INFORMATION MEASURES AND BAYESIAN HIERARCHICAL MODELS

by

Prem K. Goel*
Purdue University

Technical Report #81-41

Department of Statistics
Purdue University

Revised July 1982

*This research was supported by the National Science Foundation under grant MCS-8002336A01.

Revised 7/13/82

INFORMATION MEASURES AND BAYESIAN HIERARCHICAL MODELS

by

Prem K. Goel*

ABSTRACT

We consider the Bayesian statistical models in which the prior distribution of the parameter vector $\underset{\sim}{\theta}_1$ in the distribution of an observable random vector $\underset{\sim}{X}$ is to be specified in a hierarchical fashion and one wants to learn about the hyperparameters at each level of this prior distribution. It is shown that for a wide class of information measures, based on the so-called f-divergence, the information decreases as one moves to higher levels of hyperparameters. This result unifies all the theorems in Goel and DeGroot (1981) and provides several other information measures for which the above desirable property holds.

Key Words and Phrases: Hyperparameters; f-divergence; Amount of Sample Information

*Prem K. Goel is Associate Professor, Department of Statistics, Purdue University, W. Lafayette, IN 47907. This research was supported by the National Science Foundation under grant MCS-8002336A01. I would like to thank the referees and the editors for helpful suggestions.

1. INTRODUCTION AND SUMMARY

In the Bayesian hierarchical models the prior distribution of $\underset{\sim}{\theta}_1$, the parameter involved in the probability distribution of the observable random vector $\underset{\sim}{X}$, is specified by the Decision Maker (DM) in several stages. Specifically, let the prior distribution of $\underset{\sim}{\theta}_1$ belong to a family of distributions indexed by the 'hyperparameter' vector $\underset{\sim}{\theta}_2$. Instead of assigning a value to $\underset{\sim}{\theta}_2$ in order to make an optimal decision, the DM is willing to assign another prior distribution to the possible values of $\underset{\sim}{\theta}_2$, which in turn may be indexed by an hyperparameter vector $\underset{\sim}{\theta}_3$, etc. This mode of expressing prior information was advocated by Bayesians in the early fifties (see Good, 1950). Since the early seventies, the Bayesians and empirical-Bayes statisticians have taken a vigorous interest in hierarchical models. Good (1980) provides a historical perspective and the usefulness of the hierarchical Bayesian methodology. In this article we are specifically interested in studying the behavior of the information about various hyperparameters in the observation $\underset{\sim}{X}$. The results obtained here show that for a wide class of information measures, which includes all measures discussed in Goel and DeGroot (1981), [referred to as G&D from here on], the information about the hyperparameters decreases as one moves away from the data through the various levels. Of course, this concept of decreasing information does not hold true for all measures as shown by examples in G&D. This article presents a more general and unified approach to the problem considered in G&D.

In Section 2, we consider a general hierarchical model, and show that information about $\underset{\sim}{\theta}_i$ in $\underset{\sim}{X}$ decreases as one moves to higher levels

of hyperparameters. This result is proved in terms of the class of information measures based on the f-divergence between either (1) the posterior and the prior distribution of $\underset{\sim}{\theta}_1$ or (2) posterior distributions of $\underset{\sim}{\theta}_1$ corresponding to two different observations. In Section 3, we illustrate the results of Section 2 in terms of a linear hierarchical model.

2. INFORMATION IN BAYESIAN HIERARCHICAL MODELS

Let $\mathcal{E}$ denote an experiment in which the value of the random vector $\underset{\sim}{X}$ is to be observed. Suppose that the gpdf $g(\underset{\sim}{x}|\underset{\sim}{\theta}_1)$ of $\underset{\sim}{X}$ involves an unknown parameter $\underset{\sim}{\theta}_1$. By information about the unknown parameter $\underset{\sim}{\theta}_1$ in the experiment $\mathcal{E}$, or a posteriori in the observed value $\underset{\sim}{x}$ of $\underset{\sim}{X}$, one may mean anything which changes the distribution of $\underset{\sim}{\theta}_1$. Hence there is no unique yardstick for measuring information and a measure which is very general in definition, in that most other measures in use are its special cases, will be most useful. In the Bayesian formulation, measures of information are defined either in terms of both utility and probability or in terms of probability alone.

