Ann. Inst. Statist. Math.
Vol. 44, No. 3, 405–416 (1992)

BAYESIAN PRIORS BASED ON A PARAMETER TRANSFORMATION USING THE DISTRIBUTION FUNCTION

MARTIN CROWDER

Department of Mathematical and Computing Sciences, University of Surrey, Guildford, Surrey, GU2 5XH, U.K.

(Received January 30, 1989; revised May 20, 1991)

Abstract. One of the tasks of the Bayesian consulting statistician is to elicit prior information from his client who may be unfamiliar with parametric statistical models. In some cases it may be more illuminating to base a prior distribution for parameter θ on the transformed version $F(v|\theta)$, where F is the data distribution function and v is a designated reference value, rather than on θ directly. This approach is outlined and explored in various directions to assess its implications. Some applications are given, including general linear regression and transformed linear models.

Key words and phrases: Bayesian priors, priors for location-scale regression models, priors for transformed linear models, proper priors.

1. Introduction

Suppose that a Bayesian statistician is involved with an experimental scientist in planning a study and must construct a prior distribution $p(\theta)$ for parameter θ. It is not useful for the statistician to ask the client for his views on θ directly if the latter has not made it his business to become familiar with the detailed workings of parametric statistical models. In this case questions concerning θ itself are not likely to produce a useful exchange of information unless θ happens to have a meaningful interpretation in the context. However, the client will often have fairly well defined beliefs about the sort of data he expects to obtain. He may have had to construct a well reasoned argument in support of such expectations in order to gain financial backing for the work. It is natural then to try to use this aspect of the client's knowledge to construct $p(\theta)$ from his assessed probabilities for his future data. In other words, the discussion is about observable data rather than statistical parameters.

There is now a vast literature on the complex problem of eliciting prior distributions and many approaches have been discussed (e.g. Winkler (1967), Savage (1971), Tversky (1974), Hogarth (1975), Bernardo (1979), Dickey (1980), Kadane (1980), Kadane *et al.* (1980)). There has been much emphasis also on constructing non-informative priors (e.g. Hartigan (1964), Novick (1969)) but the present

concern is to characterize "prior knowledge rather than prior ignorance" (Dawid *et al.* (1973)).

This paper is concerned with generating $p(\theta)$ from prior information quantified in terms of the particular transformation $u = F(v \mid \theta)$ of θ, where F is the distribution function for the observations and v is some designated reference value. (Actually, everything here could equally be done in terms of the survivor function $\bar{F}(v \mid \theta) = P(y > v \mid \theta)$, and in some cases, particularly for certain multivariate distributions, this may be much tidier.) The intention here is to explore this suggestion to see where it might provide a useful aid to a Bayesian analysis.

2. Basic outline

Let $F(y \mid \theta)$ be the distribution function of y (univariate or multivariate, continuous or discrete) depending on a continuous parameter θ which is assumed to be scalar for now. A prior $p(\theta)$ will induce a distribution $p(u)$ on $(0,1)$ for $u = F(v \mid \theta)$, where v is any fixed reference value. Conversely, provided that the relation $u = F(v \mid \theta)$ can be inverted to give θ in terms of u and v, v and $p(u)$ will together determine $p(\theta)$. Thus, one way of implementing the approach described above is to ask the client to designate a value v for which he can assign probabilities or odds to u.

Example 1. Suppose that y is the finishing position of a particular horse in a particular race. The reference value $v = 3$ corresponds to the horse's being "placed". Many statistical laymen are much practised in the potentially profitable art of assessing their prior estimate for $u = P(y \leq 3)$ in that context. What we are asking for is an extension of this from an estimate of u to a prior distribution for u.

Example 2. Reliability Testing. On the basis of experience with standard components suppose that it is possible to predict the chance of failure of a modified version before $v = 20$ hours as about 50%, and fairly surely not less than 10% and not more than 75%. The prior distribution for $u = F(v \mid \theta)$ should then peak around 0.5 and tail off suitably at 0.1 and 0.75.

The specification so far is rather broad but can be narrowed via the following informal considerations. For given θ, $F(y \mid \theta)$ has the standard uniform distribution $\mathrm{U}(0,1)$ when y is a continuous variate. An assumption $p(u) = 1$ on $(0,1)$ for $u = F(v \mid \theta)$ would then roughly correspond to a belief that the reference value v had been generated from the y-distribution, i.e. that v is a typical y-value. Again, if v were one of a sample of m independent y's, say the r-th in order, it would have probability density

$$p_{rm}(v) = r \binom{m}{r} u^{r-1}(1-u)^{m-r} F'(v \mid \theta), \tag{2.1}$$

where $F' = \partial F / \partial y$, and then u would have a beta distribution $\mathrm{Be}(r, m-r+1)$. The larger m, with r/m roughly constant, the greater is the precision attributed

to the selection of v; the uniform distribution for u corresponds to $m = r = 1$. Thus, in the absence other considerations, the client might be asked to nominate a value of v and then to select, with the help of the statistician, a member of the beta family for $p(u)$ based on the above properties.

There is an echo here of the fiducial argument. If v is a randomly generated y-value, and θ is fixed, then $u = F(v \mid \theta)$ has distribution U(0, 1). According to that argument, when v becomes fixed (as data) and θ becomes random, the uniform is retained in form as the fiducial distribution of u.

