ROBUST BAYESIAN BOUNDS FOR OUTLIER DETECTION

by

M. J. Bayarri and James O. Berger
University of Valencia Purdue University

Technical Report #92-43C

Department of Statistics
Purdue University

September 1992

◇

ROBUST BAYESIAN BOUNDS FOR OUTLIER DETECTION

M. J. BAYARRI AND JAMES O. BERGER

University of Valencia and Purdue University

ABSTRACT. Outlier detection is a statistical problem that has received considerable attention and has originated a large body of research. Under suitable formulation, testing whether an observation is an outlier can be reduced to a testing problem concerning a parameter of a contaminating distribution. As in Bayesian testing of a point null, Bayes factors and posterior probabilities of the hypotheses cannot be computed here using non-informative priors. Thus we consider wide classes of priors for the relevant parameter and explore bounds on the Bayes factor as the prior varies in these classes.

1. INTRODUCTION

A universal, precise definition of the term "outlier" seems rather difficult. A generally accepted definition is the one used, for example, by Box and Tiao (1968) (also used in earlier papers, as mentioned there):

> "an outlier being an observation which is suspected to be partially or wholly irrelevant because it is *not* generated by the stochastic model assumed."

Freeman (1980) adopts a similar definition without the requirement for the outlier to be irrelevant:

> "The word 'outlier' here will mean any observation that has not been generated by the mechanism that generated the majority of observations in the data set."

The most common approach to outlier detection is to assess a model to explain the behavior of the (possible) outliers. This contaminating model is usually taken to be a generalization of the original model, involving an extra parameter (or parameter vector). The original model is then typically a particular case of the contaminating model, corresponding to some specific value of the extra parameter. Testing for outliers can then be reduced to testing for this specific value of the extra parameters.

Previous Bayesian work along these lines can be found in Box and Tiao (1968), Guttman (1973), Abraham and Box (1978), Guttman, Dutter and Freeman (1978), Freeman (1980), Pettit and Smith (1983, 1985), Pettit (1988, 1990, 1992), Verdinelli

*Research supported by the National Science Foundation, grant DMS–8923071, and by the Universitat de Valencia, ajuda de la convocatoria de 1992 d'Estades en Centres d'Investigacio Nacionals y Estrangers.

Key words and phrases. Bayes factor; Location contamination; Normal models; Scale contamination.

and Wasserman (1991), and Peña and Tiao (1992). Alternative approaches have also been explored, usually requiring either the outlier to actually "outlie" or to be poorly predicted by the rest of the data. These approaches do not require a model for the outlying observations and they have been taken, for instance, by Zellner (1975), Geisser (1980, 1985, 1987, 1989, 1990), Chaloner and Brant (1988), Guttman and Peña (1988), and Giron, Martínez and Morcillo (1992). A third approach is to develop an overall model, usually with heavy tails, that automatically handles outliers in suitable ways without requiring their identification. This approach is taken by O'Hagan (1979, 1988) and West (1984, 1985); related work can be found in DeFinetti (1961), Dawid (1973), and Gómez-Villegas and Maín (1992). Reviews of previous work and discussions of the meaning of the term "outlier" can be found in Freeman (1980) and Pettit and Smith (1985).

We shall follow the usual formulation and assume that the data consists of $n+1$ independent random variables $X_0, X_1, \ldots, X_n$. In the outlier scenario, although we feel quite confident that $X_i \sim f(x_i|\theta)$, $i = 0, 1, \ldots, n$, we recognize (or fear) that we might get some (hopefully few) outlying observations. In other words, we recognize that, for any X_i, there is a (usually small) prior probability ε that it is generated by some contaminating distribution with density g. (A more general formulation would specify a possibly different g_i for each X_i.) Usually g is taken to be a density $f^*(x|\theta, \tau)$ such that the original model $f(X|\theta)$ is the special case corresponding to a specific value τ_0, so that $f(x|\theta) = f^*(x|\theta, \tau_0)$.

If the goal of the statistical analysis is inference about θ, then it can be carried out using the implied mixture model, $X_i \sim (1-\varepsilon)f(x_i|\theta) + \varepsilon f^*(x_i|\theta, \tau)$. Our goal, however, will be to detect the outliers, perhaps because this is actually one of the main goals of the analysis (as in Raccine-Poon, 1992), or more usually, because we feel that our information about f^* is too vague to be of much help in inference about θ.

In this paper, we limit ourselves to study of the simplest case, that is, to detect whether a specific observation, x_0, is an outlier provided that all the rest are not. Hence, in the rest of the paper we assume the following model:

$$\begin{aligned} &X_1, X_2, \ldots, X_n \text{ i.i.d., } X_i \sim f(x_i|\theta), \\ &X_0 \sim (1-\varepsilon)f(x_0|\theta) + \varepsilon f^*(x_0|\theta, \tau). \end{aligned} \tag{1.1}$$

Also, we shall let $\underset{\sim}{x} = (x_1, \ldots, x_n)$ denote the non-outlying observations and $\ell(\theta)$ denote the likelihood function for θ based solely on $\underset{\sim}{x}$, that is

$$\ell(\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{1.2}$$

Prior densities for (θ, τ) are denoted by $\pi(\theta|\tau)g(\tau)$. Usually θ and τ are taken to be a priori independent, so that their joint prior would then be $\pi(\theta)g(\tau)$.

The posterior probability that x_0 is an outlier, given the data and given that $\underset{\sim}{x}$ is non-outlier, can be expressed as

$$\Pr(x_0 \text{ outlier}|x_0, \underset{\sim}{x}, \underset{\sim}{x} \text{ non-outlier}) = (1 + \frac{(1-\varepsilon)}{\varepsilon} B(g))^{-1}, \tag{1.3}$$

where

$$B(g) = \frac{\int \ell(\theta) f(x_0|\theta)\pi(\theta|\tau_0)d\theta}{\int\int \ell(\theta) f^*(x_0|\theta,\tau)\pi(\theta|\tau)g(\tau)d\theta d\tau} \tag{1.4}$$

is the Bayes factor in favor of x_0 being non-outlier when the prior g for τ is used. An alternative expression for $B(g)$ is

$$B(g) = \frac{m^f(x_0|\underset{\sim}{x})}{m^{f^*}(x_0|\underset{\sim}{x})}, \tag{1.5}$$

where $m^h(x_0|\underset{\sim}{x})$ is the (posterior) predictive density at x_0 assuming that the model generating x_0 is h. That is, $m^f(x_0|\underset{\sim}{x}) = \int f(x_0|\theta)\pi(\theta|\underset{\sim}{x},\tau_0)d\theta$ and $m^{f^*}(x_0|\underset{\sim}{x}) = \int\int f^*(x_0|\theta,\tau)\pi(\theta|\underset{\sim}{x},\tau)g(\tau)d\theta d\tau$. (For simplicity, the dependence of $m^{f^*}(x_0|\underset{\sim}{x})$ on g is left implicit in the notation.) Note that Bayesian measures of "surprise" that do not require specification of f^* are typically based solely on the numerator of (1.5).

