in a problem where it is applicable, the group analysis leads to a more efficient calculation, or a better intuitive understanding, of the result. It seems clear that group analysis does accomplish both; and often brilliantly. Therefore, by all means, let us take advantage of the DSZ group analysis whenever we can. (D) Nevertheless, whether or not any group structure exists, the necessary and sufficient condition for agreement of B_1 and B_2 is always the set of integral equations (30). For a general understanding of marginalization, then, it appears that we should appeal to the integral equations rather than the group structure.

Now an entirely different kind of group analysis (Jaynes, 1968;1971;1976) has also been proposed and illustrated in several applications. Since I believe it to be closer to the spirit of what one means intuitively by "ignorance", and also more widely applicable mathematically, let us look at it in the context of the DSZ Example 2 [Eqs. (54)-(62) above].

What prior probability element $\pi(\mu_1\mu_2\sigma)d\mu_1\, d\mu_2\, d\sigma$ expresses "complete ignorance" of two location parameters associated with a common scale parameter? We have a sampling distribution of the form

$$p(dxdy|\mu_1\mu_2\sigma) = h\left(\frac{x-\mu_1}{\sigma}\,;\,\frac{y-\mu_2}{\sigma}\right)\frac{dx}{\sigma}\,\frac{dy}{\sigma} \qquad (A.1)$$

and we consider

Problem 1: Given the data $D \equiv \{(x_1,y_1);(x_2,y_2),\ldots(x_n,y_n)\}$, estimate (μ_1,μ_2,σ).

Complete initial ignorance means, intuitively, that having no other basis for inference, our estimates are obliged to follow the data; i.e., a noninformative prior is the means by which one achieves Fisher's goal of letting the data speak for themselves. As noted in the text, it is also the necessary starting point for construction of an informative prior.

Of course, a mere verbal statement such as "complete initial ignorance" is too vague to determine any mathematically well-posed problem. However, there is a rather obvious and basic desideratum of consistency: In two problems where we have the same prior information we should assign the same prior probabilities. Surely, any method for assigning priors which was found to violate this requirement would be rejected as self-contradictory.

Yet, as noted before (Jaynes, 1968), in many cases this desideratum is already sufficient to determine a unique solution. For, given the above Problem 1, we can carry out a transformation of all variables: $\{x_i,y_i,\mu_1\mu_2\sigma\} \to \{x_i'\, y_i'\, \mu_1'\, \mu_2'\sigma'\}$ which involves a mapping $\theta \to \theta'$ of the parameter space onto itself, and consider:

Problem 2: Given the data D', estimate $(\mu_1', \mu_2', \sigma')$.

Any proposed prior $f(\mu_1\ \mu_2\ \sigma)\, d\mu_1\ d\mu_2\ d\sigma$ will be transformed into $g(\mu_1'\ \mu_2'\ \sigma) d\mu_1'\ d\mu_2'\ d\sigma'$ according to the Jacobian of the transformation:

$$g(\mu_1'\ \mu_2'\ \sigma') = J^{-1}\ f(\mu_1\ \mu_2\ \sigma) \tag{A.2}$$

where $J(\mu_1\ \mu_2\ \sigma) = \partial(\mu_1'\ \mu_2'\ \sigma')/\partial(\mu_1\ \mu_2\ \sigma)$; and of course the transformation rule (A.2) will hold whatever the function $f(\mu_1\ \mu_2\ \sigma)$. But now the transformation may be such that we recognize Problems 1 and 2 as entirely equivalent problems; i.e., they have the same sampling distribution and if initially we were "completely ignorant" of $(\mu_1\ \mu_2\ \sigma)$ in Problem 1--whatever that means--we are at least in the same state of knowledge about $(\mu_1'\ \mu_2'\ \sigma')$ in Problem 2. But our desideratum of consistency then demands that f and g must be the same function; i.e., the prior representing complete ignorance must satisfy the functional equation

$$f(\mu_1'\ \mu_2'\ \sigma') = J^{-1}\ f(\mu_1\ \mu_2\ \sigma) \tag{A.3}$$

which determines the ratio of prior density at any two points θ, θ' of the parameter space that are connected by the mapping.

If then the mapping $\theta \to \theta'$ is one of a group of transformations that is transitive on the parameter space (i.e., from any point θ any other point θ' can be reached by some transformation of the group), then (A.3) determines the prior, to within a multiplicative constant, everywhere.

Note that, in this method the prior is determined by the Jacobian of the transformation on the parameter space; and this remains true whether the group is Abelian or non-Abelian, compact or non-compact. Therefore, considerations of right Haar measure or left Haar measure do not arise. Haar measure is defined on the group manifold; and not on our parameter space.

