Posterior Convergence under Incomplete Information

by

A. Ben-Tal
Technion Technical University
Haiffa, Israel
and
Dept. of Industrial and Operations Engineering
University of Michigan
Ann Arbor, Michigan 48109 USA

D.E. Brown
Department of Systems Engineering
University of Virginia
Charlottesville, Virginia 22901 USA

R.L. Smith
Dept. of Industrial and Operations Engineering
University of Michigan
Ann Arbor, Michigan 48109 USA

Technical Report 87-23
October 1987

Posterior Convergence under Incomplete Information

by

A. Ben-Tal

Technion Technical University
Haiffa, Israel
and
Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, Michigan 48109

D.E. Brown*

Department of Systems Engineering
University of Virginia
Charlottesville, Virginia 22901

R.L. Smith

Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, Michigan 48109

---

*The research of this author was partially funded by a grant from The Jet Propulsion Laboratory under Grant #JPL957721.

ABSTRACT

It is well known that under complete information the posterior distribution in Bayesian inference converges as the sample size grows to a degenerate distribution concentrated at the true probability distribution. This paper examines a class of problems in which the sample possesses average properties not shared in expectation by any member of the support of the prior. We show that the posterior distribution converges as the sample size grows to a degenerate distribution in the support of the prior, concentrated at the probability distribution closest in relative entropy to the set of distributions sharing these average properties. We show, moreover, that the empirical distribution converges in probability to that empirical distribution closest in relative entropy to the support of the prior.

1. Introduction

The essence of Bayesian inference is the adjustment of a prior probability distribution based on data from observations or measurement. The appropriate representation of the data to facilitate the adjustment or updating of the prior is in the form of a likelihood function. This function expresses the probability of observing the data given values of the random variable of interest. In most cases the exact value of the empirical distribution used in the likelihood function is assumed known and to lie for large sample sizes within the support of the prior.

However, many inference problems do not involve exact knowledge of the empirical distribution. Instead, this distribution is simply known to reside within a set, typically defined by linear equality or inequality constraints. The problem is to update initial knowledge expressed as a prior distribution to account for all the information contained in the contraints on the empirical distribution but no more. The problem is further complicated by the fact that the number of trials used to generate the empirical distribution is unknown.

This is a class of problems amenable to inference procedures based on information theory. These procedures grew out of the work of Shannon [1] in communications. In general, information theoretic approaches to inference view lack of information and uncertainty as related concepts. Hence, these approaches account for the amount of information available for inference as a measure of the uncertainty in the problem.

Two measures of information are frequently used in inference procedures. The more common is entropy, which has been employed in numerous applications, to include: statistical mechanics [2], reliability [3], urban modeling [4], stock market pricing [5], and image restoration [6]. A

second, and more general measure, is relative entropy (directed divergence, discrimination information, Kullback-Liebler discriminator). Relative entropy has been applied to problems in statistics [7], risk assessment [8], legal inference [9], and software reliability [10]. The work of Charnes et al. [11,12,13] is a prominent example of applications in the OR/MS field.

This paper discusses results that demonstrate for a discrete sample space that the asymptotic points of concentration within the constraint set for the empirical distribution and within the support of the prior are equal to the distributions in each set which minimize the relative entropy.

The following section describes the inference procedure based on relative entropy, which is known as the principle of minimum relative entropy. Section 3 summarizes related research concerning the asymptotic properties of the principle of minimum relative entropy. Section 4 formally presents the equivalence results between Bayesian inference based on long run averages outside the support of the prior and relative entropy minimization. Finally, section 5 gives our conclusions.

2. The Principle of Minimum Relative Entropy

Relative entropy is the basis for a general procedure for updating the distribution function of one or several random variables. This information theoretic procedure has its origins in a related inference rule known as the principle of maximum entropy. This rule is applied to estimate a distribution function given contraints on that function. The contraints represent the information available to the decision maker. The principle of maximum entropy then prescribes selecting the distribution which is maximally uniformative but still satisfies the constraints.

