Some Thoughts about S. James Press and Bayesian Analysis·

By

Arnold Zellner
University of Chicago

I. Introduction

It is indeed an honor and a pleasure to have this opportunity to address a Conference honoring Professor S. James Press on the occasion of his formal retirement. I say "formal" retirement because from my knowledge of Jim's amazing productivity and very inquisitive nature over the many years that I have known him, it is impossible for me to imagine him in a state of full retirement in the usual sense of the word. I believe, with probability 0.99 (subjective or objective; see Press (2003) and Press and Tanur (2001) for clarification) that he will continue to pursue his research and other interests vigorously and make additional important contributions in the years ahead. As prior information for this assessment, note from his University of California at Riverside web page, he lists "only" the following as his Research Areas: "Image classification and reconstruction (The project involves developing multivariate Bayesian statistical methods for reconstructing scenes on the ground based on noisy signals received by sensors in the satellite.), Statistical analysis of microarrays, Cognitive modeling in sample surveys, Bayesian factor analysis, and Data mining: Bayesian analysis in data mining." With a research agenda like this, who can imagine him spending much time playing golf on the course alongside his home in Riverside?

As you all know, Jim has had a very productive career in lecturing, mentoring students, research, service to the universities with which he has been associated and to the Statistics profession and creating a wonderful family. And he has played a key role in ushering in the current Bayesian Era with his many Bayesian research contributions, his fine Bayesian books and papers and his key role in the formation of the American

· Prepared for presentation at Conference in Honor of Professor S. James Press, held at University of California at Riverside, May 14, 2005.

Statistical Association's Section on Bayesian Statistical Science (SBSS, http://www.amstat.org) and the International Society for Bayesian Analysis (ISBA, http://www.bayesian.org). See the indicated home pages for detailed information regarding his contributions to the founding of these organizations. In what follows, some of these topics and others will be discussed in sections titled, with great "originality", The Past, The Present, and The Future.

II. The Past

From the very first time that I met Jim when he arrived in Chicago in the late 1960s after completing his doctoral degree in Statistics at Stanford, I have been impressed with his breadth and depth of knowledge, creativity, productivity and persistent curiosity. Soon after his arrival in Chicago, he asked me for permission to sit in my graduate Bayesian Econometrics course. Of course I answered positively. The thought of having a bright new PhD in Statistics in my course who had worked and studied with such statistical luminaries as Olkin, Stein, Anderson and others at Stanford was exciting. However, I underestimated the extent to which Jim would contribute to the course's content. During each meeting of the course, he asked imaginative, relevant, fundamental questions in a most creative and constructive manner. For example, when I was discussing Bayesian solutions to the multicollinearity problem in multiple regression analysis, he asked, "Why not consider using a generalized inverse to solve the multicollinearity problem?" I responded that use of a generalized inverse must involve use of added prior information and suggested that it would be useful to describe it in detail. The result was a paper, S.J. Press and A. Zellner, "On Generalized Inverses and Prior Information in Regression Analysis," (1968) that revealed the hidden information after some algebraic analysis and produced a Bayesian interpretation of the sampling theory generalized inverse approach to solving multicollinearity problems.

At another point, he inquired about the properties of a ratio of random variables that are distributed as Student t, as in the case of the posterior distribution of the ratio of two regression coefficients or a structural parameter that is equal to a ratio of two reduced form equation coefficients, ratios that I explained do not have finite posterior moments.

But that did not completely satisfy him. The result was Jim's paper, "The t-Ratio Distribution" (1969) in which he analyzed the non-existence of moments and pointed to the possibility of encountering bimodal posterior distributions under certain conditions. Further, his questions about the finite sample properties of the sampling distribution of R^2, the squared multiple correlation coefficient, led to our joint paper, "Posterior Distribution for the Multiple Correlation Coefficient with Fixed Regressors" (1978). There were also many deep and useful philosophical, methodological and statistical questions that Jim raised throughout the quarter that we all enjoyed discussing and trying to answer. His contribution to my course that year was indeed significant, appreciated by all and indicative of a brilliant future for him.

This interaction with Jim in my course, our joint NSF research grant and research workshops and meetings at Chicago and elsewhere reflected his earlier work not only in Statistics but also his studies for his MS degree in Mathematics and his BA degree in Physics and some experience gained from his employment in research organizations. This background plus an inquisitive nature and brilliant intellect help to explain the origins, breadth and depth of his comments and questions. Then too, I learned that he interacted a good deal with my former University of Chicago colleague Harry Roberts, a famous Bayesian statistician. Indeed in Jim's most recent book, *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*, Wiley 2003, he writes in the preface, "I am deeply grateful to Drs. Harry Roberts and Arnold Zellner for exposing me to Bayesian ideas." (p. xxvii). I believe that Harry's deep appreciation and understanding of Bayesian ideas and his ability to apply them in solving practical statistical, quality control, business and forecasting problems appealed very much to Jim. Here was a case of theory with application, not just theory, and the Bayesian solutions to practical problems compared very favorably to non-Bayesian solutions. Also, Jim enjoyed learning about the philosophical aspects of Bayesian analysis and statistical research results of Harold Jeffreys, Leonard Savage, Bruno de Finetti, George Box, George Tiao, Dennis Lindley, Morris DeGroot, James Berger and other Bayesians.

In the late 1960s and early 1970s at Chicago, Jim enriched the environment by interacting on Bayesian and other statistical issues with Harry Roberts, Gordon

Antelman, George Tiao, David Wallace, Milton Friedman, myself and other Chicago Bayesians, as well as with non-Bayesians, or neutralists Albert Madansky, Stephen Stigler, Henri Theil, et al. In the advance copy of his 1972 book, *Applied Multivariate Analysis,* that he gave to me, he wrote an inscription that I value greatly, namely, "To Arnold, In gratitude for years of intellectual stimulation, and for exposing me to some of the solutions, and also some of the problems a Bayesian must face. Jim" Also, in the preface to this volume he graciously acknowledged others who made contributions to his development and his book in the following words, "The author's interest in multivariate analysis was inspired by Ingram Olkin, whose enthusiasm for the subject is contagious and whose ideas are tacitly in evidence, especially in Chapter 2. Jacques Drèze, Thomas Ferguson, Seymour Geisser, Divakar Sharma, Henri Theil, and Robert Winkler provided helpful comments on various sections of the manuscript. Discussions with D.V. Lindley, M. Stone, and G. Tiao were helpful in putting various Bayesian concepts into perspective." (pp.x-xi).

