Local Prior Influence

by

Robert McCulloch

University of Minnesota
School of Statistics
Technical Report No. 477
June 1986

Summary

A simple procedure is developed for assessing the influence of the choice of prior on the resulting posterior and predictive distributions.

This work was supported by NIGMS-25271.

1. Introduction

In statistics we make decisions or inferences based on the data in hand and the model we have constructed. The model describes how particular aspects of the data relate to each other and how the data relate to parameters. In a Bayesian analysis probability enters as a method of quantifying our uncertainty about the model's relationships and parameter values.

In order to proceed we must specify precisely the model and data. Unfortunately elicitation of the probability measure and model features which truly represent current knowledge and beliefs can be difficult if not impossible. Even if exact determination were possible it may be reasonable to use a simpler and less exact rendering if the resulting inaccuracies are not too costly. In short there is often great uncertainty about the uncertainty.

In frequentist statistics robust and nonparametric methods are used to get around the difficulties of precise model specification. There is literature on nonparametric Bayes methods with Ferguson(1967) being a seminal work. However these approaches suffer from the same drawbacks as their frequentist counterparts in that they do not adapt easily to more complicated situations and violate the principle of parsimony. Trying to apply the ideas of robust statistics to the Bayes approach seems to be fraught with technical and conceptual difficulties. According to Huber(1981) the Bayesian approach to robustness "still lacks reliable guidelines on how to select the supermodel and the prior so that we end up with something robust". See Berger(1980) for some discussion and results. See also Krasker(1984) for an interesting result which extends the weak*

concept of robustness (again see Huber(1980)) to the Bayes setup.

We take a pragmatic approach. Taking reasonable care, a model is tentatively assumed and afterwards checked to determine whether minor perturbations of questionable model features influence the results in an important way. This idea of influence is one which easily extends to Bayesian analysis (see Johnson and Geisser 1982,1983,1985, and Geisser 1985). The idea is very simple and really consists of performing a sensitivity analysis. The assumptions about model and data which serve as inputs to the process are perturbed and the resulting changes in ouput are monitored. One well known application of this approach to linear regression is Cook's distance, see Cook(1977).

There are two basic parts of any sensitivity analysis of a process: (i) what are the interesting perturbations of the inputs? and (ii) what are meaningful measures of the consequent changes in ouput? In the context of frequentist statistics Cook(1986) describes a general measure for changes in ouput based upon the likelihood function and its asymptotic properties and notes that Cook's distance may be considered to be a special case of this procedure. Weisberg(1984) comments that the standard use of the likelihood measure provides a basis for a unified theory of diagnostic procedures and suggests that the resulting conceptual simplicity makes for a more useful approach than a toolbox of various robust procedures.

We apply these ideas in a Bayesian setup by considering the model and data assumptions to be input and the predictive and posterior distributions to be outputs. Traditionally statistics has emphasized parameters so the the posterior is an output of interest. Many authors have stressed the importance of the predictive distribution.

While parameters may in some cases be useful devices for describing the structure of the model, their values and the model itself are ultimately only of interest if they make useful predictions about potentially observable quantities. See for example Geisser(1971), Geisser(1980), and Zellner(1985).

Following Johnson and Geisser(1982), different predictive or posterior distributions resulting from perturbing model assumptions are compared by means of the Kullback-Leibler divergence (see chapter 1 of Kullback(1959)). Given the close relationship between the likelihood and the Kullback distance discussed in Akaike(1973), we may think of this procedure as analogous to Cook's use of the likelihood. Both the likelihood and the Kullback distance are thought of as being all purpose measures to be used in a variety of problems. For any particular problem there may be a natural measure that truly reflects the concerns of the investigator. The particular measure, like the model itself, may be difficult to determine so that the notion of a reasonable all purpose measure is useful.

The perturbation scheme must still be chosen. Cook(1986) notes that care must be taken in choosing the perturbation scheme if the results are to be reasonable. In this paper we shall concentrate on perturbing our choice of prior, although in general the approach may be applied to any model feature.

For convenience we often choose priors from standard families of distributions which are themselves indexed by parameters which we then call hyperparameters. Choosing a prior involves choosing a value for the hyperparameter. There is usually uncertainty associated with this choice. The Bayesian approach calls for a distribution to be placed on the hyperparameter. This may be more work than the investigator

wants to do. More importantly this additional refinement may have little impact on the final result given the data at hand. The approach taken here is to choose a reasonable value for the hyperparameter and then check afterwards that this choice is not overly influential. In this paper we develop a simple and general method for making this assessment.

In section 2 we introduce a method for calibrating the Kullback divergence. In section 3 we develop our method. Section 4 presents some simple examples and section 5 concludes the paper.

2. Calibration of the Kullback Divergence

We want to use the Kullback divergence to measure the difference between different posterior and predictive distributions arising from the use of different priors. Given that the Kullback divergence between two distributions is denoted by k, it is useful to have some scheme for deciding whether k is big or small.

Given two distributions P and Q with corresponding densities p and q with respect to the measure μ, the Kullback divergence between P and Q, $K(P,Q) = \int p \log(p/q)\, d\mu$.

Note that K depends on the order of its arguments, $K(P,Q) \neq K(Q,P)$ in general. The asymmetry is important with the interpretation that K(P,Q) is a canonical measure of the cost of predicting outcomes using Q when P is a more accurate description of the situation. Intuitively, we see that is a measure of the discrepancy between P and Q, log(p/q), averaged over the sample space where we average using the distribution P. See Kullback(1959).

If P and Q are discrete then K(P,Q) becomes $\Sigma\, p_i \log(p_i/q_i)$ where

p_i is the probability of the ith outcome under P and q_i is the probability of the ith outcome under Q.

In general, let (Ω,F) be a sample space and sigma field and P, Q be probability measures on (Ω,F). The following fundamental fact (see for example Renyi(1970), the appendix on information theory) underlies our calibration. Let $\Gamma = \{E_1,E_2,\ldots E_m\}$ be a measurable partition of Ω. That is each of the E_i belongs to F, their union is Ω, and the intersection of any two is the null set. Then if P is absolutely continuous with respect to Q so that there exists a function g such that dP = g dQ, we have,

$$\sup_{\Gamma} \sum_{i=1}^{m} P(E_i)\log(P(E_i)/Q(E_i)) = \int_{\Omega} \log(g)\, dP.$$

If $P = p\, d\mu$ and $Q = q\, d\mu$ with P absolutely continuous with respect to Q then $P = (p/q)\, dQ$ so that the integral becomes $\int p \log(p/q)\, d\mu$. Thus the meaning of the Kullback divergence is independent of the structure of the space since we may take any space and chop it up into discrete pieces and still have virtually the same Kullback divergence. Note that this is not true of the related concept of entropy. Since the Kullback divergence is an average, its meaning is also independent of the number of possible outcomes in the discrete setup.

Given the above background material we can now describe our calibration. Let A = { (P,Q) such that P and Q are probability measures on some pair (Ω,F) }. Now define a relation on A by $(P_1,Q_1) \sim (P_2,Q_2)$ if $K(P_1,Q_1) = K(P_2,Q_2)$, where K stands for the Kullback divergence. Clearly the relation is an equivalence relation so that we have a partition of A into equivalence classes. A pair of distributions may differ in many ways. When we choose a distance we choose a particular aspect or summary of all the ways a pair may

differ to key in on and ignore the rest. Thus, any two pairs in the same equivalence class are equally representative of the difference. Notice that here is where we need our fundamental fact given above: the meaning of the Kullback divergence is not dependent on the structrure of the underlying space. To calibrate the Kullback divergence we simply pick a pair from each equivalence class in such a way that all the chosen pairs differ in the same, simple, and meaningful way. Given any pair of distributions we simply look at the simple pair chosen from the same equivalence class to get an idea of how far apart the original pair are.

