Zbornik radova Filozofskog fakulteta u Nišu
Serija Matematika 6:2 (1992), 353–366

FILOMAT '92, Niš, October 8–10, 1992

THE PURE BAYES AND EMPIRICAL BAYES APPROACH

DRAGANA MILOJIĆ

ABSTRACT. *In this paper we present the kernel of the Bayesian statistics, its development and some problems which appear in the theory and practice. Also, I will to discuss the theory that complements the different topics, with particular emphasis on the implications of the theory in practical situations.*

1. Introduction.

Bayesian statistics bears name of English mathematician Tomas Bayes (1702-1761), author of the first expression in precise, quantitative form of a mode of inductive inference. He was a fellow of the Royal society and one of the first six nonconformist ministers to be publicly ordained as such in England. Educated privately, he was living in Leather Lane, near Holborn, London. Between 1720 and 1731 Bayes went to minister at the Presbyterian chapel in Tunbridge Wells, where he remained till his death on april 17, 1761, having retired from the ministry in 1752. In 1731 he publiched a tract entitled: Divine Benevolence, or an attempt to prove that the Principle End of the Drivine Providence and Government is the Happiness of His Creatures, and in 1736 another entitle An Introduction to The doctrine of Fluxions, and a Defence of the Mathematicians against the objections of the Author of the Analyst. He was elected to fellowship of the Royal society in 1742. Bayes's main works, the paper wich contains the theorem wich bears his name, and a paper on asymptotic series, were published posthumously, 1763., in the Philosophical Transactions by his friend the Rev. Richard Price. The Essay with new notation was publiched by Bernardo, 1958., he sead: "Bayes's mathematical work, though has a little comprehension, is the most quality and contains the settled mind which seems not as a pure expression neither now when two centuries are passed"

Thomas Bayes proved that, if $m : n$ is the relative frequency of on event on n indepedent occasions, then $m : n$ is also the most probable value of the event's probability, provided that any value of this probability is initially (a priori) as probable as any other value.

The Genealogical Tree of Bayesians has its roots in the eigtheenth century with the Bayesian trio Bernoulli, Bayes and Laplace. Its trunk is represented by the neo-Bayesian trio de Finetti, Savage and Morgenstern-Neumann. It branches out

1991 Mathematics subject classification: 62A15, 62C10, 62F15
Supported by Grant 0401A of FNS through Math. Inst. SANU

with the disciples of the neo-Bayesian trio, Harold Jeffreys being one of the known among them.

Bayesian statistics has undoubtedly seen considerable developement following the five papers delivered by the eminent professor Bruno de Finetti at the Henri Poincaré Institute in May 1935. These papers were subsequently published in the Annals of the Institute in 1937, under the general title: "La prevision, ses lois logiques, ses sources subjectives". No less important are the now famous publications of the late lamented professor Leonard J. Savage, "The Foundations of Statistics" (1954), and of Professors Oskar Morgenstern and John von Neumann: "Theory of Games and Economic Behavoir" (modern utility theory). We also discover in the Bayesian tree the names of authors having written at least five communications, differentiating books, articles in periodicals, communications to the learned societies, research centers and doctoral dissertations. Many of these autors show very specific interest in Bayesian statistics.

The Bayesian approch to data analysis covers not only a vast scope of subjects that we might qualify as standard ones, such as medicine, education, psychology... but is also appropriate for subjects as original as space ships, hydraulic resources, kinetics, history, nuclear energy logistics, road systems, botany and biology.

If it is relatively easy to pinpoint the roots of "The Genealogical Tree of Bayesians" it is not as easy to determine its trunk and its numerous ramifications added in the course of time.

In conclusion, the fruits of the Bayesian genealogical tree are ripe, fully developed and fit to use.

Much of the work in empirical Bayes methods over past two decades or so has been stimulated by Robbins in papers begining with the reference Robbins (1955), where the terminology 'empeircal Bayes' was introduced.

Neyman (1962) referred to the empirical Bayes aproach as *breakthrough* in the theory of statistical decision making. A measure of the importance of a theory is its impact on the practice of Statistics. The empirical Bayes approach has not revolutionized the practice of statistics, but there can be little argument that it has had a telling influence on the thinking of many statisticians, and on their practice in certain areas of application.

Empirical Bayes methods use some of the apparatus of the pure Bayes approach, but an actual prior distribution is assumed to generate the data sequence.

It may be noted that empirical Bayes ideas and techniques are applicable in many problems, especially in social and medical science.

2. The kernel of Bayesian statistics.

Bayesian statistics is based on one, simple idea: *the only satisfactory description of uncertainty is by means of probability,* (1983, Lindley). We are, all of us, surrounded by uncertainty; it plays a dominant role in all our lives. The Bayesian paradigm provides, in probability, powerful tool for understanding, manipulating and controlling this pervasive, and often unpleasant, feature of our appreciation of our environment. The practical import is immediate: any unknown quantity should be described probabilistically. A more description of the paradigm follows.

