Diagnostics, Divergences and Perturbation Analysis

by

Seymour Geisser*
University of Minnesota

Technical Report No. 530

August 1989

*This work was sponsored in part by an NIH grant GMS-25271.

0. Introduction

Case deletion diagnostics for influential observations were introduced by Cook (1977, 1979) for linear regression problems. Previously, among other things case deletion procedures had been used in a predictive sample reuse context for the purpose of selecting or verifying the "best" of several entertained models for data Geisser (1974, 1975), Lee (1971), Lee and Geisser (1975) and Stone (1974) under the rubric of cross-validatory choice.

In a frequentist context the case deletion procedure of Cook became a special case of perturbation diagnostics Cook (1986) which now used twice the difference of the maximized log likelihood of a standard model with that of a perturbed model with particular attention paid to local behavior of this diagnostic.

In this paper we shall review a Bayesian approach to case deletion analysis and provide a complete Bayesian decision theoretic approach to perturbation analysis when such is possible and when not feasible suggest other methods closely associated that can be implemented.

1. Bayesian Case Deletion Diagnostics

Case deletion diagnostics in a Bayesian analysis were proposed by Johnson and Geisser (1982) to ascertain the influence on prediction of both individual observations and groups of them. The Kullback-Leibler divergence for ranking influential observations was found to be useful in this context. An application to normal linear regression was given by Johnson and Geisser (1983). These methods which include multiple case deletions were also used for the multivariate normal general linear model for influential observations with respect to parameter estimation Johnson and Geisser (1985). Geisser (1980) suggested a conditional predictive ordinate diagnostic (see also Geisser (1985), and Smith and Pettit (1985) in this regard) as well as several other discordancy diagnostics, Geisser (1987) based on predictive distributions.

A simple Bayesian approach to case deletion is as follows for parametric estimation:

First find the posterior distribution with respect to the set of unknown parameters θ, with and with x_i the i^{th} observation

$$\mathcal{P}(\theta|x^{(N)}) \propto L(\theta|x^{(N)})\mathcal{P}(\theta)$$

$$\mathcal{P}_i(\theta|x_i^{(N-1)} \propto L(\theta|x_i^{(N-1)})\mathcal{P}(\theta)$$

where $X^{(N)} = (X_1,\ldots,X_N)$ and $X_i^{(N-1)}$ is $X^{(N)}$ with X_i deleted and $\mathcal{P}(\theta)$ the prior density for θ.

If θ_1 is of interest where $\theta = (\theta_1,\theta_2)$ then find

$$\mathcal{P}(\theta_1|x^{(N)}) = \int\mathcal{P}(\theta|x^{(N)})d\theta_2$$

and similarly

$$\wp_{(i)}(\theta_1|x_i^{(N-1)}) = \int\wp_{(i)}(\theta|x_i^{(N-1)})d\theta_2.$$

Next define a diagnostic which indicates the change in posteriors when the observation x_i is added. For variations see Johnson and Geisser (1982). One such diagnostic which has been useful is the Kullback divergence

$$K_i = K(\wp_{(i)},\wp) = E\left[\log \frac{\wp_{(i)}}{\wp}\right]$$

where the expectation is with respect to the first argument in the parenthesis. This ranks the observations in order of influence the largest being the most influential. If interest is focused on all of the parameters jointly rather than a subset of them one uses

$$K_i(\wp_{(i)}(\theta|x_i^{(N-1)}),\wp(\theta|x^{(N)}))$$

to rank the influence of the observations.

When interest is focused on prediction then one calculates the predictive density of a future set $X_{(M)}$

$$f(x_{(M)}|x^{(N)}) = \int f(x_{(M)}|\theta)\wp(\theta|x^{(N)})d\theta$$

and the analogue with X_i deleted

$$f_{(i)}(x_{(M)}|x_i^{(N-1)}).$$

Then one can use

$$K_i = K(f_{(i)},f)$$

to rank the relative influence of the observations with respect to the prediction of $X_{(M)}$.

