

**H.G.B. Alexander Research Foundation
Graduate School of Business
University of Chicago**

**Some Aspects of the History of Bayesian Information Processing
By**

**Arnold Zellner
University of Chicago
1101 East 58th Street
Chicago, IL 60637**

Paper 0303

July 2003

Some Aspects of the History of Bayesian Information Processing

Arnold Zellner^{*}

University of Chicago

Abstract

For many years, traditional Bayesian (TB) and information theoretic (IT) procedures for learning from data were viewed as distinctly different approaches. Derivations of the TB and IT learning models are reviewed and compared. Then the 1988 synthesis of the TB and IT learning models and generalizations of them are described along with descriptions of selected applications. Included are learning procedures that do not require use of likelihood functions and/or priors. Works by leading Bayesians and information theorists are cited and related to TB/IT issues.

Key words: Bayes's theorem, information theory, optimal information processing rules, learning from data, statistical inference.

I. Introduction

In this paper, I present some observations on aspects of Bayesian information processing or learning. As is widely appreciated, learning from data and experience is a fundamental objective of all the sciences. Thus it is important to have good learning models available for scientific use. It is a fact, that many non-Bayesians do not use a formal learning model in their work and learn informally. Bayesians, who generally use Bayes's Theorem as a fundamental learning model have shown that it works well in analyzing a broad range of problems. An information theoretic procedure for deriving Bayes's Theorem and generalizations of it will be reviewed below. In particular, it will be shown how to perform inverse inference without the use of a prior density and likelihood function, or with just a likelihood function and no prior, or with quality adjusted priors and likelihood functions. Various applications of these information processing approaches will be cited and procedures for formally comparing them with traditional

^{*}Research financed in part by the National Science Foundation and by income from the H.G.B. Alexander Endowment Fund, Graduate School of Business, University of Chicago. Contact information: e-mail: arnold.zellner@gsb.uchicago.edu web page: <http://gsbwww.uchicago.edu/fac/arnold.zellner/more>

Bayesian approaches will be described. Last, references to works in the literature in which these new approaches have been applied will be provided.

In Section II, derivations of traditional Bayesian and new information processing rules will be reviewed. Then in Section III, works in the literature using the new information processing rules will be described. In some of these works, the new procedures have been compared to traditional Bayesian procedures using posterior odds, predictive tests, etc. Last, in Section IV, a summary and some thoughts on possible future developments are presented.

II. Derivation of Learning Models

2.1 The Traditional Bayesian Learning Model

In Bayesian texts, the Bayesian learning model, Bayes's theorem is usually derived by use of the product rule of probability. That is, the joint probability density function (pdf) for an observation vector, y , and a parameter vector θ , is given by:

$$(1a) \quad p(y, \theta) = f(y|\theta)\pi(\theta) = g(\theta|y)h(y).$$

Then, we have,

$$(1b) \quad g(\theta|y) = \pi(\theta)f(y|\theta) / h(y)$$

where g is the posterior density, π is the prior density, f viewed as function of θ is the likelihood function and h is the marginal density of the observations. To obtain (1b), the product rule of probability (1a) has been employed. Note that Jeffreys (1998, pp. 24-25) in his discussion of the proof of the product rule states that, "The proof has assumed that the alternatives considered are equally probable . . . It has not been found possible to prove the theorem without using this condition . . . But it is necessary to further developments of the theory that we shall have some way of relating probabilities on different data, and Theorem 9 suggests the simplest general rule that they can follow if there is one at all. We therefore take the more general form as an axiom . . ." (p. 25). In the proof of the product rule, to which Jeffreys refers, the elements of the sets A, B and of the intersection of A and B are assumed equally likely to be drawn in deriving $\Pr(AB) = \Pr(A)\Pr(B \text{ given } A) = \Pr(B)\Pr(A \text{ given } B)$. Jeffreys mentioned that he was unsuccessful

in his attempts to prove Bayes' theorem without this "equally likely to be drawn" assumption.

It is clear from this discussion of the proof of the product rule of probability, as with all proofs, that assumptions are being made that may not be satisfied in all circumstances. Thus some years ago, it occurred to me that it would be desirable to have a new way of producing learning models, such as Bayes's theorem. In this effort, I decided to proceed in a pragmatic way, pretty much the way an engineer might, to consider informational inputs and outputs and to derive a relation between them that would result in the output information to be as close as possible to the input information in order not to lose any information. Using the traditional inputs, a prior density, denoted by π above and a likelihood function, f and outputs g and h , the problem was how to obtain an optimal rule relating the outputs to the inputs. Using information theoretic measures of information in probability density functions, namely the negative entropy relative to uniform measure, that is, the expectation of the logarithm of a density function, $E \ln f(x) = \int f(x) \ln f(x) dx$, that I reinterpreted as the expected \ln height of f , a very descriptive measure of the information in a density function, I formulated the following criterion functional:

$$(2) \quad \Delta(g) = \int g \ln g d\theta + \int g \ln h d\theta - \int g \ln f d\theta - \int g \ln \pi d\theta \\ = \int g \ln [g / (\pi f / h)] d\theta \geq 0.$$

The problem is to minimize the difference, $\Delta(g)$, between the output and input information measures in (2) with respect to the output density g subject to it being a proper density. In a 1988 paper, I solved this problem using a calculus of variations approach. I also mentioned in my response to the discussants that independently Bruce Hill, Udi Makov and Robert McCulloch pointed out to me that the terms in the first line of (2) could be collected, as shown in the second line of (2), to provide a form of the non-negative Jeffreys-Kullback-Leibler measure of the distance between g and $\pi f / h$. See, e.g., Kullback (1959, p.14ff) for a proof of the non-negativity of this distance measure.

The solution to the minimization problem using either the calculus of variations or the non-negative distance measure approach is:

$$(3) \quad g^* = \pi f / h$$

that is precisely Bayes's Theorem. Further, when the optimal solution in (3) is substituted in (2), it is the case that the input information is exactly equal to the output information and thus no information is lost in this information processing procedure. That is, it is 100% efficient.