You have a quantity, or a set of quantities, θ wich is of interest to you but whose value is unknown to you. There are available some data D bearing on the uncertain value. Notice that D, unlike θ, is know. In addition you possess background knowledge pertaining to the situacion under study. Denote this by H(for history?). The Bayesian view says that the appropriate description of your knowledge of θ in the presence of D and H is by the probability of θ, given D and H; and writen $p(\theta/D, H)$.

It is not necessary to spend much time on theoretical issues but at least we ought to recognize the reason for the description of uncertainty through probability, and in no other way. The reason is that the only way different judgements of uncertainty can fit together satisfactorily is to do so in exactly the same way as probabilities. If we progress beyond *single* statements of uncertainty and consider *sets* of statements, and if these are not to contradict each other, probability results. This basic idea is called *coherence*. There are two approaches: one based on decision-making, due to Ramsey (1926) and Savage (1954), and admirably discussed by De Groot (1970); and another based on scoring rules due to de Finetti (1974). In passing, notice that coherence is not discussed in non-Bayesian statistics: a significance test at one sample size does not cohere with the same test at a different size.

It is not necessary to rehearse before this audience the rules by which probabilities cohere. The convexity (lying in the unit interval with zero for impossibility) addition and multiplication laws are the basic ones from which all other derive.

It is vitally important when considering practical applications to recognize that the Bayesian paradigm provides rules of procedure to be followed. I like to think of it as providing a *recipe*: a set of rules for attaining the final product. The recipe goes like this. What is uncertain and of interest to you? Call it θ. What do you know? Call it D, specific to the problem, and H, general. (In some situations D may be vacuous; there are no data. The recipe still applies. H is never vacuous.) Then calculate $p(\theta/D, H)$. How? Using the rules of probability, nothing more, nothing less.

We have said probability is the only tool; this is correct whilst we take a *passive* view of the world in which we are content to describe our uncertainties. This procedure is called *inference*. To a Bayesian, inference is the numerical expression of uncertainties. Extensions of probability ideas are needed when we pass beyond this passive role and consider *action* in an uncertain world. Action will lead to consequences and the worth of a consequence is described by its *utility*. However, utility is probability-based, for if a sure consecquence is replaced by an equivalent gamble on two reference consequences, one good and one bad, the utility of the sure consequence is the probability of the good consequence in the gamble. This immediately leads to choosing that action which *maximizes expected utility*, the expectation being with respect to the probabilities already assessed in the inference. Thus, following Ramsey (1926), inference is that procedure which is needed for *any* decision problem (concerning θ) and can therefore be performed without a specific decision problem in mind.

The basic rules of probability lead to other rules, of wich the most famous is

Bayes rule that gives its name to the subject. In the given notacion, it says

$$p(\theta/D, H) \propto p(D/\theta, H)p(\theta/H) \tag{1}$$

where all probabilities are functions of θ for fixed D and H. An assential feature of scientific method is the collection of data D, preferably by controlled experimentation, or alternatively by observation. Bayes rule is the essential accompaniment to this scientific activity, telling you, the scientist, how to revise your opinion of θ on observing D.

The only contribution the data makes in Bayes rule (1) is expressed trought the probability $p(D/\theta, H)$ considered as a function of θ called the *likelihood function*. We have the important *likelihood principle* that the totality of information about θ provided by D is given by the likelihood of θ for the observed D. Generally, the principle requites consideration of a unique D, that observed, but all possible values of θ. It is in violating the principle that non-Bayesian mathods typically come to grief. In considering data values that might have occurred but did not, as with a tail-area, significance test, they become incoherent.

A second, useful rule of probability that easily follows from basic rules has no generally accepted name and here we call it the *extension rule*. It says

$$p(D/\theta, H) = \sum_{\phi} p(D/\theta, \phi, H)p(\phi/\theta, H)$$

and extends "the conversation" from θ to include ϕ. Its usefulness lies in the fact that the jugdements about D often involve not merely the quantity of interest θ but also *nuisance* quantities ϕ. The rule allows these to be eliminated by summation (or integration). It is one of the more difficult problems of the Classical school to eliminate nuisance parameters; the Bayesian view has the single, universal method of integration. Alternative methods may lead to incoherence. Methods based solely on likelihood (Edwards, 1972), are defective in that they have no general way of eliminating nuisance parameters.

Before discussion has been confined to the *calculus* of probability: in applications it is also necessary to consider its *interpretation*. There need only be one: a person's judgement about a quantity that is unknown to him. This is the *subjective* (or personalistic) view of probability. In this view probability does not exist outside the subject; there is no true probability but rather an expression of a relationship between you and the world. The word "subjective" is unfortunate because the Bayesian view is not subjective in the sence that data analysis is, where all manipulations are open to consideration and judgements are much at the whim of the analyst. The Bayesian subject is severely constrained by coherence and by the inexorable role of data. As de Finetti (1974) has said, the only objective view of probability is the subjective one, because it can be tested by the rule it must obey.

