

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

Bayesian Analysis and Information Theory

by

Arnold Zellner

University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637

(Summary of invited paper presented at the
2nd Conference on Information and Entropy Econometrics (IEE),
Sept. 23-25, 2005,
to be published in American Statistical Association's
Business and Economics Statistics Section's Proceedings Publication)

e-mail: arnold.zellner@chicagogsb.edu
home page: <http://faculty.chicagogsb.edu/arnold.zellner/more/index.htm>

Bayesian Analysis and Information Theory

Arnold Zellner

Graduate School of Business

University of Chicago

ABSTRACT

Bayesian analysts use a formal model, Bayes' theorem to learn from their data in contrast to non-Bayesians who usually learn informally. In addition to proofs of Bayes' theorem in the literature, herein it is shown how to derive Bayes' theorem, the Bayesian learning model as a solution to an information theoretic optimization problem and that the solution is 100% efficient. Since this direct link between Bayesian analysis and information theory was established in Zellner (1988), recent work has shown how this optimization approach can be employed to produce a range of optimal learning models, all of them efficient, that have been employed to solve a wide range of "non-standard" problems, e.g., those in which likelihood functions and/or prior densities are unavailable and thus the traditional, Bayesian learning model can not be employed. These models can be compared to other available models by use of posterior odds. By having a set of optimal learning models "on the shelf" to solve a broader range of inverse inference problems, Bayesian analysis will be even more effective than it is today.

Keywords: Bayes' theorem, information theory, optimal learning models, inverse inference

1. Derivations of Bayes' Theorem

Bayes' theorem is usually proved using the product rule of probability. Jeffreys (1998, p.24ff) has serious concerns about an assumption used in the proof, namely, that the elements of the sets considered have the same probability of being drawn which may not be satisfied in all practical situations. After stating that he was unable to prove the product rule without this assumption, he introduced the product rule as an axiom in his inference system. In Zellner (2004), it is shown that the product rule can hold in special cases in which the probabilities are not all equal. Also, in Zellner (1988) it is shown that the Bayesian learning model, Bayes' theorem can be produced as a solution to an information

theoretic optimization problem that involves solving for the form of a posterior density for the parameters that minimizes the criterion functional, output information minus input information. The optimal solution density for the parameters can be employed to obtain optimal point estimates, interval estimates and predictive densities for future observations. Below these results will be presented after briefly describing and interpreting measures of the information in probability mass and density functions.

To measure the information in a probability density function, the Gibbs-Shannon measure of information in a density function will be used:

$\int g(x) \ln[g(x)/m(x)] dx$ where $g(x)$ is a proper density function, $m(x)$ is a given measure, x can be a scalar or a vector and the integral is defined over the range of x . A simple interpretation of this integral is that it is the expectation of the \ln height of $g(x)$ relative to the measure $m(x)$. Note that to measure height, we have to indicate height relative to something, e.g., a plane or a surface, etc. Here, we shall use uniform measure, that is $m(x) = \text{const.}$, throughout in measuring the information in a probability density function. For another density, say $f(x)$, the information in it relative to uniform measure is: $\int g(x) \ln f(x) dx$, that is the expectation of the \ln height of $f(x)$ relative to uniform measure. Note that we take the expectation using the probability density $g(x)$ that is assumed to contain all the available information regarding x . Below we shall use these information measures in structuring and solving information processing problems.

In the first information processing problem, it is assumed that there are two inputs, a prior density

for the parameter vector, $\pi(\theta)$ and $f(y|\theta)$, a density function for a vector of observations, y , given the parameters, which when viewed as a function of θ is the likelihood function. As outputs, we have $g(\theta|y)$, the post data density for the parameters and $h(y) = \int f(y|\theta)g(\theta|y)d\theta$, the marginal density of the observations. The criterion functional is given by:

$$\Delta(g) = \int g \ln g d\theta + \int g \ln h d\theta - \int g \ln \pi d\theta - \int g \ln f d\theta = \int g \ln [g / \{\pi f / h\}] d\theta.$$

In Zellner (1988), a calculus of variations approach was employed to minimize $\Delta(g)$ wrt choice of g and the minimizing value, denoted by g^* , is: $g^* = \pi f / h$ which is just the solution yielded by Bayes' theorem, namely take the posterior equal to the prior times the likelihood function divided by the marginal density of the observations. In discussion of this result, several pointed out that from the second line of the expression for $\Delta(g)$ above, it is in the form of the non-negative Jeffreys (1946) or Kullback (1959) distance or information measure that is known to be non-negative. When g^* is inserted in $\Delta(g)$ we have, $\Delta(g^*) = 0$ and with this value for g , we have input information = output information and thus this information processing rule, Bayes' rule, is 100% efficient as pointed out in Zellner (1988, 2000), Bernardo and Smith (1994), Hill (1988), and Jaynes (1988, 2003).

Given that g^* is available, it can be employed in connection with a density for some future observations, $f(y^f|\theta)$ to obtain a predictive density for y^f , the vector of future observations as follows:

$$p(y^f|y) = \int g^*(\theta|y)f(y^f|\theta)d\theta.$$

Such predictive densities are very useful in making probability statements about future outcomes and forming Bayes' factors, as is well known. In Zellner (2005) the analysis is

extended to produce posterior odds relating to alternative hypotheses as a solution to an information processing problem and the results are 100% efficient in the sense that output information = input information.