While his 1972 book had mainly a multivariate statistics emphasis, he did take up Bayes/non-Bayes issues quite directly in Section 1.4, titled, "Sampling Theory Versus Bayesian Approach," (p.4ff) in which he wrote, "This book does not take a dogmatic position on the sampling theory versus the Bayesian approach toward solving problems. There is no claim that there is a right and a wrong way. Rather, it is believed that cogent arguments can be made for both approaches to inference and decision making, and each may involve some subjective or technical difficulties." He goes on to expand on this theme illustrating an understanding of difficulties associated with the view that there is one and only one right way to learn from data and experience. Rather, he takes a pragmatic, thoughtful, scientific approach that involves determining the logical soundness of alternative approaches and how well they work in solving important inference and decision problems, an approach that many, including myself, find very appealing.

Of course since 1972, the historical record indicates that Bayesian methods have proven to be very valuable in providing solutions to many basic statistical problems that are better than solutions provided by other methods, as, e.g., shown in Jim Press's

important Bayesian papers on factor analysis, spatial analysis, hierarchical modeling, and other problems. Along with his many Bayesian statistical contributions, he had, and still has, a great interest in Bayesian computational problems. His 1979 article, "Bayesian Computer Programs" in which he listed and discussed many early Bayesian computer programs was widely read by and extremely useful to Bayesian researchers.

In addition to his Bayesian and other statistical research, Jim participated energetically in the activities of several Bayesian organizations. In the early 1970's he played a role in the creation of the National Bureau of Economic Research, National Science Foundation Seminar on Bayesian Inference in Econometrics and Statistics (SBIES). This seminar met two times a year beginning in 1971 at universities and other sites world-wide to hear and discuss reports on current Bayesian research. At an early meeting of the SBIES at Chicago, in our "business luncheon" meeting on Saturday, we discussed the structure of the organization that had been operating for several years rather informally. I raised the issue as to whether we needed to formulate a set of operating rules or a constitution to guide us. A moment later, Seymour Geisser moved that this matter never be brought up again. The motion was duly seconded and passed unanimously. Thus we operated for 25 years rather informally and successfully without wasting much time on bureaucratic matters and held wonderful meetings world-wide. Some referred to the Seminar on several occasions as "the longest lasting floating crap game ever." Currently, Sid Chib at Washington University in St.Louis heads up the SBIES and is planning a meeting to be held in St. Louis on Aug. 1-2, 2005; for details, see http://www.olin.wustl.edu/faculty/chib/sbies.

Jim addressed the SBIES group a number of times and co-edited a book, *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard* (1990), a collection of Bayesian research papers that the SBIES brought out in our series of books honoring famous Bayesians, e.g. Savage, Jeffreys, de Finetti, and others. Originally, when it was suggested that we produce a book honoring Barnard, who had previously attended one or two of our meetings, the following issue was brought up at one of our Saturday business luncheons. How can we bring out a book honoring Barnard? He's not a Bayesian! Immediately a person at the meeting responded, "That

doesn't matter. He's a great guy." And thus the matter was settled. In a letter, Barnard wrote that he wondered why a Bayesian group would publish a book in his honor. When I explained what happened at our business meeting, he was very pleased.

Jim also arranged two very successful meetings of the SBIES, one was held at the University of California at Riverside in 1986 and the second, with the help of Ruben Klein was held in Rio de Janeiro, Brazil in 1990 at a beautiful hotel across the street from one of Rio's famous beaches. Needless to say, in addition to learning much about Bayesian analysis in the research sessions, we all learned a good deal about priors and posteriors from visits to the beach. Joking aside, some claimed that this Rio meeting contributed significantly to the emergence of Bayesian research in Brazil.

Both the UCR and Rio meetings were very successful and indicated to all that Jim possesses exceptional organizational skills. Also, participants noted that he is conscientious, willing to work hard and interacts well with others in solving difficult problems that arise in planning meetings, forming organizations, etc. And later in the early 1990's he was most effective in the formation of the International Society for Bayesian Analysis (ISBA) and the American Statistical Association's Section on Bayesian Statistical Science (SBSS); see the websites http://www.bayesian.org. and http://www.amstat.org. for further information regarding Jim's important role in the founding and activities of these two impressive organizations that have been very successful, including his essay, "The Formation of the Section on Bayesian Statistical Science (SBSS) of the American Statistical Association."

III. The Present

So much for the past. Let us now turn to the present. Since the present is just a point of time between the past and the future, I shall try to be brief in this section.

There can be no doubt but that Jim's 2003 Wiley book, *Subjective and Objective Bayesian Statistics: Principles, Models and Applications* is one of the most important and impressive Bayesian books on the scene today. It has received very favorable reviews and been given a special designation by Wiley. The broad and deep coverage of many

Bayesian and related topics is impressive. As explained on page 13, "The book is subdivided into four parts. Part 1 includes Chapters 1 to 5 on foundations and principles, Part 2 includes Chapters 6 and 7 on numerical implementation of the Bayesian paradigm, Part 3 includes Chapters 8 to 11 on Bayesian inference and decision making and Part 4 includes Chapters 12 to 16 on models and applications." From this listing, it is clear that the book treats a broad spectrum of fundamental topics. The presentation involves clear, explicit explanations of concepts and procedures along with many helpful, understandable examples and references to the literature.

Topics covered in the book include concepts of probability, prior distributions, likelihood functions, Bayes' theorem, procedures for estimation, testing, prediction, computation, model averaging, decision-making, de Finetti transforms, entropy, data mining, factor analysis, classification and discrimination, etc. To get this broad coverage, Jim very wisely enlisted the cooperation of friends and colleagues, Sid Chib, Merlisse Clyde and Alan Zaslavsky to contribute chapters reflecting their expertise in computing, MCMC, etc., model averaging and hierachical Bayesian modeling, respectively. What a wonderful idea to achieve greater breadth and depth of coverage!