For location parameter families of the form $F(y \mid \theta) = G(y - \theta)$, $|\partial F(y \mid \theta)/\partial\theta| = |\partial F(y \mid \theta)/\partial y|$. Then, if $u = F(v \mid \theta)$ has distribution Be$(r, m - r + 1)$, θ has density

$$p(\theta) = r\binom{m}{r} u^{r-1}(1-u)^{m-r}|\partial F(v \mid \theta)/\partial\theta|$$

which is equal to $p_{rm}(v)$ in (2.1). In this special case the density $p(\theta)$ is numerically equal to the likelihood of a single observation v known to be the r-th in order from a sample of m y's. For the uniform distribution ($r = m = 1$) the suggested prior amounts to adding a single reference value to the future sample, i.e. asking for a guess of a typical future y-value on the basis of prior knowledge.

3. Single parameter case

Before getting down to the implementation of the method in particular situations, we first dispose of a few general questions and consequences in this section for the single parameter case. To reiterate, the proposal here is aimed at situations where prior information is more easily appreciated in terms of probabilities $u = F(v \mid \theta)$ than in the metric of the parameter. θ.

Since $p(\theta) = p(u)|\partial u/\partial\theta|$ and $u = F(v \mid \theta)$, $p(\theta)$ will depend on the form of F in general. It has been argued (e.g. Lindley (1972), Section 12.4) that dependence of $p(\theta)$ on the sampling rule is not sensible. Other priors which fall foul of this criticism are Jeffreys' invariant rule, based on the expected information, and natural conjugate priors. However, the dependence here of $p(\theta)$ on the sampling model is just a consequence of focussing on u rather than θ, i.e. on potential data rather than on parameters. It is $p(u)$ here rather than $p(\theta)$ which is independent of the form of F.

When $u = F(v \mid \theta)$ is assigned a particular type of distribution then that of $u' = F(v' \mid \theta)$ for $v' \neq v$ will be determined but will not be of the same type in general. However, there is invariance on a transformed scale of u. Suppose that $F(v \mid \theta)$ is a monotone function of θ for each v; if this is not so the sampling model may have identifiability problems. Then $u = F(v \mid \theta)$ can be solved for θ, say as $\theta = h(u, v)$, and for each v the transformation $h(u, v)$ of u has the same density, namely $p(\theta)$. For example, in a location parameter model, with $F(y \mid \theta) = G(y-\theta)$ for some distribution function G, $h(u, v) = v - G^{-1}(u)$. Then $v - G^{-1}(u)$ and $v' - G^{-1}(u')$ have the same distribution for all v and v'. Likewise, for a scale parameter model $F(y \mid \theta) = G(y/\theta)$ and $h(u, v) = v/G^{-1}(u)$.

In general one cannot arbitrarily assign u-distributions for more than one reference value v since they will be inconsistent. In some cases what seems to be a

reasonable choice of distribution for $u = F(v \mid \theta)$ may impose an unreasonable distribution on $u' = F(v' \mid \theta)$. One way of avoiding this problem would be to assign distributions say for $u_1, \ldots, u_r$ as initially thought fit, and then to compromise between the resulting priors $p_1(\theta), \ldots, p_r(\theta)$. A general discussion of the reconciliation of prior assessments is given by Lindley *et al.* (1979).

Suppose in particular that u is to be assigned a beta distribution $\mathrm{Be}(\alpha, \beta)$: $p(u) = \mathrm{B}(\alpha, \beta)^{-1} u^{\alpha-1}(1-u)^{\beta-1}$, where $\mathrm{B}(\alpha, \beta)$ is the beta function. Values for the prior parameters (α, β) must be found to fit the client's prior information.

Example 2 (continued). Let us take "fairly surely" to mean 90% certain. Then, for $v = 20$ hours, $P(u \geq 0.1) = 0.9$ and $P(u \leq 0.75) = 0.9$. Hence $\mathrm{I}_{0.1}(\alpha, \beta) = 0.1$ and $\mathrm{I}_{0.75}(\alpha, \beta) = 0.9$, where $\mathrm{I}_z(\alpha, \beta)$ is the incomplete beta function ratio, leading to values for α and β. If the predicted peak near 0.5 of $p(u)$ is to be accommodated also then a suitable compromise must be made towards values satisfying $(\alpha - 1)(\alpha + \beta - 2)^{-1} = 1/2$.

For the purpose of illustration here a naive, direct route has been taken to assign a beta distribution for u. A more sophisticated approach to such assignation is described by Chaloner and Duncan (1983). There the client is asked to consider his predictions for the modal and adjacent numbers of successes in Bernoulli trials with probability u.

A referee has called for clarification of the relationship between the prior derived by the present approach and the prior predictive distribution. For a future observation Y the latter is given by

$$P(Y \leq y) = \int F(y \mid \theta) p(\theta) d\theta = \int F(y \mid \theta_{uv}) p(u) du,$$

the prior mean of $F(y \mid \theta_{uv})$, where θ_{uv} is the expression of θ in terms of u and v obtained by inverting the relationship $u = F(v \mid \theta)$. In particular, $P(Y \leq v)$ equals the prior mean of u itself. The referee has also drawn attention to Good's (1950) "device of imaginary results". There, one assigns a posterior distribution which is felt to be in accordance with invented data (the "imaginary results"), and then derives the corresponding prior. That approach is quite different from the one here, but could possibly be used to appraise the resulting posteriors.