Bayes factors have become an increasingly popular way to report the experimental evidence in favor of (or against) the null hypothesis, especially in contexts of testing a sharp null hypothesis (see Berger and Delampady, 1987, and Berger and Sellke, 1987). Note that all experimental evidence for or against x_0 being an outlier enters (1.3) only through $B(g)$ and that the prior probability, ε, that x_0 is an outlier does not enter $B(g)$. This makes use of $B(g)$, rather than posterior probability, attractive in ordinary hypothesis testing.

In the outlier problem, however, the effect of ε cannot reasonably be ignored, since ε is usually quite small. As an indication of how small B needs to be in order to have reasonable belief that x_0 is an outlier, suppose $\varepsilon = 0.05$ (a typically considered value) and that we would decide that x_0 is an outlier if the posterior probability of such were larger than 0.85; this would require $B \leq .0093$. Thus, as a rough rule of thumb, $B \leq .01$ would be reasonable evidence that x_0 is an outlier. (This is compatible with the reasoning in Pettit, 1992, which suggests that reasonable evidence for an outlier is $B \leq .015$.)

A well known limitation to the use of Bayes factors for testing a point null hypothesis is that improper non-informative priors over the parameters identifying the two hypotheses cannot be used; indeed, by multiplying these *improper* priors by an arbitrary constant, the Bayes factors can be made arbitrary large or small. In our problem, this means that $g(\tau)$ in (1.4) cannot be a standard, non-informative prior. (Notice that $\pi(\theta)$ does not share this difficulty, in that multiplication of an improper $\pi(\theta)$ by a constant does not alter the Bayes factor.) This limitation is particularly disturbing in the outlier scenario since, as is almost implicit in the formulation of the problem, very little is typically known about the contaminating distribution, so that quantification of a proper, informative $g(\tau)$ is very difficult.

Pettit and Smith (1983, 1985), and Pettit (1988, 1990, 1992) deal with the problem by using the Spiegelhalter and Smith (1982) device of an imaginary training sample, in order to choose a suitable non-informative prior. In this paper, a different approach is taken. We select wide classes, Γ, of priors, g, and investigate the lower bound, $\underline{B}$, of $B(g)$ as g varies in Γ. Thus, the evidence in favor of x_0 *not* being an outlier would be *at least* $\underline{B}$, no matter what our prior g might be (provided, of course, that it belongs to the class Γ). If $\underline{B}$ is moderately large, x_0 would not be

considered an outlier and further refinements of our prior knowledge about g would not be necessary.

In this paper, we confine ourselves to the most widely studied scenario for outliers, namely that in which the original distribution is a normal $N(x|\theta,\sigma^2)$ distribution and the contaminating distribution is normal $N(x|\theta,\sigma^2,\tau)$, with an extra parameter τ. This contaminating distribution will be taken to be either $N(x|\theta,\tau\sigma^2)$ with $\tau > 1$, or $N(x|\theta+\tau,\sigma^2)$. The former is usually referred to as a *scale* contamination model, while the later is usually called a *location* or *location-shift* contamination model. In Section 2, we study the scale contamination model, both for known and unknown σ^2; the same is done in Section 3 for the location contamination model. Section 4 is devoted to comparisons and generalizations.

2. Scale Contamination

In this section we assume that $x_1,\ldots,x_n$ are non outlying observations from a $N(\theta,\sigma^2)$ distribution, and that the possible outlier, x_0, is generated by the mixture density

$$(1-\varepsilon)N(x_0|\theta,\sigma^2)+\varepsilon\, N(x_0|\theta,\tau\sigma^2) \tag{2.1}$$

with $\tau > 1$. We shall continue to let $\underset{\sim}{x}$ denote the non outlying observations $\underset{\sim}{x} = (x_1,x_2,\ldots,x_n)^t$ and let $\pi(\theta|\underset{\sim}{x})$ denote the posterior density of θ given $\underset{\sim}{x}$ (excluding x_0). Also, the posterior predictive densities (given $\underset{\sim}{x}$) at x_0 under model $N(\theta,\sigma^2)$ and under model $N(\theta,\tau\sigma^2)$ for X_0 are denoted by $m(x_0|\underset{\sim}{x})$ and $m^*(x_0|\underset{\sim}{x},\tau)$, respectively. We explore both the situations σ^2 known and σ^2 unknown.

2.1 σ^2 Known.

Here, and in the rest of the paper, we assume that θ and τ are independent a priori, so that their joint density is given by $\pi(\theta)g(\tau)$. We take $\pi(\theta)$ to be the natural conjugate density $\pi(\theta) = N(\theta|m_0,\sigma_0^2)$, so that the posterior density is $\pi(\theta|\underset{\sim}{x}) = N(\theta|m_1,\sigma_1^2)$, where

$$\begin{aligned} m_1 &= \lambda\overline{x}+(1-\lambda)m_0,\ \sigma_1^2 = \lambda\sigma^2/n \\ \lambda &= n\sigma_0^2/(\sigma^2+n\sigma_0^2). \end{aligned} \tag{2.2}$$

The non-informative prior $\pi(\theta) \propto 1$ results in the same posterior, but with $\lambda = 1$.

Testing H_0: x_0 *is not an outlier*, versus H_1: x_0 *is an outlier*, is, with the above formulation, equivalent to testing H_0: $\tau = 1$, versus H_1: $\tau \neq 1$. The Bayes factor (1.5) in favor of H_0 is, for any given prior $g(\tau)$ for τ:

$$\begin{aligned} B(g) &= \frac{m(x_0|\underset{\sim}{x})}{\int m^*(x_0|\underset{\sim}{x},\tau)g(\tau)d\tau} \\ &= \frac{\int N(x_0|\theta,\sigma^2)N(\theta|m_1,\sigma_1^2)d\theta}{\int\int N(x_0|\theta,\tau\sigma^2)N(\theta|m_1,\sigma_1^2)g(\tau)d\theta d\tau} \\ &= \frac{N(x_0|m_1,\sigma_1^2+\sigma^2)}{\int N(x_0|m_1,\sigma_1^2+\tau\sigma^2)g(\tau)d\tau}. \end{aligned} \tag{2.3}$$

Following the robust Bayesian approach, we shall derive

$$\underline{B} = \inf_{g\in\Gamma} B(g) = \frac{N(x_0|m_1, \sigma_1^2+\sigma^2)}{\sup_{g\in\Gamma} \int N(x_0|m_1, \sigma_1^2+\tau\sigma^2)g(\tau)d\tau} \tag{2.4}$$

for two classes, Γ, of priors $g(\tau)$, namely the class of all possible prior densities and the class of all densities that are non-increasing functions of τ.