Furthermore, this method is more general; for if by any means we can recognize the group on the parameter space that transforms our prior state of knowledge into an equivalent one, the same result (A.3) will follow whether there is or is not an image group on the sample space. Thus, the Mallows example (45) has no group structure of the DSZ type; yet there is a natural group induced on S_n by Bayes' theorem (Jaynes, 1968) which leads to the uninformative prior $\pi(n) \propto [n(1-n)]^{-1}$; and let me acknowledge [correcting an erroneous statement in Jaynes (1968)] that, unknown to me at the time, this result, too, had been anticipated by Jeffreys.

This will perhaps make clearer the distinction between our method and other group invariance arguments which do not appear to be motivated by the desideratum of consistency; or at least, to the best of the writer's knowledge, do not explicitly invoke it.

In this method a noninformative prior is not in general determined merely by the form of the sampling distribution; it is determined by specifying the invariance group on the parameter space. Furthermore, even if we do choose a group by the form of the sampling distribution, a given sampling distribution may be invariant under more than one group.

For the sampling distribution (A.1) perhaps the simplest transformation group is given by

$$\begin{aligned} \sigma' \quad \mu' &= a\sigma & 0 < a < \infty \\ \mu_1' &= \mu_1 + b_1 & -\infty < b_1 < \infty \\ \mu_2' &= \mu_2 + b_2 & -\infty < b_2 < \infty \end{aligned} \tag{A.4}$$

with

$$\begin{aligned} (x_i' - \mu_1') &= a(x_i - \mu_1) \\ (y_i' - \mu_2') &= a(y_i - \mu_2) \end{aligned} \quad . \tag{A.5}$$

The new sampling distribution $p(dx'\, dy' | \mu_1' \mu_2' \sigma')$ is then identical with (A.1). If our state of prior knowledge is such that this transformation

results in a Problem 2 that is entirely equivalent to Problem 1, then from the Jacobian $J = a^{-1}$ of (A.4) the uninformative prior must satisfy the functional equation

$$af(\mu_1 + b_1, \mu_2 + b_2, a\sigma) = f(\mu_1\, \mu_2\, \sigma) \tag{A.6}$$

from which we obtain the "widely recommended" prior element

$$f(\mu_1\, \mu_2\, \sigma)d\mu_1\, d\mu_2\, d\sigma = d\mu_1\, d\mu_2\, \frac{d\sigma}{\sigma} \quad . \tag{A.7}$$

However, when two "location" parameters are present, we may in some cases feel that this does not represent our prior knowledge. In (A.4) a change of scale $\sigma' = a\sigma$ affects only the accuracy of the x_i, y_i measurements. It may be that for other reasons not discernible in the sampling distribution (A.1), we know that the parameter σ not only sets the scale for the "measurement errors" $(x_i - \mu_1)$, $(y_i - \mu_2)$; it also sets the scale on which the difference of means $(\mu_2 - \mu_1)$ is to be measured.

As a concrete if oversimplified example, a spectroscopist may wish to determine the difference in magnetic moment of two atomic states by observation of the Zeeman effect, but the available magnet has uncontrollable field fluctuations. Here σ corresponds to the magnetic field strength, and μ_1, μ_2 to the resonant frequencies one is trying to measure. Doubling σ doubles the probable error in the measurements; but it also doubles the measurable difference $(\mu_2 - \mu_1)$. On the other hand, the crystalline environment of the atoms affects both their frequencies in the same unknown way independent of σ. All this prior information is in the mind of an experimenter E_1, but it does not appear at all

in the sampling distribution (A.1). Because of it, E_1 replaces (A.4) by

$$\begin{aligned} \sigma' &= a \quad , & 0 < a < \infty \\ (\mu_2' - \mu_1') &= a(\mu_2 - \mu_1) + c \quad , & -\infty < c < \infty \\ \tfrac{1}{2}(\mu_1' + \mu_2') &= \tfrac{1}{2}(\mu_1 + \mu_2) + b \quad , & -\infty < b < \infty \end{aligned} \tag{A.8}$$

This leads to the functional equation

$$a^2 f(q\mu_1 + p\mu_2 + b - c, p\mu_1 + q\mu_2 + b + c, a\sigma) = f(\mu_1\mu_2\sigma) \tag{A.9}$$

where $2q \equiv 1 + a$, $2p \equiv 1 - a$. But the LHS can be independent of both b, c only if $f(\mu_1\mu_2\sigma) = f(\sigma)$. The functional equation then collapses to $a^2 f(a\sigma) = f(\sigma)$, or $f(\sigma) \propto \sigma^{-2}$, the prior that DSZ found to avoid the paradox in Example 2.