2. Multiple Linear Regression

The initial situation for which case deletion arose was linear regression Cook (1977) and we shall take this up to give a Bayesian analogue. Consider a normal linear regression situation where

$$\begin{aligned} Y &= X\beta + e, & e &\sim N(0,\sigma^2 I) \\ Y' &= (Y_1,\ldots,Y_N), & e' &= (e_1,\ldots,e_N) \\ x_i' &= (x_{i1},\ldots,x_{ip}), & \beta' &= (\beta_1,\ldots,\beta_p) \end{aligned}$$

and

$$X = \begin{pmatrix} x_{11} & x_{12}\cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{N1} & x_{N2}\cdots & x_{Np} \end{pmatrix} = \begin{pmatrix} x_1' \\ \vdots \\ x_N' \end{pmatrix}$$

with assumed prior density for β and σ^2, say $\mathcal{P}(\beta,\sigma^2)$.

The first step in assessing the influence of individual observations with regard to the estimation of β alone, say, is the computation of the posterior densities $\mathcal{P}(\beta) = \mathcal{P}(\beta|y,X)$ and $\mathcal{P}_{(i)}(\beta) = \mathcal{P}(\beta|y_{(i)},X_{(i)})$ where $X_{(i)}$ is X with the

i^{th} row deleted and similarly for $y_{(i)}$. Next we compute,

$$K_i(\beta) = E[\ln \mathcal{P}_{(i)}(\beta) - \ln \mathcal{P}(\beta)].$$

It is easy to show that

$$K_i(\beta,\sigma^2) = K_i(\sigma^2) + E[K_i(\beta|\sigma^2)]$$

where

$$K_i(\sigma^2) = E[\ln \mathcal{P}_{(i)}(\sigma^2) - \ln \mathcal{P}(\sigma^2)],$$

and

$\mathcal{P}_{(i)}(\sigma^2)$ and $\mathcal{P}(\sigma^2)$ refer to $\mathcal{P}_{(i)}(\sigma^2|y_{(i)},X_{(i)})$ and $\mathcal{P}(\sigma^2|y,X)$ respectively; similarly

$$K_i(\beta|\sigma^2) = E[\ln \mathcal{P}_{(i)}(\beta|\sigma^2) - \ln \mathcal{P}(\beta|\sigma^2)]$$

and $K_i(\beta|\sigma^2)$ above is averaged over the density $\mathcal{P}_{(i)}(\sigma^2)$. This partition often helps to pinpoint the sources of influence.

For prediction it is necessary to calculate the predictive distribution of Z, the M×1 future vector to be observed for a given W, an M×p matrix, i.e.

$$Z = W\beta + e \quad e \sim N(0,\sigma^2 I)$$

with and without y_i. Consequently,

$$f_{(i)}(z) = f_{(i)}(z|W,y_{(i)},X_{(i)}) = \int f(z|W,\beta,\sigma^2)\wp_{(i)}(\beta,\sigma^2)d\beta d\sigma^2$$
$$f(z) = f(z|W,y,X) = \int f(z|W,\beta,\sigma^2)\wp(\beta,\sigma^2)d\beta d\sigma.$$

One then calculates

$$K_i(Z) = E(\ln f_{(i)}(Z) - \ln f(Z)].$$

If W is unknown but can be assigned probabilities over some restricted space then this can be incorporated into the assessment, or alternatively the worst possible situation may be considered. If this is not the case, it has been found useful to set W=X, i.e. to essentially ascertain the effect of predicting back on the original set of independent variables as indicative of an overall assessment. The details of this procedure are given by Johnson and Geisser (1982, 1983).