Case 1: $\Gamma_1 = \{$all priors $g(\tau)\}$.

For this case, the supremum in the denominator of (2.4) is attained at a point-mass density, so that $\underline{B}$ is

$$\underline{B}_1 = \frac{N(x_0|m_1, \sigma_1^2+\sigma^2)}{\sup_{\tau} N(x_0|m_1, \sigma_1^2+\tau\sigma^2)}, \tag{2.5}$$

which is easily seen to be

$$\begin{aligned} \underline{B}_1 &= \sqrt{e}\; z\; e^{-z^2/2}, \quad \text{for } z > 1, \\ &= 1, \quad \text{for } z \le 1, \end{aligned} \tag{2.6}$$

where $z = |x_0 - m_1|/s_1$, and m_1 and $s_1 = \sqrt{\sigma_1^2+\sigma^2}$ are the mean and standard deviation, respectively, of the predictive distribution, $m(x_0|\underline{x})$, of x_0 under the non-outlier model. For the non-informative $\pi(\theta) \propto 1$, $m_1 = \bar{x}$ and $s_1^2 = (n+1)\sigma^2/n$. Thus, z in (2.6) is the distance, measured in units of predictive standard deviation, of x_0 from its predicted value m_1 under the non-outlier assumption.

Case 2: $\Gamma_2 = \{$all non-increasing densities g on $[1,\infty)\}$.

If the class Γ_1 is deemed to be too large, a natural reduction is to consider all densities on $[1,\infty)$ that are non-increasing in τ. (Some authors maintain that τ should be much larger than 1; the analysis presented here can be trivially generalized to densities g on $[c,\infty)$, non increasing in τ, with c any positive constant.) All such densities can be represented as mixtures of uniform densities, so that now (2.4) becomes

$$\underline{B}_2 = \frac{N(x_0|m_1, \sigma_1^2+\sigma^2)}{\sup_{\Gamma} \frac{1}{\tau-1}\int_1^\tau N(x_0|m_1, \sigma_1^2+\tau\sigma^2)d\tau}, \tag{2.7}$$

which, after some algebra, can be expressed as

$$\begin{aligned} \underline{B}_2 &= \gamma^{-1/2} e^{-(1-\gamma)z^2/2}, \quad \text{for } z > 1, \\ &= 1, \quad \text{for } z \le 1, \end{aligned} \tag{2.8}$$

where $z = |x_0 - m_1|/s_1$ and γ is the unique solution to

$$\frac{1+\gamma}{2\sqrt{\gamma}} e^{-\gamma z^2/2} - e^{-z^2/2} - z\sqrt{2\pi}\{\Phi(z) - \Phi(z\sqrt{\gamma})\} = 0, \tag{2.9}$$

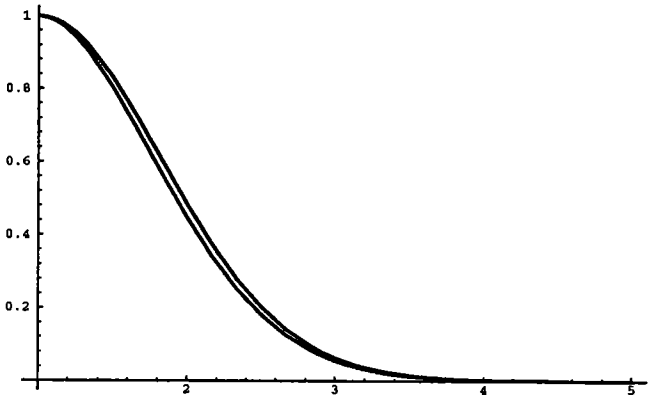
where Φ denotes, as usual, the standard normal distribution function.

Although, obviously, $\underline{B}_2$ is greater than $\underline{B}_1$ (and in fact γ can be shown to be smaller than $1/z^2$), both lower bounds are surprisingly close to each other, even for moderate values of z. Table 1 gives the values of $\underline{B}_1$ and $\underline{B}_2$ for selected values of z. Figure 1 displays $\underline{B}_1$ and $\underline{B}_2$ as functions of z. Both Table 1 and Figure 1 assume the non-informative prior, $\pi(\theta) = 1$, for θ.

TABLE 1. LOWER BOUNDS ON THE BAYES FACTORS FOR THE SCALE CONTAMINATION MODEL (σ^2 KNOWN).

z	1.5	2	2.5	3	3.5	4	4.5
$\underline{B}_1$	.8029	.4463	.1811	.0549	.0126	.0022	.0003
$\underline{B}_2$	.8305	.4832	.2026	.0628	.0146	.0026	.0003

FIGURE 1. $\underline{B}_1$ AND $\underline{B}_2$ FOR THE SCALE-CONTAMINATION MODEL (σ^2 KNOWN) AS FUNCTIONS OF z.



2.2 σ^2 Unknown.

If we define $\delta = 1/\sigma^2$, then the usual conjugate prior for (θ, δ) has $\theta|\delta \sim N(m_0, (h_0\delta)^{-1})$, and $\delta \sim Ga(a_0, b_0)$, that is, $\pi(\delta) \propto \delta^{a_0-1} \exp\{-b_0\delta\}$. The ensuing joint distribution for (θ, δ) is the normal-gamma distribution, denoted $(\theta, \delta) \sim NGa(m_0, h_0, a_0, b_0)$. The posterior distribution of (θ, δ), given the non-outlying $\underset{\sim}{x}$, is then the revised normal-gamma $(\theta, \delta|\underset{\sim}{x}) \sim NGa(m_1, h_1, a_1, b_1)$, where

$$\begin{aligned} m_1 &= \lambda\bar{x} + (1-\lambda)m_0, \ \lambda = n/(h_0+n) \\ h_1 &= h_0 + n, \ a_1 = a_0 + b/2, \\ b_1 &= b_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\lambda h_0}{2}(\bar{x} - m_0)^2. \end{aligned} \tag{2.10}$$

If the non-informative $\pi(\theta, \delta) \propto 1/\delta$ is used, then $m_1 = \bar{x}$, $h_1 = n$, $a_1 = (n-1)/2$, $b_1 = \Sigma(x_i - \bar{x})^2/2$.