Now in order to calibrate the Kullback distance we need to choose a pair of distributions (P_k,Q_k) such that $K(P_k,Q_k) = k$, for each $k \in [0,\infty)$, since that is the range of K. The trick is to choose pairs that differ along one simple "dimension". We would like to be able to imagine attempting to predict based on our choice of Q_k while P_k is actually the correct description of the uncertainty. A simple and natural choice is $P_k = (1/2,1/2)$ and $Q_k = (1 - q(k),q(k))$ where the first entry is the probability that an event will occur and the second entry is the probability that the event will not occur. The function $q(k)$ is defined by the equation $k = K(P_k,Q_k)$. Since $K((1/2,1/2),(1 - q,q)) = -\log(4q(1 - q))/2$ we invert to obtain, $q(k) = (1 + (1 - e^{-2k})^{1/2})/2$.

As an example, suppose that P is normal with mean 0 and variance 5 and that Q is normal with mean 3 and variance 7. The Kullback divergence between P and Q is .67. Is .67 big or small? $q(.67)$ is about .93. This means that, as measured by the Kullback divergence, P and Q are just as different from each other as the assessment that an event will occur with probability .5 is from the assessment that an

event will occur with probability .93. If P is normal with mean 0 and variance 100 and Q is normal with mean 3 and variance 100 the Kullback divergence is .045. q(.045) = .64.

The following two tables show a range of q values with the corresponding k values. Given a value k for the Kullback divergence you find the q value from the table which corresponds most closely to k. Then the user decides how comfortable he would be predicting the outcome of an event assuming the probability that the event will occur is q when in fact it is .5.

For comparison we consider an alternative calibration obtained from normal distributions by choosing the μ such that the univariate normal distribution with mean 0 and variance 1 has a Kullback divergence of k from that with mean μ and variance 1. We solve $K(N(0,1),N(\mu,1))=k$. This results in $\mu(k) = (2k)^{1/2}$. It is also common in statistics to calculate p-values so in the following table for each q we give the corresponding k, $\mu(k)$, and $\mathrm{Prob}\{Z > \mu(k) \mid Z \sim N(0,1)\}$. The p-values in the third column of Table 1 and the Q values in the first column do not clearly contradict each other.

Table 1

Calibration of the Kullback Divergence with Normal Comparison

q	k	Prob{Z > μ(k) | Z~N(0,1)}	μ
.60	.02	.42	.20
.70	.09	.34	.42
.75	.14	.29	.54
.80	.22	.25	.67
.85	.34	.20	.82
.88	.43	.18	.93
.90	.51	.16	1.01
.91	.56	.15	1.06
.92	.61	.13	1.11
.93	.67	.12	1.16
.94	.75	.11	1.22
.95	.83	.10	1.29
.96	.94	.09	1.37
.97	1.08	.07	1.47
.98	1.27	.06	1.60
.99	1.60	.03	1.79
.995	1.96	.02	1.98

A finer table of q,k pairs is:

Table 2

Calibration of the Kullback Divergence

q	k	q	k	q	k	q	k	q	k
.50	.000	.60	.020	.70	.087	.80	.223	.90	.511
.52	.001	.62	.030	.72	.108	.82	.264	.92	.611
.54	.003	.64	.041	.74	.131	.84	.310	.94	.745
.56	.007	.66	.054	.76	.158	.86	.365	.96	.937
.58	.013	.68	.069	.78	.188	.88	.431	.98	1.273

3. Local Prior Influence

We assume that the prior is chosen from a class of distributions indexed by a finite dimensional parameter γ. Often priors are chosen from standard classes of distributions in which case the nature of γ is obvious. In particular it is often convenient to use conjugate priors. More generally we note that the elicitation process should involve asking a finite number of questions. In fact there is often uncertainty associated with the answers so that the answers themselves may be regarded as the parameter γ. We need to check that the conclusions of the analysis are not overly sensitive to the statements elicited from the investigator, that is, to the choice of γ.

As an example of parameters like γ consider Jaynes(1968) maximum

entropy priors. Given the constraints $\int g_i(\theta)\, p(\theta)\, d\nu(\theta) = c_i$ for $i=1,2,\ldots m$ we choose the prior p which maximizes the entropy of the distribution $p \cdot d\nu$. Clearly the c_i are based on the prior knowledge of the investigator. Here we have $\gamma = (c_1, c_2, \ldots c_m)$. Usually there is some uncertainty about the choice of the numbers c_i and this is not incorporated in our model. Often the maximum entropy prior gives reasonable results so that if the outcome of the analysis is not overly sensitive to the choice of the c_i we can be fairly confident in our procedure. Zellner(1977) has also proposed a method for choosing priors based on optimizing a functional of the prior subject to constraints. Zellner calls these priors maximal data information priors. As the name suggests, this method is designed to maximize the amount of information given by the data relative to that given by the prior. Again we want to make sure that the outcome is not overly sensitive to the choice of the constraints which determine the prior.

Let $\mathrm{pri}(\gamma)$, $\mathrm{post}(\gamma)$, and $\mathrm{pred}(\gamma)$ be the prior, posterior, and predictive distributions resulting from the choice of γ. Now suppose that an initial choice of γ_0 had been made for γ. We define the following three functions:

$$I(\gamma) = K(\mathrm{pri}(\gamma), \mathrm{pri}(\gamma_0)),$$

$$E(\gamma) = K(\mathrm{pred}(\gamma), \mathrm{pred}(\gamma_0)),$$

$$S(\gamma) = K(\mathrm{post}(\gamma), \mathrm{post}(\gamma_0)),$$

where K is the Kullback divergence. We want to make sure that there are not choices of γ such that I is small while E or S is large. Note that from our discussion in section 2 we know that even though the parameter space and the space in which our observations lie may be very different, the Kullback divergences between the different priors and the different predictive or posterior distribution are comparable.

If small changes in the prior correspond to large changes in the posterior or predictive distributions then the outcome of the analysis may be essentially determined by the investigator's ability to choose between priors which are very similar. There is however, some hope that the investigator can readily decide between priors that are distant without difficulty. Conversely, if large changes in the prior cause only small changes in the predictive or posterior distribution then we are reassured.

Since we are interested in small changes of the prior we consider γ close to γ_0. Following Cook(1986), we use Taylor expansions of I, E, and S to examine their properties local to γ_0. Since I, E, and S are 0 at γ_0 and positive everywhere, we have,

$$\nabla I(\gamma_0) = \nabla E(\gamma_0) = \nabla S(\gamma_0) = 0.$$

Letting $\gamma = \gamma_0 + \delta$, we have :

$$I(\gamma) \approx \delta^T D_2 I(\gamma_0)\delta/2,$$
$$E(\gamma) \approx \delta^T D_2 E(\gamma_0)\delta/2,$$
$$S(\gamma) \approx \delta^T D_2 S(\gamma_0)\delta/2.$$

The matrices $D_2I(\gamma_0)$, $D_2E(\gamma_0)$, $D_2S(\gamma_0)$ are actually familiar objects. In general if γ indexes a family of distributions so that $P(\gamma)$ is a distribution for each γ, then if we let $k(\gamma) = K(P(\gamma),P(\gamma_0))$, the Fisher information matrix evaluated at γ_0 equalls $D_2k(\gamma_0)$. So the three matrices $D_2I(\gamma_0)$, $D_2E(\gamma_0)$, $D_2S(\gamma_0)$ are the Fisher information matrices for the prior, predictive, and posterior families of distributions all of which are indexed by γ. See Kullback(1959).