For the purpose of demonstration we use the 'non-informative' prior

$$\wp(\beta,\sigma^2) \propto \frac{1}{\sigma^2}$$

as this case is the closest to a frequentist analysis. Let x_i' be the i^{th} row of X then define

$$v_i = x_i'(X'X)^{-1}x_i, \quad (N-p)s^2 = (y-\hat{y})'(y-\hat{y}),$$

$$\hat{\beta} = (X'X)^{-1}X'y, \hat{y} = X\hat{\beta}, \hat{y}_i = x_i'\hat{\beta},$$

$$t_i^2 = \frac{(\hat{y}_i - y_i)^2}{(N-p)s^2(1-v_i)} .$$

Using these results we can calculate the various influence divergences previously defined. First we obtain $2K_i(\beta, \sigma^2)$ which is the sum of the following two expressions,

$$2K_i(\sigma^2) = C_1 + (N-1-p)t_i^2(1-t_i^2)^{-1} + (N-p)\ln(1-t_i^2)$$

which reflects a discrepancy in the estimation of σ^2 and

$$2E[K_i(\beta|\sigma^2)] = C_2 + (N-1-p)\frac{v_i t_i^2}{(1-v_i)(1-t_i^2)} + \frac{v_i}{1-v_i} + \ln(1-v_i)$$

whose second term is proportional to Cook's distance where C_1 and C_2 are constants independent of the deleted observation. Although an explicit expression of $K_i(\beta)$ is not obtainable, the following approximation, based on the Kullback-Leibler divergence of the closest multivariate normal approximation to a multivariate student distribution, Johnson and Geisser (1983) should be adequate,

$$2\hat{K}_i(\beta) = \frac{(N-p-2)v_i}{1-v_i}t_i^2 + \ln(1-v_i)$$

$$+ p\left[\frac{N-p-2}{N-p-3} + \ln\frac{N-p-3}{N-p-2} - 1 - \ln(1-t_i^2) - \frac{t_i^2(N-p-2)}{N-p-3}\right]$$

$$- \frac{v_i}{1-v_i}\left[\frac{(N-p-2)(t_i^2-1)}{(N-p-3)}\right].$$

The components of $2\hat{K}_i(\beta)$ reflect a discrepancy in the point estimate for β multiplied by a leverage measure (Cook's distance); a discrepancy in the shapes of the ellipsoidal posterior regions for β; and a discrepancy in the dispersion.

For the predictive divergence a similar closest multivariate normal approximation to a multivariate student distribution is utilised. This results in

$$2\hat{K}_i(Z) = \frac{(N-p-2)v_i t_i^2(N-p-4)}{2(1-v_i)(N-p-3)} + \frac{v_i(N-p-2)}{2(1-v_i)(N-p-3)} - \ln\left[1 + \frac{v_i}{2(1-v_i)}\right] \tag{4.1}$$

$$+ N\left[\frac{N-p-2}{N-p-3}(1-t_i^2) - \ln\frac{N-p-2}{N-p-3}(1-t_i^2) - 1\right].$$

Here the first term is proportional to Cook's distance, the second and third terms together reflect leverage while the last term reflects the discrepancy in observational error.

These results are extended to sets of k case deletions and multivariate normal regression, Johnson and Geisser (1983, 1985).

3. A General Bayesian Approach

The approach previously discussed involves only case deletions i.e. a concern with one or few observations that may greatly influence an analysis for one reason or another. In particular there is the concern or possibility of an aberrant observation. A more important concern is that the analysis itself may depend critically on the modeling assumptions which include the prior, the likelihood and in some contexts the loss function as well. Box (1980) proposed methods for testing the model and perturbing it via a discrepancy parameter. Cook (1986) proposed a frequentist perturbation diagnostic based on a likelihood displacement. Geisser (1986, 1987) proposed a general Bayesian approach for perturbation analyses and further developed it Geisser (1988).

4. Types of Perturbation

Frequentist perturbation analysis assumes that some "standard" likelihood $f(x^{(N)}|\theta,\omega_0)$ may be perturbed to $f(x^{(N)}|\theta,\omega)$ for $\omega\epsilon\Omega$ when ω is the index that governs the perturbation and $\omega=\omega_0$ is the standard. In this regard there are a number of possible types of perturbation which we shall list for convenience but with the caution that they are not necessarily mutually exclusive.

