BAYESIAN METHODS FOR OUTLIERS IN UNIFORM AND PARETO SAMPLES

by

Benjamin Charles Gaby, BSc Honours

A thesis submitted for the degree of Master of Philosophy

Queen Mary University of London, April 2012

Declaration

I declare that the work presented in the thesis is my own research.

ABSTRACT

We begin by reviewing the current literature on outliers and look at what has been done both classically and from a Bayesian viewpoint. We then extend these Bayesian ideas to model outliers in uniform and Pareto samples.

We consider the problem of deciding if there are any outliers in a sample from a uniform distribution. For a sample from a one parameter uniform distribution we show that the largest observation in the sample has the smallest conditional predictive ordinate. Hence we derive the Bayes factor for testing whether it is an outlier when the amount of contamination is known and unknown using two different outlier models. Then we investigate this problem when we have multiple outliers, assuming that our outliers are generated by the same probability distribution or by different probability distributions. Similarly for two parameter uniform samples we show that the most extreme observation in the sample has the smallest conditional predictive ordinate. Hence we derive the Bayes factors for testing whether extreme observations are outliers using the stricter outlier model that we had for the one parameter case.

We consider the problem of deciding if there are any outliers in a sample from a Pareto distribution. For a sample from a univariate Pareto distribution we show that the largest observation in the sample has the smallest conditional predictive ordinate and derive the Bayes factor for testing whether

it is an outlier when the amount of contamination is known and unknown. Then we investigate this problem when we have multiple outliers, assuming that our outliers are generated by the same probability distribution or by different probability distributions. Finally we extend these ideas to the multivariate case both when the marginal samples are independent of one another and when there are correlations/partial correlations.

CONTENTS

1. INTRODUCTION

1.1 Outliers - What they are

We first discuss what is meant by an outlier. Much has been written on the subject, for example Barnett and Lewis (1995) and Beckman and Cook (1983) and the many references therein. We may usefully distinguish between two types of observations which have at times been referred to as outliers. Firstly there are contaminants, observations which have been generated by a different probabilistic mechanism to the rest of the sample. Secondly there are extremes, observations which are away from the mass of other observations and thus cause surprise. Different authors refer to contaminants or extremes as outliers. From hereafter we define outliers as extreme contaminants, contaminated observations which are extreme.

Generally there are two approaches to dealing with outliers which Barnett and Lewis (1995) refer to as identification and accommodation. Classical methods for identification range from informal graphical techniques, often involving residuals, to formal tests of significance. Later we will discuss in detail Bayesian diagnostics which indicate surprising observations. Bayes factors can be used to test whether extreme observations are outliers, many of these are functions of classical test statistics. Accommodation is often achieved classically by the use of robust estimation techniques which have

good frequentist properties even when some of the underlying assumptions do not hold. Parameter estimates from Bayesian mixture models accommodate outliers, but the posterior weights on the models can also identify the number of outliers. In everything that follows we use the identification approach for dealing with outliers.

1.2 Some classical approaches to outliers

For univariate samples Barnett and Lewis (1995) describe the seven most common test statistics used for testing whether extreme observations are outliers. They are:

(1) **Excess/spread statistics** - These are the ratio of the difference between the suspected outlier and next most extreme observation in the sample to some measure of spread. One example of these are Dixon statistics which are discussed in Dixon (1951), other examples are discussed in Irwin (1925).

(2) **Range/spread statistics** - These are the ratio of the range of the data to some measure of spread. Examples of these are discussed in David, Hartley and Pearson (1954), and Pearson and Stephens (1964).

(3) **Deviation/spread statistics** - These are the ratio of the distance between the suspected outlier from some measure of central tendency to some measure of spread. Examples of these are discussed in Grubbs (1950).

(4) **Sums of squares statistics** - These are ratios of sums of squares for the restricted and total samples. Grubbs (1950) uses such test statistics for testing whether extreme observations are outliers in samples from a normal distribution.

(5) **Exteme/location statistics** - These are ratios of extreme observations to measures of location. Such test statistics can be used for testing whether extreme observations are outliers in samples from a gamma distribution, which are discussed in Likes (1966), Lewis and Fieller (1979), Kimber and Stevens (1981), and Kimber (1982).

(6) **Higher order moment statistics** - Statistics such as measures of skewness and kurtosis, not specifically designed for assessing outliers, can none the less be useful in this context; for example Ferguson (1961).

(7) **W statistics** - These again are not specifically designed for assessing outliers, but can none the less be useful in this context. These are the ratio of the square of a particular linear combination of the ordered sample to the sum of squares of the individual deviations about the sample mean. Examples of these are discussed in Shapiro and Wilk (1965, 1972), Shapiro, Wilk and Chen (1968), and Stephens (1978).

When there is more than one extreme observation in the sample many of the above test statistics are sensitive to both masking and swamping and so it creates problems. When testing whether a set of extreme observations are outliers and not including all of the extreme observations in the sample, we may fail to declare the extreme observations in the set as outliers which is known as masking. However in some cases we may have the opposite problem of swamping, which is declaring non extreme observations as outliers when testing whether a set containing both very extreme and non extreme observations are outliers. Suppose that the only outliers in the ordered sample $\{x_1, ..., x_n\}$ are x_n and x_{n-1}. Masking can be illustrated using the test statistic $T_1 = \frac{X_n}{\sum_{j=1}^{n} X_j}$, we see that in this case when testing whether x_n is an outlier

x_{n-1} is larger than it should be, making the observed value of T_1 smaller than it should be and hence we may possibly fail to declare x_n as an outlier. Swamping can be illustrated using the test statistic $T_3 = \frac{X_n+X_{n-1}+X_{n-2}}{\sum_{j=1}^{n} X_j}$, as in this case when testing whether x_n, x_{n-1} and x_{n-2} are outliers x_n and x_{n-1} may be sufficiently large to make T_3 large enough to declare x_{n-2} as an outlier. To overcome these problems Barnett and Lewis (1995) suggest to use the following procedure:

Consider the least extreme observation that could possibly be an outlier and delete all of the more extreme observations from the sample. Then for this new sample test whether it is an outlier. If we conclude that it is, then all of the other extreme observations are outliers. Otherwise we conclude that this observation is not an outlier and repeat the procedure until either we conclude that an observation is an outlier or that none of the observations are outliers.

Barnett and Lewis (1995) describe many tests which are based on the previous statistics for testing whether extreme observations are outliers in samples from normal and gamma type distributions when making many different assumptions. These include whether or not the distribution parameters are known or unknown, the number of outliers that we are testing for and how we define an outlier (whether we do a one sided test for upper or lower outliers or a two sided test for any outliers). The difficulty with using these tests is that all of the test statistics have very complicated test distributions and so it is hard to find critical values without using simulation techniques. They also use similar test procedures for Gumbel, Frechet and Weibull samples, as well as introducing conditional test statistics to extend these ideas

to the Poisson and binomial cases.

For a sample from a uniform(θ_1, θ_2) distribution with both θ_1 and θ_2 unknown, Barnett and Roberts (1993) construct the following test for testing whether extreme observations are outliers, based on the fact that differences between the order statistics have an exponential distribution. When we suspect that there are u upper outliers and l lower outliers in our ordered sample $\{x_1, ..., x_n\}$ the test statistic is

$$F = \frac{(n-u-l-1)\,(X_n - X_{n-u} + X_{l+1} - X_1)}{(u+l)\,(X_{n-u} - X_{l+1})}$$

and has a $F^{2(u+l)}_{2(n-u-l-1)}$ test distribution. From this test statistic the obvious tests can be derived in the special cases when either we only have observations suspected of being upper outliers, only have observations suspected of being lower outliers or have known distribution parameters. See Barnett and Roberts (1993) or Barnett and Lewis (1995) for more details. As an example to illustrate this test consider the testing problem in part (ii) of **Example 2.6** in **Section 2.5**. For this problem the observed value of F is equal to 9.596 and the test distribution is F^4_{14}. The critical value for the 0.1% significance level test is 8.622, hence there is overwhelming evidence against the null hypothesis that there are no outliers in the sample.

When we have a sample from a Pareto(θ, k) distribution and k is unknown, we transform the data to an exponential(θ) sample with origin $\log(k)$, where there are excess/spread test statistics in six different cases for exponential samples with an unknown origin. Note that if k is assumed to be known, we transform the data to a regular exponential(θ) sample and can use all of the test statistics for a particular case on such samples.

For multivariate samples Barnett (1979) derived a series of tests that were based on simulations to test whether extreme points in a sample are outliers. He did this for a variety of different probability distributions including the normal, exponential, uniform and Pareto distributions. For uniform samples he focuses on the bivariate case, here he assumes that the end points of the intervals are known and we do not see how any point which lies in a known rectangle can be described as outlying. For Pareto samples he focuses on the bivariate case, here he simulates the critical values based on finding the value that makes the cumulative distribution function of $\frac{X_{1j}}{k_1} + \frac{X_{2j}}{k_2}$ equal to $(1-\alpha)^{\frac{1}{n}}$, where α is the significance level of the test and the (X_{1j}, X_{2j}) are independent and all come from the same bivariate Pareto distribution with distinct terminals k_1 and k_2. These values are used to test whether a single observation is an outlier when assuming that the correlation structure is known, but it is not extended to the case when the correlation structure is unknown or to multiple outlier problems.

1.3 Some Bayesian approaches to outliers

The conditional predictive ordinate (CPO) was first defined by Geisser (1980). Given a sample $\mathbf{x} = \{x_1, ..., x_n\}$, let us suppose that the standard model generating the observations has the form $p(x|\theta)$, where θ represents all of the unknown parameters. Let us denote by $\mathbf{x_S}$ the elements of $\mathbf{x}$ whose labels are in $S \subset \{1, ..., n\}$ and (S) by the complement of S. Then we define the CPO for $\mathbf{x_S}$ given $\mathbf{x_{(S)}}$ by

$$p\left(\mathbf{x_S}|\mathbf{x_{(S)}}\right) = \int_{-\infty}^{\infty} p\left(\mathbf{x_S}|\theta\right) p\left(\theta|\mathbf{x_{(S)}}\right) d\theta.$$

Small values of this indicate that observations $\mathbf{x_S}$ are surprising in relation to $\mathbf{x_{(S)}}$ and the prior specification for θ. The CPO can be used to order individual observations, pairs, triples, etc. on the basis of their surprisingness compared with the other observations, and is useful in detecting potential outliers.