In the following discussion we shall deal with the predictive

distribution. An analogous discussion applies to posterior distributions.

We look for values of γ close to γ_0 which correspond to the greatest change in the predictive distribution relative to the change in the prior by choosing δ to maximize,

$$\frac{\delta^T D_2 E(\gamma_0)\ \delta}{\delta^T D_2 I(\gamma_0)\ \delta} \qquad \text{where we assume } \|\delta\| = 1.$$

Let δ^* be the optimizing δ. As is well known δ^* is the eigenvector of the matrix $(D_2 I(\gamma_0))^{-1/2}\ (D_2 E(\gamma_0))^{-1/2}\ (D_2 I(\gamma_0))^{-1/2}$ corresponding to the largest eigenvalue. Let λ^* be the largest eigenvalue. To obtain δ^* and λ^* we need only the Fisher information matrices for the relevant families and an eigen analysis.

λ^* has a useful interpretation. If λ^* is close to or greater than 1, then for γ close to γ_0 changes in the prior result in comparable or greater changes in the predictive. This is clearly a dangerous situation of which the investigator needs to be aware. On the other hand if λ^* is close to 0 then we are reassured. Nearby priors result in predictive distributions which are much the same.

We can use the calibration of the Kullback divergence given in section 2 above to calibrate λ^*. Recall that the calibration of the Kullback divergence produced a function q(k) such that the Kullback divergence between the (1/2,1/2) distribution and the (1-q(k),q(k)) distribution equals k. To calibrate λ^* we consider the function $c^*(q) = q(\lambda^* k(q))$ where by k(q) we mean the value of k corresponding to q. We interpret c^* as follows: given a pair of priors whose Kullback divergence is calibrated by q, then locally, the Kullback divergence between the corresponding predictive distributions is calibrated by

$c^*(q)$. So if, for example, $c^*(.55) = .95$ there is a problem since the diffence between a probability of .5 and .55 for an event is a lot smaller than the difference between .5 and .95, so that nearby priors correspond to predictive distributions which are relatively far apart. Figures 3.1, 3.2, and 3.3 are graphs of c^* for $\lambda^* = .05$, .8 and 2. If $\lambda^* = .05$ we see from figure 3.1 that a q value of .75 for priors corresponds to a q value of about .56 for predictive distributions. Certainly .56 indicates a much less severe difference than .75. At the other extreme we have $\lambda^* = 2$ in figure 3.3. where a q value of .75 for prior distributions corresponds to a q value of about .85 for predictive distributions.