Raiffa and Schlaifer (1961, Chapter 4) define the value of sample information in terms of both utility and probability. The conditional value of sample information (CVSI) is defined by the difference of expected terminal utilities (with respect to the posterior distribution) of the posterior Bayes action and the prior Bayes action. They also define the expected value of sample information (EVSI) as the expectation of CVSI with respect to the marginal distribution of $\underset{\sim}{X}$.

The information measures that are based on probability alone use some measure of divergence between the posterior and the prior

distribution of $\underset{\sim}{\theta}$. The idea is that if the posterior distribution is close to the prior distribution, then the DM's opinion about $\underset{\sim}{\theta}$ has not changed much after observing $\underset{\sim}{x}$, and therefore the observation $\underset{\sim}{x}$ has not provided any significant amount of information about $\underset{\sim}{\theta}$. However, if the two distributions are far apart then x has provided a lot of information about $\underset{\sim}{\theta}$. We shall call these as conditional amount of sample information (CASI) measures. Gavurin (1963) calls the expectation of CASI as the expected amount of sample information (EASI).

Some attempts have been made to relate EVSI and EASI (see Perez, 1968). The argument in favor of using EASI as an information measure is that whenever the DM does not have a clear idea of the utility function and therefore cannot choose an experiment with maximum expected utility, he or she may consider, as an alternative, maximizing EASI to choose an experiment. Bernardo (1979) argues that with an appropriate choice of the decision space and reasonable constraints on the utility function, maximizing Shannon information really amounts to maximizing EVSI. However, as shown by Perez (1968), other utility functions will not lead to this EASI.

We use several well-known CASI measures in G&D. However, a most general CASI measure is based on the concept of f-divergence (see Csiszár, 1963) for discriminating between two probability distributions.

Let μ_1 and μ_2 be two probability measures and λ denote a finite or a σ-finite measure such that both μ_1 and μ_2 are absolutely continuous with respect to λ with the corresponding gpdf's denoted by g_1 and g_2 respectively. Furthermore, let f(u) denote an arbitrary convex function

defined on the interval $(0,\infty)$.

Definition. The f-divergence of μ_1 and μ_2 is defined by

$$\mathcal{I}_f(\mu_1,\mu_2) = \int g_2(\underset{\sim}{\theta}) f\left[\frac{g_1(\underset{\sim}{\theta})}{g_2(\underset{\sim}{\theta})}\right] d\lambda(\underset{\sim}{\theta}) \tag{2.1}$$

Some properties of f-divergence are discussed in Csiszár (1977) and Ali and Silvey (1966). The f-divergence has as its special cases all the well known measures for discriminating between two distributions listed in Adhikari and Joshi (1956).

Note that $\mathcal{I}_f(\mu_1,\mu_2)$ does not depend on the dominating measure λ. However, in order to follow notations of G&D we will write $\mathcal{I}_f(g_1,g_2)$ instead of $\mathcal{I}_f(\mu_1,\mu_2)$.

In order to simplify the notation for the Bayesian hierarchical models, we shall denote the observable random vector $\underset{\sim}{X}$ by $\underset{\sim}{\theta}_0$. Let $g(\underset{\sim}{\theta}_0|\underset{\sim}{\theta}_1)$ denote the gpdf of $\underset{\sim}{\theta}_0$ with respect to some σ-finite measure, where $\underset{\sim}{\theta}_1$ is the unknown parameter vector. The prior gpdf of $\underset{\sim}{\theta}_1$ is denoted by $g(\underset{\sim}{\theta}_1|\underset{\sim}{\theta}_2)$, where $\underset{\sim}{\theta}_2$ is the 2nd-level hyperparameter vector; and in general the gpdf of the ith-level hyperparameter $\underset{\sim}{\theta}_i$, given $\underset{\sim}{\theta}_{i+1}$ is denoted by $g(\underset{\sim}{\theta}_i|\underset{\sim}{\theta}_{i+1})$. Without any loss of generality, we assume that each of the conditional distributions $g(\underset{\sim}{\theta}_i|\underset{\sim}{\theta}_{i+1})$, i=0,1,2,... is nondegenerate for each possible value of $\underset{\sim}{\theta}_{i+1}$. If for some level k, the hyperparameter $\underset{\sim}{\theta}_k$ in the distribution of $\underset{\sim}{\theta}_{k-1}$ is known, then we call it a k-1 stage hierarchical model.