For the sample $\mathbf{x} = \{x_1, ..., x_n\}$, let M_0 and M_q denote the models with no and q outliers in the sample respectively. The model generating good observations is $p(x|M_0, \theta)$, where θ represents all of the unknown parameters and $p(\theta)$ is our prior specification for θ. Suppose we suspect that q observations in the sample are outliers generated by the same model $p(x|M_q, \theta, \delta)$, where θ represents all of the unknown null model parameters, δ represents the contamination factor and $p(\theta, \delta)$ is our joint prior specification for θ and δ. Then the Bayes factor for comparing the models M_0 and M_q is defined by

$$B_{0,q} = \frac{p(\mathbf{x}|M_0)}{p(\mathbf{x}|M_q)} = \frac{\int_{-\infty}^{\infty} p(\mathbf{x}|M_0, \theta)\, p(\theta)\, d\theta}{\int_{-\infty}^{\infty} p(\mathbf{x}|M_q, \theta, \delta)\, p(\theta, \delta)\, d\theta d\delta}$$

and is used to formally test whether a set of extreme observations are outliers. It is often a function of the corresponding classical test statistic. A small Bayes factor is needed to declare the set of extreme observations as outliers. Originally when making general inferences, not necessarily about outliers, Jeffreys (1961) suggests that a Bayes factor of less than or equal to $10^{-\frac{1}{2}}$ or 10^{-1} is needed to select the alternative model over the null model. Pettit (1992) argues that these values are not appropriate for outlier problems, as we have some knowledge that we can use to give an approximation to the prior probabilities of the null and alternative models. He suggests that we can judge how small a Bayes factor is necessary to lead us to a greater posterior probability on the alternative model than the null model. Then gives a formal argument for assuming that in the case of comparing the models M_γ and

$M_{\gamma+1}$, a Bayes factor of less than or equal to 0.015 provides strong evidence for concluding that an observation is an outlier and a Bayes factor of less than or equal to 0.005 provides very strong evidence for concluding that an observation is an outlier. This is based on believing that the probability that a randomly chosen observation is an outlier in a sample of size n is 0.05 together with the scale suggested by Jeffreys. He derives the values for other cases in a similar way.

Two early Bayesian models for modelling contaminants are the ones given by Box and Tiao (1968) and Guttman, Dutter and Freeman (1978). These authors assume that, in a given data set, most observations arise from a distribution with a normal density $p(x|\theta)$, where θ represents all of the unknown parameters, but a few observations may arise from an alternative normal density $p_\delta(x|\theta)$, where δ is an additional parameter, which may be known or unknown, representing some form of alternative. The models are distinguished by the assumptions about the alternative. These models and that of Abraham and Box (1978) have been reviewed by Freeman (1980). Consider the standard linear model

$$Y = X\theta + \epsilon,$$

where Y is an $n \times 1$ vector of random variables, X is an $n \times p$ known design matrix, θ is a $p \times 1$ vector of unknown parameters and ϵ is an $n \times 1$ vector of random errors. Suppose that a particular subset $\{y_{i_1}, ..., y_{i_r}\}$ of the y's are suspected of being contaminants. Let $R = \{i_1, ..., i_r\}$ which is a subset of $\{1, ..., n\}$. Denote by y_R the vector whose components are the observations with labels in R and by $y_{(R)}$ the remaining components. We may partition the design matrix correspondingly into X_R and $X_{(R)}$. Application of Bayes

theorem gives the posterior distribution of θ as

$$p(\theta|\mathbf{y}) = \sum W_R p_R(\theta|\mathbf{y}),$$

where the summation extends over all 2^n possible subsets R, W_R denotes the posterior probability that y_R are contaminants and $y_{(R)}$ are not and $p_R(\theta|\mathbf{y})$ is the posterior distribution of θ given the assumed division of the data set into contaminants and good observations. For both models, given a normal distribution for the elements of ϵ and either conjugate priors or standard improper limits for the unknown parameters (θ, σ) it can be shown that $p_R(\theta|\mathbf{y})$ is a p-variate t-distribution with different means, dispersion matrices and degrees of freedom.

Box and Tiao (1968) assume that associated with each observation i there is a probability $1-\alpha$ that ϵ_i is normally distributed with mean zero and variance σ^2 and probability α that the variance is inflated to $\delta^2\sigma^2$. They assume that α and δ are known and use the standard improper uniform prior on θ and $\log\sigma$. They find that their results are fairly insensitive to choices of α in the range $(0.03, 0.07)$ and δ in the range $(3, 10)$.

Guttman, Dutter and Freeman (1978) consider the model

$$Y = X\theta + \delta_R + \epsilon,$$

where δ_R is a vector, exactly r of whose elements are non-zero; for example, if $R = \{1, 3\}$ then $\delta = (\delta_1, 0, \delta_3, ..., 0)^T$ represents the case when the first and third observations are contaminants and all other observations are not outliers. The non-zero elements of δ_R are given a uniform prior as are θ and $\log\sigma$. The elements of ϵ are all assumed to have independent $N(0, \sigma^2)$ distributions.

The main problem with all of the Bayesian models that we have discussed so far is that they assume that a contaminated observation will always out-lie. Other problems include not being able to use these models in situations when we have both extreme large and extreme small observations, if there is a large number of observations in R the computation gets very complicated, masking and swamping. Pettit and Smith (1983, 1985) suggest that these models can be used for allowing for contaminants, but for an outlier model there should be a high probability that a contaminant does out-lie. In the univariate normal case for both the shifted location and inflated variance models, they choose a value of δ such that it is almost certain that a contaminated observation will out-lie when δ is assumed to be known, but if δ is unknown choose a proper prior distribution that reflects this. For multivariate normal samples Pettit (1990) proves that the point in the sample with the smallest CPO (which is the most extreme point in the sample) must lie on one on the vertices of the convex hull of the observations. These ideas are then used for exponential samples in Pettit (1988) to derive the CPO to detect possible outliers and Bayes factors to declare whether or not they are outliers. Here he assumes that the good observations come from an exponential(λ) distribution and outliers come from an exponential($\delta\lambda$) distribution with $0 < \delta < 1$.

In some situations we may be unable or unwilling to specify a proper prior distribution for δ because we are unsure of the actual mechanism for generating outliers. Pettit (1992) reconsiders normal and exponential sample problems when giving δ an improper prior distribution, using the device of imaginary observations described in Spiegelhalter and Smith (1982). All

prior distributions depend on a constant that make its probability density function integrate to one over the whole distribution range, but in the case of improper prior distributions we cannot find this constant. Spiegelhalter and Smith (1982) approach this problem by considering the smallest possible experiment to distinguish between the null and alternative models that gives maximal support to the null model, then setting the Bayes factor for this experiment equal to one and solving for the unknown constant. Pettit (1994) uses this method for Poisson samples. O'Hagan (1995) suggested the method of fractional Bayes factors to overcome the problem of the undefined constant in model choice with improper priors. Pettit (1995) applied this method to outliers in Poisson samples. Sothinathan and Pettit (2005) applied both of these methods to binomial samples.

Verdinelli and Wasserman (1991) reconsider some of the previous problems using a Gibbs sampling approach by regarding outliers as unknown parameters. However Justel and Pena (1996) show that Gibbs sampling will fail in outlier problems due to strong masking in situations when there are multiple outliers.

1.4 An outline of the remaining chapters

In **Chapter 2** we extend these Bayesian methods to model outliers in uniform samples and in **Chapter 3** we extend them to model outliers in Pareto samples.

2. MODELLING OUTLIERS IN UNIFORM SAMPLES

2.1 Modelling outliers in one parameter uniform samples

2.1.1 Modelling a single outlier in a one parameter uniform sample

Suppose that $\{X_1, ..., X_n\}$ are independent uniform$(0, \theta)$ random variables, then the joint probability density function of $\{X_1, ..., X_n\}$ not including X_i is $p\left(\mathbf{x_{(i)}}|\theta\right) = \frac{1}{\theta^{n-1}}$, for $\theta > \max\{x_j : j \neq i\} > 0$. The conjugate prior for θ is a Pareto prior and for this problem θ shall be given a Pareto(α, θ_0) prior, so that $p(\theta) = \frac{\alpha {\theta_0}^{\alpha}}{\theta^{\alpha+1}}$, for $\alpha > 0$, $\theta > \theta_0$ and where α and θ_0 are both assumed to be known. We can find the posterior distribution of θ given $\{x_1, ..., x_n\}$ not including x_i, where its probability density function is written as $p\left(\theta|\mathbf{x_{(i)}}\right)$ and is such that

$$\begin{aligned} p\left(\theta|\mathbf{x_{(i)}}\right) &\propto p\left(\mathbf{x_{(i)}}|\theta\right) p(\theta) \\ &\propto \frac{1}{\theta^{\alpha+n}}, \text{ for } \theta > s = \max\{\theta_0,\ x_j : j \neq i\}. \end{aligned}$$

The constant C is such that

$$\begin{aligned} & C\int_s^{\infty} \frac{1}{\theta^{\alpha+n}} d\theta \\ = & \frac{C}{(\alpha+n-1)\, s^{\alpha+n-1}} \\ = & 1, \end{aligned}$$

thus

$$C = (\alpha + n - 1)\, s^{\alpha+n-1}$$

and so

$$\theta|\mathbf{x_{(i)}} \sim \text{Pareto}\,(\alpha + n - 1, s)\,.$$

The conditional predictive ordinate is then given by

$$\begin{aligned} p\left(x_i|\mathbf{x_{(i)}}\right) &= \int_{s'}^{\infty} p\,(x_i|\theta)\, p\left(\theta|\mathbf{x_{(i)}}\right) d\theta \\ &= \int_{s'}^{\infty} \frac{1}{\theta} \frac{(\alpha + n - 1)\, s^{\alpha+n-1}}{\theta^{\alpha+n}} d\theta \\ &= \frac{(\alpha + n - 1)\, s^{\alpha+n-1}}{(\alpha + n)\, s'^{\alpha+n}}, \end{aligned}$$

where $s' = \max\{\theta_0,\ \mathbf{x}\} = \max\{s,\ x_i\}$. We can clearly see that the largest observation in the sample has the smallest conditional predictive ordinate.