4. Multiparameter i.i.d. case

The treatment is now applied for a $q \times 1$ parameter vector θ by assigning a joint distribution for $u = (u_1, \ldots, u_q)^T$ where $u_j = F(v_j \mid \theta)$ is based on designated reference value v_j. The informal considerations which led to beta distributions in the single parameter case may be developed in various ways. Note that the distribution $\mathrm{Be}(r, m - r + 1)$ given in (2.1) would arise from regarding u there as the r-th order statistic in a random sample of size m from $\mathrm{U}(0, 1)$. A natural extension then is to take the u_j's here as order statistics, u_j having rank r_j, from a single $\mathrm{U}(0, 1)$ sample of size $m \geq q$. The r_j's must have the same ordering as

the v_j's and we may take $v_1 < v_2 < \cdots < v_q$ without loss of generality. Also, let $r_0 = 0$, $r_{q+1} = m + 1$, $u_0 = 0$ and $u_{q+1} = 1$. The resulting prior is

$$p(\theta) \propto \left\{ \prod_{j=0}^{q} (u_{j+1} - u_j)^{\alpha_{j+1}-1} \right\} |\det J| r(u), \tag{4.1}$$

where $\alpha_{j+1} = r_{j+1} - r_j$, and $r(u) = 1$ for $u_1 < u_2 < \cdots < u_q$ and 0 otherwise; $J(q \times q) = \partial u / \partial \theta$ is the Jacobian matrix of the transformation, assumed to be $1 - 1$ for identifiability. For a more general class of priors the restriction of the α_j's to integer values in (4.1) may be relaxed.

If $w_j = u_j / u_{j+1}$ $(j = 1, \ldots, q)$ then the w_j's are independent $\mathrm{Be}(r_j, r_{j+1} - r_j)$ variates. It may be easier to base prior assessment on the w_j's than on the u_j's directly, thus leading to choice of the r_j's or α_j's in (4.1).

Example 2 (continued). Suppose that the parameter θ is two-dimensional and that the previous considerations are retained for $v_2 = 20$ hours. Thus, a beta distribution $\mathrm{Be}(\alpha_1 + \alpha_2, \alpha_3)$ is assigned to $w_2 = F(v_2 \mid \theta)$ as described above; this fixes values for $\alpha_1 + \alpha_2$ and α_3. Now, say with $v_1 = 15$ hours, a beta distribution $\mathrm{Be}(\alpha_1, \alpha_2)$ for the ratio $w_1 = F(v_1 \mid \theta) / F(v_2 \mid \theta)$ is to be assigned in similar fashion. Explicity, w_1 measures how much less likely is failure by 15 hours than by 20 hours, and it is this ratio for which prior probabilities need to be elicited. For instance, the engineer might guess this ratio to be about 1/3. Equating the mode $(\alpha_1 - 1)/(\alpha_1 + \alpha_2 - 1)$ to 1/3, and using the previously fixed value of $\alpha_1 + \alpha_2$, now yields α_1 and hence α_2.

Example 3. Location-scale family: $F(y \mid \theta) = G\{(y - \beta)/\phi\}$. Here $q = 2$, $\theta = (\beta, \phi)$ with $\phi > 0$, and $\det J = \phi^{-3}(v_2 - v_1) u_1' u_2'$ where $u_j' = G'\{(v_j - \beta)/\phi\}$. For instance, with the uniform assignment $\alpha_{j+1} = 1$ for each j, (4.1) yields $p(\theta) \propto \phi^{-1} L_1 L_2 r(v)$ where $L_j = \phi^{-1} u_j'$ and $v = (v_1, v_2)$. Then $p(\theta)$ has the form of a posterior based on two ordered "observations", $v_1 < v_2$, together with the improper prior $p(\phi) \propto \phi^{-1}$ for ϕ. Since $\phi = (v_2 - v_1)/\{G^{-1}(u_2) - G^{-1}(u_1)\}$, the requirement $r(u) = r(v)$ ensures that $\phi > 0$ with probability 1 in the prior.

The example illustrates the property that β and ϕ are not generally prior-independent under this method. This is sometimes held to be a failing in the location-scale case; it occurs with the (unmodified) Jeffreys' rule and with the conjugate prior family for $N(\beta, \phi^2)$ data. In connection with the present method it means that the smaller ϕ, the smaller is the variation in β required to make $u_j = G\{(v_j - \beta)/\phi\}$ cover any sub-interval of $(0, 1)$. However: (a) If prior information has been carefully quantified by the statistician and client in terms of u then any purely mathematical consequence, such as prior non-independence of β and ϕ, should not be allowed to override it; (b) If $F(y \mid \theta) = G\{(y - \beta)/\phi\}$ and $H(z) = G(z + \alpha)$, then H is a distribution function and $F(y \mid \theta) = H\{(y - \beta - \alpha\phi)/\phi\}$. A blanket prior-independence rule for location and scale parameters would decree that β and ϕ be independent in the G-version, and that $(\beta + \alpha\phi)$ and ϕ be independent in

the H-version. This seems contradictory. The construction in (b) just shows that the location parameter is not well defined in this setting.