While it is infeasible to comment on all the above topics treated in Jim's book, I shall just mention one briefly in this section. Given past and recent intense discussions of alternative statistical testing procedures, see, e.g. Berger (2003), Ziliak and McCloskey (2004), and Zellner (2004a), I believe that many would benefit considerably from a reading of Jim's discussion of testing, pp. 217-233. He appreciates the major issues and reaches conclusions that are based on much past theoretical and empirical analysis. His carefully worded and insightful discussions probably will be appealing to most Bayesians if not to Fisherian p-value, frequentist, and mechanical 5% Neyman-Pearson types. Jim clearly knows the issues and their complexity. He appreciates that learning from data about alternative hypotheses or models and making decisions in evaluating them are difficult problems and illustrates how Bayesian posterior odds have been employed to solve them. At the end of his chapter titled "Bayesian Hypothesis Testing," he writes, "The Bayesian (Jeffreys) approach is now the preferred method of comparing scientific

theories. . . . Richard Feynman suggests that to compare contending theories (in physics) one should use the Bayesian approach." (p. 230)

With these brief remarks made about the present, now it's off to the future.

IV. The Future

I shall just say a few words about some aspects of the future of Bayesian analysis relative to its current successful state, as ably presented in Jim's book and other current Bayesian texts.

How might the Bayesian approach or approaches change in the years ahead? We must face this problem since "innovation and continual improvement" is the name of the game in science, industry and other competitive areas of life. From this point of view, since my (1988) paper, "Optimal Information Processing and Bayes' Theorem," that Jim cites in his book, I have been concerned about alternative justifications and derivations of Bayes' Theorem and generalizations of it in order to provide a number of optimal learning models that can be on the shelf for use to solve particular learning and other problems as they arise. Note that some are currently "adjusting" likelihood functions and prior densities in their use of Bayesian methods; see, e.g. the JASA paper by Ibrahim, Chen and Sinha (2003) , a departure from 'usual' Bayesian methods. In my (2000) paper, I introduced "information quality adjusted" priors and likelihood functions as contrasted to the usually assumed "standard" information quality in priors and likelihood functions. Also, Jeffreys (1998, p. 24ff) in his discussion of the standard proof of the product rule of probability has some serious concerns that an assumption used in the proof, namely that elements of the sets have the same probability of being drawn, may not always be satisfied in practice and thus introduced the product rule not as a theorem but just as an axiom in his inference system. In Zellner (2004b), is my recent effort to generalize the product rule of probability that shows that the product rule can hold in special cases in which probabilities of elements being drawn are not all the same. In psychology, there are a number of different learning models, as is well known, some incorporating "anchoring effects" or "costs of changing belief effects", etc. Indeed, there is a huge theoretical and empirical literature on interactive learning, learning in game theory, computer science, marketing, economics, and other fields. Also, many recognize that Bayesians and non-

Bayesians learn in various ways in performing data snooping, data mining, reductive inference or model formulation, diagnostic checking of models, etc. Indeed, in Hill (1988, p. 281) there is a brief discussion of a theory of data snooping that "…takes us, to some extent outside the classical version of Bayes and decision theory; however, the classical version often provides a first approximation to the data-analytic approach that I recommend…Some related discussions concerning time coherency and/or generalizations of the Bayesian approach are in Diaconis and Zabell (1986), Goldstein (1983) and Lane and Sudderth (1985)."

Recognizing all of the above and other relatively unexplored features of learning, a question is whether it is possible to create an optimizing framework within the context of which it is possible to produce optimal learning models that are useful in solving many learning problems just as use of minimum expected loss Bayesian estimation procedures produce optimal estimates for a broad range of models and problems that have good properties. Before getting into the details regarding an optimization procedure that produces Bayes' theorem and other learning models, I must address a commonly encountered question, "Why tamper with the well known and useful learning model, Bayes' theorem?" The same question might have been raised vis á vis the Model T Ford or Einstein's tampering with Newton's well-established and useful "laws of motion"? And indeed, as is well known, Einstein had many early critics. For example, as quoted in Cerf and Navasky (1998, p. 331), Ernst Mach, Professor of Physics at the University of Vienna remarked in 1913, "I can accept the theory of relativity as little as I can accept the existence of atoms and other such dogma." Of course empirical evidence supporting some predictions of Einstein's theory changed many individuals' degree of belief or confidence in the validity of his theory, a learning experience. Since the Michaelson-Morley experiments showing the non-existence of ether drift, there were some empirical "correction factors" introduced in attempts to patch up the Newtonian laws, perhaps indicating the need for a more general theory that Einstein produced and that was quite "subjective" in its early stages. However, importantly, it was a testable theory and indeed some future observations were in accord with the theory's predictions and not in accord with the predictions of Newton's laws, thereby causing many to increase their degree of belief in Einstein's theory that was on its way to becoming an "objective" law. However,

astrophysicists have recently reported that observations indicate that the universe is expanding at an increasing rate, not at a decreasing rate as predicted by Einstein's theory. Thus there is a need for some new theory that explains this "anomaly" and works well in explaining and predicting old and new data. And so on, and so on in the iterative process that is called science.

Similarly with respect to the famous Bayes' learning model, apparently the only formal learning model that explains how initial beliefs are transformed into post data beliefs in Statistics, it has worked very well over a broad range of problems but there are other problems for which it does not provide answers. These are problems in which the forms of likelihood functions and/or prior density functions are unknown, problems in which some information is of poor quality and some of high quality, etc. These and many other problems can be cited, as recognized by Hill (1988), Jaynes (1988), and others, to indicate the need for broader learning models that work well in providing information about models' parameters, future values of variables and in comparing and/or combining alternative models for observations. Needless to say, just as with Einstein's theory, it will have to be shown that the new learning models actually do perform well in solving learning problems encountered in practice before they will be accepted for use, as is normal in science.