Suppose that $s' = x_i$ and it is suspected of being an outlier. We can derive the Bayes factor to test whether the model M_0 that all of the X_j have a uniform$(0, \theta)$ distribution or the model M_1 that all of the X_j except for X_i have a uniform$(0, \theta)$ distribution and X_i has a uniform$(0, \delta\theta)$ distribution, is more appropriate, where $\delta > 1$ and is known. The Bayes factor is denoted by $B_{0,1} = \frac{p(\mathbf{x}|M_0)}{p(\mathbf{x}|M_1)}$, where

$$\begin{aligned} p\,(\mathbf{x}|M_0) &= \int_{x_i}^{\infty} \frac{1}{\theta^n} \frac{\alpha {\theta_0}^{\alpha}}{\theta^{\alpha+1}} d\theta \\ &= \frac{\alpha {\theta_0}^{\alpha}}{(\alpha + n)\, {x_i}^{\alpha+n}}, \\ p\,(\mathbf{x}|M_1) &= \int_{s^*}^{\infty} \frac{1}{\delta\theta^n} \frac{\alpha {\theta_0}^{\alpha}}{\theta^{\alpha+1}} d\theta \\ &= \frac{\alpha {\theta_0}^{\alpha}}{\delta\,(\alpha + n)\, s^{*\alpha+n}} \end{aligned}$$

and

$$s^* = \max\left\{\theta_0,\ \frac{x_i}{\delta},\ x_j : j \neq i\right\} = \max\left\{s,\ \frac{x_i}{\delta}\right\}.$$

Therefore

$$B_{0,1} = \delta \left(\frac{s^*}{x_i} \right)^{\alpha+n},$$

which is minimized when x_i is very large compared to the other observations, noting that $\delta^{1-\alpha-n}$ is very small for any sensible sample size such as $n \geq 5$.

Now consider the previous testing problem when δ is unknown. We shall give δ a Pareto$(\beta, 1)$ prior distribution, as the prior probability that δ is larger than any constant greater than one gets small rather quickly. Also $p(\delta) = \frac{\beta}{\delta^{\beta+1}}$, for $\delta > 1$ and where β is known. In order for it to make sense to test for outliers we assume that $n \geq 5$ and in order to have $E(\delta) \geq 2$ we assume that $1 < \beta \leq 2$. The Bayes factor is denoted by $B_{0,1} = \frac{p(\mathbf{x}|M_0)}{p(\mathbf{x}|M_1)}$, where $p(\mathbf{x}|M_0)$ is given by the same expression as we had for this problem when δ was known and

$$\begin{aligned} p(\mathbf{x}|M_1) &= \int_1^\infty \int_{s^*}^\infty \frac{1}{\delta\theta^n} \frac{\alpha {\theta_0}^\alpha}{\theta^{\alpha+1}} \frac{\beta}{\delta^{\beta+1}} d\theta d\delta \\ &= \frac{\alpha {\theta_0}^\alpha \beta}{\alpha+n} \int_1^\infty \frac{1}{{s^*}^{\alpha+n}\delta^{\beta+2}} d\delta. \end{aligned}$$

By splitting the previous integral up:

If $s^* = s$, we have

$$\begin{aligned} & \frac{\alpha {\theta_0}^\alpha \beta}{\alpha+n} \int_{\frac{x_i}{s}}^\infty \frac{1}{s^{\alpha+n}\delta^{\beta+2}} d\delta \\ = & \frac{\alpha {\theta_0}^\alpha \beta}{(\alpha+n)(\beta+1) {x_i}^{\beta+1} s^{\alpha+n-\beta-1}}; \end{aligned}$$

If $s^* = \frac{x_i}{\delta}$, we have

$$\begin{aligned} & \frac{\alpha {\theta_0}^\alpha \beta}{(\alpha+n) {x_i}^{\alpha+n}} \int_1^{\frac{x_i}{s}} \delta^{\alpha+n-\beta-2} d\delta \\ = & \frac{\alpha {\theta_0}^\alpha \beta}{(\alpha+n)(\alpha+n-\beta-1) {x_i}^{\alpha+n}} \left(\left(\frac{x_i}{s} \right)^{\alpha+n-\beta-1} - 1 \right). \end{aligned}$$

Hence

$$\begin{aligned} p\left(\mathbf{x}|M_1\right) &= \frac{\alpha {\theta_0}^{\alpha} \beta}{\alpha+n}\left(\frac{1}{\left(\beta+1\right){x_i}^{\beta+1} s^{\alpha+n-\beta-1}}+\frac{\left(\frac{x_i}{s}\right)^{\alpha+n-\beta-1}-1}{\left(\alpha+n-\beta-1\right){x_i}^{\alpha+n}}\right) \\ &= \frac{\alpha {\theta_0}^{\alpha} \beta}{\left(\alpha+n\right)\left(\alpha+n-\beta-1\right)}\left(\frac{\alpha+n}{\left(\beta+1\right) s^{\alpha+n-\beta-1} {x_i}^{\beta+1}}-\frac{1}{{x_i}^{\alpha+n}}\right) \end{aligned}$$

and therefore

$$B_{0,1}=\frac{\alpha+n-\beta-1}{\beta\left(\frac{\alpha+n}{\beta+1}\left(\frac{x_i}{s}\right)^{\alpha+n-\beta-1}-1\right)}.$$

When x_i is very large compared to the other observations and θ_0, the Bayes factor will be close to zero because $\lim_{x_i\to\infty}\left(B_{0,1}\right)=0$ and in such cases we should conclude that x_i is an outlier. Note that when both $\alpha\to 0$ and $\theta_0\to 0$ in our current formulae for $B_{0,1}$, we can obtain the corresponding Bayes factors using the noninformative prior $p\left(\theta\right)=\frac{\kappa}{\theta}$, where κ is an unknown constant whose exact value does not matter because it cancels out in the calculations.

Example 2.1 As an example to illustrate the previous methods, we use the following data, where the sample has been obtained by simulating ten uniform$(0,1)$ observations in R. We have displayed the data correct to three decimal places, but will use the true values when performing all of the calculations that follow.

0.561 0.770 0.125 0.352 0.647 0.847 0.327 0.622 0.515 0.333

Since we have simulated a uniform$(0,1)$ sample, one possible way of choosing α and θ_0 is to take values that make the prior mean of θ equal to one. When α is small the prior distribution for θ is less informative than when it is large and so for this example we shall use $\alpha=2$ and $\theta_0=\frac{1}{2}$. We have randomly

chosen to multiply observation one by five so that it is now 2.806 and can be suspected of being an outlier. Note that if we would have multiplied another observation by five and the resulting value was not larger than all of the other observations in the sample, then obviously there would have been no evidence to suspect that this value was extreme. We shall test whether 2.806 is an outlier, firstly for the case when δ is known to be five and then for the case when δ is unknown. It is argued in Pettit (1992) that a Bayes factor of less than or equal to 0.015 provides strong evidence for concluding that an observation is an outlier and a Bayes factor of less than or equal to 0.005 provides very strong evidence for concluding that an observation is an outlier, based on assuming that outliers occur randomly in a sample with probability 0.05. Therefore we shall conclude that x_i is an outlier if our Bayes factor is less than or equal to 0.015.

(i) When δ is known to be five it follows that $s^* = 0.847$, hence

$$\begin{aligned} B_{0,1} &= 5\left(\frac{0.847}{2.806}\right)^{12} \\ &= 2.86 \times 10^{-6} \end{aligned}$$

and so we should certainly conclude that 2.806 is an outlier. **Table 2.1** shows the Bayes factors for this sample using various different combinations of α and δ when $\theta_0 = \frac{1}{2}$. These values will be exactly the same for any choice of θ_0 which is less than or equal to 0.847, as then s is always equal to 0.847. Note that if we had simulated a larger uniform$(0, 1)$ sample, then the largest observation not suspected of being an outlier would have probably been closer to one than 0.847 is. **Table 2.2** shows the critical values for this sample, where these were calculated by finding the value of x_i that makes $B_{0,1}$ equal to 0.015 based on all of the other nine observations in this sample

when $s^* = 0.847$. We see that these decrease as the prior mean of θ gets smaller and increase when we know that an observation has to be larger in order to be an outlier. Therefore we conclude that this method in general is hardly at all sensitive to any reasonable choices of θ_0 and δ, but very sensitive to our choices of α and n, where our formula for $B_{0,1}$ shows that increasing the sample size has exactly the same effect as increasing α.

Tab. 2.1: Bayes factors when δ is known and $\theta_0 = \frac{1}{2}$

	δ		
α	3	5	10
2	5.65×10^{-6}	2.86×10^{-6}	5.72×10^{-6}
3	1.88×10^{-6}	8.63×10^{-7}	1.73×10^{-6}
4	6.27×10^{-7}	2.60×10^{-7}	5.21×10^{-7}
5	2.09×10^{-7}	7.86×10^{-8}	1.57×10^{-7}
10	8.60×10^{-10}	1.97×10^{-10}	3.94×10^{-10}

Tab. 2.2: Critical values for this sample when δ is known and $\theta_0 = \frac{1}{2}$

	δ		
α	2	3	5
2	1.27	1.32	1.37
3	1.23	1.27	1.32
4	1.20	1.24	1.28
5	1.17	1.21	1.25
10	1.08	1.10	1.13

(ii) When δ is unknown, we choose $\beta = \frac{5}{4}$, which is the unique value of β that makes the prior mean of δ equal to five. Therefore

$$\begin{aligned} B_{0,1} &= \frac{9.75}{1.25\left(\frac{12}{2.25}\left(\frac{2.806}{0.847}\right)^{9.75} - 1\right)} \\ &= 1.24 \times 10^{-5} \end{aligned}$$

and hence we should certainly conclude that 2.806 is an outlier, noting that this is larger than for the case when δ is assumed to be known because not knowing δ adds extra uncertainty to the problem. Similarly **Table 2.3** and **Table 2.4** show the Bayes factors and critical values for this sample using various different combinations of α and β when $\theta_0 = \frac{1}{2}$. Therefore we conclude that this method in general is hardly at all sensitive to any reasonable choices of θ_0 and β, but very sensitive to our choices of α and n. This is because of the same reasons as before and as the values of β are required to be so small, where it is shown by our formula for $B_{0,1}$ that $B_{0,1}$ gets smaller as β gets smaller and hence the critical values get smaller as this happens.