Before proceeding to discuss variants of the above problem, it is useful to review various reactions to the result in (3). The eminent physicist Edwin T. Jaynes (1988) commented as follows in his discussion of my result, “. . .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 in numerous new applications including statistical mechanics, spectrum analysis, image reconstruction, and biological macromolecular structure determination. . . . 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 this 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.” (pp. 280-281).

As Jaynes points out, the usefulness of entropy maximization as a tool in generating probability distributions has been recognized by many. In economics and econometrics, Davis (1941) was an early pioneer in using maxent to produce income distributions, firm size distributions and a model of consumer behavior. In addition see Lisman (1949) for comments on Davis's entropic theory of the household, Maasoumi (1990) and Zellner (1991) for reviews of research on entropy in economics and econometrics by many leading workers, and Golan (2002), Mittelhammer, Judge and Miller (2002) and Soofi (1996, 2000) for descriptions of new developments in information theory as it relates to economics, econometrics and statistics. Note also, that

entropic procedures have been used by many to produce prior densities for parameters as well as density functions for observations or likelihood functions in many fields of science. Last, Barnard (1951) provides a discussion of Shannon's and Fisher's measures of information, information theory and statistics in a paper with many invited discussants and his reply. Bartlett commented, "...he [Barnard] has done the Society a service by discussing the communication engineer's modern use of the word "information" and its implications for mathematicians and statisticians."

Further, the statistician Hill (1988) commented on the need to consider time coherence and referred to some of his and others' related work. See Zellner (2000) for results on dynamic information processing that show that it is optimal to update using Bayes's theorem, a procedure that is a solution to a dynamic programming problem, a dynamic version of the optimization problem described in equation (2). In addition, Hill commented as follows, "Zellner is to be congratulated for clearly formulating the conservation property implicit in Bayes's theorem, and holding it up for our careful scrutiny. ...If one does not wish to conserve this property, as is the case in all strictly non-Bayesian analyses of data, then it should be incumbent upon the statistician to state explicitly from whence the violation arises." (p.281) He goes on to illustrate this point using the procedure of "size-biased sampling" as an example.

Bernardo (1988) in his interesting discussion of my paper pointed to his and others' earlier work that involved using utility theory to derive optimal or proper scoring rules. See, e.g. Bernardo's (1979) article, "Expected Information as Expected Utility," and related work by I. J. Good and L.J. Savage that he cites. In this work, the utility of a density function, denoted by g , is employed, say $u(g) = a \ln g + b(\theta)$ and inference is viewed as an expected utility maximizing process that has yielded the Bayesian learning model as a solution. My response to Bernardo's thoughtful comments was as follows: "As regards Savage's and others' proofs that Bayes's theorem or the Bayesian IPR [Information Processing Rule] is an expected utility-maximizing solution, this is a fundamentally different result from that in my article, where no utility considerations enter and there is no assumption that the expected utility hypothesis is in some sense "valid" or "rational." My result deals with information-processing, not utility-maximizing behavior." (p. 284). However, it is interesting to note that the utility function, $u(g)$ shown

above is a monotonic function of $\ln g$, the log height of the density function g that is also an argument of the standard Gibbs-Shannon entropy information measure, as shown explicitly above. More general utility functions and/or information measures are available, as is well known, and use of them in formulating and solving the above optimization problem will lead to alternatives to the traditional Bayesian learning model, Bayes's Theorem.

Of course this response to Bernardo's comments does not rule out the use of the concepts of the utility of information or the price per unit of information. Indeed, in some economic models of firms, consumers and investors, information has been viewed as an input and decision-makers have been modeled as Bayesian learners in a number of studies. For some early work, see e.g. Zellner and Chetty (1965), Grossman (1975), Bawa, Brown and Klein (1979), Boyer and Kihlstrom (1984) and Cyert and DeGroot (1987). For more recent work on Bayesian portfolio analysis, see Quintana, Chopra and Putnam (1995). Whether these economic information processing models are useful in modeling the production of scientific research output is of course a difficult issue. In this regard, note that there is still much controversy regarding the axiomatic foundations of utility theory, as noted intuitively by Jeffreys (1998, p. 30ff) who declined to base his axiom system for probability theory on utility considerations or "expectation of benefit" as Bayes, Ramsey and others had done but was not against use of his theory in analyzing problems involving utility considerations. See also Machina and Schmeidler (1992) for work on an axiom system for probability theory that involves "the separation of an individual's preferences from their beliefs." (p. 748). Also, their axiom system is a "choice-theoretic axiomatization of classical subjective probability which neither assumes nor implies the expected utility hypothesis." (p. 748). As I mentioned to Machina some years ago, this separation was just what Jeffreys adopted in his *Theory of Probability* many years ago. Thus it appears "reasonable" and useful to entertain the production of information as a "technical" process and to characterize and design optimal or good technical information production and processing procedures that do not involve utility considerations. Then, given these technical results, they can be employed, just as economists employ production functions, in dealing with analyses involving the utility of information, e.g. in decision-making contexts, and possibly in analyses of the price of

information and how it is determined, say, in a market for information. There is indeed room for much more work to be done in these areas. However, below, I shall just concentrate on the “technical” aspects and not the utility aspects of information processing.

In the last sentence of my 1988 paper, I stated, “Further research to consider extended variants of the criterion functional used in this study as well as alternative measures of information would be valuable.” In subsequent work, described below, I have pursued these and other topics to produce a broader range of information processing rules that can be implemented in practice; see also the innovative information processing rules described by a former student in my course, David Just (2001) that he used to explain paradoxical behavior in nine psychological learning experiments.