5. Multiparameter i.n.i.d. case: regression model

Suppose now that the observations y_i are independent and non-identically distributed with distribution functions $F_i(y \mid \beta)$ of form $G(y - x_i^T\beta)$; $\beta = (\beta_1, \ldots, \beta_q)^T$ comprises the regression coefficients and $x_i(q \times 1)$ the explanatory variables. To proceed as before one might think of fixing on a reference distribution of the class, say $F_0(y \mid \beta) = G(y - a_0^T\beta)$ for some designated $a_0(q \times 1)$, and taking $u_j = F_0(v_j \mid \beta)$ for $j = 1, \ldots, q$. However, this does not work because the resulting transformation $\beta \to u = (u_1, \ldots, u_q)$ is not invertible, only $a_0^T\beta$ being identifiable in this case. The problem can be resolved by taking a full set of reference distributions and values as follows.

Take $u_j = G(v_j - a_j^T\beta)$ $(j = 1, \ldots, q)$ for designated $v = (v_1, \ldots, v_q)^T$ and $q \times 1$ vectors $a_1, \ldots, a_q$. Let $A(q \times q)$ have j-th row a_j^T and $g = (g_1, \ldots, g_q)^T$ where $g_j = G^{-1}(u_j)$. Then we have $g = v - A\beta$ which is invertible as $\beta = A^{-1}(v - g)$ provided that the a_j are linearly independent. The Jacobian matrix $J = \partial u/\partial\beta$ of the transformation has determinant

$$\det J = (-1)^q(\det A)\prod_{j=1}^{q} G'(v_j - a_j^T\beta).$$

Once a joint distribution for the u_j's is assigned the prior $p(\beta)$ follows by transformation.

A particular, special situation is as follows. Suppose that $a_1, \ldots, a_q$ can be identified so that $a_1^T\beta, \ldots, a_q^T\beta$ are prior-independent. For instance, in a treatment/yield experiment, in which x_i records the levels of treatments applied, one might take $a_j = (0, \ldots, 1, \ldots, 0)^T$ (with a 1 in the j-th position) if prior beliefs about the q different treatments are unrelated. Thus, after designating reference values $v_1, \ldots, v_q$, distributions on $(0, 1)$ may be assigned independently to the u_j. Notice that prior-independence of the $a_j^T\beta$ implies nonsingularity of A; if A were singular there would exist a non-zero $q \times 1$ vector c such that $c^TA\beta = 0$ for all β, implying the linear relation $\sum c_j(a_j^T\beta) = 0$ connecting the $a_j^T\beta$. In the following examples this special situation is not assumed.

Example 2 (continued). Suppose that the failure time distribution of the components is taken to be exponential, $F_i(y \mid \beta) = 1 - \exp(-\lambda_i y)$, with loglinear model $\log\lambda_i = -x_i^T\beta$. Then $F_i(y \mid \beta) = G(\log y - x_i^T\beta)$ with $G(y) = 1 - \exp(-e^y)$, a linear regression model for the log-data. To be specific, suppose that $q = 2$ and $x_i = (1, x_i')^T$ where x_i' is a measure of production quality higher values of which are known to enhance the component lifetime. Thus, in $x_i^T\beta = \beta_1 + \beta_2 x_i'$, there is the additional complication of a constraint $\beta_2 > 0$ to cope with. Take $v_1 = 20$ hours, as before, and $a_1 = (1, 1)^T$. A distribution $p(u_1)$ on $(0, 1)$ can be assigned on the basis previously described for this example. Next take $v_2 = 20$ hours and $a_2 = (1, 2)^T$, and suppose that the prior expectation of the engineer is that the

probability of failure by 20 hours should be roughly halved when x' increases from 1 to 2. Now $u_2 < u_1$, since $\beta_2 > 0$, so $p(u_2 \mid u_1)$ can be taken as a (beta) distribution on $(0, u_1)$ with peak around $u_1/2$, and spread in accordance with the supposed accuracy of the engineer's expectation. Having thus assigned $p(u_1)$ and $p(u_2 \mid u_1)$, the prior for β follows as $p(\beta) = p(u_1)p(u_2 \mid u_1)|\det J|$.

The treatment will be extended now to encompass a scale parameter ϕ. Thus $F_i(y \mid \theta)$ has form $G\{(y - x_i^T\beta)/\phi\}$ and the parameter set $\theta = (\beta, \phi)$ has $q + 1$ elements. To v and A, designated as before, an extra pair must be added, say v_{q+1} (scalar) and a_{q+1} ($q \times 1$ vector), and a corresponding extra distribution $p(u_{q+1} \mid u_1, \ldots, u_q)$ must be assigned. The transformation becomes $g = (v - A\beta)/\phi$ and $g_{q+1} = (v_{q+1} - a_{q+1}^T\beta)/\phi$, where $g_{q+1} = G^{-1}(u_{q+1})$, from which $A\beta = v - \phi g$ and $a_{q+1}^T\beta = v_{q+1} - \phi g_{q+1}$. Hence $a_{q+1}^T A^{-1}(v - \phi g) = v_{q+1} - \phi g_{q+1}$ so

$$\phi = (v_{q+1} - a_{q+1}^T A^{-1} v)/(g_{q+1} - a_{q+1}^T A^{-1} g) \tag{5.1}$$

and the expression of β as a function of (u, u_{q+1}) follows. The Jacobian matrix $J = \partial(u, u_{q+1})/\partial(\beta, \phi)$ has determinant

$$\det J = (-1)^{q+1}\phi^{-(q+2)}(v_{q+1} - a_{q+1}^T A^{-1} v)(\det A) \prod_{j=1}^{q+1} G'\{(v_j - a_j^T\beta)/\phi\}. \tag{5.2}$$

It is true that, as before, one now just needs to assign a joint distribution $p(u, u_{q+1})$ and then transform to $p(\theta)$. However, there is at least one constraint, this being required to ensure that $P(\phi > 0) = 1$ in (5.1). Thus, if $v_{q+1} > a_{q+1}^T A^{-1} v$ then $g_{q+1} > a_{q+1}^T A^{-1} g$ must hold with probability 1, so once a joint distribution $p(u)$ has been assigned the further conditional distribution $p(u_{q+1} \mid u)$ must have support $\{G(a_{q+1}^T A^{-1} g), 1\}$. Similarly, if $v_{q+1} < a_{q+1}^T A^{-1} v$ then $p(u_{q+1} \mid u)$ must be restricted to the interval $\{0, G(a_{q+1}^T A^{-1} g)\}$, and (5.1) rules out $v_{q+1} = a_{q+1}^T A^{-1} v$.