Tab. 2.3: Bayes factors when δ is unknown and $\theta_0 = \frac{1}{2}$

	β		
α	$\frac{3}{2}$	$\frac{5}{4}$	$\frac{10}{9}$
2	1.51×10^{-5}	1.24×10^{-5}	1.12×10^{-5}
3	4.64×10^{-6}	3.80×10^{-6}	3.44×10^{-6}
4	1.42×10^{-6}	1.17×10^{-6}	1.05×10^{-6}
5	4.36×10^{-7}	3.56×10^{-7}	3.22×10^{-7}
10	1.15×10^{-9}	9.32×10^{-10}	8.39×10^{-10}

Tab. 2.4: Critical values for this sample when δ is unknown and $\theta_0 = \frac{1}{2}$

	β		
α	2	$\frac{3}{2}$	$\frac{5}{4}$
2	1.37	1.36	1.35
3	1.31	1.30	1.30
4	1.26	1.25	1.25
5	1.22	1.22	1.22
10	1.10	1.10	1.10

A truncated exponential(λ) prior distribution is an alternative distribution such that the prior probability that δ is larger than any constant greater than one gets small rather quickly, where $p(\delta) = \lambda e^{-\lambda(\delta-1)}$, for $\delta > 1$ and $\lambda > 0$. Therefore

$$\begin{aligned} p(\mathbf{x}|M_1) &= \int_1^\infty \int_{s^*}^\infty \frac{1}{\delta\theta^n} \frac{\alpha {\theta_0}^\alpha}{\theta^{\alpha+1}} \lambda e^{-\lambda(\delta-1)} d\theta d\delta \\ &= \frac{\alpha {\theta_0}^\alpha \lambda e^\lambda}{\alpha+n} \int_1^\infty \frac{e^{-\lambda\delta}}{\delta {s^*}^{\alpha+n}} d\delta \\ &= \frac{\alpha {\theta_0}^\alpha \lambda e^\lambda}{(\alpha+n)\, s^{\alpha+n}} \int_{\frac{x_i}{s}}^\infty \frac{e^{-\lambda\delta}}{\delta} d\delta + \frac{\alpha {\theta_0}^\alpha \lambda e^\lambda}{(\alpha+n)\, {x_i}^{\alpha+n}} \int_1^{\frac{x_i}{s}} \delta^{\alpha+n-1} e^{-\lambda\delta} d\delta, \end{aligned}$$

where

$$\int_{\frac{x_i}{s}}^\infty \frac{e^{-\lambda\delta}}{\delta} d\delta \text{ and } \int_1^{\frac{x_i}{s}} \delta^{\alpha+n-1} e^{-\lambda\delta} d\delta$$

can both be evaluated numerically. Hence

$$B_{0,1} = \frac{1}{\lambda e^\lambda \left(\left(\frac{x_i}{s}\right)^{\alpha+n} \int_{\frac{x_i}{s}}^\infty \frac{e^{-\lambda\delta}}{\delta} d\delta + \int_1^{\frac{x_i}{s}} \delta^{\alpha+n-1} e^{-\lambda\delta} d\delta \right)}.$$

We now recalculate the critical values for the data given in **Example 2.1** using this Bayes factor, which are shown in **Table 2.5**, where λ is chosen

such that $E(\delta)$ is equal to the values of δ used in **Table 2.2**. We see that the values given in **Table 2.5** are not much smaller than the values given in **Table 2.4**. Therefore the Pareto prior is the better prior distribution to use because it produces analytical and simpler Bayes factors.

Tab. 2.5: Critical values for this sample when δ has a truncated exponential(λ) prior distribution and $\theta_0 = \frac{1}{2}$

	λ		
α	1	$\frac{1}{2}$	$\frac{1}{4}$
2	1.33	1.33	1.35
3	1.28	1.28	1.30
4	1.23	1.24	1.26
5	1.20	1.21	1.22
10	1.09	1.10	1.11

2.1.2 Modelling multiple outliers in a one parameter uniform sample

We now consider the problem when it is believed that more than one observation in the sample is an outlier, where it is assumed that $s' = z_i$. Suppose that $\{z_1, ..., z_q\}$ are the q largest observations in the sample and are suspected of being outliers generated by the same probability distribution, where q is the number of observations that we suspect of being outliers, $q < n$, $\{Z_1, ..., Z_q\} \subset \{X_1, ..., X_n\}$ and $\{Z_{[1]}, ..., Z_{[n-q]}\}$ denote the random variables corresponding to the observations not suspected of being outliers. We can derive the Bayes factor to test whether the model M_0 that all of the X_j have a uniform$(0, \theta)$ distribution or the model M_q that all of the $Z_{[h]}$ have

a uniform$(0,\theta)$ distribution and all of the Z_g have a uniform$(0,\delta\theta)$ distribution, is more appropriate. It is assumed that δ is unknown and is again given a Pareto$(\beta,1)$ prior distribution, for $1<\beta\leq 2$. In what follows we again have $n\geq 5$ and assume that $q<\frac{n}{2}$, as if $q\geq\frac{n}{2}$ it might imply that the $Z_{[h]}$ should be suspected of being outliers rather than the Z_g. We write the Bayes factor as $B_{0,q}=\frac{p(\mathbf{x}|M_0)}{p(\mathbf{x}|M_q)}$ to compare these models. Therefore

$$\begin{aligned} p\left(\mathbf{x}|M_q\right) &= \int_1^\infty \int_{t^*}^\infty \frac{1}{\delta^q\theta^n}\frac{\alpha{\theta_0}^\alpha}{\theta^{\alpha+1}}\frac{\beta}{\delta^{\beta+1}}d\theta d\delta \\ &= \frac{\alpha{\theta_0}^\alpha\beta}{(\alpha+n)(\alpha+n-\beta-q)}\left(\frac{\alpha+n}{(\beta+q)\,t^{\alpha+n-\beta-q}{z_i}^{\beta+q}}-\frac{1}{{z_i}^{\alpha+n}}\right), \end{aligned}$$

where $t=\max\left\{\theta_0,\ z_{[h]}\right\}$ and $t^*=\max\left\{\theta_0,\ \frac{z_i}{\delta},\ z_{[h]}\right\}$. Hence the Bayes factor for comparing the models M_0 and M_q is

$$B_{0,q}=\frac{\alpha+n-\beta-q}{\beta\left(\frac{\alpha+n}{\beta+q}\left(\frac{z_i}{t}\right)^{\alpha+n-\beta-q}-1\right)}.$$

We can see that $B_{0,q}$ only depends on the most extreme of the z_g, but not the rest of the z_g. For this reason when z_i is very large compared to the $z_{[h]}$ and θ_0, the Bayes factor will be close to zero because $\lim_{z_i\to\infty}\left(B_{0,q}\right)=0$ and in such cases we should conclude that $\{z_1,...,z_q\}$ are outliers generated by the same probability distribution. Note that if we were to assume that δ is known, then the corresponding Bayes factor is equal to

$$B_{0,q}=\delta^q\left(\frac{t^*}{z_i}\right)^{\alpha+n}.$$

We can see that the previous methods are sensitive to both masking and swamping and so it creates problems. To overcome these problems, we repeatedly calculate Bayes factors of the form $B_{\gamma,\gamma+1}=\frac{p(\mathbf{x}|M_\gamma)}{p(\mathbf{x}|M_{\gamma+1})}$, starting

with $\gamma = 0$, until we can see that γ is equal to some value such that the $(\gamma+1)^{\text{th}}$ largest observation in the sample is not extreme. If on all of the iterations we select the model M_γ over the model $M_{\gamma+1}$, then we should select M_0 as our final model. Otherwise our final model is $M_{\eta+1}$, which is the last model that we selected over the model suspected of having one less outlier. Using a similar derivation as we did for finding $B_{0,q}$ it can be shown that the Bayes factor for comparing the models M_γ and $M_{\gamma+1}$ is

$$B_{\gamma,\gamma+1} = \frac{(\alpha+n-\beta-\gamma-1)\left(\frac{\alpha+n}{(\beta+\gamma){t_\gamma}^{\alpha+n-\beta-\gamma}{z_i}^{\beta+\gamma}} - \frac{1}{{z_i}^{\alpha+n}}\right)}{(\alpha+n-\beta-\gamma)\left(\frac{\alpha+n}{(\beta+\gamma+1){t_{\gamma+1}}^{\alpha+n-\beta-\gamma-1}{z_i}^{\beta+\gamma+1}} - \frac{1}{{z_i}^{\alpha+n}}\right)},$$

where $\gamma \geq 1$, t_γ is the $(\gamma+1)^{\text{th}}$ largest observation in the sample and $t_{\gamma+1}$ is the $(\gamma+2)^{\text{th}}$ largest observation in the sample. Note that if we strongly believe that there are q outliers in the sample, we could use the previous method by putting γ equal to $q-1$ instead of zero on the first iteration, so that we can save time from not having to perform as many iterations to arrive at the final model. Also if we were to assume that δ is known, then

$$B_{\gamma,\gamma+1} = \delta\left(\frac{{t^*}_{\gamma+1}}{{t^*}_\gamma}\right)^{\alpha+n},$$

where ${t^*}_\gamma = \max\left\{\theta_0,\ \frac{z_i}{\delta},\ z_{[1]},...,z_{[n-\gamma]}\right\}$, ${t^*}_{\gamma+1} = \max\left\{\theta_0,\ \frac{z_i}{\delta},\ \tilde{z}_{[1]},...,\tilde{z}_{[n-\gamma-1]}\right\}$ and the sample $\left\{\tilde{z}_{[1]},...,\tilde{z}_{[n-\gamma-1]}\right\}$ is the same as the sample $\left\{z_{[1]},...,z_{[n-\gamma]}\right\}$ with $\max\left\{z_{[h]}\right\}$ removed from it.