Subjective probability, in its interpretation, contains no element of repetition; it has no frequency basis. In Classical school, statisticians are restricting themselves to exchangeable (or partially exchangeable) cases, where the notation of *chance* arises.

This is a severe and unnecessary restriction, for the paradigm equally applies to the unique occasion.

In describing the theory it has been emphasized that *only* probability is required. The practical relevance of this remark is apparent when Bayesian and classical view are contrasted. The Bayesian only requires $p(\theta/D)$. The probability distribution is a *complete* statement and point or interval estimates, or significance tests, are not necessary. Estimates may by calculated, for example, $E(\theta/D)$ is a possible point estimate, but they are derived from the full specification and information is necessarily lost in using them. They may help in appreciating the distribution. Note that one of the most challanging of the many, difficult, technical problems in Bayesian statistics is to find ways of appreciating distributions when θ has high dimensionality.

Until now, I have writen as though the Bayesian paradigm was a unique, well-defined and agreed position. This is properly not so: there are important, undecided issues to consider. All agree on the probabilistic description, but not on the exact rules of probability, nor on their interpretacion. Could the improper distributions, for instance, ones that are not finitely integrable, be used? The paradoxes particulary associated with Stone (1976) ilustrate the pitfalls.

Attracted by the apparent objectivity of Classical statistics, many workers have sought for an objective "prior" $p(\theta/H)$, especially when H is almost vacuous so that there is little information about θ. The originator is Jeffreys (1961). The idea is that if we have a random sample $x^{(n)}$ from a distribution with density $p(x_i/\theta)$ then there is natural probability for θ associated with it. An alternative view (Bernardo, 1979), is to regard this distribution for θ as a reference distribution from which the information in other distributions can be measured. Either way, we obtain a useful probabilistic description for θ wich can be combined with the agreed likelihood and an "objective" analysis of the data is obtained. Lindley's view is opposed to this and he sees the argument as an example of technique overriding standpoint in which Greek letter takes precedence over its meaning. But the point is by no means settled and has some importance in the analisis of data.

Another point of discussion amongst Bayesians is whether the "only" aspect of paradigm is corect: do we need techniques that go outside the strict, probabilistic argument used above? Dempster (1980), Good (1967) and Box (1980) have argued that tail-area, significance test of Classical school have to play a role in inference. The point arises because there are situations in which a "natural" hypothesis occurs, but alternatives are not easy to specify. (It is easy to say that two distributions are the same; more difficult to say how they might differ.) The difficulty with the approach is that *whatever* alternatives were to be considered, only the data actually observed are needed. It is closely related to the need for models, more in the Bayesian than in Classical techniques.

The whole nature of model is obscure. The Bayesian sees a model as *s* helpful way of specifyng the probabilities that are essential to his method for studying uncertainty. But part of statistician's job does not involve uncertainty. Statisticians spend some of their time studying data. The book by Anscombe (1981) provides many illustrations where the question of Bayes or Classical does not arise because

appreciation of data is all that is initially required.

One feature possessed by model is best explained in the context of Bayes theorem (1). It is usually written as

$$p(\theta/D, H) \propto p(D/\theta)p(\theta/H)$$

in wich, given θ, D and H are assumed independent. In other words, the model separates D from H, with θ acting as the intermediary; or θ gives a Markov structure to the sequence (H, θ, D). With this admitted, Bayesians might agree on the model specification even if their knowledges differ.

The most important practical problem in the inplementation of the Bayesian paradigm is determination of the numerical values for the probabilities. Classical statisticians say, correctly, that we are unable to do this and then, erroneously, infer the Bayesian method is useless.

Ramsey's discovery that the laws of probability govern uncertainty parallels Newton's discovery of laws of motion; laws that required for their exploitation methods of measuring speeds, forces, ect. Ramsey's laws require us to measure probabilities. Physicists built apparatus to measure the quantities required for Newtonian mechanics, they did not sit back and say force cannot be measured, so Newtonian machanics is useless. So ways have to be found to measure probabilities. We cannot expect the "prior" to come naturally any more then arithmetic does: we have to learn probability assessement. Even those who favour reference priors have to admit the problem because their methods fail without echangeability.

3. Bayes and empirical Bayes methods.

In the pure Bayesian approach to the decision problem the parameter value itself is regarded as a realization of a random variable Λ with distribution function $G(\lambda)$. The distribution of Λ is called the prior distribution. The probabilities defined by $G(\lambda)$ are not necessarily interpretable in terms of relative frequencies.

As above, a fundamental problem in the pure Bayes approach is the specification of G a prior function satisfying the relationship

$$H(x) = \int F(x/\lambda)\, dG(\lambda)$$

where $H(x)$ and $F(x/\lambda)$ are given d.f.s. The Bayes solution of decision problem generally depends on G, and the Bayes decision function is denoted by $\delta_G(x)$ to show this dependence.