A referee has pointed out that the joint distribution $p(u, u_{q+1})$ can be considered symmetrically in $u_1, \ldots, u_{q+1}$. It is just a consequence of the development here that u_{q+1} is separated from $u_1, \ldots, u_q$. Symmetrically expressed, the transformation is $g_+ = (v_+ - A_+\beta)/\phi$, where g_+ is $(q+1) \times 1$ with j-th element $G^{-1}(u_j)$, $v_+ = (v_1, \ldots, v_{q+1})^T$ and A_+ is $(q + 1) \times q$ with j-th row a_j^T. For invertibility β must be identifiable which means that there must be q linearly independent vectors among the a_j's, i.e. A_+ must have full rank q. Assuming, without loss of generality, that $a_1, \ldots, a_q$ are linearly independent, A_+ may be written as $(A^T, a_{q+1})^T$ where $A(q \times q)$ is nonsingular. Expressions for the inverse transformation and the Jacobian now follow as before.

Example 2 (continued). The failure time distribution is now generalized to the Weibull family: $F_i(y \mid \theta) = G\{(\log y - x_i^T\beta)/\phi\}$ with $G(y) = 1 - \exp(-e^y)$. First $p(u)$ is assigned in steps, $p(u_1)$ then $p(u_2 \mid u_1)$, as described above in this section. Now take $v_3 = 15$ hours and $a_3 = (1, 1)^T$, and suppose that the engineer expects the probability of failure by 15 hours with $x' = 1$ to be about 1/3 of that

by 20 hours, i.e. a prior expectation that $u_3 \approx u_1/3$. Note also that $u_3 < u_1$ with probability 1 since $15 < 20$. Thus $p(u_3 \mid u)$ is to have support $(0, u_1)$, peak around $u_1/3$, and have spread to conform with the degree of prior certainty. There are other implied relationships between the u's in the joint distribution $p(u, u_3)$ so constructed, and these must be checked for conformity with prior knowledge and modifications made if necessary. An illustration of this process is given below. The resulting prior for θ is $p(\theta) = p(u, u_3)|\det J|$ with $\det J$ given by (5.2).

To give a very simple specific example consider an assignment in which all three distributions, $p(u_1)$, $p(u_2 \mid u_1)$ and $p(u_3 \mid u_1, u_2)$, are uniform. This yields

$$p(\theta) \propto \phi^{-4} G\{(v_1' - \beta_1 - \beta_2)/\phi\}^{-2} G'\{(v_1' - \beta_1 - \beta_2)/\phi\}$$
$$\cdot G'\{(v_2' - \beta_1 - 2\beta_2)/\phi\} G'\{(v_3' - \beta_1 - \beta_2)/\phi\}$$

with $G'(y) = \exp(y - e^y)$, $v_1' = \log v_1 = 3.00$ and $v_3' = \log v_3 = 2.71$; the support region $\{u_2 < u_1, u_3 < u_1\}$ correctly transforms back to $\{\beta_2 > 0, \phi > 0\}$. Just one conformity check will now be illustrated. Note that $p(u_1, u_2) = u_1^{-1}$ on $0 \le u_2 \le u_1 \le 1$, so $p(u_2) = -\log u_2$ on $(0, 1)$ and $p(u_1 \mid u_2) = -1/(u_1 \log u_2)$ on $(u_2, 1)$. Also, $p(u_3 \mid u_1, u_2) = u_1^{-1}$ on $0 \le u_3 \le u_1 \le 1$, so

$$p(u_3 \mid u_2) = \int p(u_3 \mid u_1, u_2) p(u_1 \mid u_2) du_1 = (\log u_2)^{-1}\{1 - 1/\max(u_2, u_3)\}$$

on $(0, 1)$. Hence, $p(u_3 \mid u_2)$ is constant on $(0, u_2)$ and then swings down to 0 on $(u_2, 1)$. This function must be critically assessed to judge its reasonableness in the light of the engineer's prior knowledge. For instance, it implies that $P(u_3 < 1/2 \mid u_2 = 1/2) = 1/(2 \log 2) = 0.72$. Thus, reducing the equality from 2 to 1 and the lifetime assessment from 20 to 15 hours makes the failure probability change from $1/2$ to something more likely than not to be less than $1/2$. Roughly speaking, the lifetime reduction has had more effect than the quality reduction in this case. If this is not reasonable some revisions to $p(u, u_3)$ are called for.