Use of this principle requires a definition of "informativeness." Suppose X assumes values in $X = (x_0, x_1, \cdots, x_m)$. Let $P(X=x_i) = p_i$. Let $I(x_i)$ be a measure of the information contained in a message that $X = x_i$. If for two experiments on X represented by X_1 and X_2 it is true that $I(x_i \wedge x_j) = I(x_i) + I(x_j)$ when $P(X_1 = x_i, X_2 = x_j) = p_i \cdot p_j$ and $I(x_i) \geq 0$ for $i = 0,1,\ldots,m$ then $I(x_i) = -k \log p_i$. k is a constant and for our purposes $I(x_i) = -\ln p_i$ where ln is the natural logarithm. For this discrete distribution a measure of "uninformativeness" or uncertainty is the expected information gained when it is learned that $X = x_i$ for some $i = 1,\cdots,m$ or

$$H(p) = -\sum_{i=0}^{m} p_i \ln p_i \tag{1}$$

The expression in (1) is known as the entropy of $p = (p_0,\ldots,p_m)$ If the inference problem is to estimate a distribution which corresponds to a convex constraint set, Λ, then the principle of maximum entropy prescribes the argmax H(p) in (1) subject to $p \in \Lambda$, and

$$\sum_{i=1}^{m} p_i = 1.$$

The principle of minimum relative entropy is a more general inference procedure. Suppose a decision maker initially believes $P(X=x_i) = p_i$ $(i=1,\cdots,m)$. Additional information in the form of a convex constraint set Λ becomes available and $p \notin \Lambda$. The principle of minimum relative entropy requires selecting $q = (q_0,\ldots,q_m)$ which minimizes

$$I(q,p) = \sum_{i=0}^{m} q_i \ln q_i/p_i \tag{2}$$

subject to $q \in \Lambda$ and $\sum_{i=0}^{m} q_i = 1$. Clearly, the principles of minimum relative entropy and maximum entropy are equivalent when $p = 1/(m+1)$. Relative entropy minimization is the more general procedures which admits an initial or prior distribution.

Justifications for using the principle of minimum relative entropy as an inference procedure are based on both axiomatic and asymptotic arguments. Axiomatic results are contained in [14] and [15]. The asymptotic properties are discussed in the next two sections.

3. Asymptotic Properties of the Principle of Minimum Relative Entropy

The major approaches to statistical inference -- classical, Bayesian, and information theoretic -- are related. A number of authors have linked classical and Bayesian inference in a variety of circumstances. The existence of prior distributions that produce classical results is well known [16]. The asymptotic normality of the posterior and its convergence to the true value of the parameter for long run observations is also known [17]. This last result assumes the values attained by the empirical distribution at each trial are known (complete informtion) and that these values are allowed to reside in the support of the prior. Section 4 presents results when these two conditions are relaxed. Specifically, the empirical distributions are known only through averages (limited information) not possessed by distributions in the support of the prior. Inference under these conditions is referred to throughout the rest of this paper as limited information Bayesian inference. These results effectively establish the relationship between the principle of minimum relative entropy and limited information Bayesian inference.

In related work, the relationship between relative entropy minimization and classical inference was recently established. [18] showed the equivalence of the minimum relative entropy result and conditional probabilities given information in the form of averages. The convergence in probability of an empirical distribution under linear contraints to the minimum relative entropy distribution is given in [19], and more general covergence is described in [20]. Finally, [21] showed a correspondence between the long run empirical distribution constrained to lie within a convex set and the distribution that minimizes relative entropy.

The relationship between information theoretic procedures and Bayesian inference has also been explored. [22-24] describe how Bayesian inference can be viewed as a special case of the principle of maximum entropy. This result is debated by [25]. Significantly, [22-24] do not base their claims on asymptotic results but assume the standard Bayesian paradigm involving the existence of complete information for the empirical distribution and consistency with the support of the prior. None of these previous researchers have investigated the important case examined in section 4.

Duality relations and computational issues of the relative entropy minimization problem under both linear equality and inequality constraints, have been investigated in [26], [27], and [28].

4. Convergence Properties of Limited Information Bayesian Inference

This section establishes the relationship between limited information, Bayesian inference, and the principle of minimum relative entropy through asymptotic arguments. Except where specifically noted otherwise, the proofs of the results given can be found in [29]. The notation in this section

follows the definitions given in section 2.

A random variable X takes on values in a discrete space $X = \{x_0, x_1, \ldots, x_m\}$. The true probability mass function on X is unknown and is modeled by the random vector $P = (P_0, P_1, \ldots, P_m)$. The initial uncertainty regarding P is modeled by a prior distribution $\mathcal{P}$.