Before turning to these results, it is relevant to note that a fuller characterization of the information in a density function occurred to me several years ago that I reported in a lecture at the U. of Wisconsin and discussed at length with Ehsan Soofi. Above the information in a density function relative to uniform measure was defined to be the negative entropy or $E \ln g$ that I interpreted as the expected log height of the density. It occurred to me that not only the first but also higher order moments of $\ln g$ might be considered, that is $E(\ln g)^n = \int (\ln g)^n g d\theta, n = 1, 2, \dots$. Then with these given moments, a maxent density for the log height of the density, $\ln g$ can be produced. This is very operational, as Soofi mentioned to me after he worked a few problems in the evening following my lecture. E.g., in the simple case, if $E \ln g = a$ and $\text{Var} \ln g = b$, the maxent density for $\ln g$ is $N(a, b)$ and g is log normally distributed.

As another example of the use of higher order moments of $\ln g$, consider Bayes’s Theorem given in (3). On logging both sides and taking the expectation with respect to g , we have: $E \ln g = E \ln \pi + E \ln f - \ln h$. Then, $\ln g - E \ln g = \ln \pi - E \ln \pi + \ln f - E \ln f$. On squaring both sides of this last expression and taking expectations of the terms with respect to g , the result is

$$(4) \quad \begin{aligned} \text{Var}(\ln g) &= E(\ln g - E \ln g)^2 \\ &= E(\ln \pi - E \ln \pi)^2 + E(\ln f - E \ln f)^2 + 2E(\ln \pi - E \ln \pi)(\ln f - E \ln f) \end{aligned}$$

Thus the variance of $\ln g$ is decomposed into posterior second moments of $\ln \pi$, the log height of the prior, and $\ln f$, the log height of the likelihood function. In addition this analysis provides a measure of covariance or correlation between $\ln \pi$ and $\ln f$ that quantifies the extent to which there is dependence between information in a prior density and that in a likelihood function, a topic that has been discussed qualitatively in the literature for many years. Note that if the prior is uniform, the covariance between the information in the prior and that in the likelihood function is equal to zero.

Second, other measures of the information in a density may be employed as alternatives to $E \ln g$, the expected \ln height. For example, Silver (1991, 1999) suggested using what he calls the Fisher information, namely $E[\partial g / \partial x / g]^2 = \int g [\partial g / \partial x / g]^2 dx$, the expectation of the squared relative slope of the density g as a measure of information in both univariate and multivariate densities. He notes that minimizing this criterion functional subject to certain side conditions results in a solution in the form of a Schrödinger-like partial differential equation. He suggests that such solution densities may be more general and useful than those produced by minimizing the usual $E \ln g$ subject to given side conditions. In addition, some have considered the Rényi information measure that includes the Shannon $E \ln g$ measure as a special case; see, e.g., the illuminating paper by Jizba and Arimitsu (2002) for comparative analyses of the Shannon and Rényi measures of information. How use of the Fisher and Rényi measures of information affects solutions to the information processing problem in (2) above, namely minimize the difference between the output and input information with respect to the choice of the form of g is a problem at the top of my “to do” list. In this connection, note that Jizba and Arimitsu (2002) remark, “Although Rényi’s information measure offers a very natural ...setting for entropy, it has not found so far as much applicability as Shannon’s (or Gibbs’s) entropy.”(p. 1). However, see Golan and Perloff (2002) for an interesting study that extensively uses Rényi’s entropy measure.

Last, it is important to note that some researchers seem to be adverse to the use of information theory on grounds that information theoretic or maximum entropy procedures lack “order invariance.” In Zellner (1998), included as Appendix A of this

paper, it is shown that this argument is fallacious. Indeed, when the same side conditions, e.g. given zero'th, first and second moments, are employed throughout, then maximum entropy procedures are order invariant. That is the same results are obtained when, e.g., sample 1 is analyzed followed by sample 2 or sample 2 is analyzed before sample 1 or the two samples are analyzed simultaneously. With respect to the critical literature, it is pointed out that authors have changed the number and/or nature of the side conditions, e.g. in going from sample 1 to sample 2, etc. Under such conditions, maxent procedures should not be invariant, as pointed out dramatically by Jaynes in his remark that if they were invariant we wouldn't have the laws of physics. See Appendix A for a more detailed analysis of this invariance issue with references to the literature.

2.2 Some Alternative Optimal Information Processing Rules

Clearly there are many variants of the optimization problem described above. For example, there are situations in which one might wish to input just a likelihood function and not a prior density, as R.A. Fisher did in his fiducial approach. If in the above problem, we omit the prior density input and just input a likelihood function, the solution is to take the post data density for the parameters proportional to the likelihood function, that is, $g^* \propto f$. This solution is equivalent to employing a uniform prior for the parameters but there is no need to introduce it in obtaining the above solution. Another problem, analyzed in Zellner (2000) involves adjusting the prior and likelihood functions for quality by raising each of them to fractional powers, i.e. π^{w_1} and f^{w_2} , with the w 's having values in the closed interval zero to one. Raising densities to fractional powers usually spreads them out, as noted in the literature on "power priors"; see, e.g., Ibrahim, Chen and Sinha (2003), a paper dealing with "power priors" and the use of information processing analysis to rationalize them. Using the "quality adjusted" input likelihood function and prior density in the criterion functional in (2) and minimizing it with respect to the choice of g subject to it being a proper density yields the following solution:

$$(5) \quad g^{**} = c\pi^{w_1} f^{w_2}$$

where c is a normalizing constant, as shown in Zellner (2000). Thus with these quality adjustments introduced, the optimal solution is in a form different from (3), Bayes's Theorem. Also, the solution in (5) satisfies what Hill (1988) called the "conservation principle", namely information in = information out and thus the information processing rule in (5) is 100 per cent efficient.