4.1 Hyperparameters

Some examples are:

a. $$f(x^{(N)}|\theta,\omega) \propto \prod_{j=1}^{N} \left(1 + \frac{(x_j-\mu)^2}{\omega\,\sigma^2}\right)^{-(\omega+1)/2}, \qquad \omega \geq 1,\ \theta = (\mu,\sigma^2)$$

where the standard is $\omega\to\infty$, i.e. the normal distribution and the most deviant $\omega=1$, the Cauchy distribution.

b. $$f(x^{(N)}|\theta,\omega) \propto \prod_{j=1}^{N} \frac{k(\omega)}{\sigma} e^{-\frac{1}{2}\left|\frac{x_j-\mu}{\sigma}\right|^{\frac{2}{1+\omega}}} \qquad -1 < \omega \le 1$$

$\omega_0 = 0$ is the standard normal.

c. $$f(x^{(N)}|\theta,\omega) \propto \prod_{j=1}^{N} \omega\, x_j^{\omega-1}\, e^{-\theta x_j^{\omega}} \qquad \omega > 0$$

$\omega_0 = 1$ is the standard exponential.

4.2 Mixtures

Here

$$f(x^{(N)}|\theta,\omega) \propto \prod_{j=1}^{N} [\omega f_1(x_j|\alpha) + (1-\omega) f_2(x_j|\beta)], \quad 0 \le \omega \le 1$$

and $\omega=1$ represents the standard while α and β are subsets of θ not necessarily mutually exclusive e.g. $\alpha=(\mu_1,\sigma^2)$, $\beta=(\mu_2,\sigma^2)$ and $\theta=(\mu_1,\mu_2,\sigma^2)$.

4.3 Distribution Indicator

Here ω is merely used to indicate a change in the likelihood to different families of distributions, e.g. ω_0 indicates a log normal distribution and ω_1 a gamma distribution. In such cases there needn't be any common parameter. When this obtains a parametric diagnostic is irrelevant but a predictive diagnostic is always relevant. Although this can often be regarded as a special case of either of the first two methods it is best to consider it separately.

4.4 Exclusion Indicator

Here we assume that $X^{(N)} = (X_1,...,X_N)$ has a standard distribution $\omega=\omega_0$ but for $\omega\neq\omega_0$ one or more of the X_i's have either another distributional form or a completely unspecifiable distribution. In the former case this could mean for example that an observation's variance or mean differs from the others or more generally that a parameter set not under scrutiny differs for a few of the observations. The latter situation is typically reflected in problems with outliers and aberrant observations that defy satisfactory alternative modeling.

4.5 Periparametric models

Here $\omega=\omega_0$ specifies a standard density while $\omega\neq\omega_0$ specifies all model densities $f(x^{(N)}|\omega)$ that are within a given neighborhood of $f(x^{(N)}|\omega_0)$ determined by varying ω. This may be considerably more useful in the Bayesian context when applied to a subjective prior density for θ to add further uncertainty. For applications of these possibilities see Lavine (1987).

4.6 Inaccurate Measurements

Here, presumably, covariates or concomitants may be inaccurately measured and the responses themselves may be subject to errors.

5. <u>Bayesian Decision Framework</u>

We now delineate a formal Bayesian perturbation framework for a relevant parameter θ (or some subset of θ). Now the Bayesian model consists of

$$f(x^{(N)},\theta|\omega) = f(x^{(N)}|\theta,\omega)\mathcal{P}(\theta|\omega)$$

where $\mathcal{P}(\theta|\omega)$ is the perturbed prior and the quantity on the left is then the perturbed model. Note also the quantity on the left is also proportional to $\mathcal{P}(\theta|x^{(N)},\omega)$ the posterior of θ. We shall first deal with a given loss function $L(a,\theta)$ for taking action $a(X^{(N)}) \in \underset{\sim}{A}$ upon observing $X^{(N)}$, given θ is the true value (the loss function itself may also be perturbed but we shall hold it fixed for the time being). The average loss

$$\bar{L}_\omega(a) = \int L(a,\theta)\mathcal{P}(\theta|x^{(N)},\omega)d\theta,$$