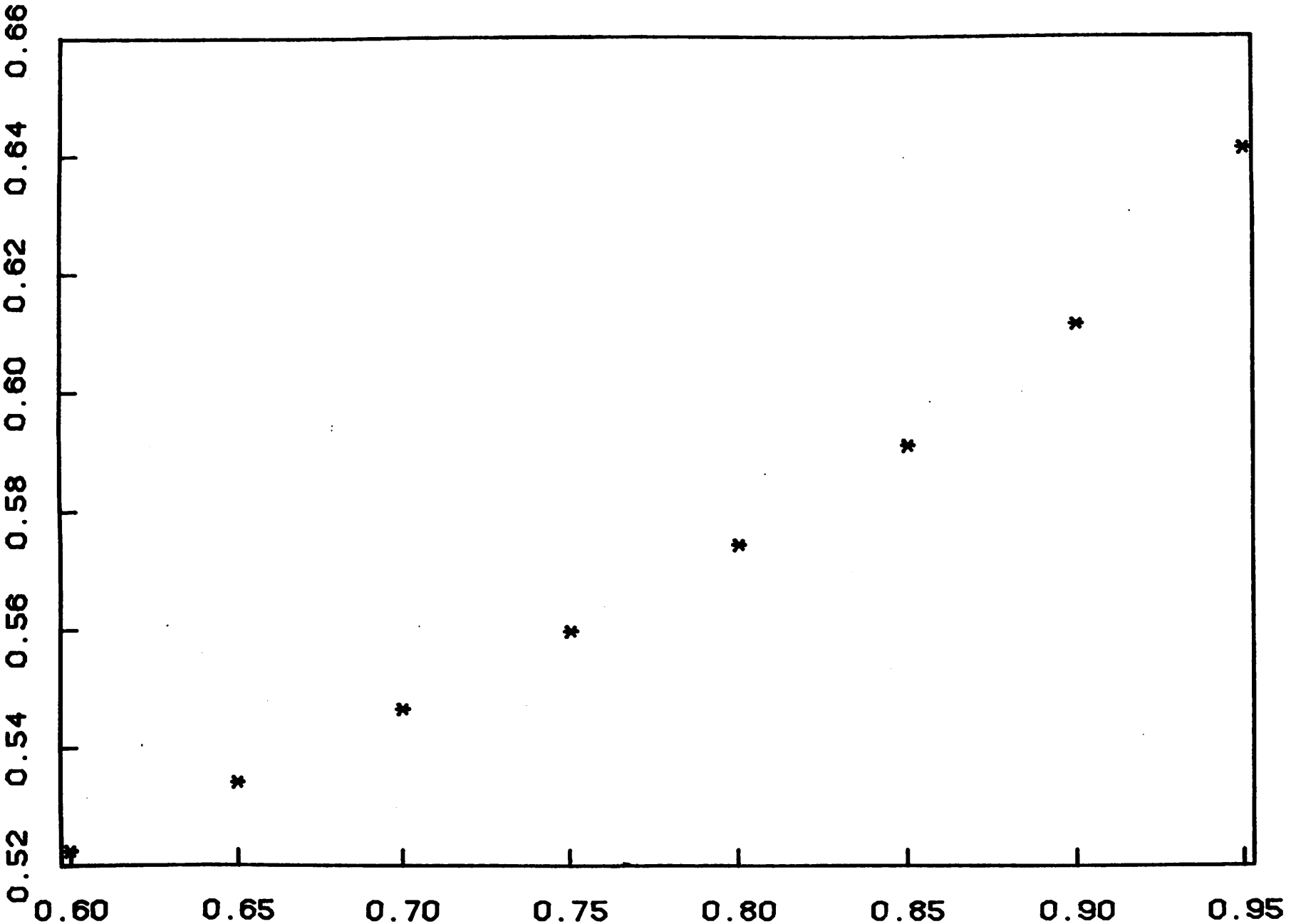

Figure 3.1: lamda star = .05

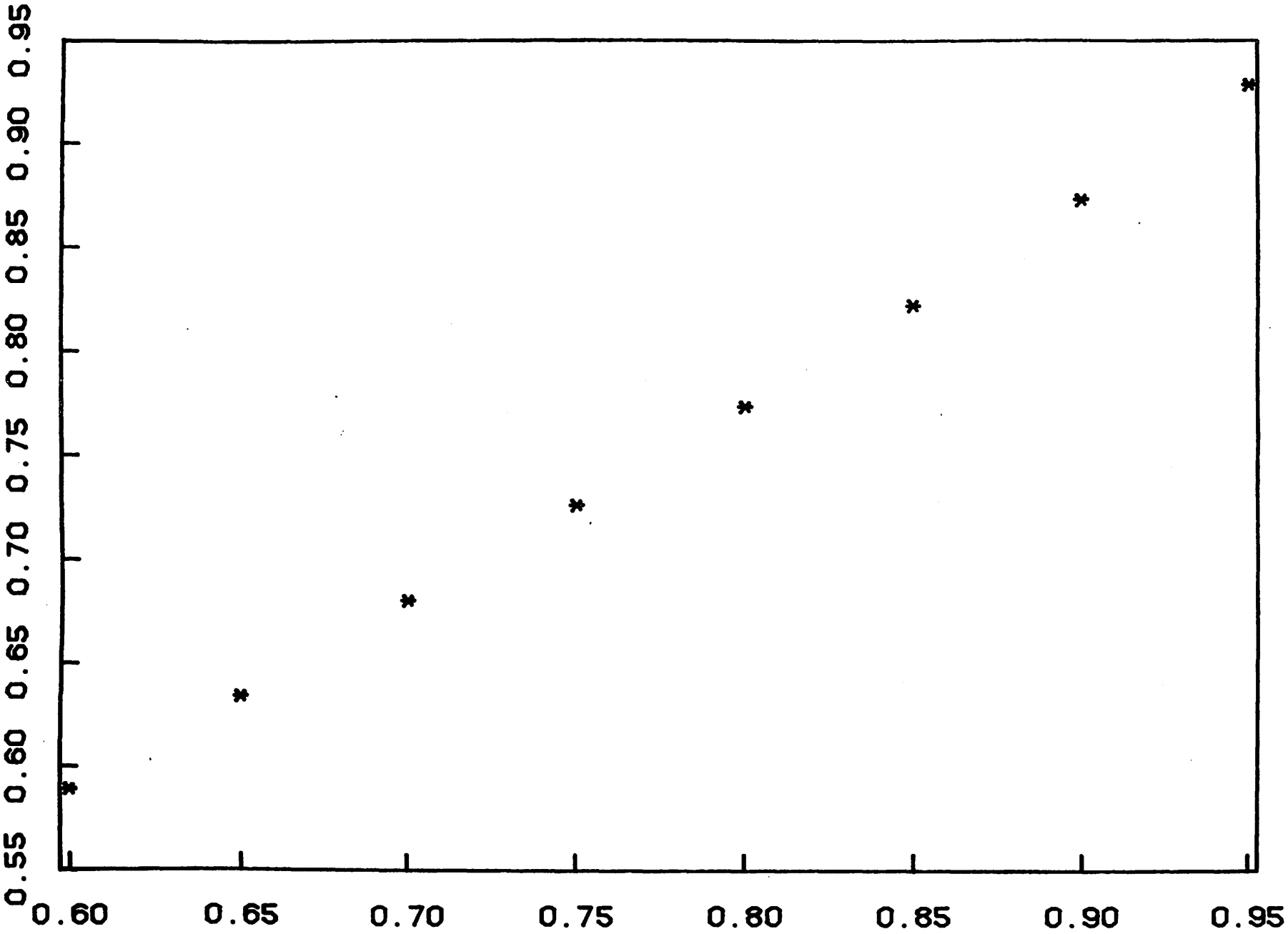

Figure 3.2: lamda star = .8

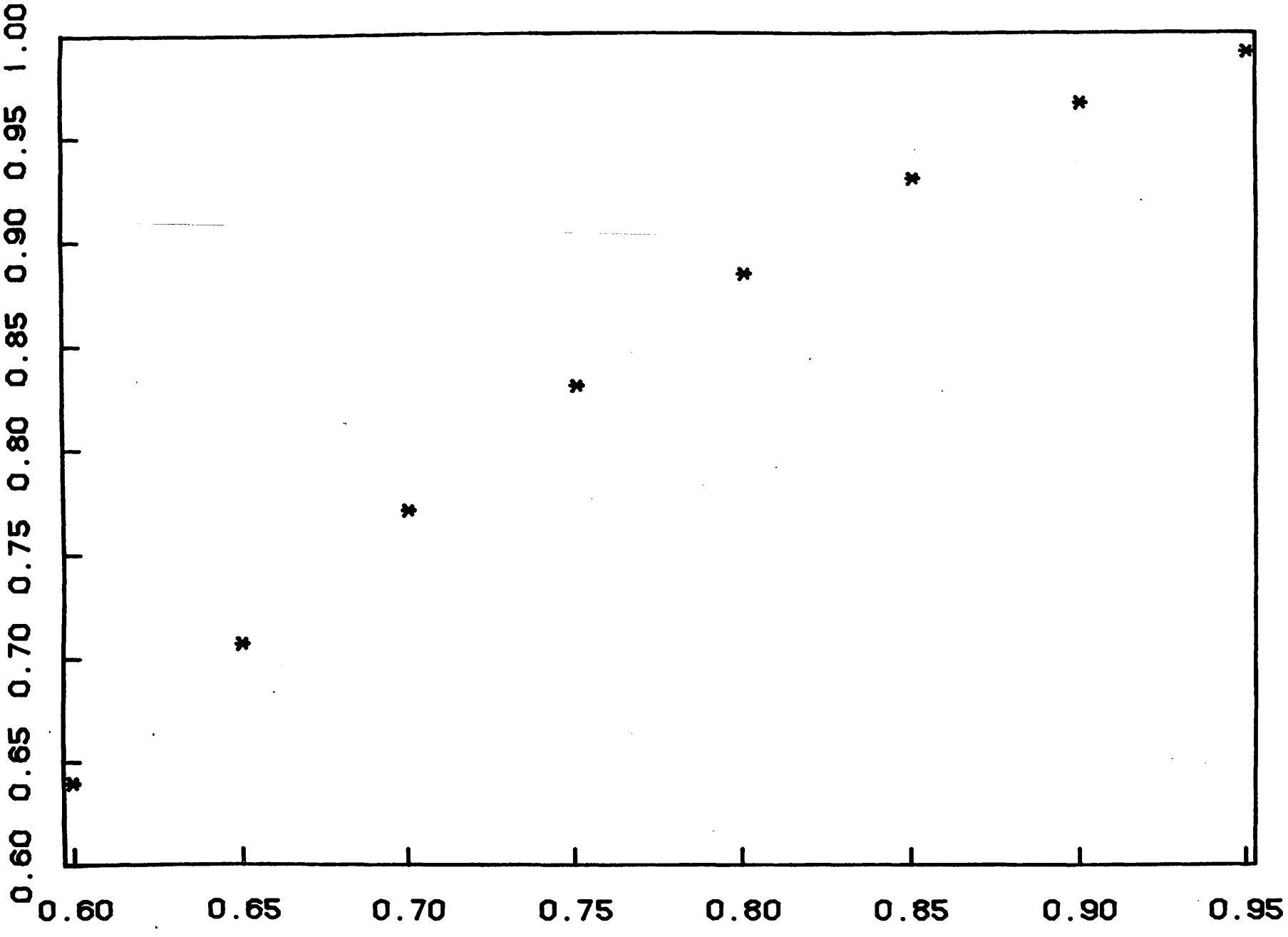

Figure 3.3: lamda star = 2

The vector δ^* is also of interest. For example if $\delta^* = (1,0,0)$ then the predictive distribution is most sensitive to the choice of the first component of γ. Any further elicitation should concentrate on this parameter. Also a graph of points of the form $(I(\gamma_0+t\delta^*),E(\gamma_0+t\delta^*))$ for t in an interval of 0 would be of interest. One might also graph points of the form $(q(I(\gamma_0+t\delta^*)),q(E(\gamma_0+t\delta^*)))$. In either case large values of the first coordinate should correspond to relatively small values of the second coordinate.

The above discussion has referred to the predictive distribution and the corresponding function E. Analogous remarks apply to the posterior distribution and the function S. We shall call the eigenvector and eigenvalue resulting from the analysis of the influence of the prior on the posterior δ^{**} and λ^{**} respectively. We also define $c^{**}(q) = q(\lambda^{**}k(q))$.

4. Some Examples

In this section we present two simple examples of the methods discussed in section 3. We will consider the exponential distribution and the multinomial distribution.