In the empirical Bayes approach the existence of a prior distribution is postulated, but it is taken to be susceptible to a frequency interpretation. Further, the availability of previous data, suitable for estimation of the prior distribution G is assumed. The mathematical derivations associated with the Bayes method are used to obtein a decision function $\delta_G(x)$, generally dependent on G, but then $\delta_G(x)$ is replaced by an estimate based on the previous data. Such an estimated $\delta_G(x)$ is called an empirical decision rule.

Much of the work in empirical Bayes (EB) methods over past two decades or so has been stimulated by Robbins in papers beginning with the reference Robbins

(1955) where the terminology 'empirical Bayes' was introduced. However, it has become clear that the applicability of EB ideas is much wider than might have been suggested by the earlier writings. It may be noted especially that EB ideas are applicable in many problems involving mixtures of distributions.

3.1 An Intoduction to Bayes techniques.

Since the EB approach uses the techniques and results of the Bayes approach some of the standard results are reviewed in this section. Applications of the EB methods are envisaged as occurring in repetitive experimentation with parameters varying from experiment to experiment. The notion of *expected loss* seems rather natural in this context, hence the Bayes method is based on the notion of a loss function.

Let a *loss*, $L(\delta(x), \lambda) \geq 0$, be incurred when the parameter value is λ and a decision $\delta(x)$ is made. For example, if $\delta(x)$ is point estimate of λ it is common to put $L(\delta(x), \lambda) = (\delta(x) - \lambda)^2$. Or, if $\delta(x)$ is an interval estimate one may put $L(\delta(x), \lambda) = 0$ or 1 according as the interval does or does not contain λ. The expected loss for fixed λ is the *risk*

$$R_\delta(\lambda) = \int L\{\delta(x), \lambda)\} f(x/\lambda) dx \tag{3.1.1}$$

where $f(x/\lambda)$ is the probability density function (p.d.f.) of X. Modification of (3.1.1) for discrete X is obvious. The selection of a decision function now becomes a matter of choosing $a\delta(x)$ whose $R_\delta(\lambda)$ has acceptable properties. Clearly, the smaller $R_\delta(\lambda)$ for any λ, the better, but it is trivially true that there is generally no $\delta^*(x)$ such that $R_{\delta^*}(\lambda) \leq R_\delta(\lambda)$ for all λ and every δ. Thus there is no uniformly best δ, and an additional criterion for selecting a δ has to be invoked. One of these is provided in the Bayes approach, in which the goodness of $a\delta$ is judged by the overall expected loss, or the average risk, with respect to the prior distribution $G(\lambda)$. It is given by

$$W(\delta) = \iint L\{\delta(x), \lambda\} f(x/\lambda)\, dx\, dG(\lambda). \tag{3.1.2}$$

Now δ is chosen to minimize W. The satisfaction of this condition will depend on G, and δ will be denoted by δ_G to indicate the dependence. We shall call $W(\delta_G)$ the *Bayes risk*. If we put in (3.1.2)

$$W = E(L) = EE(L/x)$$

we shall get

$$E(L/x) = \int L(\delta, \lambda) f(x/\lambda)\, dG(\lambda) \;/\; \int f(x/\lambda)\, dG(\lambda), \tag{3.1.3}$$

so that we choose δ to minimize $E(L/x)$ for every x.

3.2 Bayes point estimation: one parameter.

Let $\delta(x)$ be any point estimate of λ. If δ_G is the Bayes point estimate, we have

$$W(\delta) \geq W(\delta_G) \tag{3.2.1}$$

by definition. Now, with $L(\delta, \lambda) = (\delta - \lambda)^2$,

$$\begin{aligned} W(\delta) &= \iint L\{\delta(x), \lambda\} f(x/\lambda)\, dx\, dG(\lambda) \\ &= W(\delta_G) + \iint \{\delta(x) - \delta_G(x)\}^2 f(x/\lambda)\, dx\, dG(\lambda) \\ &\quad + 2 \iint \{\delta(x) - \delta_G(x)\}\{\delta_G(x) - \lambda\} f(x/\lambda)\, dx\, dG(\lambda) \end{aligned} \tag{3.3.2}$$

Condition (3.2.2) will be satisfied if the third term in (3.2.2) is zero, which can be arranged by putting

$$\int \{\delta_G(x) - \lambda\} f(x/\lambda)\, dG(\lambda) = 0$$

for every x. This gives

$$\delta_G(x) = \frac{\int \lambda f(x/\lambda) dG(\lambda)}{f(x/\lambda) dG(\lambda)} \tag{3.2.3}$$

Thus $\delta_G(x)$ is the mean of the posterior distribution of Λ for given $X = x$In the denominator of the right-hand side of (3.2.3) we have the marginal p.d.f. of X,

$$f_G(x) = \int f(x/\lambda) dG(\lambda). \tag{3.2.4}$$

— The correspoding marginal distribution function is $F_G(x)$ and sometimes it will be convenient to refer to the marginal random variable whose cumulative distribution function is $F_G(x)$ as X_G.