It is direct to apply the analysis associated with equation (4) to obtain an expression for the variance of the log of the density in (5), $\ln g^{**}$ and to compare it to that for the variance of $\ln g^*$ given in (4). From this calculation, it is the case that raising densities to fractional powers does indeed lower their informational content, as measured by expected log height. In addition, it is the case that only one side condition, the condition that the solution density be proper, was used to produce the optimal information processing rules in (3) and (5). Other possible side conditions have been mentioned in the literature namely moment side conditions, e.g. a given mean for the parameter vector, inequality constraints on parameters' values, differential equation side conditions restricting the solution density to belong to a certain family of densities, e.g. the Pearson system of densities, etc. Thus a rich range of side conditions reflecting given input information can be introduced and will modify forms of derived optimal information processing rules. Alternative rules, as well as combinations of rules can be evaluated using data as has been done in past work by van der Merwe et al (2001), Zellner and Tobias (2001) and other papers listed in the annotated bibliography in Appendix B. It is clearly desirable to use information in data, as well as analytical tools, to evaluate the performance of alternative learning models.

III. Learning without Likelihood Functions and Prior Densities

It has long been appreciated that in some circumstances dependable likelihood functions and prior densities may not be available. Without these two inputs, the solution to the optimization problem in (2) is to have g be a uniform density, a not very informative result. Is there anything that can be done to introduce additional information that may be available to working scientists? In lecturing some years ago, it occurred to me that "stories" have been made up about the sampling properties of error terms in relations, e.g. the errors are iid $N(0, \sigma^2)$ and raised the question, "Why not make up

“stories” about the realized error terms?” When measurements are made in science, as with Millikan and his oil drop experiments, each observation is obtained with a lot of background information regarding its quality, error, etc.; see, e.g., Press (2003) for an intriguing description of Millikan’s evaluations of individual data points. Thus, I thought it would be good to make assumptions about the realized error terms’ properties which, given a mathematical model for the observations, would imply information about properties of the subjectively random parameters of the model. Note that this is a reversal of the process that was employed in Zellner (1975) and Chaloner and Brant (1988) in which traditional Bayesian posterior densities for parameters were employed to calculate posterior densities for realized error terms and functions of realized error terms. For example, in terms of a standard regression equation, the observed, given observation $y_i = x_i' \beta + u_i$ where x_i is a given input vector of independent variables, β is a vector of regression coefficients with a given posterior density, say multivariate Student t, and u_i is a realized error term that is regarded as being subjectively random. While lecturing many years ago, it dawned on me that the realized error term u_i is a linear function of the elements of β and thus has a univariate Student t density. It didn’t take long to work out the details and to derive or compute the posterior densities of various functions of the realized error terms, as reported in my (1975) paper and work by Chaloner and Brant (1988), Hong (1989), Albert and Chib (1993) and others, that are very valuable in the diagnostic checking of models’s assumptions. Note that all of this work went forward in a traditional Bayesian framework with a given likelihood function and a given prior for the parameters.

A question arose in the early 1990s, namely can I reverse the process described in the previous paragraph by making prior assumptions about the *realized* error terms and/or functions of them that would imply post data moments and other properties of the parameters given the observed data? As usual when I have a new idea, I tried it out on relatively simple examples. One of them involves given time to failure data, $y_i, i = 1, 2, \dots, n, 0 < y_i < \infty$, that are assumed to satisfy the following relation:

$$(6) \quad y_i = \theta + u_i \quad i = 1, 2, \dots, n$$

where θ is a parameter with an unknown value. On summing both sides of (6) and dividing by n , the result is that the observed sample mean time to failure, \bar{y} is given by:

$$(7) \quad \bar{y} = \theta + \bar{u}$$

where $\bar{u} = \sum_{i=1}^n u_i / n$. Now to make inference about the parameter θ that we view as

subjectively random we have to make some assumptions in view of the old adage, nothing in, nothing out. Let's assume that there are no variables left out of equation (6), that its algebraic form is appropriate, that there are no outliers, and no systematic biases in the measurements. With all of these assumptions, that are usually made, but with no sampling assumptions regarding the errors in (6), we may further be willing to assume that the mean of the realized error terms satisfies, $E\bar{u}|D = 0$, where E is the subjective mean operator and D stands for the given data and background prior assumptions, stated above. Given that we have made this zero mean assumption, then from (7) we have:

$$(8) \quad E\theta|D = \bar{y} - E\bar{u}|D = \bar{y}.$$

Thus, without a likelihood function, without a prior density and without the use of Bayes's Theorem, we have the result that the post data mean of θ is \bar{y} , that is,

$E\theta|D = \bar{y}$. And it is well known that this mean is an optimal estimate of θ relative to a quadratic estimation loss function, $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$.

Further, to obtain a post data density for θ , $0 < \theta < \infty$, we can easily find the form of the density for the parameter θ , $g(\theta|D)$ that minimizes the criterion functional,

$$(9) \quad \min \Delta(g) = \int g \ln g d\theta$$

subject to the side condition given in (8) and that g be a proper density. That is, we seek the least informative density for g in terms of expected \ln height subject to the two side conditions mentioned above. The solution to this problem, obtained by standard calculus of variations procedures, is well known to be the exponential density:

$$(10) \quad g(\theta|D) = (1/\bar{y}) \exp(-\theta/\bar{y}) \quad 0 < \theta < \infty$$

Using this density, it is possible to compute the probability that θ lies between any two values, say a and b by evaluating $\int_a^b g(\theta|D) d\theta = \Pr(a < \theta < b|D)$. Since this procedure

solves the problem originally posed by Bayes many years ago, it was named the Bayesian Method of Moments (BMOM) in my first, 1994 paper on this topic. In a 1993 U. of Chicago workshop talk on this work, I mentioned that the idea for it came to me about Mothers' Day in 1993 and that I called it the BMOM approach in honor of Bayesian MOMs and MOMs of Bayesians.

Further, note that if we consider a future, as yet unobserved value of the time to failure, $y_f = \theta + u_f$, and assume that there are no measurement biases, that the functional form is satisfactory, etc., we can make assumptions regarding the properties of u_f and deduce the implications for possible values of y_f given that we have a post data density for θ . For example, if we assume $E u_f | D = 0$, we have $E y_f | D = E \theta | D = \bar{y}$. Then the maxent density for y_f is an exponential density, namely, $f(y_f | D) = (1/\bar{y}) \exp[-y_f / \bar{y}]$, $0 < y_f < \infty$. In addition, note by use of a conceptual sample, $y_c = \theta t + u_c$, where $t' = (1, 1, \dots, 1)$, additional prior information can be introduced; see Zellner (1997b) for details.