If δ_1 and δ_2 are both assumed to be known and $\delta_1 > \delta_2$, then it can be shown that the Bayes factor for testing whether the model M_0 that all of the X_j have a uniform$(0,\theta)$ distribution or the model M_{q+q^*} that all of the $Z_{[h]}$ have a uniform$(0,\theta)$ distribution, all of the Z_g have a uniform$(0,\delta_1\theta)$ distribution and all of the Z_{g^*} have a uniform$(0,\delta_2\theta)$ distribution, is more

appropriate is

$$B_{0,q} = {\delta_1}^q {\delta_2}^{q^*} \left(\frac{u}{z_i}\right)^{\alpha+n},$$

where z_i and z_{i^*} are the most extreme of the z_g and z_{g^*} respectively and

$$u = \max\left\{\theta_0,\ \frac{z_i}{\delta_1},\ \frac{z_{i^*}}{\delta_2},\ z_{[h]}\right\}.$$

Similar Bayes factors can be derived for case when we have any number of sets of outliers, although in practice this is rarely necessary when using a stretched uniform model for modelling outliers. To deal with masking and swamping, we use the same procedure that we did for the case of comparing the models M_0 and M_q, except that on each iteration we compare the current model with all of the reasonable models containing one more outlier and use the one which gives the smallest Bayes factor as the current model for the next iteration. If on all of the iterations we select the current model over the best model with one more outlier, then we should select M_0 as our final model. Otherwise our final model is the last one that we selected over the model suspected of having one less outlier.

When δ is unknown and we do not conclude that a set of extreme observations are outliers generated by the same probability distribution, but still suspect that they are outliers, we use the following method:

(i) Consider the smallest observation that could possibly be an outlier and delete all of the more extreme observations from the sample. Then for this new sample test whether it is an outlier. If we conclude that it is, then all of the other extreme observations are outliers. Otherwise we conclude that this observation is not an outlier and repeat the procedure until either we conclude that an observation is an outlier or that none of the observations are outliers.

If we are only interested in declaring whether or not extreme observations are outliers, then this is sufficient, otherwise we continue by using the following method to see which outliers are generated by the same probability distribution.

(ii) Consider the two smallest outliers and delete the rest of them. If we find out that these two outliers are generated by the same probability distribution, then we consider adding in a third outlier. Otherwise we conclude that the first and second smallest outliers are generated by different probability distributions, but then consider if the second and third smallest outliers are generated by the same probability distribution while deleting the first smallest outlier as well as the fourth smallest to the largest outliers. This is done until we have found out which probability distribution the largest outlier has been generated by relative to the other outliers.

Example 2.2 As an example to illustrate the methods in this subsection, we return to our uniform$(0, 1)$ data which was used in **Example 2.1** and again assume that $\alpha = 2$ and $\theta_0 = \frac{1}{2}$.

(i) We have multiplied observations one and nine by five so that they are now 2.806 and 2.575 respectively, hence it is assumed that $\beta = \frac{5}{4}$ because of the same reason which was given in part (ii) of **Example 2.1**. Suppose we only suspect that 2.806 is an outlier. The Bayes factor for comparing the models M_0 and M_1 is

$$\begin{aligned} B_{0,1} &= \frac{9.75}{1.25\left(\frac{12}{2.25}\left(\frac{2.806}{2.575}\right)^{9.75} - 1\right)} \\ &= 0.690, \end{aligned}$$

by comparing this to our answer from part (ii) of **Example 2.1**, we can

see that masking has definitely occurred. Therefore the Bayes factor for comparing the models M_1 and M_2 is

$$\begin{aligned} B_{1,2} &= \frac{8.75\left(\frac{12}{2.25(2.575)^{9.75}(2.806)^{2.25}} - \frac{1}{(2.806)^{12}}\right)}{9.75\left(\frac{12}{3.25(0.847)^{8.75}(2.806)^{3.25}} - \frac{1}{(2.806)^{12}}\right)} \\ &= 7.71 \times 10^{-5} \end{aligned}$$

and so we should certainly select M_2 over M_1. We do not have to calculate any more Bayes factors because we can clearly see that 0.847 is not extreme, therefore we select M_2 as our final model and conclude that 2.806 and 2.575 are outliers generated by the same probability distribution. The Bayes factors for comparing the models M_1 and M_2 using various different values of α when $\theta_0 = \frac{1}{2}$ and $\beta = \frac{5}{4}$ are given in **Table 2.6**. We can see from **Table 2.6** that $B_{1,2}$ is still very sensitive to our choice of α and hence we conclude in general that the $B_{\gamma,\gamma+1}$ are very sensitive to our choices of α and n, but not to any reasonable choices of θ_0 and β.

Tab. 2.6: Bayes factors for comparing M_1 and M_2 when $\theta_0 = \frac{1}{2}$ and $\beta = \frac{5}{4}$

α				
2	3	4	5	10
7.71×10^{-5}	2.60×10^{-5}	8.71×10^{-6}	2.91×10^{-6}	1.18×10^{-8}

(ii) We have multiplied observation one by ten and observation nine by three so that they are now 5.611 and 1.545 respectively and can be suspected of being outliers. When $\beta = \frac{10}{9}$ it follows that $B_{1,2} = 2.30 \times 10^{-2}$, where if we would of had the exact same sample except with 1.545 replaced by 1.613 it would have been concluded that 5.611 and 1.613 are outliers generated

by the same probability distribution. We can clearly see that 0.847 is not extreme and therefore start by deleting observation one from the sample and testing whether 1.545 is an outlier. When $\beta = \frac{3}{2}$ it can be shown that $B_{0,1} = 7.78 \times 10^{-3}$, hence 1.545 is an outlier and therefore we definitely conclude that 5.611 and 1.545 are outliers generated by different probability distributions.

2.1.3 An alternative way of testing for outliers

Consider a single outlier problem for a one parameter uniform sample. Another way of finding out whether x_i is an outlier is to derive the Bayes factor for testing whether the model M_0 that all of the X_j have a uniform$(0, \theta)$ distribution or the model M_1 that all of the X_j except for X_i have a uniform$(0, \theta)$ distribution and X_i has a uniform$(\epsilon, \theta + \epsilon)$ distribution, is more appropriate, where θ is given a Pareto(α, θ_0) prior distribution as before and $\epsilon > 0$. Firstly we do this for the case when ϵ is assumed to be known, then it is done for the case when ϵ is unknown and given an exponential(λ) prior distribution, as the prior probability that ϵ is larger than any constant greater than zero gets small rather quickly, where λ is known. Also we assume that $n \geq 5$ for the same reason as before.

When ϵ is assumed to be known it follows that

$$\begin{aligned} p\left(\mathbf{x}|M_1\right) &= \int_v^\infty \frac{1}{\theta^n} \frac{\alpha {\theta_0}^\alpha}{\theta^{\alpha+1}} d\theta \\ &= \frac{\alpha {\theta_0}^\alpha}{\left(\alpha + n\right) v^{\alpha+n}}, \end{aligned}$$

where

$$v = \max\left\{\theta_0,\ x_i - \epsilon,\ x_j : j \neq i\right\}.$$

Hence

$$B_{0,1} = \left(\frac{v}{x_i}\right)^{\alpha+n}$$

and is minimized when x_i is very large compared to the other observations for any reasonable choice of ϵ.

When ϵ is unknown it follows that

$$\begin{aligned} p\left(\mathbf{x}|M_1\right) &= \int_0^\infty \int_v^\infty \frac{1}{\theta^n} \frac{\alpha {\theta_0}^\alpha}{\theta^{\alpha+1}} \lambda e^{-\lambda\epsilon} d\theta d\epsilon \\ &= \int_0^\infty \frac{\alpha {\theta_0}^\alpha}{(\alpha+n)\, v^{\alpha+n}} \lambda e^{-\lambda\epsilon} d\epsilon \\ &= \frac{\alpha {\theta_0}^\alpha}{(\alpha+n)} \left(\frac{e^{-\lambda(x_i-s)}}{s^{\alpha+n}} + I\right), \end{aligned}$$

where

$$I = \int_0^{x_i-s} \frac{\lambda e^{-\lambda\epsilon}}{(x_i-\epsilon)^{\alpha+n}} d\epsilon$$

and can be evaluated numerically because of the following reason, noting that $s = \max\{\theta_0,\ x_j : j \neq i\}$ as before. To evaluate this integral, we change the variable to $\phi = x_i - \epsilon$, so that

$$I = \lambda e^{-\lambda x_i} \int_s^{x_i} \frac{e^{\lambda\phi}}{\phi^{\alpha+n}} d\phi.$$

Let

$$I_k = \int_s^{x_i} \frac{e^{\lambda\phi}}{\phi^k} d\phi,$$

then integration by parts gives a recurrence relation for I_k in terms of I_{k-1}. Since

$$I = \lambda e^{-\lambda x_i} I_{\alpha+n}$$

it reduces to an expression in terms of

$$I_1 = \int_s^{x_i} \frac{e^{\lambda\phi}}{\phi} d\phi$$

which can be evaluated numerically. Therefore

$$B_{0,1} = \frac{1}{\left(\frac{e^{-\lambda(x_i-s)}}{s^{\alpha+n}} + I\right) {x_i}^{\alpha+n}}$$

and is shown in part (ii) of **Example 2.3** that this is minimal when x_i is very large compared to the other observations for any reasonable choice of λ.

Example 2.3 We now recalculate the Bayes factors using this approach for our uniform$(0, 1)$ data given in **Example 2.1** when $\alpha = 2$ and $\theta_0 = \frac{1}{2}$.

(i) When $\epsilon = 2.806 - 0.561 = 2.245$ it follows that

$$\begin{aligned} B_{0,1} &= \left(\frac{0.847}{2.806}\right)^{12} \\ &= 5.72 \times 10^{-7}, \end{aligned}$$

where we have used the true value of ϵ to make our answers consistent. **Table 2.7** shows the Bayes factors for this sample using various different combinations of α and ϵ when $\theta_0 = \frac{1}{2}$, noting that $B_{0,1}$ is hardly at all sensitive to our choice of θ_0 for the same reason as before. For $\alpha = 2$ and $\theta_0 = \frac{1}{2}$ it can be shown that the critical value for this sample is equal to 1.20 when $0.353 < \epsilon \leq 1.200$, which is smaller than the critical values that were given in **Table 2.2** in **Example 2.1**. Therefore this method in general is hardly at all sensitive to any reasonable choices of θ_0 and ϵ, but very sensitive to our choices of α and n.