For the general hierarchical model specified above, it was proved in G&D that the CASI $\mathcal{I}_f[g(\underset{\sim}{\theta}_i|\underset{\sim}{\theta}_0,\underset{\sim}{\theta}_k), g(\underset{\sim}{\theta}_i|\underset{\sim}{\theta}_k)]$, based on f-divergence

functions $f(u)=u \log u$; $f(u)=(u-1)\log u$; $f(u)=1-u^{\alpha}$, $0<\alpha<1$; and $f(u)=u^{\alpha}\mathrm{sgn}(\alpha-1)$, $0<\alpha\neq 1$, decreases as the level of the hyperparameter i increases from i=1 to i=k-1, for every observation $\underset{\sim}{\theta}_0$, every value of $\underset{\sim}{\theta}_k$ and all k, k=3,4,... . Thus the corresponding EASI also decrease as i increases for every value of $\underset{\sim}{\theta}_k$ and k=3,... . We shall now unify the results of Theorems 3.1 - 3.5 in G&D by proving this property for a much wider class of CASI and EASI measures. These results will be a direct consequence of the following theorem.

Theorem 3.1. Let (U,V,W) be a Markov chain. Then for any convex function $f(\cdot)$ on $(0,\infty)$,

$$\int g(u) f\left[\frac{g(u|w)}{g(u)}\right] d\lambda(u) \leq \int g(v) f\left[\frac{g(v|w)}{g(v)}\right] d\nu(v). \tag{3.1}$$

Proof. Because of the Markov property, we have $g(u|v,w)/g(u) = g(v|u)/g(v)$, and therefore,

$$\begin{aligned}\frac{g(u|w)}{g(u)} &= \int \frac{g(v|w)g(u|v,w)}{g(u)}\, d\nu(v) \\ &= \int \frac{g(v|w)}{g(v)}\, g(v|u)\, d\nu(v).\end{aligned} \tag{3.2}$$

Now, using Jensen's inequality, it follows from (3.2) that

$$\int g(u) f\left[\frac{g(u|w)}{g(u)}\right] d\lambda(u) \leq \iint g(u)g(v|u) f\left[\frac{g(v|w)}{g(v)}\right] d\nu(v)d\lambda(u). \tag{3.3}$$

Changing the order of integration on the right hand side, (3.3) reduces to (3.1).

For the Bayesian hierarchical model $\underset{\sim}{\theta}_{i-1}$ and $\underset{\sim}{\theta}_{i+1}$ are conditionally independent given $\underset{\sim}{\theta}_i$, i=1,2,... . Therefore $\{\underset{\sim}{\theta}_0,\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_2,\ldots\}$ form a Markov Sequence. Now, conditionally on $\underset{\sim}{\theta}_k$, if we identify U,V,W by $\underset{\sim}{\theta}_{i+1}$, $\underset{\sim}{\theta}_i$ and $\underset{\sim}{\theta}_0$ respectively, then the next theorem is a direct consequence of the above result.

Theorem 3.2. For any convex function $f(\cdot)$ on $(0,\infty)$, let the CASI for the ith level hyperparameter of the hierarchical model, given the observation $\underset{\sim}{\theta}_0$ and the hyperparameter $\underset{\sim}{\theta}_k$ (i<k), be denoted by $I_f(i|\underset{\sim}{\theta}_k)$, i.e.,

$$I_f(i|\underset{\sim}{\theta}_k) = \mathcal{I}_f[g(\underset{\sim}{\theta}_i|\underset{\sim}{\theta}_0,\underset{\sim}{\theta}_k), g(\underset{\sim}{\theta}_i|\underset{\sim}{\theta}_k)], \tag{3.4}$$

where $\mathcal{I}_f(g_1,g_2)$ is defined by (2.1). Then $I_f(i|\underset{\sim}{\theta}_k)$ is a decreasing function of i ($i\in\{1,2,\ldots,k-1\}$) for every $\underset{\sim}{\theta}_0,\underset{\sim}{\theta}_k$ and $k\geq 3$.