4.1 The Exponential Distribution

The exponential distribution has density $f(y|\theta) = \theta e^{-y\theta}$ with respect to Lebesgue measure on the positive real line. θ is a positive real parameter. We consider the conjugate family of priors having density ,

$$p(\theta|\rho,t) = e^{-\rho\theta}\,\theta^{t-1}\,\frac{\rho^t}{\Gamma(t)}\,,$$

where Γ is the gamma function and our parameter γ of the previous section is (ρ,t), a vector in R^2. Thus θ has a Gamma distribution with scale parameter ρ and degrees of freedom t.

Given the observation that $Y_i = y_i$ for $i=1,2...n$ where the Y_i are i.i.d. from our exponential distribution, let $s = \Sigma\, y_i$.

The posterior density is,

$$p(\theta|\rho,t,s) = \frac{(s+\rho)^{(n+t)}}{\Gamma(n+t)}\quad e^{-\theta(s+\rho)}\;\theta^{(n+t-1)}\;.$$

We shall consider the influence of the prior on the the posterior only. We need the Fisher information matrix of our prior family.

We shall denote the Fisher information matrix for the family of prior distributions at γ by $I_{pri}(\gamma)$. $I_{post}(\gamma)$ shall denote the Fisher information matrix for the family of posterior distributions at γ. Referring to the notation of section 3 we have, $I_{pri}(\gamma) = D_2 I(\gamma)$ and $I_{post}(\gamma) = D_2 S(\gamma)$.

$$\log p(\theta|\rho,t) = -\rho\theta + (t-1)\log\theta + t\log\theta - \log\Gamma(t).$$

$$\frac{\partial \log p(\theta|\rho,t)}{\partial\rho} = -\theta + \frac{t}{\rho}\quad,\quad \frac{\partial^2 \log p(\theta|\rho,t)}{\partial\rho^2} = -\frac{t}{\rho^2}$$

$$\frac{\partial^2 \log p(\theta|\rho,t)}{\partial t\,\partial\rho} = \frac{1}{\rho}\quad,\quad \frac{\partial \log p(\theta|\rho,t)}{\partial t} = \log\theta + \log\rho - \Psi(t)$$

where Ψ is the well known Psi function.

$\dfrac{\partial^2 \log p(\theta|,\rho,t)}{\partial t^2} = -\dot{\Psi}\,(t)$. To approximate $\dot{\Psi}$ we use the asymptotic expansion given in Abramowitz and Stegun(1972),

$$\dot{\psi}(t) \approx \frac{1}{t} + \frac{1}{2t^2} + \frac{1}{6t^3} - \frac{1}{30t^5} + \frac{1}{42t^7} - \frac{1}{30t^9}.$$

So,

$$I_{pri}(\rho_\diamond, t_\diamond) = \begin{bmatrix} \frac{t_\diamond}{\rho_\diamond^2} & -\frac{1}{\rho_\diamond} \\ -\frac{1}{\rho_\diamond} & \dot{\psi}(t_\diamond) \end{bmatrix}.$$

Clearly, $I_{post}(\rho_\diamond, t_\diamond) = I_{pri}(s + \rho_\diamond, n + t_\diamond)$.

<u>Example 1</u>: In this example we have $t_\diamond = 30$, sample size $n = 5$, y_i $i=1,2,3,4,5$ such that the sum of the y_i is 25, and $\rho_\diamond = 10$.

$$I_{post}(10,30) = \begin{bmatrix} .02857 & -.02857 \\ -.02857 & .02898 \end{bmatrix}$$

$$I_{pri}(10,30) = \begin{bmatrix} .3 & -.1 \\ -.1 & .03389 \end{bmatrix}$$

The matrix $I_{pri}(10,30)^{-1}\, I_{post}(10,30)\, I_{pri}(10,30)^{-1}$ has eigenvalues 23.43 and .0029818, so $\lambda^{**} = 23.43$. The eigenvector corresponding to 23.43 is (.24397,.96978) . λ^{**} is larger than 1, suggesting that the predictive distribution is sensitive to the choice of $\gamma_\diamond$. $c^{**}(.55) = .73$ and $c^{**}(.65) = .97$. This implies that having chosen the prior corresponding to $\gamma_\diamond$, there is a γ whose corresponding prior has a Kullback divergence from the original prior which is calibrated by .65. However, the corresponding posteriors have a Kullback divergence which is calibrated by .97. The investigator needs to be extremely sure of his prior since nearby priors correspond to significantly different results.

As suggested in section 3, we now plot the the Kullback distance between the posterior distributions against the Kullback distance between the priors where we change (ρ,t) by moving along the

eigenvector (.24397,.96978). This is figure 4.1. Clearly small changes in the prior result in large changes in the posterior. For comparison we plot the Kullback distances obtained by moving (ρ,t) along the eigenvector corresponding to the eigenvalue .003. This is figure 4.2. In this graph large changes in the prior are needed to obtain significant changes in the posterior as we would anticipate from a value of .003 for the eigenvalue. Note that in the case of two or more hyperparamters the eigenvectors give information about what apsects of the prior have the largest impact on the posterior.