Let $\mathcal{P}$ be an absolutely continuous probability measure with support $\Omega \subset S$. For $q = (q_0, \ldots, q_m)$,

$$S = \{q : \sum_{i=0}^{m} q_i = 1;\ q_i > 0 \text{ for } i = 0, \ldots, m\}$$

is the interior of the m dimensional standard simplex in R^{m+1}.

Suppose P is randomly selected according to $\mathcal{P}$ and fixed at the value chosen. X is then repeatedly observed in an experiment with n trials to yield a sample $X_1, X_2, \ldots, X_n$. Thus, $X_1, X_2, \ldots, X_n$ are identically distributed with probability mass function P and conditionally independent when given P. Let N_i be the random variable equal to the number of times outcome x_i occurs in the n trials for $i = 0, 1, \ldots, m$. The <u>empirical</u> distribution V^n is the random vector in R^{m+1} whose <u>ith</u> component represents the relative frequency of occurence of outcome x_i in n trials, that is, $V^n = (1/n)(N_0, N_1, \ldots, N_m)$.

Let Λ be a Borel measurable subset of S with $\Pr\{V^n \in \Lambda) > 0$ for sufficiently large n. The <u>posterior distribution</u> $\mathcal{P}^n$ of P when given the information that $V^n \in \Lambda$ is $\mathcal{P}^n(B) = \Pr\{P \in B | V^n \in \Lambda)$ for Borel sets $B \subset S$. The minimum relative entropy between Λ and $P = p$ is:

$$I(\Lambda, p) = \inf_{q \in \Lambda} I(q, p)$$

for $I(q,p)$ as defined in (2), $\Lambda \subset S$, and $p \in S$. $I(\Lambda, p) = \infty$ if $\Lambda = \phi$; $I(\Lambda, p) = 0$ if $p \in \Lambda$; and $I(\Lambda, p) > 0$ if $p \notin \Lambda$.

The first lemma shows that the empirical distribution obtainable in n

trials in Λ which minimizes the relative entropy with p converges uniformly to the minimum relative entropy distribution in Λ.

Lemma 1: Let Ω and Λ be subsets of S with Ω closed and Λ a closed body (i.e. Λ is the closure of its interior). Then $I(\Lambda^n,p) \to I(\Lambda,p)$ uniformly over $p \in \Omega$ as $n \to \infty$ where $\Lambda^n = \Lambda \cap S^n$ with $S^n = \{q \in S: nq_i \geq 0 \text{ integer for } i = 0,\ldots,m\}$.

The next result establishes bounds on the probability of observing an empirical distribution in Λ. The bounds are in terms of the relative entropy with p.

Lemma 2: For $A \subseteq S$ and $p \in S$, there exits a positive constant $\gamma(m)$, depending only on m, such that for all n

$$\exp\{-n[I(A^n,p) + (m-1)(\ln n)/(2n) - (\ln \gamma(m))/n]\} \leq$$
$$\Pr\{V^n \in \Lambda \mid P=p\} \leq \exp\{-n[I(A^n,p) - m(\ln (n+1))/n]\},$$

where $A^n = A \cap S^n$.

The proof of this result is in [30].

Lemmas 1 and 2 provided the machinery to establish uniform convergence of the probability of observing the empirical distribution in Λ given $P = p$ to an exponential function of the minimum relative entropy between Λ and p.

Lemma 3: For Ω and Λ as defined in lemma 1,

$$(1/n) \ln \Pr\{V^n \in \Lambda \mid P=p\} \to -I(\Lambda,p)$$

uniformly over $p \in \Omega$ as $n \to \infty$.

One additional result is required to establish convergence of the limited information Bayesian posterior to the minimum relative entropy distribution in Ω. The next lemma presents this result and shows that the empirical distribution in Λ which uniquely minimizes the relative entropy with p is a continuous function of p.

Lemma 4: Let Λ be a closed subset of S. Suppose

$$\nu(p) = \underset{\nu \in \Lambda}{\operatorname{argmin}}\ I(\nu,p) \tag{3}$$

is unique for all $p \in S$. Then $\nu(p)$ is a continuous function of p.

These lemmas provide the foundation for two results describing the convergence properties of minimum relative entropy. The first, Theorem 1, shows the convergence of the limited information Bayesian posterior in the support of the prior, Ω, to the minimum relative entropy distribution in the information set, Λ.