With such considerations in mind, some years ago in a 1988 paper, I put forward an information theoretic optimization approach for producing optimal learning models, including the Bayesian learning model, Bayes' theorem that have the property that input information = output information and thus the optimal learning models are 100% efficient, as noted and discussed by Jaynes (1988, 2003), Hill (1988), Kullback (1988), Bernardo(1988), Soofi (1996, 2000), Bernardo and Smith (1994) and others. In later work, Zellner (1991, 1997, 2000, 2003), new variants of the original optimization problem were formulated and solved to obtain a battery of optimal learning models, all of which are 100% efficient. Some of these are briefly reviewed below and then optimal information procedures for evaluating alternative models, including the traditional Bayesian posterior odds procedure, will be derived as optimal information processing procedures and their properties and uses indicated. In particular, it will be shown how to evaluate an initial "subjective" model vis á vis an established "objective" model, a

problem that may appeal to those, including Judy, Jim and others who have worked hard to understand the "subjectivity" and "objectivity" of science. As mentioned above, it appears that "subjective" theories become "objective" when it is shown convincingly that they work better in explanation, prediction and policy-making than currently utilized "objective" theories or mechanical empirical procedures.

To structure an information processing problem, there is of course a need to choose an information measure. In the past, I have used the Gibbs-Shannon (GS) measure, and noted that it would be interesting to extend my analyses to use of the Rényi (1961) and Silver (1991) measures of information. Below, I shall show that my solutions are invariant to the use of the GS and a form of the Rényi information measures. In an effort to simplify the interpretation of the GS measure of information in a probability density function, e.g. $\int g(x)\ln[g(x)/m(x)]dx$, where $g(x)$ is a probability density function with x either a scalar or vector, I interpret this as the "expected ln height of the density g(x) relative to the measure m(x)." Of course to measure height, there has to be some reference level, say height relative to a plane or line or some given surface. In what follows, I shall, along with many others, e.g. Press (2003, p.246), take m(x) = constant throughout the analysis. Note too, that generally, $0 < g < +\infty$ and $-\infty < \ln g < +\infty$. That a scale of measurement extends over negative values is not unusual. Note that Fahrenheit , Centigrade and Absolute temperature scales include negative values as anyone who has lived in Madison, Wisconsin or Oslo, Norway knows. Whether or not there are lower and/or upper limits to the information scale as with the Absolute temperature scale is an issue that has not as yet been considered.

Now if we consider another density, say p(x), we can consider its height and ln height, use uniform measure, and consider the expected value of the ln height of p(x), that is, Elnp(x) = $\int$ g(x)lnp(x)dx. Note that this provides a measure of the information in $p(x)$ relative to uniform measure and it is recognized that this measure is dependent on the form of g. We want a form for g that incorporates all the relevant information regarding x in order to get a good measure of the information in a density function

involving x. Given this procedure, we now wish to consider a "standard" inference problem in which we have two inputs, a prior density for the parameters, θ, $\pi(\theta)$ and a likelihood function for θ given the observations, denoted by y, $f(y|\theta)$, the probability density function for the observations, evaluated at the observed data, y viewed as a function of θ, that is, the likelihood function. The outputs of the information processing problem are a post data density for the parameters given the observations, y and prior information, and a marginal density for the observations, evaluated at the observed observations, y, denoted by $h(y) = \int f(y|\theta)\pi(\theta)d\theta$. Now we write down information measures on the outputs, g and h and subtract information measures on the inputs, the prior density π and the likelihood function, f, as follows, denoting the difference by $\Delta(g)$:

(1) $\Delta(g)$ = Output Information - Input Information

$$= \int g \ln g d\theta + \int g \ln h d\theta - \int g \ln f d\theta - \int g \ln \pi d\theta$$

$$= \int g \ln[g/(\pi f/h)]d\theta \geq 0$$

We wish to minimize $\Delta(g)$ with respect to the choice of g subject to it's being a proper density, namely, $\int g d\theta = 1$, in order to keep the output information as close as possible to the input information so as not to lose any information. The solution to the problem, denoted by g*, using a calculus of variations approach, first given in Zellner (1988) is precisely in the form provided by Bayes' theorem, namely,

(2) $g^* = \pi f / h$

Some, including Hill (1988), Kullback (1988), Robert McCulloch and Udi Makov, pointed out that the second line of (1) is in the form of the Jeffreys-Kullback-Leibler distance or cross entropy measure, shown in the third line of (1) that is known to be non-negative, see, e.g. Kullback (1959. p. 14ff) for a proof and Jeffreys (1946) for his

distance measures, and thus g* in (2) is the form of g that minimizes (1). It is interesting that if we used a more general form for the information in a density, e.g. the expectation of the log of a density raised to the power lambda, with $0 < \lambda < \infty$, that is $E \ln f^{\lambda}$, it is the case that the solution to the above optimization problem is unchanged. Thus the solution is "invariant" to the use of this variant of Rényi's information measure.

The form of g^* in (2) is such that $\Delta\left(g^*\right) = 0$, that is the input information = the output information and thus the information processing procedure, Bayes' theorem is 100% efficient, as noted in my 1988 paper and by Bernardo and Smith (1994) or obeys an "information conservation principle" as Hill (1988) puts it after explaining that many sampling theory procedures do not satisfy this principle. Further, Jaynes (1988, pp. 280-281) commented, ". . . entropy has been a recognized part of probability theory since the work of Shannon 40 years ago, and the usefulness of entropy maximization as a tool in generating probability distributions is thoroughly established. . . . This makes it seem scandalous that the exact relation of entropy to the other principles of probability is still rather obscure and confused. But now we see that there is, after all, a close connection between entropy and Bayes's theorem. Having seen a start, other such connections may be found, leading to a more unified theory of inference in general. Thus, in my view, Zellner's work is probably not the end of an old story but the beginning of a new one." See also, Jaynes' comments on the result in (2) in his recently published book, Jaynes (2003).

As part of the "new story," in Table 1, I have provided several post data densities for parameters that minimize the difference between output information and input information and are 100% efficient in the sense that input information = output information in each case. In line 1, we have the "standard" Bayesian inputs, a prior density and a likelihood function and the solution, namely Bayes' theorem, as mentioned above. In the second line, we have inputted just a likelihood function and no prior, as R.A. Fisher wished to do in his fiducial approach. The minimizing solution is to take the post data density for the parameters proportional to the likelihood function and when this is done, information in = information out and the procedure is 100% efficient. In the third line the input information is in the form of moment side conditions, as in the Bayesian

method of moments procedure, see references and discussion in Zellner (2003), and the efficient post data density function for the parameters is in the form of an exponential function of a linear combination of the powers of the parameter. Similar solutions are available in the case of moment and other side conditions involving vectors of parameters. See, e.g. Green and Strawderman (1996), La France (1999), Zellner and Tobias (2001) and Zellner (1997) for examples involving vectors of parameters in regression and other models.