It is clear that Theorems 3.1 - 3.4 in G&D are special cases of Theorem 3.2. Here, the information is being measured in terms of the divergence between the posterior and the prior distributions of the hyperparameters. However, one may also be interested in measuring information in a local sense, as with Bayes-Fisher information, defined in G&D, which measures the effect of a small change in the observation vector $\underset{\sim}{\theta}_0$ on the posterior distribution on $\underset{\sim}{\theta}_i$. In Theorem 3.5 of G&D, it is shown that the Bayes-Fisher information about $\underset{\sim}{\theta}_i$ decreases as

i increases. However, the Bayes-Fisher information can be shown to be the limit, as $\Delta\underset{\sim}{\theta}_0 \to 0$, of the f-divergence between $g(\underset{\sim}{\theta}_i|\underset{\sim}{\theta}_0+\Delta\underset{\sim}{\theta}_0,\underset{\sim}{\theta}_k)$ and $g(\underset{\sim}{\theta}_i|\underset{\sim}{\theta}_0,\underset{\sim}{\theta}_k)$, for $f(u)=u \log u$. So Theorem 3.5 of G&D can be obtained as a limiting case of the following general result, which can be proved by tracing the steps in the proof of Theorem 3.1.

Theorem 3.3. For any convex function $f(\cdot)$ on $(0,\infty)$, the f-divergence,

$$I_f(i,\underset{\sim}{\theta}_0^*,\underset{\sim}{\theta}_0|\underset{\sim}{\theta}_k) = \mathcal{I}_f[g(\underset{\sim}{\theta}_i|\underset{\sim}{\theta}_0^*,\underset{\sim}{\theta}_k),g(\underset{\sim}{\theta}_i|\underset{\sim}{\theta}_0,\underset{\sim}{\theta}_k)] \tag{3.5}$$

between the posterior distributions of $\underset{\sim}{\theta}_i$ corresponding to two different observations $\underset{\sim}{\theta}_0^*$ and $\underset{\sim}{\theta}_0$ decreases as i increases.

Theorem 3.3 indicates that as the level of the hyperparameter moves away from the data, the posterior distributions corresponding to two different observations get closer and closer. Thus, the information about the hyperparameters decreases as the level i increases in this sense also.

3. THE LINEAR HIERARCHICAL MODEL

In this section, we shall consider the linear hierarchical model given by

$$\underset{\sim}{\theta}_{i-1} = \underset{\sim}{A}_i\underset{\sim}{\theta}_i + \underset{\sim}{\varepsilon}_i, \quad i=1,2,\ldots . \tag{3.1}$$

Here $\underset{\sim}{\theta}_0$ is an observable p_0-dimensional random vector, $\underset{\sim}{\theta}_i$ is a p_i-dimensional vector of hyperparameters; ε_i are independent normal $N(0,\underset{\sim}{C}_i)$ random vectors which are independent of $\underset{\sim}{\theta}_i$'s. This model has been widely used in the Bayesian literature and is discussed in G&D.

It is well known that if the prior distribution of $\underset{\sim}{\theta}_i$ given $\underset{\sim}{\theta}_k$ is $N(\underset{\sim}{B}_{ik}\underset{\sim}{\theta}_k, \underset{\sim}{P}_i)$, then the posterior distribution of $\underset{\sim}{\theta}_i$ given $\underset{\sim}{\theta}_0$ and $\underset{\sim}{\theta}_k$ is $N(\underset{\sim}{H}_i\underset{\sim}{h}_i, \underset{\sim}{H}_i)$ where $\underset{\sim}{B}_{ik}, \underset{\sim}{P}_i$ and $\underset{\sim}{H}_i, \underset{\sim}{h}_i$ are given by (4.3) and (4.4) respectively in G&D.

We know from Theorems 3.2 and 3.3 that the f-divergence information measures decrease as one moves to higher level of hyperparameters. The CASI and the corresponding EASI values for some f-functions are given in G&D. The CASI corresponding to $f^*(u)=(u-1)^2$, i.e., the χ^2-divergence between the posterior and the prior distributions of $\underset{\sim}{\theta}_i$ can be shown to be equal to

$$I_{f^*}(i|\underset{\sim}{\theta}_k) = \frac{|\underset{\sim}{B}_i|}{|2\underset{\sim}{B}_i - \underset{\sim}{I}|^{\frac{1}{2}}} \exp[\tfrac{1}{2}\, tr\{\underset{\sim}{P}_i^{-1}(\underset{\sim}{I}+(2\underset{\sim}{B}_i-\underset{\sim}{I})^{-1})\underset{\sim}{\gamma}_i\underset{\sim}{\gamma}_i'\}] - 1 \tag{3.2}$$

where $\underset{\sim}{B}_i=\underset{\sim}{H}_i^{-1}\underset{\sim}{P}_i$ and $\underset{\sim}{\gamma}_i=\underset{\sim}{H}_i\underset{\sim}{h}_i-\underset{\sim}{B}_{ik}\underset{\sim}{\theta}_k$.