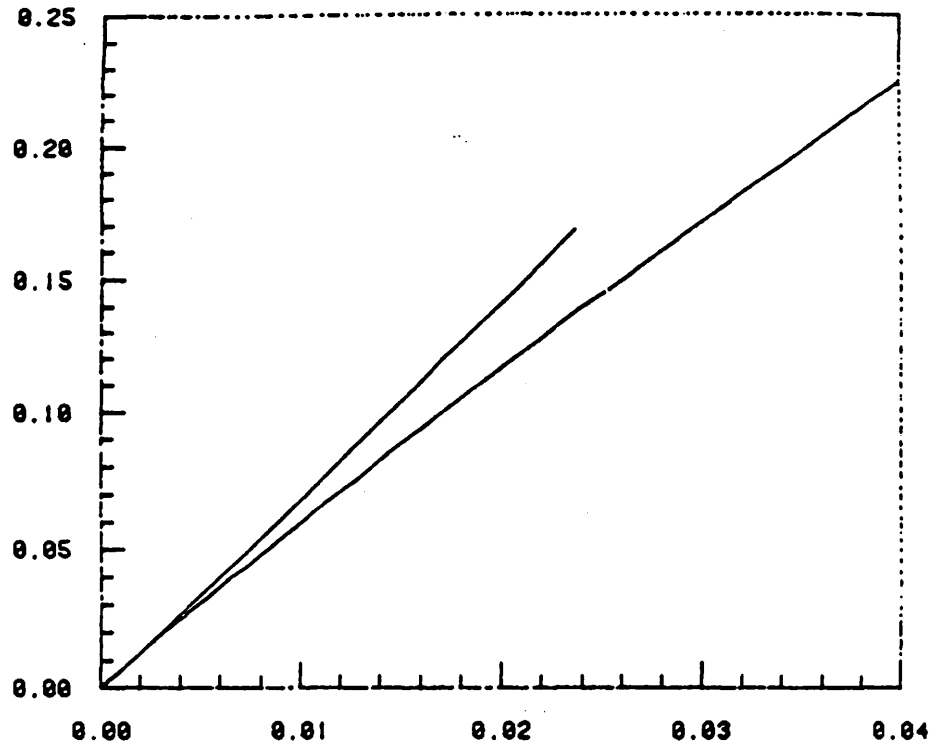

Figure 4.1 The effect of prior perturbation on the posterior distribution

Vertical axis: Kullback distance between posteriors

Horizontal axis: Kullback distance between priors

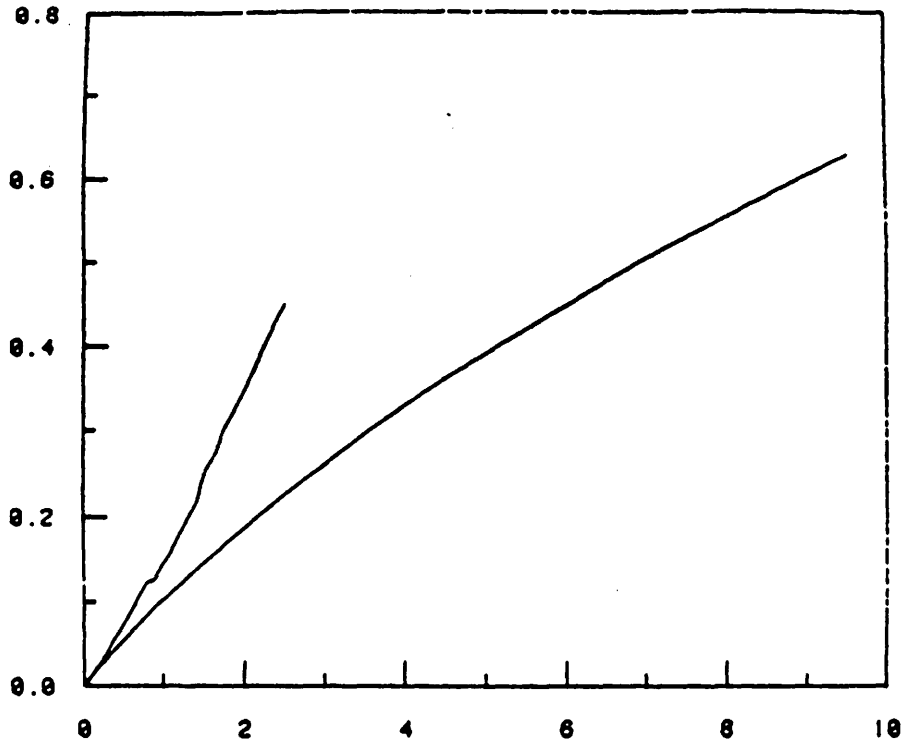

Figure 4.2: **The effect of prior perturbation on the posterior distribution**

Vertical axis: Kullback distance between posteriors

Horizontal axis: Kullback distance between priors

<u>Example 2</u>: In this example we have $t_o = 5$, sample size $n = 20$, y_i such that their sum is 80, and $\rho_o = 20$.

$$I_{post}(20,5) = \begin{bmatrix} .0025 & -.01 \\ -.01 & .04081 \end{bmatrix}$$

$$I_{pri}(20,5) = \begin{bmatrix} .0125 & -.05 \\ -.05 & .22132 \end{bmatrix}$$

The matrix $I_{pri}(20,5)^{-1} I_{post}(20,5) I_{pri}(20,5)^{-1}$ has eigenvalues .2 and .038. Since .2 is not close to 1 we expect that prior influence will not be a problem. Figure 4.3 is a plot obtained by moving (ρ,t) along (−.49946,.86634) which is the eigenvector corresponding to .2. Figure 4.4 is the plot obtained by moving (ρ,t) along (.86634,.49946) which is the eigenvector corresponding to .038. The graphs behave as we expect with neither indicating a problem with prior influence. A more formal evaluation of $\lambda^{**} = .2$ is provided by $c^{**}(.65) = .57$, which again indicates that the posterior is not overly sensitive to the choice of prior.

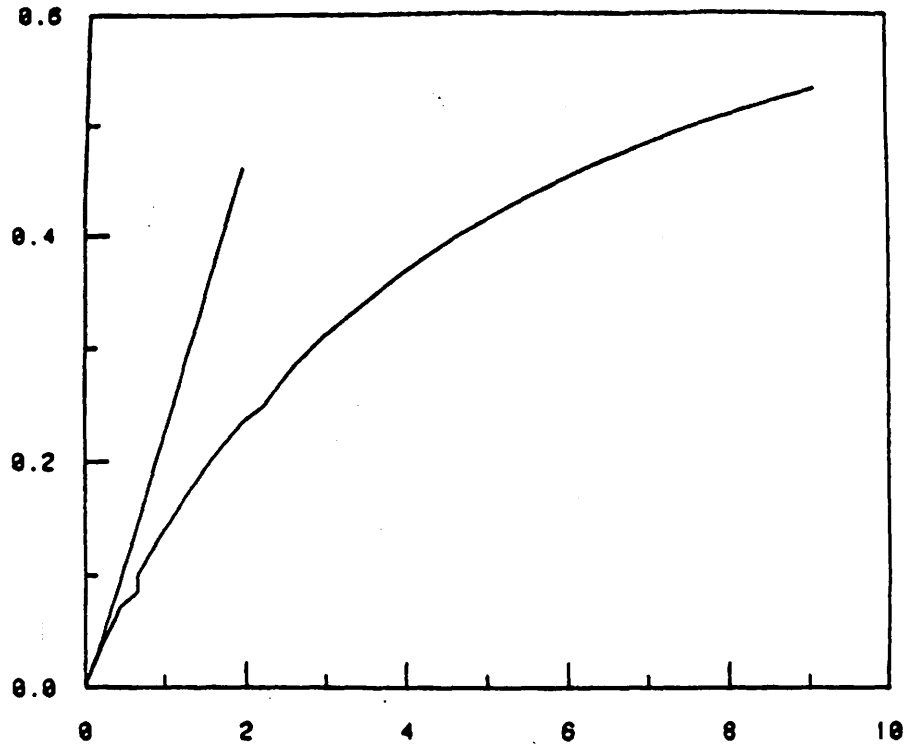

Figure 4.3: **The effect of prior perturbation on the posterior distribution**

Vertical axis: Kullback distance between posteriors

Horizontal axis: Kullback distance between priors

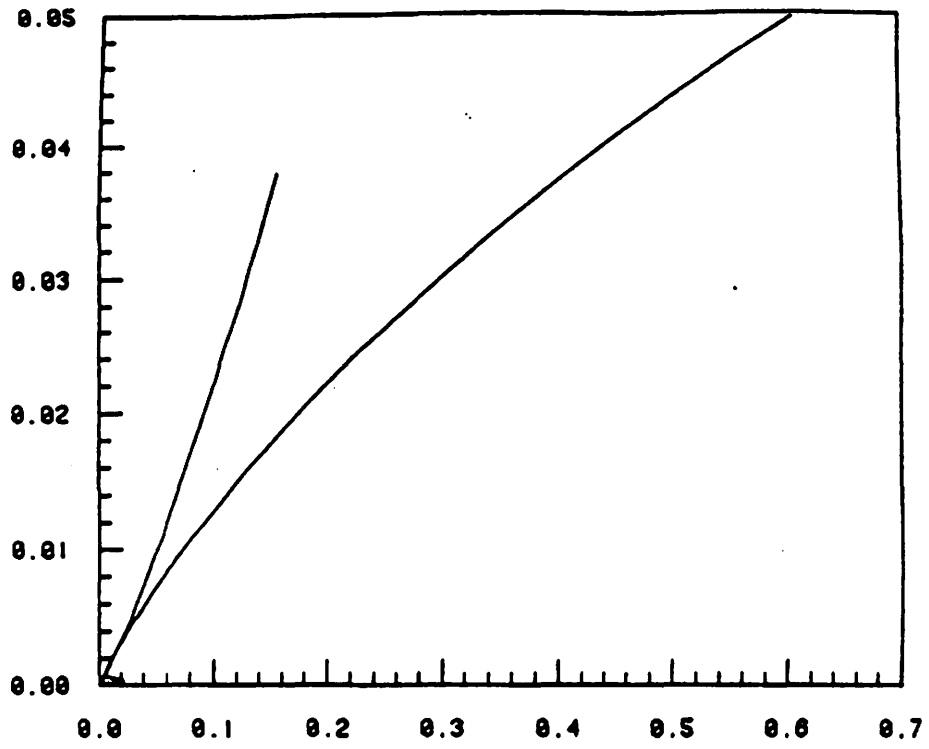

Figure 4.4: The effect of prior perturbation on the posterior distribution

Vertical axis: Kullback distance between posteriors

Horizontal axis: Kullback distance between priors

Section 4.2. Discrete Distributions

Let the random variable take on one of m possible values and p_i be the probability of the ith value. Given n i.i.d. observations let t_j be the number of observations taking on the jth value.