— In the joint distribution of X_G and Λ, $\delta_G(x)$ is the regresion of Λ on X_G.

— The marginal distribution of X_G is also called a compound or a *mixed* distribution.

3.3 Bayes decision between k simple hypotheses.

We consider first the case of two simple hypotheses H_1: $\lambda = \lambda_1$ and H_2: $\lambda = \lambda_2$, $\lambda_1 < \lambda_2$. The prior probabilities are $P(\Lambda = \lambda_j) = \theta_j, j = 1, 2$ with $\theta_1 + \theta_2 = 1$. Thus $G(\lambda)$ is a step function whith jumps at λ_1 and λ_2 of size θ_1 and θ_2 respectively. The decision function $\delta(x)$ is defined in terms of a partition of sample space into two regions A_1 and A_2 such that H_j is accepted when $x \in A_j, j = 1, 2$. In this context the loss function, $L\{\delta(x), \lambda\}$, will be defined as $L = 0$ when the correct decision is made, and $L = 1$ otherwise. Then

$$W(\delta) = \theta_1 \int_{A_2} f(x/\lambda) dx + \theta_2 \int_{A_1} f(x/\lambda) dx,$$

with the obvious modifications for discrete X. We see that $W(\delta)$ is the overall expected proportion of wrong decisions.

To minimize $W(\delta)$ we define A_1 and A_2 such that $x \in A_1$ when $\theta_2 f(x/\lambda_2) < \theta_1 f(x/\lambda_1)$. Hence the Bayes decision rule can be stated as follows: choose H_1 if the posterior probability of H_1 exceeds $1/2$. Equivalently, we choose whichever H_1 and H_2 has the greater posterior probability.

When $f(x/\lambda)$ is such that $f(x/\lambda_1)/f(x/\lambda_2)$ is monotonic in x, it follows immediately that A_1comprises all values of $x < \xi_G$, where $x = \xi_G$ is the solution of

$$\theta_2 f(x/\lambda_2) = \theta_1 f(x/\lambda_1), \tag{3.3.1}$$

with suitable modification when X is discrete.

3.4 Bayes estimation of vector parameters.

Much of what has gone before can be generalized quite easily when X is replased by the vector r.v. $X = (X_1 ... X_p)^T$ and Λ by $\Lambda = (\Lambda_1, \Lambda_2, ..., \Lambda_k)^T$. Standard examples are the multivariate normal and multinominal distributions. Let $L\{\delta(x), \lambda\}$ be the loss in making decision $\delta(x)$ about λ. The Bayes decision rule δ_G can be obtained by using (3.1.3) where x will be replaced by x, ect. Of course, G is now a k -variate distribution.

For some problems of point estimation a natural generalization of squared error loss is

$$L(\delta, \lambda) = (\delta - \lambda)^T A(\delta - \lambda)$$

where A is positive definite matrix. The Bayes point estimate $\delta_G(x)$ can be obtained from the vector version of (3.2.2),

$$\begin{aligned} W(\delta) = W(\delta_G) &\\ &+ \iint \{\delta(x) - \delta_G(x)\}^T A\{\delta(x) - \delta_G(x)\} f(x/\lambda) dG(\lambda) dx \\ &+ 2 \iint \{\delta(x) - \delta_G(x)\}^T A\{\delta_G(x) - \lambda\} f(x/\lambda) dG(\lambda) dx. \end{aligned}$$

The third term in the rigth-hand side of the above expression can be made equal to zero, thus ensuring $W(\delta) > W(\delta_G)$, by letting the ith element $\delta_{G_i}(x)$ of $\delta_G(x)$ be

$$\delta_{G_i}(x) = \frac{\int \lambda_i f(x/\lambda) dG(\lambda)}{f(x/\lambda) dG(\lambda)}$$

And that is the posterior mean of Λ_i. Remarkably this Bayes point estimate is not dependent on A. Of course, the value of $W(\lambda)$ for any δ, including δ_G, will depend on A. If a single function $w(\lambda)$ of the parameters is to be estimed subject to quadratic loss, its Bayes point estimate is readily seen to be the posterior mean of $w(\Lambda)$.

3.5 Bayes decision and multiple independent observations.

Concentrating for now on the univariate single parameter case, suppose that m independent observations are made on X. Then $f(x/\lambda)$ in (3.1.3) is replaced by the likelihood $\Pi_{i=1}^{m} f(x_i/\lambda)$, the method otherwise remaining unchanged. If a one-dimensional sufficient statistic $t(x_1, ..., x_m)$ exists we have $\Pi_{i=1}^{m} f(x_i/\lambda) = g(x_1, ..., x_m)h(t/\lambda)$. Therefore by substitution in (3.1.3), the problem is essentially reduced to the one-sample case.