which depends on ω, is now minimized

$$\min_a \bar{L}_\omega(a) = \bar{L}_\omega(a^*_\omega)$$

yielding optimal action a^*_ω when ω is "true". We then consider the difference in the loss when taking action $a^*_{\omega_0} = a^*$, the optimal action under the standard and when $\omega\neq\omega_0$ is true. We define the differential loss as

$$d(\omega) = \bar{L}_\omega(a^*) - \bar{L}_\omega(a^*_\omega) \geq 0.$$

One then can examine this loss over a possible range of ω to assess its importance with regard to the action taken under ω_0 and in particular $d^* = \max_{\omega \in \Omega} d(\omega)$. We could also assess its local significance by examining $d(\omega)$ in a neighborhood about ω_0. In fact if ω is a scalar and the second derivative of $d(\omega)$ exists and is continuous the calculation of the curvature at $\omega = \omega_0$ i.e. $d''(\omega_0)$, since $d'(\omega_0) = 0$, could be rather informative regarding local perturbations. For example a large curvature would indicate that the actions taken could be highly sensitive to a slight variation in the standard model.

Of course $d(\omega)$ depends on the standard loss function $L(\theta, a)$ which also may be varied so that $L^{\tau}(\theta, a)$ and τ represent a perturbation for the loss function where $L^{\tau_0}(\theta, a) = L(\theta, a)$ then we could obtain $d_{\tau}(\omega)$ and find, globally

$$\max_{\tau, \omega} d_{\tau}(\omega)$$

to assess the importance of varying the loss function (all variations assumed to be in the same units) as well as the model. For a vector ω, the matrix of second derivatives will govern the local curvature and one could assess the maximum curvature i.e. in the direction of the normed vector associated with the largest root of the matrix of second derivatives evaluated at the standard $\omega = \omega_0$. Cook (1986) has proposed probing local curvature with regard to the displacement of maximized log-likelihoods.

For a predictive analysis we consider the perturbed model to be

$$f(x^{(N)}, x_{(M)}, \theta | \omega) = f(x_{(M)} | x^{(N)}, \theta, \omega) f(x^{(N)} | \theta, \omega) g(\theta | \omega)$$

whence we obtain

$$f(x_{(M)}|x^{(N)},\omega) = \frac{\int f(x^{(N)},x_{(M)},\theta|\omega)d\theta}{\int\int f(x^{(N)},x_{(M)},\theta|\omega)d\theta dx_{(M)}} \quad .$$

Now assume that $L(a,x_{(M)})$ is the standard loss incurred in taking action a when observing $x^{(N)}$ given a future realization $x_{(M)}$. The average predictive loss

$$\bar{L}_\omega(a) = \int L(a,x_{(M)})f(x_{(M)}|x^{(N)},\omega)dx_{(M)}$$

is then minimized

$$\min_a \bar{L}_\omega(a) = \bar{L}_\omega(a^*_\omega)$$

where a^*_ω is the optimal action. As before letting $a^*_{\omega_0} = a^*$, we define the differential loss as $d(\omega) = \bar{L}_\omega(a^*) - \bar{L}_\omega(a^*_\omega)$ and examine globally $\max_\omega d(\omega)$ to determine the possible extent of the maximum effect of the perturbations. We may also perturb the loss function as previously discussed.

Further in regular cases one can again study locally the maximum curvature which occurs in the direction of the normed vector associated with the largest root of the Hessian matrix, say $d''(\omega_0)$. If local curvature is appreciable it would appear that the action indicated by the standard analysis is not even robust locally and a review of the standard model is in order. Of course if the perturbed ω model is deemed reasonable one possibility is to define a prior distribution for ω, if feasible, and then integrate it out to obtain

$$f(x_{(M)}|x^{(N)}) = \int g(\omega) f(x_{(M)}|x^{(N)},\omega) d\omega,$$

or to completely rethink the modeling apparatus.