Table 1

Optimal Bayesian Information Processing Results

Inputs	Output: Optimal Information Processing Rule
1. Prior density, π Likelihood function, l	$g \propto \pi l$
2. Likelihood function, l	$g \propto l$
3. Post data moments,[1] $\mu_i = \int \theta^i g d\theta \quad i = 1,...,m$	$g \propto \exp\{-\sum_1^m \lambda_i \theta^i\}$
4. Prior density, π Post data moments, $\int \theta^i g\left(\theta \middle| D\right) d\theta = m_i$ $i = 1,2,...,m$	$g \propto \pi \exp\{-\sum_1^m \lambda_i \theta^i\}$
5. Quality adjusted inputs π^{w_1} and l^{w_2}, $0 < w_1, w_2 \leq 1$	$g \propto \pi^{w_1} l^{w_2}$
6. Inputs for time period t,[1]	

[1] g denotes the post data density and λ_i's are Lagrange multipliers. Extensions to cases in which vectors and matrices of parameters are employed, as in multiple and multivariate regression are available; see references at end of paper.

$t = 1,2,...,T$ $\qquad\qquad$ $g_t \propto g_{t-1} l_t \quad t = 1,2,...,T$

g_{t-1}, l_t

(with $g_0 = \pi_0$, the initial prior density)

In line 4 of Table 1, the input information is that in a prior density and in moment side conditions on the parameters and the optimal information processing density is in the form of a prior times the maxent density for the parameters. As with line 3, these results have also been obtained for problems involving vectors of parameters.

In line 5 of Table 1, there are "quality corrected" inputs, so-called because raising a density to a fractional power usually spreads them out or reduces their average height, as noted in the literature on "power priors." See, e.g., the paper by Ibrahim, Chen and Sinha (2003) for a thorough discussion of power priors and references to the literature. With such inputs as shown in line 5, the information processing solution, that is 100% efficient is to take the post data density for the parameters proportional to the prior raised to the power w_1 times the likelihood function raised to the power w_2. It is noted that the solution post data density in line 5 of Table 1 is in the form of a well known "Cobb-Douglas" production function with returns to scale = $w_1 + w_2$ and elasticity of substitution = -1, properties discussed in standard Econometrics and Economic texts. Other side conditions can be imposed to produce solutions that have other returns to scale and elasticity of substitution properties. See Zellner (1996, p. 169ff.), Zellner and Ryu (1998) and Dorfman and Koop (2005) for information regarding other general forms for production functions and many references to the literature. Also MacKay (2004, p.471ff.) provides some functions that relate informational inputs to informational output that are being used in information theory and its applications.

In the last line of Table 1, mention is made of dynamic information processing wherein the output of one period is the input to the next along with new data input each period. See Zellner (2000) for consideration of this optimization problem, a dynamic programming problem for which the Bellman solution is such that it is optimal to update post data densities for the parameters using Bayes' theorem. And when this is done,

[1] See Zellner (2000) for discussion of the solution to this multiperiod information processing problem, a dynamic programming problem.

information in = information out period by period and thus the usual Bayesian updating procedure is 100% efficient. However, as mentioned in the paper, if there are costs of changing beliefs or costs associated with obtaining new data, the optimal solution will differ from the traditional Bayesian solution and in some cases can resemble the forms of learning models used in the psychological literature. For further examples of information processing in relation to psychological learning processes, see the doctoral dissertation by David Just (2001) in which he employs optimal information processing rules to explain paradoxical results in nine psychological experiments.

Given the results in Table 1, we have a range of optimal post data densities for the parameters of a model that can be employed in point estimation, that is to compute post data means that are optimal relative to quadratic loss functions, or post data medians that are optimal relative to absolute error loss functions, etc. And of course, post data intervals and regions can be computed to solve problems of interval estimation. Note too that various properties of the optimal densities shown in Table 1 can be easily obtained. For example, at a talk at the University of Wisconsin that I was invited to present by Ehsan Soofi, I mentioned that the first and higher moments of lng, the ln height of g, can easily be evaluated, analytically or numerically by evaluating the integrals, $\int g\left(\ln g\right)^i d\theta$, $i = 1, 2, ..., m$ and thus means, variances and higher order moments are available to characterize the properties of lng or g. For example, if $E \ln g = a$, and $Var\ \ln g = b$ the maxent density for lng is normal with mean a and variance b, and also, g has a log-normal density. This density for g can be employed to characterize its properties.

Further, as shown in Zellner (2003), if we consider Bayes' Theorem in line 1 of Table 1, it follows that $E \ln g = E \ln c + E \ln \pi + E \ln l$ and $\text{var}\left(\ln g\right) = \text{var}\left(\ln \pi\right) + \text{var}\left(\ln l\right) + 2\,\text{cov}\left(\ln \pi, \ln l\right)$. The correlation of the ln-height of the prior and the ln height of the likelihood function can then be evaluated to characterize the dependence of the prior on the likelihood function. If the prior is uniform, the correlation is zero. Also, using this type of analysis the properties of the traditional Bayesian learning model in line 1 of Table 1 can be compared to those of other

learning models. Many other comparisons of the properties of the optimal information processing rules in Table 1 can also be made.

Properties of the predictive densities associated with alternative information processing rules in Table 1 can be compared as in the case of regression models studied in Zellner and Tobias (2001). Given that we have alternative predictive pdfs for future observations associated with alternative information processing procedures described in Table 1, and the observed future observations, we can obtain the predictive pdf, $h(y_f|D) = \int g(\theta|D) g(y_f|\theta, D) d\theta$ where D stands for the past data and prior information. Such predictive densities have been employed to form Bayes' factors and posterior odds to evaluate alternative models and their associated assumptions, e.g., a Bayesian method of moments model versus a traditional Bayesian model, a procedure about which Barnard (1997) commented as follows:

> "And above all any method is welcome which, unlike nonparametrics, remains fully quantifiable without paying obeisance to a model which one knows to be false. And your proposal to compare BMOM results with a model-based one should achieve the best of both worlds."