6. Application: transformed linear model

Box and Cox (1964) analysed a model in which $y(\lambda)$, a transformation of y involving parameter λ, has distribution $N(x^T\beta, \phi^2)$. One of the difficulties for a Bayesian analysis is to generate a prior for $\theta = (\beta, \phi, \lambda)$ which sensibly reflects the variation in the $y(\lambda)$ scale as λ varies, i.e. a change in λ should cause an appropriate shift in the probability mass distribution of $p(\beta, \phi \mid \lambda)$. Box and Cox gave an approximate solution resulting in a slightly data-dependent ignorance prior. In the method proposed here convariation of the parameters in the prior is constrained by inhibiting the variation of certain probabilities, i.e. the u_j's. The prior obtained is proper, and does not involve any posterior element, but requires some input of information.

Consider the general set up in which the observations y_i are taken to be independent with distribution functions $G[\{y(\lambda) - x_i^T\beta\}/\phi]$; $y(\lambda)$ is some transformation of y, of a specified class indexed by λ, and $G(\cdot)$ is some specified distribution

function. Take $\theta = (\beta, \phi, \lambda)$, where β is $q \times 1$, and $u_j = G[\{v_j(\lambda) - a_j^T\beta\}/\phi]$ for reference values $\{v_j, a_j : j = 1, \ldots, q+2\}$, with $A(q \times q) = (a_1, \ldots, a_q)^T$ nonsingular. The Jacobian matrix J, with elements $\partial u_j/\partial\theta_k$, is found to have determinant

$$\det J = (-1)^{q+1}\phi^{-1}W_\lambda(\det A)\prod_{j=1}^{q+2} L_j$$

where $L_j = \phi^{-1}G'[\{v_j(\lambda) - a_j^T\beta\}/\phi]$ and

$$W_\lambda = \det\begin{pmatrix} v_{q+1}(\lambda) - a_{q+1}^T A^{-1}v_\lambda & v'_{q+1}(\lambda) - a_{q+1}^T A^{-1}v'_\lambda \\ v_{q+2}(\lambda) - a_{q+2}^T A^{-1}v_\lambda & v'_{q+2}(\lambda) - a_{q+2}^T A^{-1}v'_\lambda \end{pmatrix}$$

with $v_\lambda(q \times 1) = \{v_1(\lambda), \ldots, v_q(\lambda)\}^T$; W_λ is the Wronskian of the functions $\{v_{q+j}(\lambda) - a_{q+j}^T A^{-1}v_\lambda : j = 1, 2\}$ (Apostol (1957), Exercise 5–9) and so governs linear dependencies between them. The transformation $\theta \to (u_1, \ldots, u_{q+2})$ is thus invertible in the neighbourhood of any point where $W_\lambda \neq 0$.

As for (5.1) it is found that $\beta = A^{-1}(v_\lambda - \phi g)$ and then

$$\phi = \{v_{q+j}(\lambda) - a_{q+j}^T A^{-1}v_\lambda\}/\{g_{q+j} - a_{q+j}^T A^{-1}g\} \tag{6.1}$$

for $j = 1$ and 2, the g's being defined as in Section 5; (6.1) yields ϕ and λ in terms of $(u_1, \ldots, u_{q+2})$. Now, the prior $p(u_1, \ldots, u_{q+2})$ must be constructed to observe the constraint $\phi > 0$, so we require here that $v_{q+j}(\lambda) - a_{q+j}^T A^{-1}v_\lambda$ and $g_{q+j} - a_{q+j}^T A^{-1}g$ should have the same sign for $j = 1$ and 2. However, it is possible that $v_{q+j}(\lambda) - a_{q+j}^T A^{-1}v_\lambda$ may change sign as $(u_1, \ldots, u_{q+2})$, and hence λ, varies. The particular transformation treated in Box and Cox (1964) is just such a case. However, provided $y(\lambda)$ is monotone in y for each λ this potential problem can be avoided by suitable choice of v_{q+j} and a_{q+j} for $j = 1, 2$. For example, suppose that the choice $a_{q+1} = a_{q+2} = a_r$ is made for some $r \in \{1, \ldots, q\}$. Note that $a_r = A^T e_r$, so $a_{q+j}^T A^{-1} = e_r^T$, where $e_r = (0, \ldots, 1, \ldots, 0)^T$ with a 1 in the r-th position and 0's elsewhere. Then

$$v_{q+j}(\lambda) - a_{q+j}^T A^{-1}v_\lambda = v_{q+j}(\lambda) - v_r(\lambda)$$

and

$$g_{q+j} - a_{q+j}^T A^{-1}g = g_{q+j} - g_r.$$

These two quantities will have the same sign, as required, if $v_{q+j} > v_r$ since then either (a) $v_{q+j}(\lambda) > v_r(\lambda)$ and $g_{q+j} > g_r$ (for a monotone increasing transformation) or (b) $v_{q+j}(\lambda) < v_r(\lambda)$ and $g_{q+j} < g_r$ (for a monotone decreasing transformation).

For brevity, let us give the simplest possible concrete demonstration of the method. Suppose that it is possible to designate $a_1, \ldots, a_q$ so that $u_1, \ldots, u_q$ are prior-independent as discussed in Section 5. Take $u_1, \ldots, u_{r-1}, u_{r+1}, \ldots, u_q$ as independent U(0, 1), and, independently, $u_r < u_{q+1} < u_{q+2}$ as order statistics from a sample of size three from U(0, 1). Hence $p(u_1, \ldots, u_{q+2}) \propto 1$ and $p(\theta) \propto$

$\phi^{-1}|W_\lambda| \prod_{j=1}^{q+2} L_j$ based on $\det J$ as given above. Then, for this simple case, one can obtain the marginal prior $p(\phi, \lambda) \propto \phi^{-3}|W_\lambda| I_{\lambda\phi}$ where

$$I_{\lambda\phi} = \int G'(s)G'(s + d_{1\lambda}/\phi)G'(s + d_{2\lambda}/\phi)ds$$

and $d_{j\lambda} = v_{q+j}(\lambda) - v_r(\lambda)$; $p(\lambda)$ follows by integration over ϕ, and finally the function of interest $p(\beta, \phi \mid \lambda)$ as $p(\theta)/p(\lambda)$.

