REVISTA MATEMATICA de la
Universidad Complutense de Madrid
Volumen 3, números 2 y 3; 1990.
http://dx.doi.org/10.5209/rev_REMA.1990.v3.n2.18021

Three Methods for Constructing Reference Prior Distributions

EUSEBIO GÓMEZ SÁNCHEZ-MANZANO and MIGUEL A. GÓMEZ-VILLEGAS

ABSTRACT. Three methods are proposed for constructing reference prior densities for certain biparametric distribution families. These densities represent approximations to the Bayesian concept of noninformative distribution.

1. INTRODUCTION

Attempts to obtain noninformative prior distributions have been made by Jeffreys (1961), Hartigan (1964), Novick and Hall (1965), Jaynes (1968, 1983), Box and Tiao (1973), Villegas (1977, 1981), Zellner (1977), Akaike (1978) and Bernardo (1979). Basically, the idea is to obtain probability distributions that allow data to speak for themselves.

We propose three methods for constructing reference densities for certain biparametric families of prior distributions. These densities represent approximations to the Bayesian concept of noninformative distribution.

The first of these methods just uses the concept of prior distribution, whilst the remaining two in addition use the likelihood function.

Sections 2, 3 and 4 describe the methods. Each section consists of an introductory example, a definition and its formal application to the example.

Section 5 describes some applications, and shows the differences between our methods and those of Jeffreys, which historically is the first, and Bernardo, the most widely applicable. Section 6 offers possible generalizations.

2. METHOD 1

Let us introduce this method by considering the following example. Suppose that someone is going to observe a random variable whose

1980 Mathematics Subject Classification (1985 revision): 62C10.

probability distribution depends on a fixed parameter θ for which there is only partial information. The person in question decides to represent his information regarding θ by means of a prior subjective probability distribution π. In the design of his prior he has help from a statistician. Suppose that the statistician, after asking him certain questions, concludes that the prior π must be a normal distribution and that he now just needs to determine its two parameters, μ and τ^2.

In order to find the mean, μ, the statistician will possibly ask his client to indicate a number μ_0 which he believes could be close to the parameter θ, and μ will then be given this value. In order to determine the variance τ^2, the statistician will ask his client to define his degree of trust in the conjecture. This, in other words, is equivalent to asking about its precision or «informative strength». The statistician will determine the prior variance by using the general criterion that the smaller the precision or informative strength of the guess the greater has to be the dispersion of the subjective variable θ about μ_0. (The way to determine the variance can at first be difficult to perfect, but this is a point that does not interest us at this point).

If we feel the statistician's approach is reasonable, we are tacitly accepting that between two normal distributions with the same mean, the one with the greater variance is less informative. In this respect, the noninformative distribution amongst the normal distributions with a given mean, μ_0, will be that with maximum variance. This maximum variance distribution does not exist in the sub-family of normal distributions with mean μ_0, but it does exist on the «edge» of the subfamily: it is the improper uniform distribution along the whole of the real line. However, it is obvious that the same distribution is reached no matter the value of μ_0 assigned to the mean. Hence, the noninformative distribution amongst the normal distributions is the improper uniform distribution in $\mathbb{R}$.

The generalization and formalization of this question gives rise to the following definition:

Defitinion 1

Let $\mathcal{F}=[\pi_{a,s}(\theta)]$ *be a family of prior densities, the parameter* a *being a measure of centralization which varies in a subset A of* $\mathbb{R}$, *and the parameter* s *being a measure of dispersion which, for each fixed value of* a, *varies within an interval* $S(a)=(s_o(a), s_1(a))$, *where* $s_1(a)$ *can be infinite. Let us suppose that, for all the values of* a, *the (improper) limiting densities corresponding to* a *and* $s_1(a)$ *are all equal (up to a multiplicative constant that could depend on* a). *We will refer to this common limiting density as the* reference density I *of the family* $\mathcal{F}$ *with respect to the parametrization we have used.* ■

We will consider that Definition 1 can also be applied to that case in which the parameter s it not a dispersion measurement but a measurement of concentration or precision. The reference density would then correspond to the value $s_o(a)$ instead of $s_1(a)$.