(ii) When ϵ is unknown and $\lambda = \frac{1}{2.245}$ it follows that

$$\begin{aligned} I &= \int_0^{1.959} \frac{\frac{1}{2.245} e^{-\frac{\epsilon}{2.245}}}{(2.806 - \epsilon)^{12}} d\epsilon \\ &= 0.109, \end{aligned}$$

Tab. 2.7: Bayes factors when ϵ is known and $\theta_0 = \frac{1}{2}$

	ϵ		
α	1.806	2.245	2.806
2	4.20×10^{-6}	5.72×10^{-7}	5.72×10^{-7}
3	1.50×10^{-6}	1.73×10^{-7}	1.73×10^{-7}
4	5.34×10^{-7}	5.21×10^{-8}	5.21×10^{-8}
5	1.90×10^{-7}	1.57×10^{-8}	1.57×10^{-8}
10	1.09×10^{-9}	3.94×10^{-11}	3.94×10^{-11}

hence

$$\begin{aligned} B_{0,1} &= \frac{1}{\left(\frac{e^{-\frac{1.959}{2.245}}}{0.847^{12}} + 0.109\right) 2.806^{12}} \\ &= 1.32 \times 10^{-6}, \end{aligned}$$

noting that this is larger than for the case when ϵ is assumed to be known because not knowing ϵ adds extra uncertainty to the problem. **Table 2.8** shows the Bayes factors for this sample using various different combinations of α and λ when $\theta_0 = \frac{1}{2}$. **Table 2.9** shows the critical values for this sample using various different values of λ when $\alpha = 2$ and $\theta_0 = \frac{1}{2}$, which are all smaller than the critical values that were given in **Table 2.4** in **Example 2.1**. Therefore this method in general is hardly at all sensitive to any reasonable choices of θ_0 and λ, but very sensitive to our choices of α and n. Suppose that x_i is very large, for example $x_i = 10$ and say $\lambda = \frac{1}{9.5}$ (because $10 - 0.5 = 9.5$), then $B_{0,1} = 3.54 \times 10^{-13}$ and hence it is confirmed that the Bayes factor is minimized when x_i very large compared to the other observations.

Clearly this set of tests works better than the first set of tests, but there

Tab. 2.8: Bayes factors when ϵ is unknown and $\theta_0 = \frac{1}{2}$

	λ		
α	$\frac{1}{1.806}$	$\frac{1}{2.245}$	$\frac{1}{2.806}$
2	1.62×10^{-6}	1.32×10^{-6}	1.12×10^{-6}
3	4.91×10^{-7}	4.00×10^{-7}	3.38×10^{-7}
4	1.49×10^{-7}	1.21×10^{-7}	1.02×10^{-7}
5	4.50×10^{-8}	3.66×10^{-8}	3.09×10^{-8}
10	1.14×10^{-10}	9.24×10^{-11}	7.79×10^{-11}

Tab. 2.9: Critical values for this sample when ϵ is unknown and $\theta_0 = \frac{1}{2}$

λ				
$\frac{1}{0.353}$	2	$\frac{4}{3}$	1	$\frac{5}{6}$
1.31	1.27	1.24	1.23	1.23

is however one problem. This is that for multiple outlier problems the extreme observations have to be quite similar to suspect that they are outliers generated by the same probability distribution. So in general we get more complicated models for modelling outliers when using this approach. In such cases when the amounts of contamination are unknown, we use the same methods as before.

For simple multiple outlier problems it can be shown that the corresponding Bayes factors for the cases when ϵ is known and unknown are

$$B_{0,q} = \left(\frac{w_1}{z_i}\right)^{\alpha+n}$$

and

$$B_{0,q} = \frac{1}{\left(\frac{e^{-\lambda(z_i-t)}}{t^{\alpha+n}} + J\right) {z_i}^{\alpha+n}}$$

respectively, where

$$w_1 = \max\left\{\theta_0,\ z_i - \epsilon,\ z_{[h]}\right\}$$

and

$$J = \int_0^{z_i-t} \frac{\lambda e^{-\lambda\epsilon}}{(z_i-\epsilon)^{\alpha+n}} d\epsilon,$$

noting that $t = \max\left\{\theta_0,\ z_{[h]}\right\}$ as before. If ϵ_1 and ϵ_2 are both assumed to be known and $\epsilon_1 > \epsilon_2$, then it can be shown that the Bayes factor for testing whether the model M_0 that all of the X_j have a uniform$(0,\theta)$ distribution or the model M_{q+q^*} that all of the $Z_{[h]}$ have a uniform$(0,\theta)$ distribution, all of the Z_g have a uniform$(\epsilon_1, \theta+\epsilon_1)$ distribution and all of the Z_{g^*} have a uniform$(\epsilon_2, \theta+\epsilon_2)$ distribution, is more appropriate is

$$B_{0,q+q^*} = \left(\frac{w_2}{z_i}\right)^{\alpha+n},$$

where

$$w_2 = \max\left\{\theta_0,\ z_i - \epsilon_1,\ z_{i^*} - \epsilon_2,\ z_{[h]}\right\}.$$

Similar Bayes factors can be derived for case when we have any number of sets of outliers, but when this number of sets is large it is questionable as to what extreme means. Note that we use the same procedures as before to deal with masking and swamping.

Example 2.4 We now reconsider **Example 2.2** using this approach when $\alpha = 2$ and $\theta_0 = \frac{1}{2}$.

(i) When $\lambda = \frac{1}{2.245}$ it can be shown that $B_{0,1} = 0.371$ and therefore by comparing this to our answer from part (ii) of **Example 2.3**, we can see that

masking has definitely occurred, noting that we have used the same value of λ in both cases. Therefore the Bayes factor for comparing the models M_1 and M_2 is

$$\begin{aligned} B_{1,2} &= \frac{\frac{e^{-\frac{0.231}{2.245}}}{2.575^{12}} + J}{\frac{e^{-\frac{1.959}{2.245}}}{0.847^{12}} + J^*} \\ &= 3.56 \times 10^{-6}, \end{aligned}$$

where

$$J = \int_0^{0.231} \frac{\frac{1}{2.245} e^{-\frac{\epsilon}{2.245}}}{(2.806 - \epsilon)^{12}} d\epsilon$$

and

$$J^* = \int_0^{1.959} \frac{\frac{1}{2.245} e^{-\frac{\epsilon}{2.245}}}{(2.806 - \epsilon)^{12}} d\epsilon.$$

Hence we should certainly select M_2 over M_1. We do not have to calculate any more Bayes factors because we can clearly see that 0.847 is not extreme, therefore we select M_2 as our final model and conclude that 2.806 and 2.575 are outliers generated by the same probability distribution. The Bayes factors for comparing the models M_1 and M_2 using various different values of α when $\theta_0 = \frac{1}{2}$ and $\lambda = \frac{1}{2.245}$ are given in **Table 2.10**. We can see from **Table 2.10** that $B_{1,2}$ is still very sensitive to our choice of α and hence we conclude in general that the $B_{\gamma,\gamma+1}$ are very sensitive to our choices of α and n, but not to any reasonable choices of θ_0 and λ.

Tab. 2.10: Bayes factors for comparing M_1 and M_2 when $\theta_0 = \frac{1}{2}$ and $\lambda = \frac{1}{2.245}$

α				
2	3	4	5	10
3.56×10^{-6}	1.17×10^{-6}	3.86×10^{-7}	1.27×10^{-7}	4.87×10^{-10}

(ii) We can clearly see that 0.847 is not extreme and therefore start by deleting observation one from the sample and testing whether 1.545 is an outlier. When $\lambda = \frac{1}{1.030}$ (again so that our answer is consistent with before) it can be shown that $B_{0,1} = 2.42 \times 10^{-3}$, hence 1.545 is an outlier and therefore 5.611 and 1.545 are definitely both outliers. We cannot possibly choose any reasonable value of λ to test whether 5.611 and 1.545 are outliers generated by the same probability distribution and hence we conclude that 5.611 and 1.545 are outliers generated by different probability distributions.

In conclusion, we should always use the stretched uniform outlier model because the tests for outliers are much simpler to perform and we get more parsimonious models for modelling outliers. However, if these tests fail to declare extreme observations as outliers, then we might use the shifted uniform outlier model for which the tests for outliers are stricter.

2.2 Modelling outliers in two parameter uniform samples

2.2.1 Modelling a single outlier in a two parameter uniform sample

Suppose that $\{X_1, ..., X_n\}$ are independent uniform(θ_1, θ_2) random variables, then the joint probability density function of $\{X_1, ..., X_n\}$ not including X_i is

$$p\left(\mathbf{x_{(i)}}|\theta_1, \theta_2\right) = \frac{1}{(\theta_2 - \theta_1)^{n-1}},$$

for $\theta_1 < \min\{x_j : j \neq i\} < \max\{x_j : j \neq i\} < \theta_2$. By letting $r = (\theta_2 - \theta_1)$ and $m = \frac{1}{2}(\theta_1 + \theta_2)$ this can be written as

$$p\left(\mathbf{x_{(i)}}|r, m\right) = \frac{1}{r^{n-1}},$$

for

$$\left(\max\left\{x_j : j \neq i\right\} - \frac{r}{2}\right) < m < \left(\min\left\{x_j : j \neq i\right\} + \frac{r}{2}\right)$$

and

$$\left(\max\left\{x_j : j \neq i\right\} - \min\left\{x_j : j \neq i\right\}\right) < r < \infty.$$

Due to the complexity of this problem, we shall give r an improper prior such that $p(r) = \frac{\alpha}{r}$ and m an improper prior such that $p(m) = \beta$, where α and β are unknown constants whose exact values do not matter in what follows. We can find the joint posterior density function of r and m given $\{x_1, ..., x_n\}$ not including x_i, which is written as $p\left(r, m|\mathbf{x_{(i)}}\right)$ and is such that

$$\begin{aligned} p\left(r, m|\mathbf{x_{(i)}}\right) &\propto p\left(\mathbf{x_{(i)}}|r, m\right) p(r)\, p(m) \\ &\propto \frac{1}{r^n}. \end{aligned}$$