The predictive densities that enter the computation of the Bayes factor are the posterior predictive densities of X_0 (at x_0) assuming that either X_0 is not an outlier:

$$\begin{aligned} m(x_0|\underset{\sim}{x}) &= \int\int N(x_0|\theta, \sigma^2) NGa(\theta, \sigma^2|m_1, h_1, a_1, b_1) d\theta d\sigma^2 \\ &= St(x_0|m_1, (1+\frac{1}{h_1})\frac{b_1}{a_1 - 1}, 2(a_1 - 1)), \end{aligned} \tag{2.11}$$

or that x_0 comes from the contaminating density whose extra-parameter has the value τ:

$$m^*(x_0|\underset{\sim}{x},\tau) = \int\int N(x_0|\theta,\tau\sigma^2)NGa(\theta,\sigma^2|m_1,h_1,a_1,b_1)d\theta d\sigma^2$$
$$= St(x_0|m_1,(\tau+\frac{1}{h_1})\frac{b_1}{a_1-1},2(a_1-1)), \tag{2.12}$$

where $St(x|m,s^2,\alpha)$ denotes a Student t density as given by

$$St(x|m,s^2,\alpha) = C[1+\frac{1}{\alpha s^2}(x-m)^2]^{-(\alpha+1)/2}, \tag{2.13}$$

where $C = \Gamma(\frac{\alpha+1}{2})/[\Gamma(\frac{\alpha}{2})\Gamma(\frac{1}{2})(\alpha s^2)^{1/2}]$.

Again we shall study the infimum of the Bayes factor (in favor of H_0: *non-outlier*)

$$\underline{B} = \inf_{g\in\Gamma}\frac{m(x_0|\underset{\sim}{x})}{\int m^*(x_0|\underset{\sim}{x},\tau)g(\tau)d\tau}, \tag{2.14}$$

for the same classes Γ_1,Γ_2 considered in Subsection 2.1.

Case 1: $\Gamma_1 = \{$all priors $g(\tau)\}$.

In this case, from (2.14), (2.11) and (2.12):

$$\underline{B}_1 = \frac{St(x_0|m_1,(1+h_1)b_1/[h_1(a_1-1)],2(a_1-1))}{\sup_\tau St(x_0|m_1,(1+\tau h_1)b_1/[h_1(a_1-1)],2(a_1-1))}. \tag{2.15}$$

To parallel the developments in the previous subsection, let s_1 denote the scale parameter of the posterior predictive of X_0 under the non-outlier model, that is:

$$s_1^2 = \frac{(1+h_1)}{h_1}\cdot\frac{b_1}{(a_1-1)}, \tag{2.16}$$

and define

$$z = |\frac{x_0-m_1}{s_1}|. \tag{2.17}$$

For the non-informative prior, $m_1 = \bar{x}$ and $s_1^2 = (n+1)\hat{\sigma}^2/(n-3)$, where $\hat{\sigma}^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2/n$ is the usual MLE of σ^2 based on the non-outlying observations. (Recall that, when σ^2 was known, $s_1^2 = (n+1)\sigma^2/n$.) With this notation, the lower bound $\underline{B}_1$ on the Bayes factor can be computed from (2.15) to be

$$\begin{aligned}\underline{B}_1 &= z(\tfrac{\alpha+1}{\alpha+z^2})^{(\alpha+1)/2} && \text{for } z>1,\\ &= 1, && \text{for } z\le 1,\end{aligned} \tag{2.18}$$

where $\alpha = z(a_1-1)$ is the degrees of freedom of both predictive distributions. For the non-informative prior, $\alpha = n-3$. It can be checked that this lower bound, $\underline{B}_1$, converges, as α goes to ∞, to the known σ^2 lower bound, given in (2.6).

Case 2: $\Gamma_2 = \{$all non-increasing densities, g, on $[1,\infty)\}$.

In this case, the same argument as in Subsection 2.1 gives

$$\underline{B}_2 = \frac{m(x_0|\underline{x})}{\sup_r \frac{1}{r-1}\int_1^r m^*(x_0|\underline{x},\tau)d\tau}, \tag{2.19}$$

where $m(x_0|\underline{x})$ and $m^*(x_0|\underline{x},\tau)$ are the t densities given in (2.11) and (2.12), respectively.

Again defining $z = |x_0 - m_1|/s_1$, with m_1 and s_1 as in *case 1*, $\underline{B}_2$ can be computed to be

$$\begin{aligned} \underline{B}_2 &= \gamma^{-1/2}\left(\tfrac{\alpha+z^2}{\alpha+\gamma z^2}\right)^{-(\alpha+1)/2} && \text{for } z > 1, \\ &= 1, && \text{for } z \le 1, \end{aligned} \tag{2.20}$$

where $\alpha = 2(a_1 - 1)$ and γ is the unique solution of

$$\begin{aligned} &\frac{2}{z}\left\{\frac{1}{\sqrt{\gamma}}\left(1+\frac{\gamma z^2}{\alpha}\right)^{-(\alpha+1)/2} - \left(1+\frac{z^2}{\alpha}\right)^{-(\alpha+1)/2}\right\} \\ &- \frac{\sqrt{\alpha\pi}\,\Gamma(\frac{\alpha}{2})}{\Gamma(\frac{\alpha+1}{2})}\left\{I_{\frac{\alpha}{\alpha+\gamma z^2}}\left(\frac{\alpha+2}{2},\frac{1}{2}\right) - I_{\frac{\alpha}{\alpha+z^2}}\left(\frac{\alpha+2}{2},\frac{1}{2}\right)\right\} \\ &= \frac{1-\gamma}{z\sqrt{\gamma}}\left[1+\frac{\gamma z^2}{\alpha}\right]^{-(\alpha+1)/2}, \end{aligned} \tag{2.21}$$

where

$$I_x(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^x u^{\alpha-1}(1-u)^{\beta-1}du \tag{2.22}$$

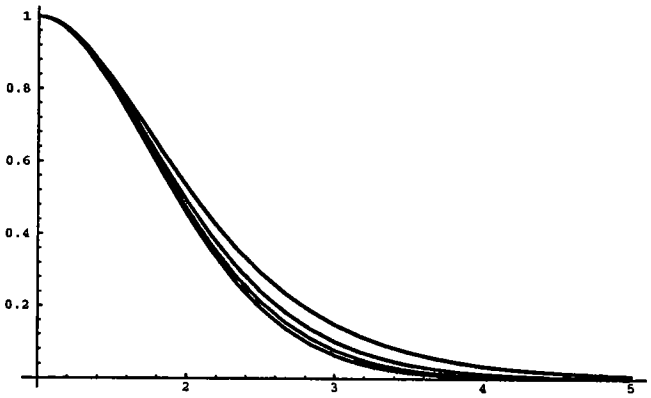
is the incomplete beta function.