It follows from (3.2) and $E_{\underset{\sim}{\theta}_0}(\underset{\sim}{\gamma}_i\underset{\sim}{\gamma}_i') = \underset{\sim}{P}_i-\underset{\sim}{H}_i$ that

$$\log|\underset{\sim}{B}_i| - \tfrac{1}{2}\log|2\underset{\sim}{B}_i-\underset{\sim}{I}| + \tfrac{1}{2}\, tr[\{\underset{\sim}{I}+(2\underset{\sim}{B}_i-\underset{\sim}{I})^{-1}\}(I-B_i^{-1})] \tag{3.3}$$

is also a decreasing function of i.

Furthermore, the χ^2-divergence between the posterior distribution of $\underset{\sim}{\theta}_i$ for two different observations $\underset{\sim}{\theta}_0^*$ and $\underset{\sim}{\theta}_0$ can be shown to be

$$I_{f^*}(i,\underset{\sim}{\theta}_0^*,\underset{\sim}{\theta}_0|\underset{\sim}{\theta}_k) = \exp\{(\underset{\sim}{\theta}_0^*-\underset{\sim}{\theta}_0)'\, \mathcal{I}(i)(\underset{\sim}{\theta}_0^*-\underset{\sim}{\theta}_0)\} - 1, \tag{3.4}$$

where $\mathcal{I}(i)$ is the Bayes-Fisher information matrix for $\underset{\sim}{\theta}_i$ given by (4.11) of G&D. In fact, it can be proved that all f-divergence measures

between these posterior distributions are monotone increasing functions of $(\underset{\sim}{\theta}_0^*-\underset{\sim}{\theta}_0)' \underset{\sim}{\mathcal{I}}(i)(\underset{\sim}{\theta}_0^*-\underset{\sim}{\theta}_0)$, which can be thought of as the Mahalonobis distance between the two posterior distributions. This result has an important implication for the classical discriminant analysis problem. Let $\underset{\sim}{Y} \sim N(\underset{\sim}{\mu},\underset{\sim}{\Sigma})$ and $\underset{\sim}{Z} \sim N(\underset{\sim}{\nu},\underset{\sim}{\Sigma})$. Then every f-divergence measure between these two distributions is a monotone increasing function of the Mahalonobis distance $D^2=(\underset{\sim}{\mu}-\underset{\sim}{\nu})'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}-\underset{\sim}{\nu})$, and therefore the classical classification procedure based on any f-divergence measure is equivalent to the one based on D^2.

References

Adhikari, B.P. and Joshi, D.D. (1956) "Distance, Discrimination et Résumé Exhaustif," Publications de l'Institut Statistics, University de Paris, 5, 57-74.

Ali, S.M. and Silvey, S.D. (1966) "A General Class of Coefficients of Divergence of One Distribution from Another," Journal of the Royal Statistical Society, Ser. B, 28, 131-142.

Bernardo, J.M. (1979) "Expected Information As Expected Utility," Annals of Statistics, 7, 686-690.

Csiszár, I. (1963) "Eine Informationtheoretiche Ungleichung unde ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten," Magyar Tud. Akad. Mat. Kutató Int. Köze., 8, 85-108.

Csiszár, I. (1977) "Information Measures: A Critical Survey," Transactions of 7th Prague Conference on Information Theory, Stat. Decision Function, Random Processes 1974, Boston: Reidel Publ.

Goel, Prem K. and DeGroot, M.H. (1981) "Information About Hyperparameters in Hierarchical Models," Journal of the American Statistical Association, 76, 140-146.

Good, I.J. (1950) Probability and Weighing of Evidence, New York: Hafner.

Good, I.J. (1980) "Some History of the Hierarchical Bayesian Methodology," Proceedings of the International Meeting on Bayesian Statistics ed. by J. Bernardo and M.H. DeGroot, Valencia, Spain: University of Valencia.

Perez, A. (1968) "Risk Estimates in Terms of Generalized f-entropies," Proceedings Colloquium on Information Theory, Debrecen 1967. Journal Bolyai Mathematical Society, Budapest, 299-315.