6. Perturbation Diagnostics

Often, decisions or actions other than to report either the posterior or predictive distribution itself or some high probability density region for θ or $X_{(M)}$ are not required. For reporting the entire posterior distribution the Kullback-Leibler estimative divergence,

$$K(\mathcal{P}_{\omega},\mathcal{P}_{\omega_0}) = E[\log \mathcal{P}_{\omega} - \log \mathcal{P}_{\omega_0}],$$

where $\mathcal{P}_{\omega} = \mathcal{P}(\theta|y^{(N)},\omega)$, appears to be a reasonable diagnostic to consider, when it exists and is finite, Geisser (1985), Johnson and Geisser (1985), McCulloch (1989) and can be investigated in a variety of paradigms. Similarly for predictive distributions a predictive divergence

$$K(\omega,\omega_0) = E[\log f_{\omega} - \log f_{\omega_0}]$$

where $f_{\omega} = f(x_{(M)}|x^{(N)},\omega)$, can serve. Divergences of this sort were already discussed in section 1 for determining influential observations, one of the particular types of perturbation mentioned in section 4.3.

Both, estimative and predictive diagnostics, are most useful in indicating the relative effect of various perturbations, although there may be some difficulty in adequately interpreting globally

$$\max_{\omega\in\Omega} K(\mathcal{P}_\omega, \mathcal{P}_{\omega_0}), \text{ or } \max_{\omega\in\Omega} K(\omega, \omega_0).$$

Again one can find the direction in which local perturbations have the greatest effect in terms of normal curvature. It can be shown that under suitable regularity conditions that the matrix of second derivatives of $K(\mathcal{P}_\omega, \mathcal{P}_{\omega_0})$ or $K(\omega, \omega_0)$ for ω a vector of perturbations, say

$$K''_{\omega=\omega_0} = I(\omega_0)$$

where $I(\omega_0)$ is the Fisher Information matrix for either the perturbed posterior or predictive distribution evaluated at $\omega=\omega_0$, Kullback (1959). The curvature in direction z where $\omega(t) = \omega_0+tz$ and $z'z = 1$ is

$$C_z = z'I(\omega_0)z$$

so that the maximum curvature C* is in the direction z*, the vector associated with the maximum root of $I(\omega_0)$, where

$$C^* = z^*I(\omega_0)z^*.$$

An examination of the components of z^* will indicate which ones, namely the larger ones, are those perturbations which relatively most alter the posterior or predictive distribution in terms of the divergence.

Once potentially significant directions are identified, an analysis involving these directions is in order to ascertain whether local departures for them are important enough to vitiate the standard analysis or support it locally.

In situations where a standard likelihood is deemed adequate it may be of interest to assess the local effect of perturbing the prior on the posterior and the predictive distribution. Here one may calculate the vector associated with the maximum root of either $I_0^{-1}(\omega_0)\ I_{\wp}(\omega_0)$, or $I_0^{-1}(\omega_0)\ I_f(\omega_0)$ where

$I_0(\omega_0)$, $I_{\wp}(\omega_0)$, and $I_f(\omega_0)$ are the Fisher information matrices associated with the prior, posterior and predictive distributions evaluated at $\omega=\omega_0$ respectively. This is then used to assess the effect of a local prior perturbation on the posterior or predictive distribution, e.g. McCulloch (1989). One can also turn this round by calculating that $\omega\epsilon\Omega$ such that $K(\wp_{\omega},\wp_{\omega_0})$ or $K(\omega,\omega_0) \le r$ where r is negligible enough to consider the analysis

(or sample) robust with respect to the subspace $\Omega_R \epsilon \Omega$.

One can utilize, as alternatives to the divergence, a whole class of distances between densities namely

$$H_n(\omega) = \int |f^{\frac{1}{n}} - g^{\frac{1}{n}}|^n dx$$

for any n. In particular the L^1 norm for n=1 and n=2, the L^2 norm between $\sqrt{f}$ and $\sqrt{g}$ have been most widely used in other contexts. The former recently by Devroye (1987) in density estimation and the latter by Pitman (1979) in a frequentist sensitivity analysis. Under suitable smoothness conditions, for ω a vector, one obtains for the L^2 norm

$$2H_2''(\omega_0) = I(\omega_0),$$

the Fisher Information matrix. Hence these two "loss" function variants $K(\omega)$ and $2H_2(\omega)$ are locally equivalent as measured by curvature at the standard.