2. Other Optimal Learning Models

Another interesting case is that in which only a likelihood function is inputted and no prior, as R. A. Fisher wished to do in his fiducial approach to perform inverse inference. Here the solution to the information processing problem, denoted by g^{**} is: $g^{**} = cf$ where c is a normalizing constant. Thus here the optimal post data density is proportional to the likelihood function, as would also be the case if one employed a uniform prior density. However, the solution g^{**} given above does not require the introduction of the uniform prior.

In Table 1, optimal solutions to a variety of information processing problems are provided. In lines 1 and 2, we have the traditional Bayesian problem and the Fisherian problem, discussed above. In line 3, we input just post data moments of the parameter or parameters, as in the Bayesian method of moments (BMOM) approach and no prior or likelihood function. For applications of the BMOM approach, see Green and Strawderman (1996), La France (1999), and van der Merwe et al (2001). In line 5 of Table 1, quality adjusted inputs, discussed in Zellner (2000) are employed to obtain an optimal learning model in which "power priors" and "discounted likelihood functions" appear. For a good discussion of power priors, see Ibrahim et al (2003). Last, in line 6 of Table 1, a dynamic learning problem in the form of a dynamic programming problem is posed with a solution that is identical to usual updating of Bayesian posterior densities. See Zellner (2000) for analysis of this and other dynamic learning problems.

As regards Bayesian posterior odds for evaluating hypotheses, it has been shown in Zellner (2005) that it is possible to derive posterior odds relating to hypotheses as a solution to an information theory optimization problem similar to those considered above. That is, if two hypotheses are exhaustive and their associated prior probabilities are Π and $1 - \Pi$ and the marginal data densities are

$h_i(y) = \int f_i(y|\theta_i)\pi_i(\theta_i)d\theta_i, i = 1, 2$, then the criterion functional, output information – input information is:

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

Minimization of $\Delta(P)$ with respect to P , leads to the solution:

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

That is, posterior odds equal to prior odds times the Bayes' factor, a traditional Bayesian result. Note also that the above solution has the property it equates input and output information and thus the solution is 100% efficient. A similar result for two non-exhaustive hypotheses is provided in Zellner (2005). Many more testing problems can be analyzed using variants of the above framework.

In summary, it has been shown that the traditional Bayesian learning model, Bayes' theorem, and variants of it can be derived as solutions to information theoretic optimization problems and that such solutions are 100% efficient. That is, these solutions are such that input information = output information and thus there is no loss of information when these rules are employed. Or as Hill (1988) puts it, the solutions satisfy an information conservation principle in contrast to many non-Bayesian solutions that do not satisfy this principle and thus are inefficient. Having a set of optimal learning models available to solve a broad range of inverse inference problems, some that can't be solved using traditional Bayesian methods, has been and will be very useful.

3. REFERENCES

- Bernardo, J.M. and Smith, A.F.M. (1994), *Bayesian Theory*, New York: Wiley.
- 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.
- Jaynes, E.T. (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.
- Jeffreys, H. (1998), *Theory of Probability*, 3rd revised ed., 1967, reprinted in *Oxford Classic Texts in the Physical Sciences*, Oxford: Oxford U. Press.
- Kullback, S. (1959), *Information Theory and Statistics*, New York: Wiley.
- La France, J. (1999), "Inferring the Nutrient Content of Food with Prior Information," *American J. of Agricultural Economics*, 81, 728-734.
- 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. (1988), "Optimal Information Processing and Bayes's Theorem," *The American Statistician*, 42, No. 4, 278-294, with discussion and the author's response.
- Zellner, A. (2000), "Information Processing and Bayesian Analysis," presented to American Statistical Association Meeting, Aug. 2001, and in *J. of Econometrics* (2002), Vol. 107, 41-50.
- Zellner, A. (2004), "Generalizing the Standard Product Rule of Probability Theory," H.G.B. Alexander Research Foundation Working Paper, available for downloading from home page.
- Zellner A. (2005), "Some Thoughts about S. James Press and Bayesian Analysis," invited keynote address for the Press Retirement Conference, May, 2005, U. of California at Riverside and available for downloading from author's web site, address given above.

TABLE 1

Optimal Information Processing Results

Inputs	Output: Optimal Information Processing Rule
1. Prior density, π Likelihood function, f	$g \propto \pi f$
2. Likelihood function, f	$g \propto f$
3. Post data moments ¹ $\mu_i = \int \theta^i g d\theta \quad i = 1, 2, \dots, m$	$g \propto \exp\{-\sum_1^m \lambda_i \theta^i\}$
4. Prior density & Post data moments	$g \propto \pi \exp\{-\sum_1^m \lambda_i \theta^i\}$
5. Quality adjusted inputs $\pi^{w_1}, f^{w_2}, 0 < w_1, w_2 \leq 1$	$g \propto \pi^{w_1} f^{w_2}$
6. Inputs for time period t , $t = 1, 2, \dots, T$, ² g_{t-1}, f_t (with $g_0 = \pi_0$ the Initial prior)	$g_t \propto g_{t-1} f_t \quad t = 1, 2, \dots, T$

¹ g denotes the post data density and the λ_i 's are Lagrange multipliers. Extensions to cases in which vectors and matrices of parameters are employed, as in multiple regression are available; see references in Zellner (2000, 2005).

² See Zellner (2000) for analysis of this and related dynamic problems.