In my first paper on the Bayesian method of moments (BMOM), the procedure was applied to location parameter and multiple regression models, using not only first order moment conditions on the error terms but also second order moment conditions. Post data moments and maxent densities for the parameters and future values of variables were derived and reported at the 1994 Valencia meeting in Alicante, Spain. Among those in the audience was Ed Green who mentioned to me that he and Bill Strawderman were working on an analysis of a forestry model with data provided by forestry scientists. Since the forestry scientists did not provide them with any information about error terms' sampling properties, they were having difficulty in formulating a likelihood function, needed for their traditional Bayesian analysis of the data. On hearing about the BMOM approach, Green remarked that it was just what he and Bill needed and would apply it on his return to the U.S. And indeed their 1996 paper, cited in Appendix B and the references, was the first published application of the BMOM approach, including some ingenious extensions of it.

Another early, innovative application of the BMOM approach was the 1999 study by La France, cited in Appendix B and the references, on inferring the nutrient content of food in which he compared the BMOM approach with other available approaches. He opted for the BMOM approach on the basis of some interesting considerations and showed that it produced useful empirical results.

In Zellner (1997b), I extended the BMOM approach to apply to dichotomous variable models and a number of other models. A reanalysis of the Laplace Rule of Succession problem yielded BMOM results that have the estimated probability of a particular outcome rising more rapidly than provided by the Laplace Rule of Succession given a sequence of outcomes all of one type, say successes in tests of a theory. Some, including Jeffreys, have argued that instead of having the estimated probability of success rise in accord with the Laplacian rule, $(n+1)/(n+2)$, where n is the number of successful outcomes in n trials, it should rise more rapidly. The BMOM solution provides a more rapidly increasing probability as successes pile up. Then in other work with J. Tobias and H. Ryu, the BMOM approach was successfully extended and applied to new problems in multiple regression, semi-parametric regression and time series estimation, prediction and other problems. The invited discussant of our BMOM time series, forecasting paper, E. de Alba wrote very favorably about our work and proposed extensions of it. Also, Soofi (2000) in his JASA review paper, "Principal Information Theoretic Approaches" commented favorably on the BMOM approach.

On delightful visits to South Africa in 1996 and 1998, it was a pleasure to present talks on the BMOM at various universities, the 1996 ISBA meeting in Cape Town and the 1998 annual South African Statistics Association meeting. In particular, during my visits to the Department of Mathematical Statistics of the University of the Free State in Bloemfontein I had the good fortune to discuss my work with Abrie van der Merwe and his colleagues. He and C. Viljoen were the first to analyze the multivariate "seemingly unrelated regression" model using the BMOM approach in a paper presented to the meeting of the South African Statistical Association in November, 1998. Also, he, A. Pretorius, J. Hugo and I in a 2001 paper, published in the Journal of the South African Statistical Association analyzed the mixed regression model using the BMOM approach

and compared the results to those provided by maximum likelihood and traditional Bayesian methods.

The BMOM approach was applied to general “reduced form,” multivariate regression and structural econometric models in my 1998 paper presented at a conference in honor of Carl F. Christ and published in the *Journal of Econometrics Annals Issue* in his honor, edited by the Nobel Prize winner, L.R. Klein. It yielded new, exact finite sample minimum expected loss (MELO) estimates for structural parameters that are very operational as well as exact, finite sample post data densities for parameters and future observations. My discussant at the Christ Conference was Adrian Pagan, the eminent Australian econometrician and economist. After studying my paper and analyzing its results, he stated that he saw some good in the Bayesian approach. I remarked in return for this kind remark that we would no longer regard him as A Pagan.

Then too, it was a pleasure to receive a deep, insightful letter from George A. Barnard (1997) in which he commented about the BMOM approach 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 general point seems to me to be that we should express prior knowledge, as far as we can, in a prior. Then our model---likelihood-producing or moment-producing, or whatever---should help us process the observed data. Then we should go back to compare what we thought we knew before with the result of our data-processing. In arriving at our (for the time being) conclusion the weight that we can attach to the three components of our inference will vary from case to case. BMOM will be specially useful when the latter two stages of the three should predominate.”

As the above, brief comments indicate, in general the new BMOM approach was given a warm reception by many. However, with anything new, as is to be expected there were critics. In particular, a referee’s report on my first BMOM paper was very critical. It took a few minutes in the evening to discover that the elaborate critical analysis of the referee was based on an assumption that I did not make. As I reported to Jack Lee, the co-editor of the volume in which the paper appeared, if one removes the referee’s

unwarranted assumption, his critique falls apart. Jack saw the point immediately and accepted my paper for publication. This practice of referees introducing unwarranted assumptions to reach negative conclusions happened not just once but three times. In one case, the editor of a journal accepted and published a paper critical of the BMOM without even sending it to me for review. When I discovered that the critical article had been published, I sent the editor my previously written working paper indicating that the critics had made an assumption that I did not make. And when this erroneous assumption was removed from their paper, there was nothing left to their critique. See the citation in Appendix B to the exchange between Geisser and Seidenfeld and myself published in the *Journal of Applied Statistics*. It seems to me that the behavior of editors in such delicate matters should be more constructive and thorough in seeking the responses of those being criticized before rushing critical papers into print.