7. Specific Perturbation Assessments

While the divergence and the distance measures are useful scalar rankings of a perturbation effect it is difficult to give a compelling calibration of them although an attempt has been made by McCulloch (1989). Their interpretation to a statistician is difficult enough, not to say to an investigator. However, when an analysis involves a fairly specific inference obtained from the posterior or the predictive distribution we shall define appropriate assessments whose values are easily understood and interpretable not only to a statistician but to an investigator as well.

If a $1-\alpha$ highest probability density region is what is going be to reported that one can assess the robustness based on the standard ω_0. Suppose this region denoted by $R_{1-\alpha}(\omega_0)$ has volume $V(\omega_0)$ and when perturbed the highest probability density region $R_{1-\alpha}(\omega)$ has volume $V(\omega)$. Let $\nu(\omega)$ be the volume of the intersection of $R_{1-\alpha}(\omega)$ and $R_{1-\alpha}(\omega_0)$ as a function of ω,

$$\nu(\omega) = \text{volume } [R_{1-\alpha}(\omega) \cap R_{1-\alpha}(\omega_0)]$$

and let

$$\Gamma_\omega = \frac{\nu(\omega)}{M(\omega)},$$

where $M(\omega) = \max(V(\omega), V(\omega_0))$, be the ratio of the volume of the intersection to whichever is larger, the standard or the perturbed for the given ω. Then calculate

$$\min_{\omega\in\Omega} \Gamma_\omega = \Gamma_{\omega^*}$$

which now yields the proportion of the region for the "worst" possible case at a given probability $1-\alpha$. Hence one has an easily interpretable value for assessing the robustness of the data set in terms of a standard analysis involving a $1-\alpha$ region in the presence of presumably anticipated perturbations. A second method focuses on the use of the standard region's $R_{1-\alpha}(\omega_0)$ perturbed probability when $\omega\neq\omega_0$. Here we use either

$$\Pr[\theta\in R_{1-\alpha}(\omega_0)|\omega] = \int_{R_{1-\alpha}(\omega_0)} \mathcal{P}(\theta|y^{(N)},\omega)d\theta = 1-\alpha_\omega$$

or

$$\Pr[X_{(M)}\in R_{1-\alpha}(\omega_0)|\omega] = \int_{R_{1-\alpha}(\omega_0)} f(x_{(M)}|x^{(N)},\omega)\ dx_{(M)} = 1-\alpha_\omega$$

and either

$$\max_{\omega\in\Omega} |1-\alpha-(1-\alpha_\omega)| = \max_{\omega\in\Omega} |\alpha_\omega-\alpha|$$

or

$$\max_{\omega\in\Omega} \frac{|\alpha_\omega-\alpha|}{1-\alpha_\omega}$$

as easily interpretable values.

A third method, which is really a variation on the second, is where the region R is fixed e.g. survival past beyond a given time point, and basically the second method is applied to determine the variation in the probability of R with respect to ω.

8. Discussion

When a standard analysis is deemed robust in the light of contemplated perturbations, we are in a secure position. When a perturbation, especially a local one, appreciably alters the analysis we have two alternatives. The first depends on a belief that although the standard is inappropriate, the expanded model is reasonable. We then would assign a prior to the perturbation, if possible, and average over ω. This is particularly appropriate when ω is a hyperparameter. In this case the problem in a sense is moved back a step to the possibility that the distribution of ω has some unknown hyperparameter and the cycle could start all over. On the other hand, if the expanded model is problematical then a review of the entire modeling apparatus is in order.

The mixture or contamination case depends largely on whether estimation of a parameter (uncontaminated) involving the standard is at issue or whether prediction is the goal. For the latter there are two possibilities. The first being that we are really interested in predicting a future value from an uncontaminated distribution or we are interested in predicting a future value from the process that generated the experiment including the contamination. Quite different approaches are required for these two goals.