In order to formalize the concept of limit that appears in Definition 1, we will accept that any density $\pi_{a,s}$ of the family $\mathscr{F}$ is continuous and positive at a and we define

$$\rho_{a,s}(\theta)=\frac{\pi_{a,s}(\theta)}{\pi_{a,s}(a)}.$$

The family $\mathscr{G}=\{\rho_{a,s}(\theta)\}$ behaves in Bayesian analysis just like the $\mathscr{F}$ family. If for each value of a there exists the limit function

$$\rho^*(\theta)=\lim_{s\to s_1(a)} \rho_{a,s}(\theta)$$

and $\rho^*(\theta)$ is independent of a (up to a multiplicative constant), then the reference density I exists and is any (improper) density independent of a and proportional to $\rho^*(\theta)$.

Definition 1 can now be formally applied to our first example. For this, let us consider the family $\mathscr{G}$ of normal densities with mean a and standard deviation s. We have $a\in A=\mathbb{R}$ and $s\in S(a)=(0,+\infty)$. We calculate $\rho_{a,s}(\theta)$, for any $\pi_{a,s}(\theta)\in\mathscr{F}$, and then $\rho^*(\theta)$:

$$\rho_{a,s}(\theta)=exp\left\{-\frac{1}{2}\left[\frac{\theta-a}{s}\right]^2\right\},$$

$$\rho^*(\theta)=\lim_{s\to+\infty} \rho_{a,s}(\theta)=1.$$

Since this limit is independent of a, we can take the improper density $\pi(\theta)=1$ as the reference density I.

3. METHOD 2

As in the previous section, we introduce the method by analyzing a particular case. Let us suppose that we have a prior distribution $N(\mu_o, \tau_o)$ and we observe a sample of size n from a random variable $N(\theta,\sigma^2)$, the variance σ^2 being known. It is known (see DeGroot (1970), p. 167) that the mean value, μ_1, of the posterior distribution is

$$\mu_1=\frac{\sigma^2}{\sigma^2+n\tau_o^2}\mu_o+\frac{n\tau_o^2}{\sigma^2+n\tau_o^2}\bar{x}, \tag{1}$$

where $\bar{x}$ represents the sample mean.

If we have accepted the procedure of the preceding section as being reasonable, we have accepted that the prior mean μ_o represents the best initial conjecture regarding the parameter θ. However, the present posterior distribution can serve as prior distribution for future experiments. Hence, the posterior mean μ_1 represents the best present conjecture regarding θ. Expression (1) shows this present conjecture as a function of the prior conjecture μ_o and of the sample data (as well as τ_o^2).

If, in (1), we make $\tau_o^2 = +\infty$, we have $\mu_1 = \bar{x}$, independently of μ_o. In this case, the final conjecture will be the same, no matter the value ascribed to the initial one. The initial conjecture has not been taken into consideration. By making $\tau_o^2 = +\infty$ we are annulling the informative strength of the prior conjecture. The corresponding distribution, the improper uniform distribution in $\mathbb{R}$, is therefore noninformative in this case.

We can therefore state the following definition:

Definition 2

Let $\mathcal{F} = \{\pi_{a,s}(\theta)\}$ be a family of prior densities, the parameter a *being a measure of centralization. Let $\mathcal{L}$ be a family of likelihood functions $f(x/\theta)$. Let us consider the posterior centralization parameter* a *as being a function of the prior parameters, of the sample observed and of its size. Finally, let us assume that there is a certain (possibly limiting) value of the prior parameter* s *for which it is found that:*

1 *the posterior centralization parameter does not depend on the prior centralization parameter, no matter the sample observed nor its size,*

2 *for all values of* a, *the (improper) densities corresponding to* a *and to this value of* s *are all equal each other (up to a multiplicative constant that could depend on* a*).*

We will refer to this common density as the reference density II *of the family $\mathcal{F}$ with respect to the parametrization used and to the likelihood family $\mathcal{L}$.* ■

The above definition particularly applies in those cases where, as in our example, the posterior parameter a is a weighted average of the prior parameter a and of some statistic and the weight of the parameter is cancelled out by a value of the parameter s.