We use the Dirichlet prior which is also a conjugate prior, having prior density $p(p_1,p_2,\ldots p_m|\gamma) \propto \prod_{i=1}^{m} p_i^{\gamma_i}$ where $\gamma = (\gamma_1,\gamma_2,\gamma_3,\ldots\gamma_m)$. The prior density is with respect to Lebesgue measure on the set of p_i between 0 and 1 and summing to 1. The normalizing constant is $\frac{\Gamma(\gamma_s + m)}{\prod \Gamma(\gamma_i + 1)}$ where $\gamma_s = \Sigma\, \gamma_i$.

The posterior density $p(p_1,p_2,\ldots p_m|\gamma,D) \propto \prod p_i^{(\gamma_i + t_i)}$,
Where D stands for the observed data.
Let $h(i|\gamma,D)$ be the predictive probability that an as yet unobserved variabale takes on the ith value. Then,

$$h(i|\gamma,D) = \frac{\gamma_i + t_i + 1}{\gamma_s + n + m}.$$

We now have our three families of distributions indexed by the parameter γ: the prior, the posterior, and the predictive. The Fisher information matrices for these three families are as follows.

$$(I_{pri}(\gamma))_{kj} = \begin{cases} -\dot{\psi}(\alpha_s + m) & k \neq j \\ \dot{\psi}(\alpha_j + 1) - \dot{\psi}(\alpha_s + m) & k = j. \end{cases}$$

Clearly, $I_{post}(\gamma) = I_{pri}(\gamma + t)$, where $t = (t_1,t_2,\ldots t_m)$.
For the predictive family we have,

$$(I_{pred}(\gamma))_{kj} = \frac{-1}{(\alpha_s + n + m)^2} \quad \text{if} \quad k \neq j$$

and,

$$\frac{1}{(\alpha_s + n + m)^2} \quad \frac{(\alpha_s - \alpha_i + n - t_i + m - 1)}{(\alpha_i + t_i + 1)} \qquad \text{if } k = j.$$

$I_{pred}(\gamma)$ is the Fisher information matrix for the family of predictive distributions indexed by γ.

<u>Example 3</u>: We consider an example where there are three possible outcomes. We have ten observations resulting in $t = (t_1, t_2, t_3) = (2,4,4)$. For our prior hyperparameter set we choose $\gamma_0 = (2,1,1)$. The resulting Fisher information matrixes are,

$$I_{pred}(\gamma_0) = \begin{bmatrix} .00830 & -.00346 & -.00346 \\ -.00346 & .00634 & -.00346 \\ -.00346 & -.00346 & .00634 \end{bmatrix}$$

$$I_{post}(\gamma_0) = \begin{bmatrix} .16074 & -.06059 & -.06059 \\ -.06059 & .12074 & -.06059 \\ -.06059 & -.06059 & .12074 \end{bmatrix}$$

$$I_{pri}(\gamma_0) = \begin{bmatrix} .24139 & -.15355 & -.15355 \\ -.15355 & .49137 & -.15355 \\ -.15355 & -.15355 & .49137 \end{bmatrix}$$

From the eigen analysis for the predictive case we obtain $\lambda^* = .04$ and $\delta^* = (.996, -.06, -.06)$ is the corresponding eigenvector. $c^*(.65) = .53$ indicating that the influence on the predictive distribution is not significant. Figure 4.5 is a plot of the Kullback divergence between predictive distributions against the Kullback divergence between the corresponding priors. As expected given the value of λ^*, the predictives are close together relative to the difference between the priors.

From the eigen analysis for the posterior case we obtain $\lambda^{**} = .84$ and $\delta^{**} = (.999, -.019, -.019)$ is the corresponding eigenvector.

$c^{**}(.65) = .64$ which suggest that the posterior distribution is sensitive to the choice of prior. Figure 4.6 is a plot of the Kullback divergence between posterior distributions against the Kullback divergence between the corresponding priors. For small values of k, the Kullback divergence, the difference between the posteriors and the difference between the priors is comparable.

Note that both δ^* and δ^{**} suggest that the first component of γ is responsible for its influence.