As in Subsection 2.1, where σ^2 was assumed known, $\underline{B}_2$ is again surprisingly close to $\underline{B}_1$. (Obviously $\underline{B}_2 > \underline{B}_1$.) In Table 2, we display the values of $\underline{B}_1$ and $\underline{B}_2$ for certain values of z and for $\alpha = 20$. (Smaller values of α produced larger values of both lower bounds, $\underline{B}_1$ and $\underline{B}_2$.) In Figure 2, the lower bound $\underline{B}_1$ is graphed, as a function of z, for $\alpha = 10$, 20, 40, 100. (The corresponding graphs of $\underline{B}_2$ were extremely similar.) Both Table 2 and Figure 2 were derived with the non-informative prior.

TABLE 2. LOWER BOUNDS ON THE BAYES FACTORS FOR THE SCALE CONTAMINATION MODEL (σ^2 UNKNOWN).

z	1.5	2	2.5	3	3.5	4	4.5	5
$\underline{B}_1$	.8174	.4922	.2401	.1012	.0387	.0139	.0049	.0017
$\underline{B}_2$	.8437	.5300	.2665	.1146	.0444	.0161	.0056	.0020

FIGURE 2. $\underline{B}_1$ FOR THE SCALE-CONTAMINATION MODEL (σ^2 UNKNOWN) AS A FUNCTION OF z FOR $\alpha = 10, 20, 40, 100$.



Comparison of Tables 1 and 2 or Figures 1 and 2 show that, for moderate or large α, there is little difference between the known σ^2 and unknown σ^2 cases, as long as $z < 2.5$. However, for $z > 2.5$ the difference can be significant, with the lower bounds being substantially larger in the case of unknown σ^2. (It is, of course, not surprising that unknown σ^2 makes it less certain that large z implies that x_0 is an outlier). The stability of $\underline{B}_1$ to varying α, that is exhibited in Figure 2, actually held to remarkably small values (e.g., $\alpha = 3$).

3. LOCATION CONTAMINATION

In this section, we assume that the non-outliers, $\underset{\sim}{x} = (x_1, \ldots, x_n)$, come from a $N(\theta, \sigma^2)$ distribution and that the possible outlier, x_0, is generated by the location contamination mixture

$$X_0 \sim (1 - \varepsilon)N(\theta, \sigma^2) + \varepsilon N(\theta + \tau, \sigma^2). \tag{3.1}$$

It should be noted that this outlier model has the scale contamination model as a particular case, corresponding to certain prior distributions for τ. Indeed, if $X_0 \sim N(\theta + \tau, \sigma^2)$ and τ has the $N(0, (\nu - 1)\sigma^2)$ prior distribution (with $\nu > 1$), then, conditional on θ, ν and σ^2, the distribution of X_0 is given by the scale contamination model $X_0 \sim N(\theta, \nu\sigma^2)$ with $\nu > 1$; putting prior distributions on ν would thus yield the Section 2 analyses. It follows that the lower bounds to be computed in this section will all be smaller than those obtained in Section 2.

Here $\pi(\theta|\underset{\sim}{x})$ and $m(x_0|\underset{\sim}{x})$ are as in Section 2, while $m^*(x_0|\underset{\sim}{x})$ will denote the (posterior) predictive density of X_0, assuming now that $X_0 \sim N(\theta + \tau, \sigma^2)$. Note that the symbol τ is still used to denote the contaminating parameter, though, of course, the restriction $\tau \geq 1$ no longer applies.

3.1 σ Known.

Under a conjugate analysis, $\pi(\theta)$ and $\pi(\theta|\underset{\sim}{x})$ are the same as in Subsection 2.1,

and the Bayes factor in favor of H_0: *non-outlier* ($\tau = 0$) is

$$B(g) = \frac{\int N(x_0|\theta,\sigma^2)N(\theta|m_1,\sigma_1^2)d\theta}{\int\int N(x_0|\theta+\tau,\sigma^2)N(\theta|m_1,\sigma_1^2)g(\tau)d\theta d\tau} = \frac{N(x_0|m_1,s_1^2)}{\int N(x_0|m_1+\tau,s_1^2)g(\tau)d\tau}, \tag{3.2}$$

where $s_1^2 = \sigma^2 + \sigma_1^2$, and m_1 and σ_1^2 are given in (2.2). We shall again study

$$\underline{B} = \inf_{g\in\Gamma} B(g) = \frac{N(x_0|m_1,s_1^2)}{\sup_{g\in\Gamma}\int N(x_0|m_1+\tau,s_1^2)g(\tau)d\tau},$$

for two classes, Γ, of priors for τ.

Case 1: $\Gamma_1 = \{$all priors $g(\tau)\}$.

In this case,

$$\underline{B}_1 = \frac{N(x_0|m_1,s_1^2)}{\sup_\tau N(x_0|m_1+\tau,s_1^2)}, \tag{3.3}$$

which is trivially seen to be

$$\underline{B}_1 = e^{-z^2/2}, \tag{3.4}$$

where, again, $z = |x_0 - m_1|/s_1$, the standardized distance of x_0 to its predicted value under the non-outlier model.

Case 2: $\Gamma_2 = \{$all densities $g(\tau) = h(|\tau|)$, h non-increasing$\}$.

Since τ represents a shift of the mean θ, a natural way to restrict the class is to consider all densities that are symmetric about $\tau = 0$, and that are non-increasing in $|\tau|$. These densities can, again, be expressed as mixtures of uniform densities, so that the lower bound on the Bayes factor is

$$\underline{B}_2 = \frac{N(x_0|m_1,s_1^2)}{\sup_r \frac{1}{2r}\int_{-r}^{r} N(x_0|m_1+\tau,s_1^2)d\tau}, \tag{3.5}$$

which can be calculated to be

$$\begin{aligned} \underline{B}_2 &= \frac{2\exp\{-z^2/2\}}{\exp\{-(z+\gamma)^2/2\} - \exp\{-(z-\gamma)^2/2\}} && \text{for } z > 1, \\ &= 1, && \text{for } z \le 1, \end{aligned} \tag{3.6}$$

where $z = |x_0 - m_1|/s_1$, and γ is the unique solution of

$$\gamma[\phi(z+\gamma) + \phi(z-\gamma)] = \Phi(z+\gamma) - \Phi(z-\gamma), \tag{3.7}$$

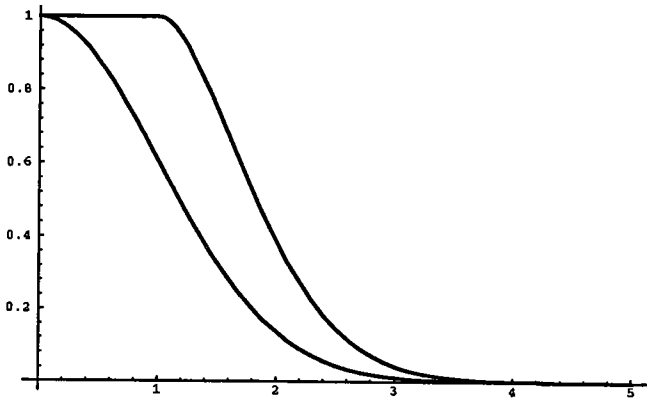
where ϕ, Φ represent the standard normal p.d.f. and c.d.f., respectively.