3.6 Empirical Bayes methods.

Empirical Bayes methods rely on the existence of a prior distribution $G(\lambda)$ which can be given a frequency interpretation, and which can be estimated using suitable observations. Thus the EB approach can be essentially non-Bayesian in the sense of not involving subjective probabilities. In the simplest case the EB sampling scheme is as follows: a current observation x is made when the parameter value is λ, *a* realization of Λ, and x is to be used in a decision about λ. At the time of making the current observation there are available past observations $x_1, ..., x_n$ obtained with independent past realizations $\lambda_1, ..., \lambda_n$ of Λ. In this scheme every x_i is a realization of X_i, and the $X_i's$ are mutually independent. It is useful to represent the EB sampling scheme as in (3.6.1)

EB sampling scheme

(3.6.1)		Previous stages	Curent stage
	unknowns	$\lambda_1, \ldots \lambda_n$	λ
	observables	$x_1, \ldots x_n$	x

The words 'current' and 'past' are not necessarily to be taken in a strictly temporal sense. Usually it is assumed that the actual values $\lambda_1, \lambda_2, ..., \lambda_n$ never become known.

The possibility of obtaining an estimate of G arises through the fact that $x_1, ..., x_n$ may be regarded as an independent sequence of observations on X_G whose distribution function F_G is given in (3.2.4) with f replaced by F. The empirical c.d.f. of these x-values, $F_n(x)$, is an estimate of $F_G(x)$ such that $F_n(x) -> F_G(x)$ in probability (P), as $n -> \infty$ for every x. This suggests that it might be possible to find a c.d.f. $G(\lambda)$ such that

$$F_n(x) \cong \int F(x/\lambda) d\hat{G}(\lambda)$$

with the property that $\hat{G}(\lambda) -> G(\lambda), (P)$, for all λ as $n -> \infty$.

If such an empirical $G(\lambda)$ can be found, substituing it for $G(\lambda)$ in the derivation of a Bayes decision rule will yield an *empirical Bayes* rule, $\delta_n(x_1, ..., x_n; x)$. This notation emphasizes that the EB rule will generally depend on all past x-values as well as the current x. We may regard it in a broad sense as an estimate of the Bayes rule. In the case of point estimation, $\delta_n(x_1, ..., x_n; x)$ is the point estimate of $\delta_G(x)$. EB rules need not necessarily be obtained by directly exploiting the relation $F_G(x) = \int F(x/\lambda) dG(\lambda)$ to obtain an estimate of G.

To conclude this introduction to EB methods where the term 'empirical Bayes' is implied by the tipical EB sampling scheme, we can say at somewhat broader interpretation, that any decision rule about the parameter θ of a prior distribution derived from observed data, we may regard as an empirical Bayes rule.

3.7 The goodness of EB procedures.

The Bayes decision rule, $\delta_G(x)$, is defined as that $\delta(x)$ which minimizes $W(\delta)$ so that $W(\delta_G) \leq W(\delta)$ for all δ. Now, for any δ the value of $W(\delta)$ is a measure of its goodness and in the Bayes sense δ_G is the best, or optimal. If δ_n is an EB rule derived from a particular set of past observations, $W(\delta_n)$ is a measure of its goodness. With respect to the past observations $W(\delta_n)$ is, of course, a random variable. Therefore an assessment of the overall goodness of an EB mathod should pay attention to the distribution of $W(\delta_n)$.

A natural measure of performance of an EB method, in the light of the preceding discussion, is the expecation of $W(\delta_n)$ with respect to previous samples of size n, i.e. $E_nW(\delta_n)$. We could then say that δ_n is asymptotically optimal (a.o.) if $E_nW(\delta_n) - > W(\delta_G)$ as $n - > \infty$(Robbins,1964). Even if δ_n is a.o., $E_nW(\delta_n)$ may be considerably greater than $W(\delta_G)$ for finite n values, and it may be greater than $W(T)$ where T is a non-Bayes rule. For example, in point estimation T may be a maximum likelihood estimator, and δ_n would not necessarily be preferred to it unless $E_nW(\delta_n) < W(T)$. Usually there will be a value of n, say n_T, such that $W(T) < E_nW(\delta_n)$ for $n < n_T$.

In choosing whether to use an EB method in preference to a non-Bayes method, criteria other than $E_nW(\delta_n)$ could be used. For example, if one was concerned that the realized δ_n shoul be better than T one may focuses attention on $P_n\{W(\delta_n) < W(T)\}$, where P_n indicates a probability calculated with reference to previous samples of size n. Another definition of asymptotic optimality is a.o.(P):$W(\delta_n) - > W(\delta_G), (P)$, as $n - > \infty$. The property $E_nW(\delta_n) - > W(\delta_G)$ as $n - > \infty$ can be called a.o.(E). With some restrictions a.o.(P) will imply a.o.(E). Also a.o.(P) will imply $P_n\{W(\delta_n) < W(T)\} > 1 - \epsilon$ for n large enough. In practice the choice bitween δ_n and T has to be made on the basis of known results for $E_nW(\delta_n)$ or $P_n\{W(\delta_n) < W(T)\}$ for cases similar to the problem in hand, or by trying to estimate these quantities.