After the BMOM approach appeared, many wondered about its exact relation to the traditional Bayesian approach based on Bayes' theorem involving use of a prior density and a likelihood function. Barnard's remarks, presented above, do much to help clarify the situation. Further, it is clear from what has been presented, that both approaches can be derived as solutions to well-defined information processing problems. In one problem there are two informational inputs, the information in a prior and in a likelihood function. In the BMOM problem where it is assumed that the likelihood function and prior density are not available, the input is the information in moment side conditions. Also, if a prior density for the parameters is available as an input, along with moment side conditions for the parameters, the solution to the information processing problem is to take the output density for the parameters proportional to the prior times the maxent density for the parameters given the data. This solution appears in the form of Bayes' theorem with the maxent density for the parameters given the data replacing the likelihood function. Thus, the information processing problem can be formulated and solved for the variety of situations that Barnard described in the excerpt from his letter, presented above, in which we may find ourselves in analyzing data. Formulating, solving and testing solutions for a wide range of information processing problems appears to me to be a good way to make progress.

IV. Summary and Conclusions

In this paper some historical issues surrounding the process of information processing, Bayes's Theorem, the Bayesian method of moments and related topics have been considered. From what has been accomplished in the last few decades, it seems clear that the synthesis of traditional Bayesian and information processing procedures is a productive one that has already led to fruitful, new approaches for learning from data, formulation of explanatory and predictive models and enlarging the capabilities of data analysts. With respect to this last point, it is now possible to perform inverse inference and to derive predictive densities when the likelihood function's and/or the prior density's forms are unknown. Also, with predictive densities available, it is possible to use them in predictive testing of alternative models and/or in combining them and their predictions. That such Bayesian information processing procedures have been applied in published studies by a number of researchers worldwide indicates a need for them and the profession's appreciation of their value. Last, past axiom systems relating to optimization problems involving utility considerations and learning will probably have to be generalized to allow for the fact that various optimal learning models are now available.

H.G.B. ALEXANDER
RESEARCH
FOUNDATION

On Order Invariance of Maximum

Entropy Procedures

by

Arnold Zellner

U. of Chicago

Abstract

In this paper, it is shown that procedures for producing maximum entropy distributions are order invariant. That is, distributional results are invariant with respect to the order in which new data are processed as long as a given number of moments are updated to reflect information in new data. In addition, some examples in the literature purporting to show that maxent updating is not order invariant are reviewed with the finding that for these examples any good information processing rule should not be order invariant. This result was expressed in earlier remarks by Edwin T. Jaynes.

Graduate School of Business
U. of Chicago
Chicago, IL 60637, USA

May, 1998

On Order Invariance of Maximum Entropy Procedures

by

Arnold Zellner^{*}

Herein we demonstrate the order invariance of maximum entropy procedures, an extension of remarks made at seminars at the U. of Toronto and the Hebrew University of Jerusalem in 1997 and of Zellner (1997). After demonstrating the order invariance of maximum entropy procedures, an analysis of some examples that have appeared in the literature will be provided in which the hypothesis space has not been preserved. For example, new moment or other restrictions have been added to an original set and a lack of invariance has been shown in such circumstances. Several years ago such a demonstration led E. T. Jaynes to remark at a seminar presentation, "If maxent were invariant in such circumstances, we would not have the laws of physics." That is, e.g., the various gas laws are generated by changing the side conditions associated with maxent solutions. If such laws were invariant in the sense of implying logically equivalent results, as Jaynes remarked we would not have the various gas laws which have logically and empirically different implications.

To illustrate the order invariance of maximum entropy densities derived subject to moment side conditions, consider sample observations $D(1) = y(1)$ where $y(1)$ is an $n \times 1$ vector of observations with sample mean $m(1)$ and sample variance $v(1)$. Then the proper maxent density with these given two moments is well known to be a normal density, $N[m(1), v(1)]$. Now if we observe an independent sample of observations $D(2) = y(2)$, we can combine it with the sample $D(1)$ to obtain the combined sample, $D(1,2)$ with mean $m(1,2)$ variance $v(1,2)$ and proper normal maxent density, $N[m(1,2), v(1,2)]$. If alternatively, we observed $D(2)$ first and then $D(1)$, the mean and variance for the combined sample, $D(2,1)$ will be identical to the mean and variance for the combined sample $D(1,2)$ and thus the associated normal maxent density, $N[m(2,1), v(2,1)] = N[m(1,2), v(1,2)]$ since $m(2,1) = m(1,2)$ and $v(2,1) = v(1,2)$ and clearly we have order invariance when the same moments are updated for two or more independent sets of data.

^{*} Research financed in part by the National Science Foundation and by income from the H.G.B. Alexander Endowment Fund, Graduate School of Business, U. of Chicago.

While we have demonstrated order invariance using two moments above, the result applies to any given number of moment side conditions. Note that we do not permit the number of moment side conditions to change from sample to sample. Usually a given subject matter theory or law indicates which moments to employ as side conditions for all data sets.

As a second example, consider a Bayesian method of moments (BMOM) analysis, Zellner (1997a,b) of data relating to a mean time to failure parameter μ and a sample of failure times, $y_i = \mu + u_i$, $i = 1, 2, \dots, n$. If we assume that the y_i 's are given and $E\bar{u}|y = 0$, where $\bar{u} = \sum_i^n u_i / n$ and E denotes the subjective post data expectation operator, we have $E\mu|y = m(y) = \sum_1^n y_i / n$.

Then the proper maxent post data density for μ given that its mean is $m(y)$ is the following exponential density, $f[\mu|m(y)] = [1/m(y)]\exp\{-\mu/m(y)\}$. Now if another sample of data becomes available, say $w_i = \mu + v_i$, $i = 1, 2, \dots, n$ and we compute the mean of the y 's and w 's, denoted by $m(y, w)$ which is equal to $E\mu|y, w$, the proper maxent post data density for μ is $f(\mu|y, w) = [1/m(y, w)]\exp\{-\mu/m(y, w)\}$. Further, if w is observed first and then y is observed, $m(w, y) = m(y, w)$, that is sample means are the same for both orderings of the data, and thus the proper maxent exponential densities subject to either of these means will be the same. That is, the procedure is order invariant. It should be emphasized that in this example the form of the likelihood function is assumed to be unknown and thus the traditional Bayesian analysis and updating procedures, using Bayes' Theorem, can not be utilized.