Note that the parameter s now does not need to be a dispersion or precision measurement.

In order to apply this definition to our example, let $\mathcal{F}$ be the family of densities $N(a,s^2)$. Let $f(x/\theta)$ be a normal distribution $N(\theta,\sigma^2)$ with a known value of σ^2. Indicating the prior parameters by a subindex 0 and the posterior parameters by a subindex 1, we have

$$a_1=\frac{\sigma^2}{\sigma^2+ns_o^2}a_o+\frac{ns_o^2}{\sigma^2+ns_o^2}\bar{x}. \tag{2}$$

For $s_o=+\infty$ we obtain $a_1=\bar{x}$, irrespective of a_0. Hence the reference density II is the improper density $\pi(\theta)=1$.

4. METHOD 3

Once again we shall use an example to introduce the method. Let us suppose that a person consults a statistician in order to determine the prior distribution of the parameter θ of a random variable X with a distribution $N(\theta,\sigma^2)$, the variance σ^2 being known. Let us consider the moment in which the statistician, having decided that the prior distribution is normal and having assigned to its mean the value μ_o of the conjecture supplied by his client, attempts to assign a value to the prior variance. Let us assume that for this, the statistician proposes to use the technique of equivalent sample sizes (see Good (1950)).

There are two versions of this tecnique and both, as we shall now see, seem to be equally reasonable. The first or «classic» technique (see Berger (1985), p. 80) compares the informative strength of the initial conjecture with that of a future sample, and does not use noninformative distribution. The second version, which we designed, considers the initial conjecture to have come from a past sample, and it uses a noninformative distribution. The noninformative distribution evidently has to be such that both methods produce the same result, i.e., the same determination of σ^2.

In either case, the statistician begins by asking his client for the «equivalent sample size», i.e., the size n that a sample of the variable X should be, in order to have the same informative strength regarding θ as his initial conjecture.

If the statistician uses the first version of the technique, his reasoning will continue as follows. If, after assigning the initial values μ_o and τ_o^2 to the parameters, a sample of size n were observed, the new conjecture μ_1 regarding θ would depend on the initial conjecture μ_o and on the statistic $\bar{x}$, as shown in formula (1). μ_1 is a weighted average of μ_o and of $\bar{x}$, and the respective weights are good measure of the informative strength of the initial conjecture and of the sample. These weights depend on the value τ_o^2 assigned to the

parameter τ^2. A suitable value τ_o^2 would therefore be that which makes the two weights equal, i.e.,

$$\tau_o^2 = \frac{\sigma^2}{n} \tag{3}$$

On the other hand, if the statistician uses the second version of the technique of equivalent sample size, his reasoning will be as follows. The client behaves as if he had observed a hypothetical sample of size n from a distribution $N(\theta, \sigma^2)$ and as if he had obtained from the sample the conjecture μ_o for θ. We can assume that μ_o is the sample mean. The client's present information is the sum of his information prior to observing the sample (no information), represented by the noninformative distribution, and the information supplied by the sample. The prior distribution, which is the one we want to determine, has to reflect the present information and has therefore to coincide with the posterior information corresponding to the noninformative prior distribution and to the observation of a sample of size n and sample mean μ_o. We accept that the noninformative distribution is $N(\mu_i, \tau_i^2)$. The prior variance, which has to coincide with the variance of the present distribution, is expressed by (see DeGroot (1970), p. 167)

$$\tau_o^2 = \frac{1}{\dfrac{1}{\tau_i^2} + \dfrac{n}{\sigma^2}}. \tag{4}$$

Hence, according to this second version, the prior variance has to take the value τ_o^2 given by (4).

The noninformative distribution has to be such that values (3) and (4) are equal. This implies that $\tau_i^2 = +\infty$, and the value of μ_i is therefore irrelevant. The noninformative distribution is therefore the normal distribution of infinite variance, i.e., the improper uniform distribution in $\mathbb{R}$.