We are far from having exhausted the number of transitive groups under which the sampling distribution (A.1) is invariant. For example,

$$\begin{aligned} \sigma' &= a\sigma & 0 < a < \infty \\ \mu_1' &= a\mu_1 + b & -\infty < b < \infty \\ \mu_2' &= \mu_2 + c & -\infty < c < \infty \end{aligned}$$

leads again to $f(\mu_1 \mu_2 \sigma) \propto \sigma^{-2}$; while

$$\begin{aligned} \sigma' &= a\sigma \\ \mu_1' &= a\mu_1 + b \\ \mu_2' &= a\mu_2 + c \end{aligned}$$

leads to $f(\mu_1 \mu_2 \sigma) \propto \sigma^{-3}$, which DSZ noted as avoiding the paradox in Example 4a, Version 1.

All of these correspond to different possible kinds of prior knowledge about the physical meaning of the parameters. These differences cannot be seen in the sampling distribution, which describes only the measurement errors. Thus, when we pass beyond pedagogical examples to real life problems, a further aspect of the quoted remark of Lindley (1971) becomes apparent.

As we see from this, group analysis does not answer questions of uniqueness. A given group leads to a definite prior, but there may be more than one group; and in any event group analysis--at least in any form yet visualized--does not tell us whether other solutions of the integral equations (30) may exist beyond those resulting from the group structure. However, it may be that new theorems bearing on this are waiting to be discovered.

APPENDIX B — HISTORICAL NOTE

Since statistical theory is returning to the original viewpoint of Laplace on the relation of inference and probability, we follow Laplace's example also in concluding with two remarks on the background of the marginalization problem, in addition to those noted by DSZ.

The mathematical facts underlying marginalization were fully recognized--and in the writer's view correctly interpreted--by Geisser and Cornfield (1963). Their equations (3.10) and (3.26) are just what we now call B_1's result and B_2's result; but instead of seeing a paradox in the difference, they very wisely termed the latter a "pseudoposterior distribution."

And inevitably, when we search for the origin of a Bayesian result, we turn to Jeffreys (1939). His §3.8 considers the bivariate normal case, and although the sample correlation coefficient r is a sufficient statistic for ρ, the posterior distributions (10), (24) again reveal the slight difference caused by different prior information about the location parameters (a,b). The comparison is reminiscent of our Equations (50), (51) above.

REFERENCES

A. P. Dawid, M. Stone, and J. V. Zidek (1973), "Marginalization Paradoxes in Bayesian and Structural Inference." J. Roy Stat. Soc. B 35, 189-233.

D. A. S. Fraser (1961), "On Fiducial Inference." Ann. Math. Stat. 32, 661-676.

George Gamow, The Birth and Death of the Sun, Penguin Books, Inc. New York (1945). The concluding sentence is: "And the year 12,000,000,000 after the Creation of the Universe, or A.D. 10,000,000,000, will find infinite space sparsely filled with still receding stellar islands populated by dead or dying stars."

M. Hamermesh (1962), Group Theory. Addison-Wesley, Reading, Mass. pp. 293-295.

E. T. Jaynes (1968), "Prior Probabilities," IEEE Trans. Systems Sci. and Cybernetics SSC-4, pp. 227-241. Reprinted in Concepts and Applications of Modern Decision Models, V. M. Rao Tummala and R. C. Henshaw, Editors (Michigan State University Business Studies Series: 1975).

E. T. Jaynes (1971), "The Well-Posed Problem," in Foundation of Statistical Inference (V. P. Godambe and D. A. Sprott, Editors) Toronto: Holt, Reinhart and Winston.

E. T. Jaynes (1976), "Confidence Intervals vs. Bayesian Intervals," in Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, W. L. Harper and C. A. Hooker, Editors, D. Reidel Publishing Co. Dordrecht-Holland; Vol. II, pp. 175-257.

Lord Kelvin, "On the Age of the Sun's Heat," Macmillan's Magazine, March 1862. He concludes: "It seems, therefore, on the whole most probable that the sun has not illuminated the earth for 100,000,000 years, and almost certain that he has not done so for 500,000,000 years. As for the future, we may say, with equal certainty, that inhabitants of the earth cannot continue to enjoy the light and heat essential to their life, for many million years longer, unless sources now unknown to us are prepared in the great storehouse of creation." Those unknown sources were revealed 43 years later, when Albert Einstein wrote $E = mc^2$.

D. V. Lindley (1958), "Fiducial Distributions and Bayes' Theorem," J. Roy. Stat. Soc. (B), 20, 102.

D. V. Lindley (1971), Bayesian Statistics: A Review. Philadelphia; Society of Industrial and Applied Mathematics.

H. Poincaré (1912), Calcul des Probabilites, pp. 118-130.

C. Villegas (1971), On Haar Priors, in Foundations of Statistical Inference (V. P. Godambe and D. A. Spott, Editors), Toronto: Holt, Rinehart, and Winston.