While many more BMOM and other examples can be provided, the above suffice to indicate that as long as the moment side conditions are not changed, the associated proper maxent densities will be invariant to the order in which the data sets are analyzed. In reviewing the literature on this issue, it appears that authors have changed the side conditions, or as Jaynes might say, changed the laws, and are surprised that results are not invariant. To illustrate, Kass and Wasserman (1996, p.1349ff), citing Seidenfeld (1987),

present the following example, “Consider a six-sided die and suppose that we have information that $E(X) = 3.5$, where X is the number of dots on the uppermost face of the die. Following Seidenfeld, it is convenient to list the constraint set: $C_0 = \{E(X) = 3.5\}$. The probability that maximizes the entropy subject to this constraint is P_0 with values $(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$. Let A be the event that the die comes up odd, and suppose we learn that A has occurred. There are two ways to include this information. We can condition P_0 to obtain $P_0(X|A)$, which has values $(1/3, 0, 1/3, 0, 1/3, 0)$, or we can regard the occurrence of A as another constraint; that is $E(I_A) = 1$, where I_A is the indicator function of the event A . The probability Q maximizes the entropy subject to the constraint set $C_1 = \{E(X) = 3.5, E(I_A) = 1\}$ has values $(.22, 0, .32, 0, .47, 0)$, which conflicts with $P_0(\cdot|A)$.”

In this example, the conditioning events have been changed, namely first $E(X) = 3.5$, second $E(X) = 3.5$ given $E(I_A) = 1$ and third, $E(X) = 3.5$ and $E(I_A) = 1$. That the maxent probability mass functions for these three cases differ is to be expected since the side conditions have been changed. By changing the side conditions, different maxent probability mass functions are obtained just as in deriving various gas laws by maxent in physics. Note also, the unusual reaction to observing “the event that the die comes up odd...” To have one observation lead to the side condition or constraint, $E(I_A) = 1$ is mind-boggling. To go from one odd observation to stating that they are all odd is hardly scientific. If one assumes that only odd sides can appear upward, that indeed is quite a different “law” from that obtained just under the constraint that $E(X) = 3.5$. In line with Jaynes’ remark above, we should not expect our results to be invariant under such disparate “laws.”

The second example that Kass and Wasserman (1996, p.1350) provide also involves a change in the constraints. “Consider again the die example. After rolling a die, we typically can see two or three visible surfaces. That is, in addition to the uppermost side of the die, we can see one or two side faces depending on the orientation of the die. Thus we can record not just the value of the upper face, but also whether the sum of all visible spots on the side faces of the die is less than, equal to, or greater than the value showing. There are 14 such possible outcomes. For example, outcome (3, equal) means

the top face shows 3 and the sum of visible side faces equals 3. The original sample space can now be viewed as a partition of this larger sample space. Maximum entropy leads to a probability Q that assigns probability $1/14$ to each outcome. The marginal of Q for the six original outcomes is not P_o . The problem is, then, which probability we should use, Q or P_o ?"

Again it is clear that the hypothesis space has been changed, as explicitly recognized in the quotation. Under conditions in which we just observe the upward face of the die, we have P_o ; under conditions in which we observe more than just the upward side of the die, we have Q . That P_o and Q are different is similar to saying that Newton's and Einstein's laws are different which is as it should be since they are based on differing assumptions. Over a range of velocities, they provide similar predictions while outside this range, they provide different predictions. However, if one stays within the Newtonian realm, everything is logically consistent and similarly within the Einsteinian realm.

Maxent is a deductive procedure for producing alternative models or hypotheses that may or may not be consistent with observed data in descriptive and predictive senses. As shown by analysis of examples in Tobias and Zellner (1997), it is possible to use data and model selection techniques to choose among predictive models produced by BMOM maxent procedures and by traditional Bayesian assumptions and methods. While both approaches are logically sound, since they make different assumptions they are obviously not logically equivalent. The empirical validity of models produced by alternative approaches is of course of utmost importance.

References

Kass, R.E. and L. Wasserman (1996), "The Selection of Priors by Formal Rules," J. American Statistical Assoc., 91, 435, 1343-1370.

Seidenfeld, T. (1987), "Entropy and Uncertainty," in Foundations of Statistical Inference, eds. I.B. MacNeil and G.J. Umphrey, Boston: Reidel, 259-287.

Tobias, J. and A. Zellner (1997), "Further Results on the Bayesian Method of Moments Analysis of the Multiple Regression Model," H.G.B. Alexander Research Foundation,

Graduate School of Business, U. of Chicago. Presented at Econometric Society meeting, June, 1997 and at world meeting of the Int. Soc. For Bayesian Analysis, August, 1997.

Zellner, A. (1997), "Remarks on a 'Critique' of the Bayesian Method of Moments (BMOM)," H.G.B. Alexander Research Foundation, Graduate School of Business, U. of Chicago

Zellner, A. (1997a), "The Bayesian Method of Moments (BMOM): Theory and Applications," in T.B. Fomby and R.C. Carter, eds., *Advances in Econometrics*, Vol. 12, 85-105.

Zellner, A. (1997b), *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*, Edward Elgar Publ. Ltd., UK and US, pp. 291-304 and 308-318.

Appendix B

ARE 214 New Econometrics

1/03

Instructor: A. Zellner

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, on reserve in Giannini Hall Library and referred to below as AZ (1997):

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.]