3.8 Approximate Bayes and empirical Bayes methods.

Any point estimator derived using the prior distribution to find a best estimator within a certain class can be called an approximate Bayes estimator if it is not the actuale Bayes estimator. We consider *linear Bayes estimators* (Hartigan, 1969, Griffin and Krutckoff,1971).

The simplest case is when there is just one observation x on X when the parameter value is λ. We consider estimates of the form

$$\delta(\omega_0, \omega_1; x) = \delta(\omega; x) = \omega_0 + \omega_1 x$$

where ω_0 and ω_1 are chosen to minimize $W\{\delta(\omega; x)\}$. The terminology'linear Bayes' is explained by the form of $\delta(\omega; x)$ and the fact that $G(\lambda)$ plays a role in the

determination of $\omega_0 i\omega_1$. Now

$$W\{\delta(\omega;x)\} = \iint (\omega_0+\omega_1 x-\lambda)^2 f(x/\lambda)dG(\lambda) \tag{3.8.1}$$

and it is easily minimized by differentiation w.r.t. ω_0 and ω_1. The Bayes values ω_G will be obtained as the solutions of the following equations in ω_0, ω_1:

$$\begin{vmatrix} 1 & \int E(x/\lambda)dG(\lambda) \\ \int E(x/\lambda)dG(\lambda) & \int E(x^2/\lambda)dG(\lambda) \end{vmatrix} \cdot \begin{vmatrix} \omega_0 \\ \omega_1 \end{vmatrix} = \begin{vmatrix} \int \lambda dG(\lambda) \\ \int \lambda E(x/\lambda)dG(\lambda) \end{vmatrix} \tag{3.8.2}$$

Approximations to the prior distribution. Suppose that G^* is an approximation to G. Then δ_{G^*} is an approximation to δ_G. For the purpose of empirical Bayes inference the sense in which G^* might be an approximation to G is that G^* is that member of certain class of distributions for which the distance $D\{F_G, F_{G^*}\}$ is minimized. The distance measure D is yet to be chosen and may depend on specific applications. This definition of approximation is motivated by the fact that F_G is observable in the EB context thus making the determination of G^* feasible, at least in the sense that it can be estimated statistically.

The classes of distributions that may be considered for G^* include

1. the natural conjugate priors;
2. finite step functions.

Using non-sufficient statistics. Suppose that m independent observations are made on the one-dimensional r.v. X whose distribution depends on the single parameter λ. If a one-dimensional sufficient statistic $t(x_1 \ldots x_m)$ exists, the Bayes decision rule reduces to a function $\delta_G(t)$ of t. When a sufficient t does not exist calculations involving likelihoods $\Pi_{i=1}^m f(x_i/\lambda)$ can become complicated, especially in the EB framework, and it may be contemplated to effect a reduction by basing the decision on an estimate $\hat{x}$ of λ.

To obtain a decision rule based on $\hat{x}$ the p.d.f. $f(x/\lambda)$ is replaced by the p.d.f. $h(\hat{x}/\lambda)$ of $\hat{x}$ in the formulae for obtaining Bayes estimates. Even if the exact distribution of $\hat{x}$ is used the resulting decision rule is not the actual Bayes rule; sometimes it may be possible only to obtain an approximation for $h(\hat{x}/\lambda)$. The advantege of this approach, especially in EB decision, is that estimation of an approximation to G can be considerably simplified.

Conclusion.

The Bayes paradigm concerns uncertainty. Its only tool is coherence expressed through the three laws of probability. It will be applied to statistical, repetitive situations where a judgement of exchangeability is possible. But it is also applied to unique situations. We are uncertain about the inflation tare next year, the world's oil reserves, or the possibility of the war. All these can be handled by subjective probability. What marvellous practical possibilities this suggests.

References

[1] Aitchison, J. A. and Dunsmore, I. R., *Statistical Prediction Analysis*, Cambridge University Press, 1975.

[2] Akaike, H., *A new look at the Bayes procedure*, Biometrika 65 (1978), 53–59..

[3] Anscombe, F. J., *Computing in statistical Science Throught APL*, Springer-Verlag, New-York, 1981.

[4] Bayes, T., *An essay towards solving a problem in doctrine of chances*, Philosophical Transactions of the Royal Society of London [Reprinted in Barnard (1958) Biometrika, 293-315] 53 (1764), 370–315.

[5] Bennet, G. K. and Martz, *A continuous empirical Bayes Smoothing technique*, Biometrika 59 (1972), 361–9.

[6] Berger, J. O., *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York, 1986.

[7] Bernardo, J. M., *Non-informative prior distributions a subjectivist approach*, Bull. Internat. Stat. Inst. 46 (1975), 94–97.

[8] Bernardo, J. M., *Reference posterior distributions for Bayesian inference (with Discussion)*, J. R. Statist. Soc. 41 (1979), 113–147.