The use of Bayes' factors and prior odds to compute posterior odds is usually justified by an appeal to Bayes' theorem in Bayesian texts and applications.[1] That is, Bayes' theorem is employed to derive the result that the posterior odds is equal to the prior odds times the Bayes' factor. Recently I have shown that it is possible to derive this last relation using an information theoretic optimization approach similar to that used to produce the optimal rules shown in Table 1. See some of these results in Table 2 wherein posterior odds are related to prior odds and Bayes' factors by minimizing the difference between output information and input information for two mutually exhaustive hypotheses and for two non-exhaustive hypotheses. In both cases the optimal information processing rules are precisely in the form of the traditional Bayesian rules but have been derived using alternative assumptions. Clearly there are many variants of the problems shown in Table 2 that can be and will be analyzed in the future, some analogous to those

[1] See, e.g. Keynes (1921, p. 297) for use of Bayes' theorem to compute the posterior odds on the hypotheses, a "Final Cause" exists and a "Final Cause" does not exist after observing a miracle, also discussed in Zellner (1984, p. 39) in connection with an analysis of causality.

analyzed in Table 1 solutions to which will be in forms not the same as those provided by the traditional Bayesian approach. If these new solutions are shown to be better than the traditional solutions, say in performing various tests of alternative drugs in clinical trials, then, in accord with what was said above, these procedures will come to be viewed as "objective" and not "subjective." And of course, these comparisons can be made readily for a broad range of problems given the already great progress that has been made on the computing front with respect to numerical integration and optimization procedures.

Table 2

Information Processing and Evaluation of Alternative Hypotheses

1. Two Exhaustive Alternative Hypotheses

Inputs	Outputs

a. Prior Probabilities

$\Pi, 1-\Pi$

b. Data densities[1] $P, 1-P$

$h_1(y), h_2(y)$

c. Criterion Functional:

$$\Delta(P) = P\ln P + (1-P)\ln(1-P) -$$

$$[P\ln\Pi + (1-P)\ln(1-\Pi) + P\ln h_1(y) + (1-P)\ln h_2(y)]$$

d. min $\Delta(P)$ wrt P leads to:

e. Solution:

$$P/(1-P) = [\Pi/(1-\Pi)][h_1(y)/h_2(y)]$$

i.e.,

Post Data Odds = Prior Odds x Bayes' Factor

[1] $h_1(y) = \int f_1(y|\theta_1)\pi_1(\theta_1)d\theta_1$ and $h_2(y) = \int f_2(y|\theta_2)\pi_2(\theta_2)d\theta_2$

and

Info in = Info out

Table 2 (Continued)

Information Processing and Evaluation of Alternative Hypotheses

2. Two Non-Exhaustive Alternative Hypotheses

Inputs	Outputs
a. Prior Probabilities	
Π_1, Π_2	P_1 / P_2, Post data odds
b. Data Densities	
$h_1(y), h_2(y)$	
c. Criterion Functional:	
$\Delta(P_1, P_2) = P_1 \ln P_1 + P_2 \ln P_2 -$	
$[P_1 \ln \Pi_1 + P_2 \ln \Pi_2 + P_1 \ln h_1 + P_2 \ln h_2]$	
d. Solution:	
$P_1 / P_2 = [\Pi_1 / \Pi_2][h_1(y) / h_2(y)]$	
i.e., Post Data Odds = Prior Odds x Bayes' Factor	
and	
Info in = Info out	

In summary, in my opinion, the future looks very bright for Bayesians who use appropriate, formal learning models and methods in their analyses of statistical problems. Their solutions to problems will often be superior to those of non-Bayesians who usually use no or inappropriate learning models. E.g., the papers by Green and Strawderman (1996), LaFrance (1999), and van der Merwve et al. (2001) are outstanding examples of how new learning models have been employed to obtain good solutions to important applied problems when forms of likelihood functions are unknown. It appears that the current successful Bayesian Era will rapidly expand to incorporate various learning models and to provide a unifying framework for those concerned with learning theory and applications of it in many fields of science. Also, new and old information theory approaches for producing models for observations and prior densities for their parameters as described in Golan (2002), Jaynes (2003), Press (2003), Soofi (1996, 2000), Zellner (1996) and references cited in these sources, along with the creation of more computerized data bases, improved measurements and many more effective numerical integration and other computing techniques will be key factors in promoting the future progress of Bayesian analysis in all the sciences. With such developments, Bayesian analysis will continue to be dominant in the competitive arena that we call science and we can toast each other for the roles that we have played in creating a wonderfully productive Bayesian Era.

References

Barnard, G.A. (1997), Personal Communication.

Berger, J.O. (2003), Fisher Lecture: "Could Fisher, Jeffreys and Neyman have Agreed on Testing?" (with discussion), Statistical Science 18 (1), 1-32.

Bernardo, J.M. (1988), "Comment," The American Statistician, 42, No. 4, p. 282.

_____ and Smith, A.F.M. (1994), Bayesian Theory, New York: Wiley.

Cerf, C. and Navasky, V. (1998), The Experts Speak, first revised ed., New York: Villard Books, Random House.

Diaconis, P. and Zabell, S. (1986), "Some Alternatives to Bayes's Rule," in Grofman, B. and Owen, G., eds., Proceedings of the Second U. of California, Irvine Conference, Greenwich, CT: JAI Press, 25-38.

Dorfman, J.H. and Koop, G., eds., (2005), Current Developments in Productivity and Efficiency Measurement, Annals Issue, J. of Econometrics, 126, Issue 2, 233-588.

Golan, A. ed., (2002), "Information Theory and Entropy Econometrics," Annals Issue of the J. of Econometrics, Vol. 107, Nos. 1-2, 374 pp.

Goldstein, M. (1983), "The Prevision of a Prevision," J. of the American Statistical Association, 78, 817-819.

Green, E. and Strawderman, W. (1996), "A Bayesian Growth and Yield Model for Slash Pine Plantations," J. of Applied Statistics, 23, 285-299.

Hill, B.M. (1988), "Comment," The American Statistician, 42, No. 4, 281-282.