Theorem 1 (Convergence of the Limited Information Bayesian Posterior):

For Λ and Ω as defined in lemma 1, and $\nu(p)$ as given in (3) and unique, let

$$p^* = \underset{p \in \Omega}{\operatorname{argmin}}\ I(\nu(p),p)$$

be unique. Then

$$\wp^n \underset{\Rightarrow}{d} p^* \text{ as } n \to \infty$$

in the sense that

$$\wp^n(N(p^*)) \to 1 \text{ as } n \to \infty$$

for all neighborhoods $N(p^*)$ of p^*.

Hence, as the sample size grows the posterior unambiguously concentrates around the distribution which minimizes the relative entropy from Ω to Λ. As noted in section 3 when the empirical distribution is unconstrained for each trial and $\Lambda \cap \Omega \neq \phi$, the posterior distribution is asymptotically normal with mean equal to the true distribution. Theorem 1 provides an extension when the long run empirical distribution resides outside of the support of the prior. In this case, the point of concentration within the support of the prior is determined by minimum relative entropy, rather than the mean of the asymptotic normal density function. Sufficient conditions for $\nu(p)$ and p^* to be unique are that Λ be convex and Ω be strictly convex.

The second major result, Theorem 2, examines the convergence of the empirical distribution. However, before Theorem 2 can be presented the following lemma is required.

Lemma 5: If

$$\nu(p) = \underset{\nu \in \Lambda}{\operatorname{argmin}}\ I(\nu(p),p)$$

is unique for all $p \in \Omega$ then

$$\Pr\{ V^n \notin N(\nu(p)) \mid V^n \in \Lambda,\ P=p\} \to 0$$

uniformly over $p \in \Omega$ as $n \to \infty$, where $N(\nu(p))$ is an open neighborhood of $\nu(p)$.

Lemma 5 establishes convergence in probability of the empirical distribution given fixed p and Λ to the unique minimum relative entropy solution. Using this lemma it is possible to establish the convergence in distribution of the empirical outcome to the minimum relative entropy

distribution in Λ.

Theorem 2 (Convergence of the Empirical Distribution):

For P, Ω, and Λ as previously defined, let

$$V^n(B) = \Pr\{ V^n \in B \mid V^n \in \Lambda \}$$

for all Borel sets $B \subseteq S$. Also let $\nu^* = \nu(p^*)$.

Then

$$V^n \underset{\Rightarrow}{d} \nu^* \text{ as } n \to \infty$$

in the sense that $V^n(N(\nu^*)) \to 1$ as $n \to \infty$ for all open neighborhoods $N(\nu^*)$ of ν^*.

Theorems 1 and 2 provide the linkage between limited information Bayesian inference and the principle of minimum relative entropy. Theorem 1 shows the convergence of the Bayesian posterior under limited and conflicting information to the relative entropy minimum in Ω. Theorem 2 completes the relationship by showing that the distribution prescribed by the principle of minimum relative entropy is equivalent to the most likely empirical distribution observed in Λ in the long run.

5. Conclusions

The results of section 4 are important to a number of inference scenarios. The most basic is a class of problems in which population statistics are known. However, a decision maker is required to make decisions involving an atypical (i.e. not randomly selected) subpopulation of the total popluation. Frequently, the decision maker in these cases has access to limited information about the subpopulation, but not complete

information.

The major results of section 4 are directly applicable to this class of problems. The support of the prior, Ω, corresponds to the subspace of S defined by the constraints on the population statistics. Λ is defined by the limited information available to the decision maker on the subpopulation. For the results in section 4 to apply, these constraints need only define the empirical distributions in terms of a convex set in S. Typically, the constraints are inequality constraints (e.g. mean value or moment constraints).

Inferences regarding the subpopulation are available through the principle of minimum relative entropy. An updated distribution is obtained in Λ and a focus for the prior is found in Ω. Inferences using these distributions are justified based on the convergence properties demonstrated in section 4. Hence, although the principle of minimum relative entropy is used in precisely those situations where Bayesian inference is not directly applicable (because of the use of constraint information vice a likelihood function), it is interesting to know that the results obtained from both methods are in long run agreement.

REFERENCES

1. Shannon, C.E., "A mathematical theory of communication," Bell System Technical Journal, 27 (1948), 379-423 and 623-656.