[9] Box, G. E. P. and Tiao G. C., *Multiparameter problems from a Bayesian point view*, Ann. Math. Statist. 36 (1965), 1468-1482.

[10] Box, G. E. P. and Tiao, G. C., *Bayesian Inference in Statistical Analysis*, Reasing, Massachusetts: Addison-Wesley, 1973.

[11] Box, G. E. P., *Sampling and Bayes inference in scientific modelling and robustness*, J. R. Statist. Soc. **A**, 143 (1980), 384-404.

[12] Confield, J., *The Bayesian outlook and its application*, Biometrics 25, 617-637 (1969), 617-637.

[13] Copas, J. B., *Compound decisions and empirical Bayes*, J. Roy. Statist. Soc. **B**, 31 (1969), 397-425.

[14] Cressie, N., *A quick and easy empirical Bayes estimate of true scores*, Sankhya **B**,41 (1979), 101-108.

[15] De Finetti,B., *Theory of probability, Vol. I.*, London, Wiley, 1974.

[16] De Finetti, B., *Bayesianism: its unifying role for both the foundations and applications of statistics*, Inter. Statist. Rev. 42 (1974), 117-30.

[17] De Groot, M. H., *Optimal Statistical Decisions*, McGraw-Hill, New-York, 1970.

[18] Deely, J. J. and Lindley, D. V., *Bayes empirical Bayes*, J. Amer. Statist. Assoc. 76 (1981), 833-841.

[19] Edwards, A. W. F., *Likelihood*, Cambridge University Press, 1972.

[20] Gruffin, B. S. and Krutchkoff, R. G., *Optimal Linear estimators: an empirical Bayes version with application to the Binomal distribution*, Biometrika 58 (1971), 195-201.

[21] Hartigan. J. A., *Linear Bayes methods*, J. Roj. Statist. Soc. **B**,31 (1069), 446-56.

[22] Holland, J. D., *The reverend Thomas Bayes, F. R. S.*, J. R. Statist. Soc. **A**, 125 (1962), 451-61.

[23] Jeffreys, H., *Theory of Probability, 3rd edn.*, Clarendon Press, Oxford, 1961.

[24] Lindley, D. V., *The use of prior probability distributions in statistical inference and decision*, Proc. Fourth. Berkeley Symposium on Math. Statist. and prob. 1 (1962), 453-68.

[25] Lindley, D. V., *Introduction to probability and statistics from a Bayesian viewpoint, Part2, introduction*, Cambridge University Press, 1965.

[26] Lindley, D. V., *Bayesian least squaresInst.*, Inst. Internat. Statist. 43(2) (1969), 152-53.

[27] Lindley, D. V. and Smith, A. F. M., *Bayes estimates for the linear model (with Discussion)*, J. Roy. Statist. Soc. **B**, 34 (1972), 1-42.

[28] Maritz, J. S., *Distribution free statistical methods Chapman and Hall, London*, Chapman and Hall, London, 1981.

[29] Morris, C. N., *Parametric empirical Bayes inference: theory and applications*, J. Amer. Statist. Assoc. 78 (1983), 47-59.

[30] Nichols, W. G. and Tsokos, C. P., *Empirical Bayes point estimation in a family of probability distributions*, Inter. Statist. Rev. 40 (1972), 147-51.
[31] Ramsey, F. P., *Truth and probability [Reprinted in Studies in Subjective Probability (1964)]*, (eds. H. E. Kyburg, Jr and H. E. Smokler) Wiley, New York, 1926.
[32] Rao, C. R., *Simultaneous estimation of parameters in different linear models and applications to biometrics problems*, Biometrics 31, 545-554 **31** (1975), 545-554.
[33] Robbins, H., *An empirical Bayes approach to statistics*, Proc. Third Berkeley Symposium on Math. Statist and Prob. **1** (1955), 157-64.
[34] Robbins, H., *The empirical Bayes approach to statistical problems*, Ann. Math. Statist. **35** (1964), 1-20.
[35] Rolph, J. E., *Bayes estimation mixing distributions*, Ann.Math.statist. **39** (1968), 1289-1302.
[36] Rutherford, J. R. and Krutchkoff, R. G., *ε-asymptotic optimality of empirical Bayes estimators*, Biometrica **56** (1969), 220-23.
[37] Samuel, E., *An empirical Bayes approach to the testing of certain parametric hypotheses*, Ann. Math. Statist. **34** (1963), 1370-1385.
[38] Savage, L. J., *The foundations of statistics*, Wiley, New York, 1954.
[39] Stein, C., *Inadmissibility of the usual estimators for the mean of a multivariate normal distribution*, Proc. Third Berkeley Symp. Math. Stat. Prob. **1** (1955), 197-206.
[40] Stone, M., *Strong inconsistency from uniform priors (with Discussion)*, J.Amer.Statist.Ass. **71** (1976), 119-25.
[41] Zellner, A., *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York, 1971.

Institute of Economic Sciences
Zmaj Jovina 12, Beograd
Yugoslavia