Ibrahim, J.G., Chen, M-H., and Sinha, D. (2003), "On Optimality of the Power Prior," J. of the American Statistical Association, 98, No. 461, 204-213.

Jaynes, E.T. (1988), "Comment," The American Statistician, 42, No. 4, 280-281.

_____ (2003), Probability Theory, Cambridge: Cambridge U. Press.

Jeffreys, H. (1946), "An Invariant Form for the Prior Probability in Estimation Problems," Proc. of the Royal Statistical Society (London), Series A., 186, 453-461.

_____ (1998), Theory of Probability, 3rd revised edition, 1967, reprinted in Oxford Classic Texts in the Physical Sciences, Oxford: Oxford U. Press.

Just, D. (2001), "Information and Learning," PhD thesis, Dept of Agricultural and Economic Resources, U. of California, Berkeley.

Keynes, J.M. (1921), A Treatise on Probability, London: Macmillan & Co.

Kullback, S. (1959), Information Theory and Statistics, New York: Wiley.

_____ (1988), "Comment," The American Statistician, 42, No. 4, 282-283.

La France, J. (1999), "Inferring the Nutrient Content of Food with Prior Information," American J. of Agricultural Economics, 81, 728-734.

Lane, D. and Sudderth, W. (1985), "Coherent Predictions are Strategic," The Annals of Statistics, 13, 1244-1248.

Mackay, D.J.C. (2004), Information Theory, Inference and Learning Algorithms, Cambridge: Cambridge U. Press.

Press, S.J. (1969), "The t-Ratio Distribution," J. of the American Statistical Association, 64, 242-252.

_____ (1972), Applied Multivariate Analysis, New York: Holt, Rinehart and Winston, Inc.

_____(1979), "Bayesian Computer Programs" in Zellner, A., ed., Studies in Bayesian Econometrics and Statistics in Honor of Harold Jeffreys, Amsterdam: North-Holland Publishing Co., 429-442.

_____(1990), co-editor, Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard, Amsterdam: North-Holland Publishing Co.

_____(2003), Subjective and Objective Bayesian Statistics: Principles, Models, and Applications, New York: Wiley.

_____ and Tanur, J.M. (2001), The Subjectivity of Scientists and the Bayesian Approach, New York: Wiley.

_____ and Zellner, A. (1968), "On Generalized Inverses and Prior Information in Regression Analysis," ms., Center for Mathematical Economics and Econometrics, Dept. of Economics and Grad. School of Business, U. of Chicago.

_____ and _____, (1978), "Posterior Distribution for the Multiple Correlation Coefficient with Fixed Regressors," J. of Econometrics, 307-322.

Rényi, A. (1961), "On Measures of Entropy and Information," Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability, 1, 547-561.

Silver, R. N. (1991), "Quantum Statistical Inference," ms., Los Alamos National Laboratory, N.M., 15pp., published in Grandy, J.W.T. and Milonni, P.W., eds., Physics & Probability: Essays in Honor of Edwin T. Jaynes, Cambridge, UK: Cambridge U. Press.

Soofi, E.S. (1996), "Information Theory and Bayesian Statistics," in Berry, D.A., Chaloner, K.M. and Geweke, J.K., eds., Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner, New York: Wiley, 179-189.

_____ (2000), "Principal Information Theoretic Approaches," J. of the American Statistical Association, 95, 1349-1353.

van der Merwve, A.J., Pretorius, A.L., Hugo, J. and Zellner, A. (2001), "Traditional Bayes and the Bayesian Method of Moments Analysis for the Mixed Linear

Model with an Application to Animal Breeding," South African Statistical Journal, 35, 19-68.

Zellner, A. (1984), Basic Issues in Econometrics, Chicago: U. of Chicago Press.

_____ (1988), "Optimal Information Processing and Bayes's Theorem," The American Statistician, 42, No. 4, 278-294, with discussion and the author's reply.

_____(1991), "Bayesian Methods and Entropy in Economics and Econometrics," in Grandy, W.T. and Schick, L.H., eds., Maximum Entropy and Bayesian Methods, Dordrecht: Kluwer Academic Publishers, 17-31.

_____ (1996), An Introduction to Bayesian Inference in Econometrics, in Wiley Classics Series, New York: Wiley.

_____ (1997), Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers, invited contribution to Perlman, M. and Blaugh, M., eds., Economists of the Twentieth Century Series, Cheltenham, UK: Edward Elgar Publishing Ltd.

_____ (2000), "Information Processing and Bayesian Analysis," presented to ASA meeting Aug. 2001 and published in J. of Econometrics, 107 (2002), 41-50.

_____ (2003), "Some Aspects of the History of Bayesian Information Processing," presented to ASA meeting Aug. 2003 and to appear in Annals Issue of the J. of Econometrics, edited by A. Golan.

_____ (2004a), "To Test or Not to Test, and If So, How? Comments on Size Matters: The Standard Error of Regression in the American Economic Review," J. of Socio-Economics, Vol. 1, 2, 581-586.

_____ (2004b), "Generalizing the Standard Product Rule of Probability Theory," H.G.B. Alexander Research Foundation Working Paper, July, 2004.

_____ and Ryu, H. (1998), "Alternative Functional Forms for Production, Cost and Returns to Scale Functions," J. of Applied Econometrics, 13, 101-127.

_____ and Tobias, J. (2001), "Further Results on Bayesian Method of Moments Analysis of the Multiple Regression Model," International Economic Review, 42, No. 1, 121-140.

Ziliak, S.J. and McCloskey, D.M. (2004), "Size Matters: The Standard Error of Regression in the American Economic Review," J. of Socio-Economics, Vol. 1, 2, 331-358.

Appendix

Selected References on New Information Processing and Bayesian Method of Moments (BMOM) Methods

I. General Information Processing Results: Producing Models, Priors and Information Processing Rules (including Bayes' Theorem)

See A. Zellner, Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers, Elgar, 1997, Part III, "Bayesian Priors, Models and Information Processing," pp. 97-175.