2. Jaynes, E.T., "Information theory and statistical mechanics," Physical Review, 106 (1956), 620-630.

3. Tribus, M., Rational Descriptions, Decisions, and Designs, Pergamon, New York, 1969.

4. Wilson, A.G., Entropy and Urban Modeling, Pion Limited, London 1970.

5. Cozzolino, J.M. and M.J. Zahner, "The maximum-entropy distribution of the future market price of a stock," Operations Research, 21 (1973), 1200-1211.

6. Frieden, B.R., "Restoring with maximum likelihood and maximum entropy," Journal of the Optical Society of America, 62 (1972), 511-518.

7. Kullback, S. Information Theory and Statistics, John Wiley and Sons, New York, 1959.

8. Sampson, A.R. and R.L. Smith, "Assessing risks through the determination of rare event probabilities," Operations Research, 30 (1982), 839-866.

9. Sampson, A.R. and R.L. Smith, "An information theory model for the evaluation of circumstantial evidence," IEEE Trans. Systems, Man and Cybernetics, SMC-15 (1984), 9-16.

10. Brown, D.E. "A method for obtaining software reliability measures during development," IEEE Trans. Reliability, forthcoming.

11. Charnes, A., K. Haynes, S. Phillips and G. White, "Dual Extended Geometric Programming Problems and the Gravity Model," Journal of Regional Science, 17, 1977, 71-76.

12. Charnes, A., W.W. Cooper and D.B. Learner, "Constrained Information Theoretic Characteristics in Consumer Purchase Behaviour," Journal of Operational Research Society, 1978, 833-842.

13. Charnes, A., W.W. Cooper, D.B. Learner and S. Phillips, "An MDI Model and an Algorithm for Composite Hypothesis Testing in Marketing," Marketing Science, 3, 1984, 55-72.

14. Shore, J.E. and R.W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," IEEE Trans. on Information Theory, IT-26 (1980), 26-37.

15. Hobson, A. and B.K. Cheung, "A comparison of the Shannon and Kullback information measures," Journal of Statistical Physics, 7 (1973), 301-310.

16. Lindley, D.V., "The use of prior probability distributions in statistical inference and decisions," Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, 1 (1961), University of California Press, Berkeley, Ca., 453-468.

17. Von Mises, R. Mathematical Theory of Probability and Statistics, Academic Press, New York, 1964.

18. Van Campenhout, J. and T. Cover, "Maximum entropy and conditional probability," IEEE Trans. on Information Theory, IT-27 (1981), 483-489.

19. Vasicek, O.A. "A conditional law of large numbers," Annals of Probability, 8 (1980), 142-147.

20. Csisar, I., "Sanov property, generalized I-projection, and a conditional limit theorem," Annals of Probability, 12 (1984), 768-793.

21. Brown, D.E. and R.L. Smith, "A weak law of large numbers for rare events," Technical Report 86-4 (1986), Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan.

22. Jaynes, E.T., "Where do we stand on maximum entropy?" in The Maximum Entropy Formalism R.D. Levine and M. Tribus (eds.), The MIT Press, Cambridge, Ma. 1978.

23. Jaynes, E.T., "What is the question?" in Bayesian Statistics, J.M. Bernardo, et al. (eds.) University of Valencia Press, Valencia, Spain (1981).

24. Williams, P.M., "Bayesian conditionalisation and the principle of minimum information," Philosophy of Science, 31 (1980), 131-144.

25. Seidenfeld, T., "Entropy and uncertainty," Philosophy of Science, 53 (1986), 467-491.

26. Charnes, A. and W.W. Cooper, "Constrained Kullback-Liebler estimation: general Cobb-Douglas balance and unconstrained convex programming," Rediconti di Accedemia Nazionale dei Lincei, VIII, vol. LVIII (1975), 568-576.

27. Charnes, A., W.W. Cooper, and L. Seiford, "Extrema principles and optimization dualities for Khinchin-Kullback-Liebler estimation," Mathematische Operationforshung und Statistik, 9 (1978), 21-29.

28. Brockett, P.L., A. Charnes, and W.W. Cooper, "MDI estimation via unconstrained convex programming," Research Report CCS 326, 1978,Center for Cybernetic Studies, University of Texas, Austin.

29. Ben-Tal, A., D.E. Brown, and R.L. Smith, "Limited information Bayesian inference and the principle of minimum relative entropy," Technical Report 87- (1987), Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan.

30. Bahadur, R.R., Some Limit Theorems in Statistics, SIAM, Philadelphia, Pa.