References

- Albert, J. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *J. of the Am. Stat. Assoc.*, 88, 669-679.
- Barnard, G.A. (1951), "The Theory of Information," *J. of the Royal Statistical Society, Series B*, 46-59 with invited discussion by M.S. Bartlett, P.A. Moran, N. Wiener, D. Gabor, I.J. Good, F. J. Anscombe, R.L. Plackett and C.A.B. Smith and the author's response, 59-64.
- _____ (1997), Personal Communication.
- Bawa, V.S., Brown, S.J. and Klein, R.W. (1979), *Estimation Risk and Optimal Portfolio Analysis*, Amsterdam: North-Holland Publishing Company.
- Bernardo, J.M. (1988), "Comment," *The American Statistician*, 42, No. 4, p. 158, reprinted in Zellner (1997a).
- _____ (1979), "Expected Information as Expected Utility," *The Annals of Statistics*, 7, 686-690.
- Berry, D., Chaloner, K. and Geweke, J. (1996), *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, New York: John Wiley & Sons, Inc.
- Boyer, M. and Kihlstrom, R.E. (1984), *Bayesian Models in Economic Theory*, Amsterdam: North-Holland Publishing Company.
- Chaloner, K. and Brant, R. (1988), "A Bayesian Approach to Outlier Detection and Residual Analysis," *Biometrika*, 75, 651-659.
- Cyert, R.M. and DeGroot, M.H. (1987), *Bayesian Analysis and Uncertainty in Economic Analysis*, London: Chapman and Hall, Ltd.

Davis, H.T. (1941), *The Theory of Econometrics*, Bloomington, Indiana: Principia Press.

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

Golan, A. and Perloff, J.M., (2002), "Comparison of Maximum Entropy and Higher-Order Entropy Estimators," in Golan (2002), 195-211.

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

Grossman, S. (1975), *Essays on Rational Expectations, the Informational Role of Futures Markets and Equilibrium Bayesian Experimentation*, Ph.D. Thesis, Dept. of Economics, U. of Chicago.

Hill, B. M. (1988), "Comment," *The American Statistician*, 42, No.4, 281-282, reprinted in Zellner (1997a).

Hong, C. (1989), *Forecasting Real Output Growth Rates and Cyclical Properties of Models: A Bayesian Approach*, Ph.D. Thesis, Dept. of Economics, U. of Chicago.

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, reprinted in Zellner (1997a).

Jeffreys, H. (1998), *Theory of Probability*, 3rd revised 1967 edition, reprinted in Oxford Classics Series, Oxford: Oxford U. Press.

Jizba, P. and Arimitsu, T. (2002), "The World According to Rényi: Thermodynamics of Multifractal Systems," ms., Institute of Physics, U. of Tsukuba, Japan, 24pp.

Just, D.R. (2001), *Information and Learning*, Ph.D. Thesis, Dept. of Agricultural and Resource Economics, U. of California at Berkeley.

Kullback, S.(1959), *Information Theory and Statistics*, New York: John Wiley & Sons, Inc.

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

Lisman, J.H.C. (1949), "Economics and Thermodynamics: A Remark on Davis' Theory of Budgets," *Econometrica*, 17, 59-62.

Maasoumi, E. (1990), "Information Theory" in *The New Palgrave Econometrics*, Eatwell, J., Millgate, M. and Newman, P. (eds.), New York: W.W. Norton & Co., 101-112.

Machina, M. and Schmeidler, D. (1992), "A More Robust Definition of Subjective Probability," *Econometrica* 60, 745-780.

Min, C-k, (1992), *Economic Analysis and Forecasting of International Growth Rates Using Bayesian Techniques*, Ph.D. Thesis, Dept. of Economics, U. of Chicago.

Mittelhammer, R.C., Judge, G.G., and Miller, D.J. (2000), *Econometric Foundations*, Cambridge: Cambridge U. Press.

Press, S. J. (2003), "A Note on Modeling Subjectivity," ms, Dept. of Statistics, U. of California at Riverside, 9pp.

Quintana, J., Chopra, V. and Putnam, B. (1995), "Global Asset Allocation: Stretching Returns by Shrinking Forecasts," *Proceedings Volume. of the Section on Bayesian Statistical Science*, American Statistical Association.

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

_____ (1999), "Quantum Entropy Regularization," in von der Linden, W., Dose, V., Fischer, R. and Preuss, R.(eds.), *Maximum Entropy and Bayesian Methods*, Dordrecht/Boston/London: Kluwer Academic Publishers, 91-98.

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: John Wiley & Sons, Inc., 179-189.

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

van der Merwe, A.J. 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.

van der Merwe, A.J., Pretorius, A.L., Hugo, J. and Zellner, A. (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.

Zellner, A. (1975), "Bayesian Analysis of Regression Error Terms," *J. of the American Statistical Association*, 70, 138-144.

_____ (1988), "Optimal Information Processing and Bayes's Theorem," *The American Statistician*, 42, No. 4, 278-294, with invited 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/Boston/London: Kluwer Academic Publishers, 17-31.

_____ (1997a), *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*, Cheltenham, UK & Lyme, US: Edward Elgar Publishing Ltd.

_____ (1997b), "The Bayesian Method of Moments (BMOM): Theory and Applications," *Advances in Econometrics*, 12, 85-105.

_____ (1998), "The Finite Sample Properties of Simultaneous Equations' Estimates and Estimators: Bayesian and Non-Bayesian Approaches," invited paper presented at research conference in honor of Prof. Carl F. Christ and published in Klein, L.R. (ed.) *Annals Issue of the J. of Econometrics*, 83, 185-212.

_____ (2000), "Information Processing and Bayesian Analysis," presented to the American Statistical Association Meeting, Aug., 2001 and published in A. Golan (ed.), *Annals Issue of the J. of Econometrics*, 107 (2002), 41-50.

Zellner, A. and Chetty, V.K. (1965), "Prediction and Decision Problems in Regression Models from the Bayesian Point of View," *J. of the American Statistical Association*, 60, 608-616.

_____ 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.

_____, Tobias, J. and Ryu, H. (1997), "Bayesian Method of Moments Analysis of Time Series Models with an Application to Forecasting Turning Points in Output Growth Rates," *Estadistica, J. of the Inter-American Statistical Association*, 49-51, Nos. 152-157, (1997-1999), 3-63 with invited discussion by Prof. Enrique de Alba.

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