We state the following definition:

Definition 3

Let $\mathscr{F} = \{\pi_{a,s}(\theta)\}$ be a family of prior densities, the parameter a *being a measure of centralization. Let $\mathscr{L}$ be a family of likelihood functions $f(x/\theta)$. Let us consider the posterior parameters* a *and* s *as functions of the prior parameters, of the sample and of its size. Let us assume that the posterior parameter* a *is a weighted average of the prior parameter* a *and of some statistic, and that the weights depend only on the prior parameter* s *and on the sample size. Let* ŝ *be the value of the prior parameter* s *(expressed as a function of the sample size) so that the two weights are equal. Let us assume*

that there exists a value (possibly limiting) of the prior parameter s *for which we find that*

1 *the posterior parameter* s *coincides with* $\hat{s}$, *irrespective of the prior parameter* a, *the sample observed and its size,*

2 *for all values of* a, *the (improper) densities corresponding to* a *and to this value of* s *are all equal each other (up to a multiplicative constant that could depend on* a*).*

We will refer to this common density as reference density III *of the family* $\mathcal{F}$ *with respect to the parametrization used and to the likelihood family* $\mathcal{S}$. ■

To formalize our example, we once again consider the family $N(a, s^2)$ and the likelihood $f(x/\theta)$, normal $N(\theta, \sigma^2)$ with a known value of σ^2.

Equation (2) adapts to Definition 3. The value of s_o for which the weights of a_o and $\bar{x}$ become equal each other is $\hat{s} = \sigma/\sqrt{n}$.

The transformation of the parameter s is expressed b

$$s_1^2 = \frac{1}{\dfrac{1}{s_o^2} + \dfrac{n}{\sigma^2}}.$$

If $s_o = +\infty$ then $s_1 = \hat{s} = \sigma/\sqrt{n}$, irrespective of the value of the prior parameter a, the sample observed and its size. Since $\pi_{a,+\infty}(\theta) = 1$, this expression does not depend on a. Hence the reference density III is the improper density $\pi(\theta) = 1$.

5. APPLICATIONS

Let $\mathcal{F} = \{\pi_{a,s}(\theta)\}$ be the family of beta densities. We can consider for this family three distinct parametrizations. The parameter a is the mean, $a = p/(p+q)$, in all three cases. In the first parametrization, s is the variance; in the second, s is a precision measurement: $s = p+q$; while in the third, s is a dispersion measurement whose range of variation is independent of a: $s = (p+q)^{-1}$. When applying methods 2 and 3, the Bernoulli distribution is taken as the likelihood.

All three methods can be applied to the three parametrizations, with the exception of the first pametrization which can only take method 1 because of the form of the parametric space. In all viable cases, the improper density $\pi(\theta) = \theta^{-1}(1-\theta)^{-1} I_{(0,1)}(\theta)$ is obtained as the reference density.

For the beta distributions, Jeffreys and Bernardo obtain $\pi(\theta) = \theta^{-1/2}(1-\theta)^{-1/2} I_{(0,1)}(\theta)$.

Let us now consider the family $\mathcal{F}$ of gamma densities with mean a and variance s. Let the likelihood be a Poisson distribution. Methods 1 and 2 produce the same reference density, the improper density $\pi(\theta) = \theta^{-1} I_{(0,+\infty)}(\theta)$. Method 3 cannot be applied to this case (see section 6). Jeffreys and Bernardo procedures yield $\pi(\theta) = \theta^{-1/2} I_{(0,+\infty)}(\theta)$.

In the normal case of known variance, all methods obtain the same distribution: $\pi(\theta) = 1$.

6. GENERALIZATIONS

I. The k-dimensional case

The three methods can be extended to the k-dimensional case, when the parameters are a set of marginal centralization measurements and a joint dispersion measurement.

This extension can be illustrated by considering the family $\mathcal{S}$ of Dirichlet densities, in which the parameters are the marginal means and the sum of the usual parameters. Definition 1 can be applied to this family. If we take the multinomial distribution as the likelihood, Definitions 2 and 3 can also be applied. In the three cases, the improper density

$$\pi(\theta_1, \ldots, \theta_k) = \prod_{i=1}^{k} \theta_i^{-1} \left(1 - \sum_{i=1}^{k} \theta_i\right)^{-1} I_{(0,1)}(\theta_i)$$

is obtained as the reference density.