In Table 3 we give $\underline{B}_1$ and $\underline{B}_2$ for certain values of z. In Figure 3, they are graphed as functions of z. Both Table 3 and Figure 3 are, as before, derived for the non-informative prior $\pi(\theta) = 1$. In contrast to the situation in Section 2 with the

TABLE 3. LOWER BOUNDS ON THE BAYES FACTOR FOR THE LOCATION CONTAMINATION MODEL (σ^2 KNOWN).

z	1.5	2	2.5	3	3.5	4	4.5
$\underline{B}_1$	.3247	.1353	.0439	.0111	.0022	.0003	.0000
$\underline{B}_2$	.7493	.3835	.1458	.0420	.0093	.0016	.0002

FIGURE 3. $\underline{B}_1$ AND $\underline{B}_2$ FOR THE LOCATION-CONTAMINATION MODEL (σ^2 KNOWN) AS FUNCTIONS OF z.



scale contamination model, the differences between $\underline{B}_1$ and $\underline{B}_2$ can here be quite substantial.

3.2 σ^2 Unknown.

Letting $\delta = 1/\sigma^2$, the joint prior and the joint posterior (given only $\underline{x}$) of (θ, δ) are the same as those in Subsection 2.2. Furthermore, the posterior predictive under the non-outlier $m(x_0|\underline{x})$ is as in (2.11), while the predictive under the outlier model for x_0, $m^*(x_0|\underline{x}, \tau)$, is

$$m^*(x_0|\underline{x}, \tau) = St(m_1 + \tau, \frac{(h_1+1)}{h_1}\frac{b_1}{(a_1-1)}, 2(a_1-1)). \tag{3.7}$$

Again define $z = |x_0 - m_1|/s_1$, where m_1 and s_1 are the location and scale parameter of the non-outlier predictive, $m(x_0|\underline{x})$, and are given in (2.10), together with h_1, a_1, b_1 that appear in (3.7) above. Our goal is to study

$$\underline{B} = \inf_{g\in\Gamma} B(g) = \frac{St(x_0|m_1, s_1^2, \alpha)}{\sup_{g\in\Gamma} \int St(x_0|m_1+\tau, s_1^2, \alpha)g(\tau)d\tau}, \tag{3.8}$$

where $\alpha = 2(a_1 - 1)$.

Case 1: $\Gamma_1 = \{\text{all priors } g(\tau)\}$.
In this case, (3.8) becomes

$$\underline{B}_1 = \frac{St(x_0|m_1, s_1^2, \alpha)}{\sup_\tau St(x_0|m_1+\tau, s_1^2, \alpha)}, \tag{3.9}$$

which is trivially seen to be

$$\underline{B}_1 = \left(\frac{\alpha}{\alpha+z^2}\right)^{(\alpha+1)/2}. \tag{3.10}$$

Note that, as $\alpha \to \infty$ (recall that, with the non-informative prior, $\alpha = n-3$), (3.10) goes to the $\underline{B}_1$ given in (3.4) for known σ^2.

Case 2: $\Gamma_2 = \{\text{all densities } g(\tau) = h(|\tau|),\ h \text{ non increasing}\}$.
In this case, (3.8) becomes

$$\underline{B}_2 = \frac{St(x_0|m_1, s_1^2, \alpha)}{\sup_r \frac{1}{2r}\int_{-r}^{r} St(x_0|m_1+\tau, s_1^2, \alpha)d\tau}, \tag{3.11}$$

which can be calculated to be

$$\underline{B}_2 = \frac{2[1+z^2/\alpha]^{-(\alpha+1)/2}}{[1+(z+\gamma)^2/\alpha]^{-(\alpha+1)/2} + [1+(z-\gamma)^2/\alpha]^{-(\alpha+1)/2}}, \tag{3.12}$$

where γ is the solution of

$$\gamma[t_\alpha(z+\gamma) + t_\alpha(z-\gamma)] = T_\alpha(z+\gamma) - T_\alpha(z-\gamma), \tag{3.13}$$

with t_α and T_α being the p.d.f. and c.d.f., respectively, of a $St(0,1,\alpha)$ distribution.

TABLE 4. LOWER BOUNDS ON THE BAYES FACTORS FOR THE LOCATION CONTAMINATION MODEL WITH σ^2 UNKNOWN AND $\alpha = 1, 10, 50$.

	$\alpha = 1$		$\alpha = 10$		$\alpha = 50$	
z	$\underline{B}_1$	$\underline{B}_2$	$\underline{B}_1$	$\underline{B}_2$	$\underline{B}_1$	$\underline{B}_2$
1.5	.3077	.7008	.3275	.7461	.3255	.7490
2	.2000	.5476	.1571	.4398	1405	.3972
2.5	.1379	.4387	.0692	.2273	.0496	.1643
3	.1000	.3611	.0293	.1098	.0147	.0555
3.5	.0755	.3043	.0123	.0516	.0037	.0158
4	.0588	.2615	.0052	.0242	.0008	.0039
4.5	.0471	.2284	.0023	.0115	.0002	.0009
5	.0385	.2021	.0010	.0056	.0000	.0002

Table 4 gives the values of $\underline{B}_1$ and $\underline{B}_2$ for $\alpha = 1$, 10 and 50 (corresponding to $n = 4$, 13 and 53 non-outliers in the sample). Figure 4 graphs $\underline{B}_1$, as a function of z, for $\alpha = 1$ and $\alpha = \infty$ (which corresponds to the known σ^2 case). Figure 5 graphs $\underline{B}_2$ for $\alpha = 1, 3, 5, 10$. All figures and tables use the non-informative prior $\pi(\theta, \delta) \propto 1/\delta$.

As in the scale contamination scenario, there is remarkably little difference between the lower bounds for the known σ^2 and unknown σ^2 cases when $z < 2.5$ (unless α is very small, e.g. $\alpha = 1$ or $\alpha = 2$). For $z > 2.5$, however, the difference can be appreciable, with the lower bounds for the unknown σ^2 case being substantially larger.