Here the problem of model formulation is discussed with many examples. In particular it is shown how information theory can be employed to derive univariate and multivariate regression and many other commonly employed models and prior densities for their parameters. Also, in the 1988 American Statistician article, "Optimal Information Processing and Bayes's Theorem," with discussion by E.T Jaynes, B.M. Hill, S. Kullback and J. Bernardo and the author's response, pp. 154-160 in AZ (1997), it is shown how to derive Bayes's Theorem as a solution to an information theory optimization problem. In a later article, A. Zellner, "Information Processing and Bayesian Analysis," (2000) presented to the Am. Stat. Assoc. in 2001 and published in J. of Econometrics, Vol. 107 (2002), 41-50, Bayes's Theorem and other learning models, including the Bayesian Method of Moments (BMOM) model are derived as solutions to optimization problems. See also the 2001 doctoral dissertation "Information and Learning," by D.R. Just, Dept. of Agricultural and Resource Economics, U. of California, Berkeley for additional information processing rules and their use in explaining anomalous behavior in psychological learning experiments. It should be appreciated that the BMOM model permits investigators to obtain posterior and predictive densities when likelihood functions and prior densities are not available.

II. References for the Theory and Applications of BMOM

1. Zellner, A. (1994), "Bayesian method of moments (BMOM) analysis of mean and regression models," in J.C. Lee, W.D. Johnson and A. Zellner (eds.), Prediction and Modeling Honoring Seymour Geisser, New York: Springer-Verlag, 61-74, reprinted in AZ (1997), pp. 291-304.

2. Green, E. and W. Strawderman, "A Bayesian Growth and Yield Model for Slash Pine Plantations," J. of Applied Statistics, 23 (1996), 285-299. [The authors did not have enough information to specify a likelihood function and thus used the BMOM in the first serious application of the method.]

3. Zellner, A. (1997), "The Bayesian Method of Moments (BMOM): Theory and Applications," Advances in Econometrics, 12, 85-105. [The BMOM approach is applied to a wide range of models.]

4. Zellner, A., J. Tobias and H. Ryu, "Bayesian Method of Moments (BMOM) Analysis of Parametric and Semi-Parametric Regression Models," in 1997 Proceedings of the Section on Bayesian Statistical Science, Am. Stat. Assoc., 211-216 and in South African Statistical Journal, 31 (1999), 41-69.

5. Zellner, A. (1998), "The finite sample properties of simultaneous equations' estimates and estimators: Bayesian and non-Bayesian approaches," J. of Econometrics, 83, 185-212. [The BMOM approach is applied to multivariate regression, unrestricted reduced form and structural estimation problems and results are compared to those yielded by traditional Bayesian and non-Bayesian estimation approaches, e.g. ML, 2SLS, etc.]

6. Zellner, A. (1998), "On Order Invariance of Maximum Entropy Procedures," ms., 5pp., H.G.B. Alexander Research Foundation, Grad. School of Business, U. of Chicago. [It is shown that maximum entropy procedures are order invariant. Arguments to the contrary in the literature are shown to be defective.]

7. La France, J. (1999), "Inferring the nutrient content of food with prior information," American J. of Agricultural Economics, 81,728-734. [An impressive analysis of an important problem using the BMOM approach and comparing it to other possible approaches.]

8. Zellner, A., J. Tobias and H. Ryu (1997), "Bayesian Method of Moments Analysis of Time Series Models with an Application to Forecasting Turning Points in Output Growth Rates," published in Estadistica, J. of the Inter-American Statistical Institute with discussion by Prof. Enrique de Alba, Vols. 49-51, Nos. 152-157, 1997-1999, 3-63.
9. van der Merwe, A. and Viljoen, C. (1998), "Bayesian Analysis of the Seemingly Unrelated Regression Model," ms., Dept. of Mathematical Statistics, U. of the Free State, Bloemfontein, S.A., presented to the annual meeting of the S.A. Statistical Association, November, 1998.
10. Geisser, S. and T. Seidenfeld (1999), "Remarks on the 'Bayesian' method of moments," J. of Applied Statistics, 26, 97-101 and Zellner, A. (2001), "Remarks on a 'critique' of the Bayesian Method of Moments," J. of Applied Statistics, 28, No. 6, 775-778, published version of my 1997 working paper. [It is pointed out that Geisser and Seidenfeld introduced an erroneous assumption that led to their negative conclusion.]
11. Soofi, E. (2000), "Principal information theoretic approaches," J. of the American Statistical Association, 95, 1349-1353. [Comments on information processing derivations of learning models and the BMOM.]
12. Mittelhammer, R.C., Judge, G.G. and Miller, D.J. (2000) Econometric Foundations, Cambridge: Cambridge U. Press, pp. 688-693. [A brief introduction to the BMOM analysis of the multiple regression model.]
13. Zellner, A. and J. Tobias (2001), "Further Results on Bayesian Method of Moments Analysis of the Multiple Regression Model," International Economic Review, 42, No. 1, 121-140.
14. van der Merwe, A.J., A.L. Pretorius, J. Hugo and A. Zellner (2001), "Traditional Bayes and the Bayesian Method of Moment Analysis for the Mixed Linear Model with an Application to Animal Breeding," South African Statistical Journal, 35, 19-68.
15. Zellner, A. and B. Chen (2001), "Bayesian Modeling of Economies and Data Requirements," Macroeconomic Dynamics, 5, 673-700. [BMOM estimation and forecasting techniques are employed, along with others, to forecast annual output growth rates for 11 sectors of the U.S. economy. Sector forecasts are aggregated to produce forecasts of aggregate U.S. GDP growth rates and such forecasts are compared with those derived from aggregate data and models. See also, Zellner, A. and J. Tobias (2000), "A

Note on Disaggregation and Forecasting Performance," J. of Forecasting, 19, 457-469, and Zellner, A. (2003), "Bayesian Shrinkage Estimates and Forecasts of Individual and Total or Aggregate Outcomes," ms, 25pp., H.G.B. Alexander Research Foundation, Grad. School of Business, U. of Chicago, for additional results on the effects of disaggregation on forecasting accuracy.]

16. Ibrahim, J.G., Chen, M-H., and Sinha, D. (2003), "On Optimality of the Power Prior," J. of the American Statistical Assoc. 98, No. 461 (March), 204-213. [Discusses the properties of power priors and their relation to earlier work on information processing with "quality corrected" prior densities and likelihood functions.]