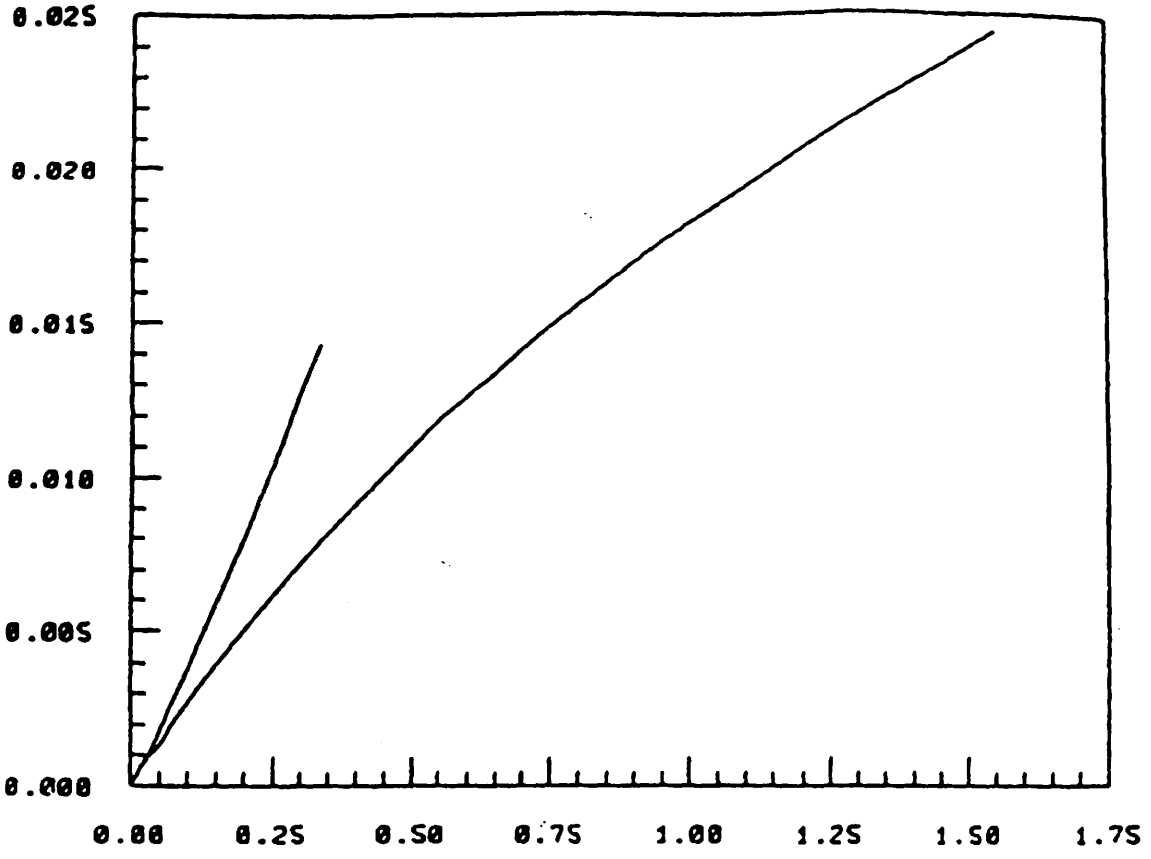

Figure 4.5: **The effect of prior perturbation on The predictive distribution**

Vertical axis: Kullback distance between predictives

Horizontal axis: Kullback distance between priors

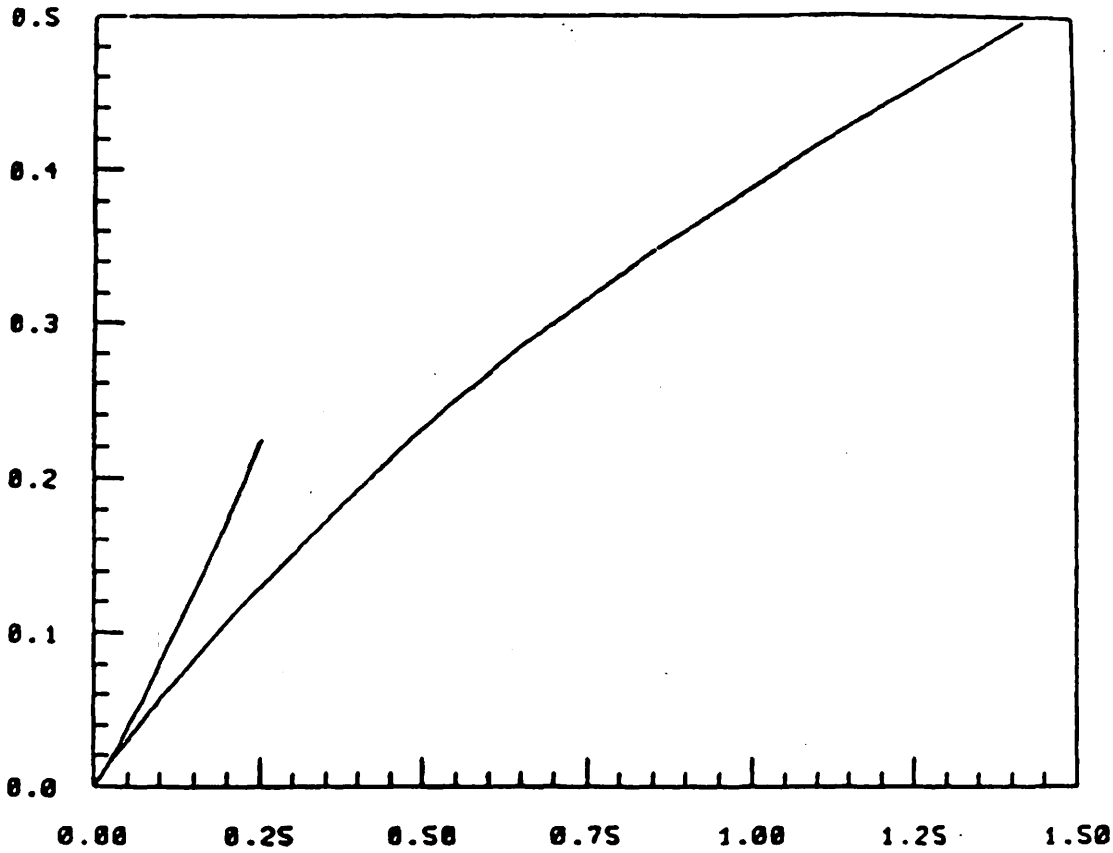

Figure 4.6: **The effect of prior perturbation on The posterior distribution**

Vertical axis: Kullback distance between posteriors

Horizontal axis: Kullback distance between priors

5. Conclusion

For some models general noninformative priors have been developed which give reasonable results. The seminal work is Jeffreys(1939). In general however, the Bayesian approach suggests that the prior should reflect the insight of the investigator. On the other hand exhaustive elicitation of priors for more than a few parameters (and perhaps only one) does not seem to be practical. A compromise is nessessary. Elicit an amount of information that the investigator is willing and able to provide. Using the provided information as constraints, construct a reasonable prior. Finally, conduct a sensitivity analysis to make sure that the information provided by the investigator is not locally influential in the sense of section 3. This final step is nessessary because there is usually uncertainty associated with the elicited input which is not incorporated into the model. The method introduced in section 3 is a simple and general tool for assessing the local influence of the elicited input. The method requires only Fisher information matrices and an eigen analysis. Two simple examples have been presented. In section 2 a calibration of the Kullback divergence has been given. This calibration makes the Kullback divergence more useful for sensitivity analyses.

Acknowledgment

The author thanks Seymour Geisser, Dennis Cook, and Jim Hodges for many helpful comments. The author thanks Professor Geisser for many things.

This work was funded by the Social Sciences and Humanities Research Council of Canada and NIH grant NIGMS 25271.

References

Abramowitz, M. and Stegun, I. (1972). Handbook of Mathematical Functions. New York, Dover

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. 2-nd International Symposium on Information Theory. Edited by Petrov and Csaki. Publishing House of the Hungarian Academy of Sciences. 267-281.

Berger, J. (1980). Statistical Decision Theory. New York, Springer-Verlag

Cook, R.D. (1977). Detection of influential observations in linear regression. Technometrics. 19, 15-18.

Cook, R.D. (1986). Assessment of local influence. J. Royal Stat. Soc. Series B, forthcoming.

Ferguson, T.S. (1973). A bayesian analysis of some non-parametric problems. Annals of Math. Stat. 1, 209-230.

Geisser, S. (1971). The inferential use of predictive distributions. Foundations of Statistical Inference. Edited by Godambe and Sprott. Toronto, Holt, Rinehart, and Winston. 456-466.

Geisser, S. (1980). A predictivistic primer. Bayesian Analysis in Econometrics and Statistics. chapter 24, edited by Arnold Zellner. North-Holland.

Geisser, S. (1985). On the prediction of observables: a selective update. Bayesian Statistics 2. 203-230. edited by J.M. Bernando, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith. North-Holland.

Huber, P.J. (1981). Robust Statistics. New York, Wiley.

Jaynes, E.T. (1968). Prior Probabilities. IEEE Transactions on Systems Science and Cybernetics. SSC-4, pp. 227-41. New York Institute of Electrical and Electronic Engineers.

Jeffreys, H. (1939). Theory of Probability. London. Oxford University Press

Johnson, W. and Geisser, S. (1982). Assessing the predictive influence of observations. Statistics and Probability: Essays in Honor of C. R. Rao. 343-358. edited by Kallianpur, Krishnaiah, and Ghosh. Amsterdam:North-Holland.

Johnson, W. and Geisser S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. J. Amer. Statist. Assoc. 78, 137-144.

Johnson, W. and Geisser, S. (1985). Estimative influences measures for the multivariate general linear model. J. of Staist. Planning and Inference. 11, 33-56.

Krasker, W. S. (1984). A note on selecting parametric models in bayesian inference. Annals of Math. Stat. 12, 751-757.

Kullback, S. (1959). Information Theory and Statistics. New York Wiley.

Renyi, A. (1970). Probability Theory. North-Holland.

Weisberg, S. (1984). Robustness and Diagnostics: Black Box or Pandora's Box? Paper presented at the American Statistical Association Annual Meeting August 14, 1984.

Zellner, A. (1977). Maximal data information prior distributions. New Developments in the Applications of Bayesian Methods. edited by A. Aykac and C. Brumat. 211-232, North-Holland.

Zellner, A. (1985). Bayesian analysis in econometrics. Tech. Report, H.G.B. Alexander Research Foundation, Graduate School of Business, University of Chicago.