II. Two extensions of Definition 3

Definition 3 can also be extended by allowing the weights to depend on a_o, the value of the prior parameter a.

In this case it cannot be postulated that the posterior parameter s coincides with $\hat{s}$ for every sample (this version of the postulate is included in the definition for its simplicity, however it is not necessary and is too demanding), but only for those samples in which a certain «good» estimator of a coincides with a_o, the (variable) value of the prior parameter used in the definition of $\hat{s}$.

The extension can be illustrated by considering the family $\mathcal{F}$ of gamma densities with mean a and variance s and the family $\mathcal{L}$ of Poisson likelihoods.

Definition 3 cannot be applied directly because a_1 is not a linear function of a_o, as is shown by the transformation equation for the parameter a:

$$a_1 = \frac{a_o}{a_o + s_o n} a_o + \frac{s_o n}{a_o + s_o n} \bar{x}. \tag{5}$$

However the same philosophy of Definition 3 can be applied.

A formal application of the first version of the technique of equivalent sample size to formula (5) gives rise to the determination of s_o as $s_o = a_o/n$.

On the other hand, the transformation of the second parameter is expressed by

$$s_1 = \frac{n\bar{x}s_o^2 + a_o^2 s_o}{(ns_o + a_o)^2}.$$

If s_o is the value of the posterior parameter s corresponding to the noninformative distribution gamma (a_i, s_i) and to the observation of a sample whose mean value is a_o, it would be expressed as

$$s_o = \frac{na_o s_i^2 + a_i^2 s_i}{(ns_i + a_i)^2}.$$

When $s_i = +\infty$, then s_o takes the value of a_o/n. Hence we can take the improper density $\pi(\theta) = \theta^{-1} I_{(0,+\infty)}(\theta)$ as the reference density.

A further possible generalization of Definition 3, applied to the k-dimensional case, consists in not postulating that the posterior parameter a_1 is a convex linear combination of a_o and of the statistic, and in defining the value $\hat{s}$ as that which makes a_1 equidistant from a_o and the statistic.

References

[1] AKAIKE, H. (1978): *A new look at the Bayes procedure.* Biometrika 65, 53-59.

[2] BERGER, J. O. (1985): *Statistical decision theory and Bayesian analysis.* Springer-Verlag. New York.

[3] BERNARDO, J. M. (1979): *Reference posterior distributions for Bayesian inference (with discussion).* J. Roy. Statist. Soc. 41, 113-147.

[4] BOX, G. E. P. and TIAO, G. C. (1973): *Bayesian inference in statistical analysis.* Addison-Wesley, Reading Massachusetts.

[5] DEGROOT, M. H. (1970): *Optimal Statistical Decisions.* McGraw-Hill. New York.

[6] GOOD, I. J. (1950): *Probability and the weighing of evidence.* Charles Griffin. London.

[7] HARTIGAN, J. (1964): *Invariant prior distributions.* Ann. Math. Statist. 35, 836-845.
[8] JAYNES, E. T. (1968): *Prior probabilities.* IEEE Transactions on Systems Science and Cybernetics, SSC-4, 227-241.
[9] JAYNES, E. T. (1983): Papers on Probability, Statistics and Statistical Physics. A reprint collection. R. D. Rosenkrantz (ed.) Reidel, Dordrecht.
[10] JEFFREYS, H. (1961): *Theory of probability.* Oxford University Press. London.
[11] NOVICK, M. R. and HALL, W. J. (1965): *A Bayesian indifference procedure.* J. Amer. Statist. Assoc. 60, 1104-1117.
[12] VILLEGAS, C. (1977): *On the representation of ignorance.* J. Amer. Statist. Assoc. 72, 651-654.
[13] VILLEGAS, C. (1981): *Inner statistical inference,* II. Ann. Statist. 9, 768-776.
[14] ZELLNER, A. (1977): *Maximal data information prior distributions.* In New methods in the applications of Bayesian methods, A. Aykac and C. Brumat (eds.) North-Holland, Amsterdam.

Departamento de Estadistica
Facultad de Ciencias Matemáticas
Universidad Complutense
28040 Madrid. Spain

Recibido: 14 de marzo de 1989