FIGURE 4. $\underline{B}_1$ FOR THE LOCATION-CONTAMINATION MODEL AS A FUNCTION OF z FOR σ^2 UNKNOWN $(\alpha = 1)$ AND σ^2 KNOWN.

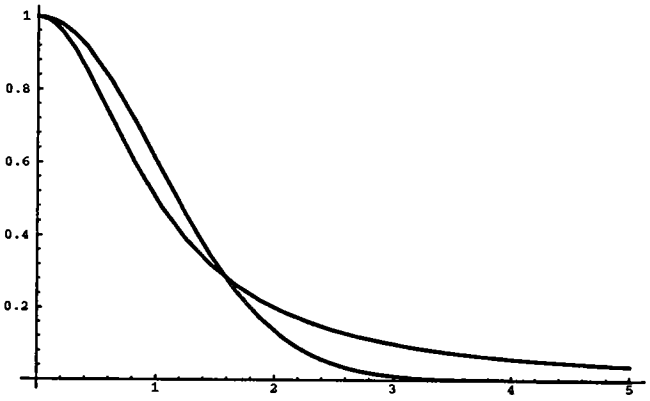
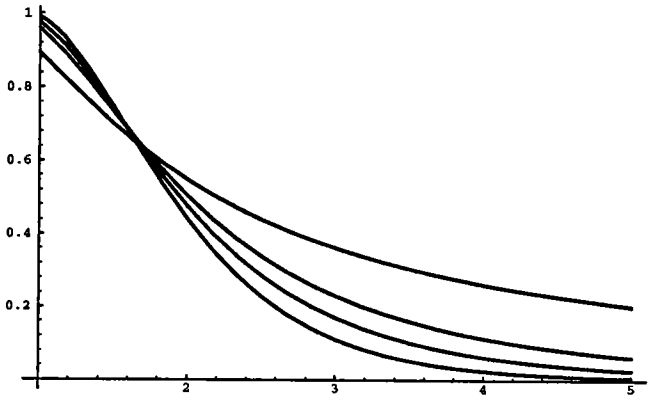


FIGURE 5. $\underline{B}_2$ FOR THE LOCATION-CONTAMINATION MODEL (σ^2 UNKNOWN) AS A FUNCTION OF z FOR $\alpha = 1, 3, 5, 10$.



4. Comparisons, Recommendations, and Generalizations

4.1 Comparisons and Recommendations.

Comparison of Tables 1 and 3 or Tables 2 and 4 reveals a rather remarkable similarity between $\underline{B}_1$ (scale contamination), $\underline{B}_2$ (scale contamination) and $\underline{B}_2$ (location contamination). These lower bounds do, however, differ substantially from $\underline{B}_1$ (location contamination). Since this last lower bound was based on the most unreasonably large class of priors (*all* priors for the contaminating location parameter), it will typically be an unreasonably small lower bound, and hence is not recommended for use.

Since the other three lower bounds are very similar, it is reasonable to choose among them according to computational simplicity. In this regard, $\underline{B}_1$ (scale contamination) is the clear winner, being given by the simple closed-form formulas (2.6) and (2.18) (for the known σ^2 and unknown σ^2 cases, respectively). It is, indeed, fortuitous (and unusual in Bayesian robustness) that a simple lower bound resulting from an "all prior" model is reasonable.

In the Introduction, it was suggested that a rough rule-of-thumb for outlier detection is to judge x_0 to be an outlier if $B \leq .01$. A *necessary* condition for this to be true, in the presence of prior uncertainty as reflected by the class Γ, is

$$\underline{B} = \inf_{g \in \Gamma} B(g) \leq .01. \tag{4.1}$$

For the recommended lower bounds (2.6) and (2.18), this necessary condition becomes (using a conservative bound for (2.18))

$$\begin{aligned} &\text{for known } \sigma^2\text{: } z \geq 3.57 \\ &\text{for unknown } \sigma^2\text{: } z \geq (3.57)(1-(.982)/\sqrt{\alpha})^{-1}, \end{aligned} \tag{4.2}$$

where $z = |x_0 - m_1|/s_1$ is the distance between x_0 and the (non-outlier) predictive mean m_1, in units of (non-outlier) predictive standard deviations. For the noninformative prior case, recall that $m_1 = \bar{x}$ and $s_1^2 = (n+1)\sigma^2/n$ (known σ^2 case) or $s_1^2 = (n+1)\sum_{i=1}^{n}(x_i - \bar{x})^2/[n(n-3)]$ (unknown σ^2 case). The second inequality in (4.2) is an approximation to (4.1) and (2.18).

The implications of this rule-of-thumb to practice are perhaps surprising. It would seem that x_0 needs to be at least 4 standard errors out (somewhat less for the known σ^2 case; somewhat more for unknown σ^2) before one can consider it to be a likely outlier. And, since $\underline{B}_1$ was a lower bound, even this would not conclusively prove that x_0 is an outlier. Our recommendation for practice is thus to use (4.2) as a preliminary screen; only observations for which z exceeds the indicated values need be investigated as possible outliers.

4.2 Generalizations.

Recall that we based our analyses on

$$\Pr(x_0 \text{ is an outlier} \mid \text{data}, \underset{\sim}{x} \text{ being non-outliers}). \tag{4.3}$$

Of perhaps more interest is

$$\Pr(x_0 \text{ is an outlier}, \underset{\sim}{x} \text{ are non-outliers} \mid \text{data}), \tag{4.4}$$

which does not assume that the remaining observations are known to be non-outliers. It is possible to show, however, that (4.3) is a good approximation to (4.4), if x_0 seems to be the only likely outlier. Here is a heuristic argument to this effect.

If $\underset{\sim}{x} = (x_1, \ldots, x_n)$ are not outlying (i.e., are reasonably compatible with $f(x_i|\theta)$), then their joint density under the full mixture model for all observations is

$$\prod_{i=1}^{n}[(1-\varepsilon)f(x_i|\theta) + \varepsilon f^*(x_i|\theta,\tau)] \cong (1-\varepsilon)^n \prod_{i=1}^{n} f(x_i|\theta),$$

the approximation based on observing that ε is typically small and that $f(x_i|\theta)$ will be comparable or larger than $f^*(x_i|\theta,\tau)$. But then

$$\begin{aligned}
&\Pr\,(x_0 \text{ is an outlier, } \underset{\sim}{x} \text{ are non-outliers} \mid \text{data}) \\
&= \frac{\varepsilon(1-\varepsilon)^n \int\int f^*(x_0|\theta,\tau)\left(\prod_{i=1}^{n} f(x_i|\theta)\right)\pi(\theta,\tau)\,d\theta\,d\tau}{\int\int \prod_{i=0}^{n}[(1-\varepsilon)f(x_i|\theta) + \varepsilon f^*(x_i|\theta,\tau)]\pi(\theta,\tau)\,d\theta\,d\tau} \\
&\cong \frac{\varepsilon \int\int f^*(x_0|\theta,\tau)\left(\prod_{i=1}^{n} f(x_i|\theta)\right)\pi(\theta,\tau)\,d\theta\,d\tau}{\int\int[(1-\varepsilon)f(x_0|\theta) + \varepsilon f^*(x_0|\theta,\tau)]\left(\prod_{i=1}^{n} f(x_i|\theta)\right)\pi(\theta,\tau)\,d\theta\,d\tau} \\
&= \Pr\,(x_0 \text{ is an outlier} \mid \text{data, } \underset{\sim}{x} \text{ being non-outliers}),
\end{aligned}$$

completing the argument.