For the normal case $G = \Phi$, the standard Gaussian distribution function, and then, for the case considered in the previous paragraph,

$$p(\theta) \propto \phi^{-(q+3)}|W_\lambda| \exp\{-T_{v\lambda}(\beta)/2\phi^2\},$$

where $T_{v\lambda}(\beta) = \sum_{j=1}^{q+2}\{v_j(\lambda) - a_j^T\beta\}^2$. Also $I_{\lambda\phi} = (2\pi\sqrt{3}\,)^{-1} \exp(-d_\lambda/2\phi^2)$ with

$$3d_\lambda = \{v_r(\lambda) - v_{q+1}(\lambda)\}^2 + \{v_r(\lambda) - v_{q+2}(\lambda)\}^2 + \{v_{q+1}(\lambda) - v_{q+2}(\lambda)\}^2.$$

Hence $p(\lambda) \propto d_\lambda^{-1}|W_\lambda|$, $p(\phi \mid \lambda) \propto \phi^{-3}\exp(-d_\lambda/\phi^2)$ and $p(\beta \mid \phi, \lambda) \propto \exp\{-T_{v\lambda}(\beta)/2\phi^2\}$, i.e. $d_\lambda/\phi^2 \mid \lambda$ has a χ^2_2 distribution and $\beta \mid \phi, \lambda$ is normal with covariance matrix proportional to ϕ^2. The covariation of β, ϕ and λ is constrained in this way at the reference values by the form of $p(\beta, \phi \mid \lambda)$: given λ, higher prior probabilities are given to (β, ϕ)-values which ensure better matching, according to the sampling model, of the $v_j(\lambda)$'s and corresponding $a_j^T\beta$'s.

The likelihood function for observations $y = (y_1, \ldots, y_n)$, corresponding to vectors $x_1, \ldots, x_n$ of explanatory variables, is proportional to $\phi^{-n}\dot{y}^{n\lambda} \cdot \exp\{-T_{y\lambda}(\beta)/2\phi^2\}$, where $T_{y\lambda}(\beta) = \sum_{i=1}^n\{y_i(\lambda) - x_i^T\beta\}^2$ and $\dot{y} = (\prod_{i=1}^n y_i)^{1/n}$. In consequence, the posterior distribution of λ is given by

$$p(\lambda \mid y) \propto |W_\lambda|\dot{y}^{n\lambda}\{T_{v\lambda}(\bar{\beta}_\lambda) + T_{y\lambda}(\bar{\beta}_\lambda)\}^{-(1+n/2)}$$

where $\bar{\beta}_\lambda = (D^TD)^{-1}D^Tz_\lambda$, D^T is $q \times (n+q+2)$ with columns $(x_1, \ldots, x_n, a_1, \ldots, a_{q+2})$, and $z_\lambda^T = \{y_1(\lambda), \ldots, y_n(\lambda), v_1(\lambda), \ldots, v_{q+2}(\lambda)\}$. The other posterior distributions of main interest are

$$p(\phi \mid \lambda, y) \propto \phi^{-(n+3)}\exp[-\{T_{v\lambda}(\bar{\beta}_\lambda) + T_{y\lambda}(\bar{\beta}_\lambda)\}/2\phi^2]$$

and

$$p(\beta \mid \phi, \lambda, y) \propto \exp[-\{T_{v\lambda}(\beta) + T_{y\lambda}(\beta)\}/2\phi^2];$$

thus $\phi^{-2}\{T_{v\lambda}(\bar{\beta}_\lambda) + T_{y\lambda}(\bar{\beta}_\lambda)\} \mid \lambda, y$ is χ^2_{n+2} and $\beta \mid \phi, \lambda, y$ is normal with mean $\bar{\beta}_\lambda$.

The preceding development will now be illustrated using some data of Box and Cox (1964) who analysed some survival times in a complete 3×4 factorial design with four replicates. Sweeting (1984) used these data to compare his own method of generating an ignorance prior with that of Box and Cox, and with that of Pericchi (1981). Figure 1 shows the λ-posteriors resulting from Box and Cox's approach, Sweeting's approach, and the one here. The former two curves are close and Pericchi's version, graphed by Sweeting, is also close to them and is omitted here for clarity. The curve to the right of them was produced as follows. Suppose