By letting $r_1 = \left(\max\left\{x_j : j \neq i\right\} - \min\left\{x_j : j \neq i\right\}\right)$, the constant C is such that

$$\begin{aligned} & C\int_{r_1}^{\infty}\int_{\max\{x_j:j\neq i\}-\frac{r}{2}}^{\min\{x_j:j\neq i\}+\frac{r}{2}} \frac{1}{r^n} dm dr \\ = \; & C\int_{r_1}^{\infty} \frac{r - r_1}{r^n} dr \\ = \; & C\left(\frac{1}{(n-2)\,{r_1}^{n-2}} - \frac{1}{(n-1)\,{r_1}^{n-2}}\right) \\ = \; & \frac{C}{(n-1)(n-2)\,{r_1}^{n-2}} \\ = \; & 1, \end{aligned}$$

thus

$$C = (n-1)(n-2)\left(\max\left\{x_j : j \neq i\right\} - \min\left\{x_j : j \neq i\right\}\right)^{n-2}$$

and so

$$p\left(r, m|\mathbf{x_{(i)}}\right) = \frac{(n-1)(n-2)\left(\max\left\{x_j : j \neq i\right\} - \min\left\{x_j : j \neq i\right\}\right)^{n-2}}{r^n},$$

for

$$\left(\max\{x_j : j \neq i\} - \frac{r}{2}\right) < m < \left(\min\{x_j : j \neq i\} + \frac{r}{2}\right)$$

and

$$(\max\{x_j : j \neq i\} - \min\{x_j : j \neq i\}) < r < \infty.$$

The conditional predictive ordinate is then given by

$$\begin{aligned} p\left(x_i|\mathbf{x_{(i)}}\right) &= \int_{r_2}^{\infty} \int_{\max\{x_j\}-\frac{r}{2}}^{\min\{x_j\}+\frac{r}{2}} \frac{1}{r} \frac{(\max\{x_j : j \neq i\} - \min\{x_j : j \neq i\})^{n-2}}{(n-1)^{-1}(n-2)^{-1}r^n} dmdr \\ &= \frac{(n-2)(\max\{x_j : j \neq i\} - \min\{x_j : j \neq i\})^{n-2}}{n(\max\{x_j\} - \min\{x_j\})^{n-1}}, \end{aligned}$$

where $r_2 = (\max\{x_j\} - \min\{x_j\})$. We can clearly see that the most extreme observation in the sample has the smallest conditional predictive ordinate.

Suppose that x_i is the most extreme observation in the sample and it is suspected of being an outlier. We can derive the Bayes factor to test whether the model M_0 that all of the X_j have a uniform(θ_1, θ_2) distribution or the model M_1 that all of the X_j except for X_i have a uniform(θ_1, θ_2) distribution and X_i has a uniform$(\theta_1 + \epsilon, \theta_2 + \epsilon)$ distribution, is more appropriate. In order for it to make sense to test for outliers we assume that $n \geq 5$. Also it is assumed that ϵ is known, where $\epsilon > 0$ when x_i is suspected of being an upper outlier and $\epsilon < 0$ when x_i is suspected of being a lower outlier. Note that we have chosen to model an outlier as something generated by a shifted uniform distribution rather than a stretched uniform distribution, as we cannot get analytical Bayes factors when we come to look at the case when δ is unknown using a stretched uniform model, which was one of the main advantages for using it before. The Bayes factor for the shifted uniform

model is $B_{0,1} = \frac{p(\mathbf{x}|M_0)}{p(\mathbf{x}|M_1)}$, where

$$\begin{aligned} p(\mathbf{x}|M_0) &= \int_{r_2}^{\infty} \int_{\max\{x_j\}-\frac{r}{2}}^{\min\{x_j\}+\frac{r}{2}} \frac{1}{r^n} \frac{\alpha\beta}{r} dmdr \\ &= \frac{\alpha\beta}{n(n-1)(\max\{x_j\} - \min\{x_j\})^{n-1}}, \\ p(\mathbf{x}|M_1) &= \int_{r_3}^{\infty} \int_{\max\{x_i-\epsilon,\ x_j:j\neq i\}-\frac{r}{2}}^{\min\{x_i-\epsilon,\ x_j:j\neq i\}+\frac{r}{2}} \frac{1}{r^n} \frac{\alpha\beta}{r} dmdr \\ &= \frac{\alpha\beta}{n(n-1)(\max\{x_i-\epsilon,\ x_j : j\neq i\} - \min\{x_i-\epsilon,\ x_j : j\neq i\})^{n-1}} \end{aligned}$$

and

$$r_3 = (\max\{x_i-\epsilon,\ x_j : j\neq i\} - \min\{x_i-\epsilon,\ x_j : j\neq i\}).$$

Therefore

$$B_{0,1} = \left(\frac{\max\{x_i-\epsilon,\ x_j : j\neq i\} - \min\{x_i-\epsilon,\ x_j : j\neq i\}}{\max\{x_j\} - \min\{x_j\}}\right)^{n-1},$$

which is minimized when x_i is very large or very small compared to the other observations for any reasonable choice of ϵ.

Now consider the previous testing problem when ϵ is unknown and x_i is suspected of being an upper outlier. We shall give ϵ an exponential(λ) prior distribution, as the prior probability that ϵ is larger than any constant greater than zero gets small rather quickly, where it is assumed that λ is known. The Bayes factor is $B_{0,1} = \frac{p(\mathbf{x}|M_0)}{p(\mathbf{x}|M_1)}$, where $p(\mathbf{x}|M_0)$ is given by the same expression as we had for this problem when ϵ was known and

$$\begin{aligned} p(\mathbf{x}|M_1) &= \int_0^{\infty} \int_{r_3}^{\infty} \int_{\max\{x_i-\epsilon,\ x_j:j\neq i\}-\frac{r}{2}}^{\min\{x_i-\epsilon,\ x_j:j\neq i\}+\frac{r}{2}} \frac{1}{r^n} \frac{\alpha\beta}{r} \lambda e^{-\lambda\epsilon} dmdrd\epsilon \\ &= \int_0^{\infty} \frac{\alpha\beta\lambda e^{-\lambda\epsilon}}{n(n-1)(\max\{x_i-\epsilon,\ x_j : j\neq i\} - \min\{x_i-\epsilon,\ x_j : j\neq i\})^{n-1}} d\epsilon. \end{aligned}$$

By splitting the previous integral up:

If $x_i - \epsilon$ is neither larger nor smaller than all the other observations, we have

$$\begin{aligned} & \frac{\alpha\beta}{n\left(n-1\right)\left(\max\left\{x_j : j \neq i\right\} - \min\left\{x_j : j \neq i\right\}\right)^{n-1}} \int_{x_i - \max\{x_j : j \neq i\}}^{x_i - \min\{x_j : j \neq i\}} \lambda e^{-\lambda\epsilon} d\epsilon \\ = & \frac{\alpha\beta\left(e^{-\lambda(x_i - \max\{x_j : j \neq i\})} - e^{-\lambda(x_i - \min\{x_j : j \neq i\})}\right)}{n\left(n-1\right)\left(\max\left\{x_j : j \neq i\right\} - \min\left\{x_j : j \neq i\right\}\right)^{n-1}} \\ = & \frac{\alpha\beta}{n\left(n-1\right)} \rho^u; \end{aligned}$$

If $x_i - \epsilon$ is larger than all the other observations, we have

$$\begin{aligned} & \int_0^{x_i - \max\{x_j : j \neq i\}} \frac{\alpha\beta}{n\left(n-1\right)\left(x_i - \min\left\{x_j : j \neq i\right\} - \epsilon\right)^{n-1}} \lambda e^{-\lambda\epsilon} d\epsilon \\ = & \frac{\alpha\beta}{n\left(n-1\right)} I^u; \end{aligned}$$

If $x_i - \epsilon$ is smaller than all the other observations, we have

$$\begin{aligned} & \int_{x_i - \min\{x_j : j \neq i\}}^{\infty} \frac{\alpha\beta}{n\left(n-1\right)\left(\max\left\{x_j : j \neq i\right\} - x_i + \epsilon\right)^{n-1}} \lambda e^{-\lambda\epsilon} d\epsilon \\ = & \frac{\alpha\beta}{n\left(n-1\right)} J^u. \end{aligned}$$

The integral I^u can be evaluated numerically because it has the same form as the integral I given in **Section 1.3**. The integral J^u can be evaluated numerically for the following reason. To evaluate J^u, we change the variable to $\phi = \left(\max\left\{x_j : j \neq i\right\} - x_i + \epsilon\right)$, so that

$$J^u = \lambda e^{\lambda(\max\{x_j : j \neq i\} - x_i)} \int_{r_1}^{\infty} \frac{e^{-\lambda\phi}}{\phi^{n-1}} d\phi,$$

where $r_1 = \left(\max\left\{x_j : j \neq i\right\} - \min\left\{x_j : j \neq i\right\}\right)$ as before. By changing the variable to $\omega = \lambda\phi$, we can simplify this expression further to

$$J^u = \lambda^{n-1} e^{\lambda(\max\{x_j : j \neq i\} - x_i)} \int_{\lambda r_1}^{\infty} \frac{e^{-\omega}}{\omega^{n-1}} d\omega.$$

For large ω we have $0 < \frac{e^{-\omega}}{\omega^{n-1}} < e^{-\omega}$ and as $e^{-\omega}$ is integrable on $(\lambda r_1, \infty)$ it follows that $\frac{e^{-\omega}}{\omega^{n-1}}$ is also integrable on $(\lambda r_1, \infty)$. Let

$$J^u{}_k = \int_{\lambda r_1}^{\infty} \frac{e^{-\omega}}{\omega^k} d\omega,$$

then integration by parts gives a recurrence relation for ${J^u}_k$ in terms of ${J^u}_{k-1}$. Since

$$J^u = \lambda^{n-1} e^{\lambda(\max\{x_j : j \neq i\} - x_i)} {J^u}_{n-1}$$

it reduces to an expression in terms of

$${J^u}_1 = \int_{\lambda r_1}^{\infty} \frac{e^{-\omega}}{\omega} d\omega$$

which can be evaluated numerically. Therefore

$$p\left(\mathbf{x}|M_1\right) = \frac{\alpha\beta}{n\left(n-1\right)}\left(\rho^u + I^u + J^u\right),$$

hence

$$B_{0,1} = \frac{1}{\left(\rho^u + I^u + J^u\right)\left(x_i - \min\left\{x_j : j \neq i\right\}\right)^{n-1}}$$

and is shown in part (ii) of **Example 2.5** that this is minimal when x_i is very large compared to the other observations for any reasonable choice of λ.