If there seems to be more than one likely outlier, the situation becomes considerably more complicated. The main complication is that the possible outliers, in combination, can perhaps provide significant information about the contaminating distribution. To properly analyze this situation, it is probably necessary to deal with the formal mixture model.

It is, however, trivial to consider the generalization to the Bayes factor for testing

H_0: $\underset{\sim}{x}_0 = (x_{01}, \ldots, x_{0k})$ *all non-outliers*, versus
H_1: $\underset{\sim}{x}_0$ *all outliers*.

Indeed, all expressions obtained for $k = 1$ (i.e., a single outlier, x_0) are valid for the above hypotheses if z is instead defined as $z = |\bar{x}_0 - m_1|/s_1^*$, where $\bar{x}_0 = \sum_{i=1}^{k} x_{0i}/k$ and

$$s_1^{*2} = \begin{cases} \sigma_1^2 + \sigma^2/k & \text{if } \sigma^2 \text{ is known,} \\ (\frac{1}{k} + \frac{1}{n_1})\frac{b_1}{(a_1-1)} & \text{if } \sigma^2 \text{ is unknown.} \end{cases}$$

A final generalization of interest is to the multiple outlier scenario when the outliers are thought to arise from *different* contaminating distributions. Then the outliers could be checked one-at-a-time, using (4.2), with the modification that $\underset{\sim}{x}$ now consist only of the clearly non-outlying observations.

REFERENCES

Abraham, B., and Box, G. E. P. (1978). Linear models and spurious observations. *Applied Statistics* **27**, 120–130.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science* **2**, 317–352.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of significance levels and evidence. *Journal of the American Statistical Association* **82**, 112–122.

Box, G. E. P., and Tiao, G. C. (1968). A Bayesian approach to some outlier problems. *Biometrika* **55**, 119–129.

Chaloner, K., and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika* **75**, 651–659.

Dawid, A. P. (1973). Posterior expectations for large observations. *Biometrika* **60**, 664–667.

DeFinetti, B. (1961). The Bayesian approach to the rejection of outliers. *Proceedings of the Fourth Berkeley Symposium on Probability and Statistics* **1**, 199–210. Berkeley: University Press.

Freeman, P. R. (1980). On the number of outliers in data from a linear model. In *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, A. F. M. Smith, D. V. Lindley, eds.), 349–365. Valencia: Valencia University Press.

Geisser, S. (1980). Discussion of Sampling and Bayes' inference in scientific modelling and robustness, by G. E. P. Box. *Journal of the Royal Statistical Society A* **143**, 416–417.

Geisser, S. (1985). On the prediction of observables: a selective update. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, A. F. M. Smith, D. V. Lindley, eds.), 203–229. Amsterdam: North-Holland.

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

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

Geisser, S. (1990). Predictive approaches to discordancy testing. In *Bayesian and Likelihood Methods in Statistics and Econometrics* (S. Geisser, J. S. Hodges, S. J. Press, A. Zellner, eds.), 321–335. Amsterdam: North Holland.

Girón, F. J., Martínez, M. L., and Morcillo, C. (1992). A Bayesian justification for the analysis of residuals and influence measures. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith, eds.), 651–660. Oxford: Clarendon Press.

Gómez-Villegas, M. A., and Maín, M. (1992). The influence of prior and likelihood tail behaviour on the posterior distribution. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith, eds.), 661–667. Oxford: Clarendon Press.

Guttman, I. (1973). Care and handling of univariate or multivariate outliers in detecting spuriosity, a Bayesian approach. *Technometrics* **15**, 723–738.

Guttman, I., Dutter, R., and Freeman, P. R. (1978). Care and handling of univariate outliers in the general linear model to detect spuriosity — a Bayesian approach. *Technometrics* **20**, 187–193.

Guttman, I., and Peña, D. (1988). Outliers and influence: evaluation by posteriors of parameters in the linear model. In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith, eds.), 631–640. Oxford: Clarendon Press.

Jeffreys, H. (1983). *Theory of Probability (Third edition)*. Oxford: Clarendon Press.

O'Hagan, A. (1979). On outlier rejection phenomena in Bayes inference. *Journal of the Royal Statistical Society B* **41**, 358–367.

O'Hagan, A. (1988). Modelling with heavy tails. In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith, eds.), 345–359.

Peña, D. and Tiao, G. C. (1992). Bayesian robustness functions for linear models. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith, eds.), 365–388. Oxford: Clarendon Press.

Pettit, L. J. (1988). Bayes methods for outliers in exponential samples. *Journal of the Royal Statistical Society B* **50**, 371–380.

Pettit, L. I. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society B* **52**, 175–184.

Pettit, L. I. (1992). Bayes factors for outlier models using the device of imaginary observations. *Journal of the American Statistical Association* **87**, 541–545.

Pettit, L. I., and Smith, A. F. M. (1983). Bayesian model comparison in the presence of outliers. *Bulletin of the International Statistical Institute* **50**, 292–309.

Pettit, L. I., and Smith, A. F. M. (1985). Outliers and influential observations in linear models. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith, eds.), 473–494. Amsterdam: North-Holland.

Racine-Poon, A. (1992). SAGA: Sample Assisted Graphical Analysis. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith, eds.), 389–404. Oxford: Clarendon Press.

Spiegelhalter, D. J., and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society B* **44**, 377–387.

Verdinelli, I., and Wasserman, L. (1991). Bayesian analysis of outlier problems using the Gibbs sampler. *Statistics and Computing* **1**, 105–117.

West, M. (1984). Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society B* **46**, 431–439.

West, M. (1985). Generalised linear models; scale parameters, outlier accommodation and prior distributions. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith, eds.), 531–558. Amsterdam: North Holland.

Zellner, A. (1975). Bayesian Analysis of regression error terms. *Journal of the American Statistical Association* **70**, 138–144.