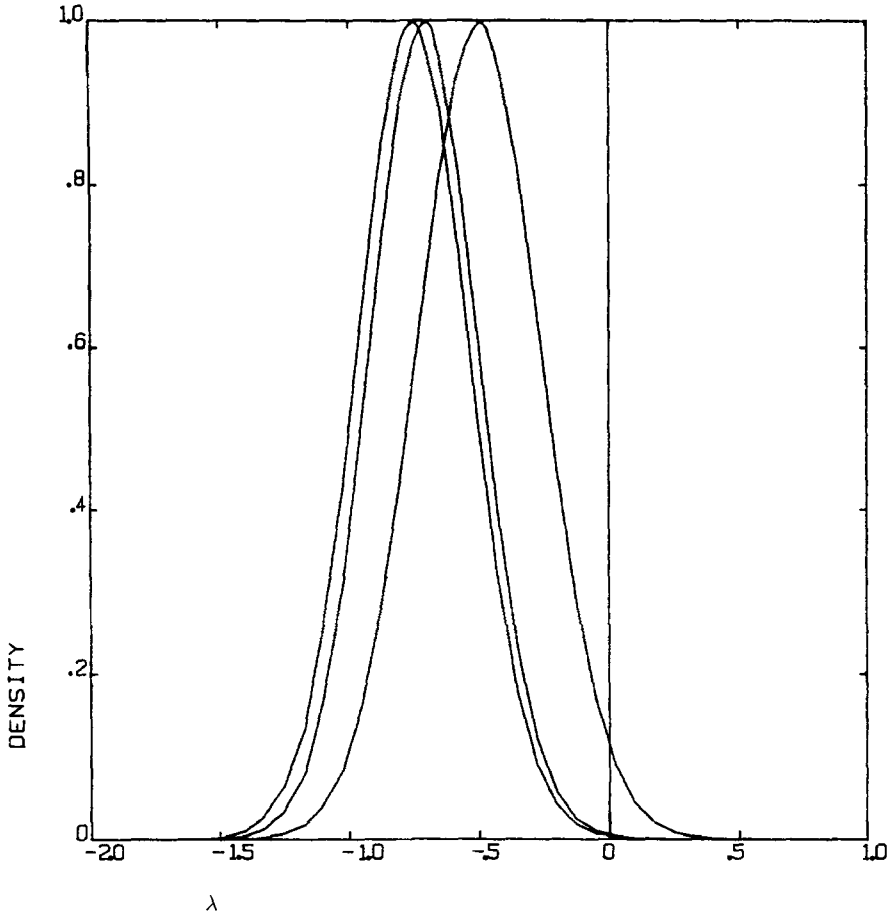

Fig. 1. Posterior densities of λ: left to right, the curves result from the Box-Cox, Sweeting, and present approaches.

that the investigator expects to see survival times of a few hours, say around five. He might then take reference values $v_1, \ldots, v_q$ each as 5, with $q = 6$ here for a main effects model, and the associated $a_1, \ldots, a_q$ as the x-vectors corresponding to the factorial effects. For v_{q+1} and v_{q+2} the values 6 and 7 will simply satisfy the monotonicity constraint with $r = 1$, say. These choices are then used in the $p(\lambda \mid y)$ from given above and the curve of Fig. 1 results. Evidently, the prior information has shifted the likely λ-values to the right a little with mode near $-1/2$. One can concoct values for the v's and a's to shift the curve back towards the others but, of course, that is not the purpose here.

Acknowledgements

I should like to thank the referees for their helpful comments which have led to many improvements.

REFERENCES

Apostol, T. M. (1957). *Mathematical Analysis*, Addison-Wesley, London.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion), *J. Roy. Statist. Soc. Ser. B*, **41**, 113–147.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion), *J. Roy. Statist. Soc. Ser. B*, **26**, 211–252.

Chaloner, K. M. and Duncan, G. T. (1983). Assessment of a Beta prior distribution: PM elicitation, *The Statistician*, **32**, 174–180.

Dawid, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion), *J. Roy. Statist. Soc. Ser. B*, **35**, 189–233.

Dickey, J. M. (1980). Beliefs about beliefs, a theory of stochastic assessments of subjective probabilities, *Bayesian Statistics* (eds. J. M. Bernardo *et al.*), University Press, Valencia.

Good, I. J. (1950). *Probability and the Weighing of Evidence*, Griffin, London.

Hartigan, J. (1964). Invariant prior distributions, *Ann. Math. Statist.*, **35**, 836–845.

Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions (with discussion), *J. Amer. Statist. Assoc.*, **70**, 271–289.

Kadane, J. B. (1980). Predictive and structural methods of eliciting prior distributions, *Bayesian Analysis in Econometrics and Statistics*, North-Holland, Amsterdam.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S. and Peters, S. C. (1980). Interactive elicitation of opinion for a normal linear model, *J. Amer. Statist. Assoc.*, **75**, 845–854.

Lindley, D. V. (1972). *Bayesian Statistics, A Review*, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

Lindley, D. V., Tversky, A. and Brown, R. V. (1979). On the reconciliation of probability assessments (with discussion), *J. Roy. Statist. Soc. Ser. A*, **142**, 146–180.

Novick, M. R. (1969). Multiparameter Bayesian indifference procedures, *J. Roy. Statist. Soc. Ser. B*, **31**, 29–64.

Pericchi, L. R. (1981). A Bayesian approach to transformations to normality, *Biometrika*, **68**, 35–43.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations, *J. Amer. Statist. Assoc.*, **66**, 783–801.

Sweeting, T. J. (1984). On the choice of prior distribution for the Box-Cox transformed linear model, *Biometrika*, **71**, 127–134.

Tversky, A. (1974). Assessing uncertainty (with discussion), *J. Roy. Statist. Soc. Ser. B*, **36**, 148–159.

Winkler, R. L. (1967). The assessment of prior distributions in Bayesian analysis, *J. Amer. Statist. Assoc.*, **62**, 776–800.