When the possibility is that one or a few aberrancies exist in the data set- then testing the candidates for discordancy by predictive significance tests may be in order, Geisser (1989a, 1989b). When one or few observations contain the parameter of interest but also a nuisance parameter, the Bayesian approach is ideal for handling that situation, Geisser (1987).

The Bayesian "robustness" point of view taken here has a narrower focus than that of say Berger (1984) or Berliner and Hill (1988). Here the focus is on a specific parametric model that is probed to ascertain whether an analysis is robust against particular anticipated expansions of the model. This view is somewhat similar in spirit to that of Box (1980). Another view is to devise nonparametric Bayesian procedures. The latter would be appropriate if no particular parametric model is even mildly compelling.

9. Acknowledgement

This work was sponsored in part by an NIH grant GMS-25271.

References

Berliner, M.L. and Hill, B. (1988). Bayesian nonparametric survival analysis (with discussion). Journal of the American Statistical Association, 83:772-784.

Berger, J. (1984). The robust Bayesian viewpoint. In Robustness of Bayesian Analysis, J. Kadane, ed., North-Holland Publishing Company, Amsterdam, 63-144.

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society, A, 143:383-430.

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

Cook, R.D. (1979). Influential observations in linear regression. Journal of the American Statistical Association, 74:169-174.

Cook, R.D. (1986). Assessment of local influence (with discussion). Journal of the Royal Statistical Society, B, 48, 2, 133-169.

Devroye, L. (1987). A Course in Density Estimation. Birkhauser.

Geisser, S. (1974). A predictive approach to the random effect model. Biometrika, 61:101-107.

Geisser, S. (1975). The predictive sample reuse method with application. Journal of the American Statistical Association, 70:320-328.

Geisser, S. (1980). Contribution to discussion. Journal of the Royal Statistical Society, A, 143: 416-417.

Geisser, S. (1985). On predicting observables: A selective update. Bayesian Statistics 2, (J.M. Bernardo et al. eds.) Amsterdam: North-Holland, 203-230. (with discussion)

Geisser, S. (1986). Contribution to discussion. Journal of the Royal Statistical Society, B, 47, 2, 163.

Geisser, S. (1987). Influential observations, diagnostics and predictive discordancy tests. Journal of Applied Statistics, 14, 2, 133-142.

Geisser, S. (1988). The future of statistics in retrospect. Bayesian Statistics 3, (J.M. Bernardo et al. eds. Oxford University Press), 147-158.

Geisser, S. (1989a). Predictive discordancy testing for exponential observations. Canadian Journal of Statistics, 17, 2, 19-26.

Geisser, S. (1989b). Approaches to discordancy testing. Bayesian and Likelihood Inference. (To appear.)

Johnson, W. and Geisser, S. (1982). Assessing the predictive influence of observations. In: G. Kallianpur, P.R. Krishnaiah and J.K. Ghosh (eds) Statistics and Probability Essays in Honor of C.R. Rao, 343-348. Amsterdam, North-Holland.

Johnson, W. and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. Journal of the American Statistical Association, 78:137-144.

Johnson, W. and Geisser, S. (1985). Estimative influence measures for the multivariate general linear model. Journal of Statistical Planning and Inference, 11:33-56.

Kullback, S. (1959). Information theory and statistics. New York, John Wiley and Sons.

Lavine, M. (1987). Prior influence in statistics. University of Minnesota Technical Report No. 504.

Lee, J.C. (1971). On the generalized growth curve model. Doctoral Dissertation, SUNY at Buffalo.

Lee, J.C. (1975). Applications of growth curve prediction. Sankhya A, 37:239-256.

McCulloch, R.E. (1989). Local model influence. Journal of the American Statistical Association, 84:473-478.

Pitman, E.J.G. (1979). Some Basic Theory for Statistical Influence. London, Chapman and Hall.

Smith, A.F.M. and Pettit, L.I. (1985). Outliers and influential observations in linear models. In: J.M. Bernardo et al. eds. Bayesian Statistics 2, 473-494.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society B, 36:111-147.