If x_i is suspected of being a lower outlier and $-\epsilon$ is given an exponential(λ) prior distribution it can be shown in a similar way that the corresponding Bayes factor is

$$B_{0,1} = \frac{1}{\left(\rho^l + J^l + I^l\right)\left(\max\left\{x_j : j \neq i\right\} - x_i\right)^{n-1}},$$

where

$$\begin{aligned}
\rho^l &= \frac{e^{\lambda(x_i - \min\{x_j : j \neq i\})} - e^{\lambda(x_i - \max\{x_j : j \neq i\})}}{\left(\max\left\{x_j : j \neq i\right\} - \min\left\{x_j : j \neq i\right\}\right)^{n-1}}, \\
I^l &= \int_{x_i - \min\{x_j : j \neq i\}}^{0} \frac{1}{\left(\max\left\{x_j : j \neq i\right\} - x_i + \epsilon\right)^{n-1}} \lambda e^{\lambda\epsilon} d\epsilon \\
\text{and } J^l &= \int_{-\infty}^{x_i - \max\{x_j : j \neq i\}} \frac{1}{\left(x_i - \min\left\{x_j : j \neq i\right\} - \epsilon\right)^{n-1}} \lambda e^{\lambda\epsilon} d\epsilon.
\end{aligned}$$

The integrals I^l and J^l have the same form as the integrals I^u and J^u respectively and therefore can be evaluated numerically. Also it is shown in part

(ii) of **Example 2.5** that $B_{0,1}$ is minimal when x_i is very small compared to the other observations for any reasonable choice of λ.

Example 2.5 As an example to illustrate the previous methods, we use the following data, where the sample has been obtained by simulating ten uniform$(-1, 1)$ observations in R. We have displayed the data correct to three decimal places, but will use the true values when performing all of the calculations that follow.

-0.225	-0.745	0.480	0.294	0.561
0.972	-0.656	-0.981	-0.334	0.713

We have randomly added two to observation five so that it is now 2.561, which is large enough to be suspected of being an upper outlier. We shall test whether 2.561 is an upper outlier, firstly for the case when ϵ is known to be two and then for the case when ϵ is unknown.

(i) If ϵ is known to be two, then $\max\{x_j\} = 2.561$, $\min\{x_j\} = -0.981$, $\max\{x_i - \epsilon,\ x_j : j \neq i\} = \max\{x_j : j \neq i\} = 0.972$ and $\min\{x_i - \epsilon,\ x_j : j \neq i\} = \min\{x_j : j \neq i\} = -0.981$, hence

$$\begin{aligned} B_{0,1} &= \left(\frac{0.972 + 0.981}{2.561 + 0.981}\right)^9 \\ &= 4.70 \times 10^{-3} \end{aligned}$$

and so we should certainly conclude that 2.561 is an upper outlier. Note that $B_{0,1}$ is hardly at all sensitive to any reasonable choice of ϵ, but very sensitive to the sample size. This is because of the same reasons that were given for the additive outlier model when ϵ is assumed to be known in the one parameter case.

(ii) When ϵ is unknown, we choose $\lambda = \frac{1}{2}$, which is the unique value of λ that makes the prior mean of ϵ equal to two. Therefore

$$\begin{aligned}
\rho^u &= \frac{e^{-\frac{1}{2}(2.561-0.972)} - e^{-\frac{1}{2}(2.561+0.981)}}{(0.972+0.981)^9} \\
&= 0.000682, \\
I^u &= \int_0^{2.561-0.972} \frac{1}{(2.561+0.981-\epsilon)^9} \frac{1}{2} e^{-\frac{1}{2}\epsilon} d\epsilon \\
&= 0.000153, \\
J^u &= \int_{2.561+0.981}^{\infty} \frac{1}{(0.972-2.561+\epsilon)^9} \frac{1}{2} e^{-\frac{1}{2}\epsilon} d\epsilon \\
&= 0.0000443
\end{aligned}$$

and

$$\rho^u + I^u + J^u = 0.000879.$$

Hence

$$\begin{aligned}
B_{0,1} &= \frac{1}{(2.561+0.981)^9 (0.000879)} \\
&= 1.30 \times 10^{-2}
\end{aligned}$$

and so we conclude that 2.561 is an upper outlier, where again the Bayes factor is larger than for the case when ϵ is assumed to be known. Note that $B_{0,1}$ is hardly at all sensitive to any reasonable choice of λ, but very sensitive to the sample size. This is because of the same reasons that were given for the additive outlier model when ϵ is unknown in the one parameter case. Suppose that x_i is very large, for example $x_i = 10$ and $\lambda = \frac{1}{10}$, then $B_{0,1} = 1.97 \times 10^{-6}$ and hence it is confirmed for upper outlier problems that the Bayes factor is minimized when x_i very large compared to the other observations. Similarly if $x_i = -10$ and $\lambda = \frac{1}{10}$, then $B_{0,1} = 1.99 \times 10^{-6}$ and hence it is confirmed for lower outlier problems that the Bayes factor is minimized when x_i very small compared to the other observations.

2.2.2 Modelling multiple outliers in a two parameter uniform sample

We now consider the problem when it is believed that more than one observation in the sample is an outlier. First we address the case when $\{z_1, ..., z_q\}$ are suspected of being outliers generated by the same probability distribution, where q is the number of observations that we suspect of being outliers, $q < n$, $\{Z_1, ..., Z_q\} \subset \{X_1, ..., X_n\}$ and $\left\{Z_{[1]}, ..., Z_{[n-q]}\right\}$ denote the random variables corresponding to the observations not suspected of being outliers. We can derive the Bayes factor to test whether the model M_0 that all of the X_j have a uniform(θ_1, θ_2) distribution or the model M_q that all of the $Z_{[h]}$ have a uniform(θ_1, θ_2) distribution and all of the Z_g have a uniform$(\theta_1 + \epsilon, \theta_2 + \epsilon)$ distribution, is more appropriate, where ϵ is unknown. In what follows we again have $n \geq 5$ and assume that $q < \frac{n}{2}$, as if $q \geq \frac{n}{2}$ it might imply that the $Z_{[h]}$ should be suspected of being outliers rather than the Z_g. The Bayes factor for comparing these models is denoted by $B_{0,q} = \frac{p(\mathbf{x}|M_0)}{p(\mathbf{x}|M_q)}$.

If $\{z_1, ..., z_q\}$ are suspected of being upper outliers and ϵ is given an exponential(λ) prior distribution it follows that

$$\begin{aligned} p\left(\mathbf{x}|M_q\right) &= \int_0^\infty \int_{r_4}^\infty \int_{\max\left\{z_g-\epsilon,\ z_{[h]}\right\}-\frac{r}{2}}^{\min\left\{z_g-\epsilon,\ z_{[h]}\right\}+\frac{r}{2}} \frac{1}{r^n}\frac{\alpha\beta}{r}\lambda e^{-\lambda\epsilon} dm dr d\epsilon \\ &= \int_0^\infty \frac{\alpha\beta\lambda e^{-\lambda\epsilon}}{n\left(n-1\right)\left(\max\left\{z_g-\epsilon,\ z_{[h]}\right\} - \min\left\{z_g-\epsilon,\ z_{[h]}\right\}\right)^{n-1}} d\epsilon, \end{aligned}$$

where

$$r_4 = \left(\max\left\{z_g-\epsilon,\ z_{[h]}\right\} - \min\left\{z_g-\epsilon,\ z_{[h]}\right\}\right).$$

Let z_i and $z_{i'}$ denote the largest and smallest of the z_g respectively, where it is assumed that $z_{i'}$ differs more from the sample median than $\min\left\{z_{[h]}\right\}$ does because otherwise we will have potential lower outliers in the sample. It can

be shown that the Bayes factor for comparing the models M_0 and M_q is

$$B_{0,q} = \frac{1}{\tau_1\left(z_i - \min\left\{z_{[h]}\right\}\right)^{n-1}} \text{ when } z_{i'} - \min\left\{z_{[h]}\right\} \geq z_i - \max\left\{z_{[h]}\right\}$$

and

$$B_{0,q} = \frac{1}{\tau_2\left(z_i - \min\left\{z_{[h]}\right\}\right)^{n-1}} \text{ when } z_{i'} - \min\left\{z_{[h]}\right\} \leq z_i - \max\left\{z_{[h]}\right\},$$

where

$$\begin{aligned}
\tau_1 &= \frac{e^{-\lambda\left(z_i - \max\left\{z_{[h]}\right\}\right)} - e^{-\lambda\left(z_{i'} - \min\left\{z_{[h]}\right\}\right)}}{\left(\max\left\{z_{[h]}\right\} - \min\left\{z_{[h]}\right\}\right)^{n-1}} \\
&+ \int_0^{z_i - \max\left\{z_{[h]}\right\}} \frac{1}{\left(z_i - \min\left\{z_{[h]}\right\} - \epsilon\right)^{n-1}} \lambda e^{-\lambda\epsilon} d\epsilon \\
&+ \int_{z_{i'} - \min\left\{z_{[h]}\right\}}^{\infty} \frac{1}{\left(\max\left\{z_{[h]}\right\} - z_{i'} + \epsilon\right)^{n-1}} \lambda e^{-\lambda\epsilon} d\epsilon
\end{aligned}$$

and

$$\begin{aligned}
\tau_2 &= \frac{e^{-\lambda\left(z_{i'} - \min\left\{z_{[h]}\right\}\right)} - e^{-\lambda\left(z_i - \max\left\{z_{[h]}\right\}\right)}}{\left(z_i - z_{i'}\right)^{n-1}} \\
&+ \int_0^{z_{i'} - \min\left\{z_{[h]}\right\}} \frac{1}{\left(z_i - \min\left\{z_{[h]}\right\} - \epsilon\right)^{n-1}} \lambda e^{-\lambda\epsilon} d\epsilon \\
&+ \int_{z_i - \max\left\{z_{[h]}\right\}}^{\infty} \frac{1}{\left(\max\left\{z_{[h]}\right\} - z_{i'} + \epsilon\right)^{n-1}} \lambda e^{-\lambda\epsilon} d\epsilon.
\end{aligned}$$

All of the integrals can be evaluated numerically because they have the same form as the integrals I^u and J^u.

If $\{z_1, ..., z_q\}$ are suspected of being lower outliers and $-\epsilon$ is given an exponential(λ) prior distribution it follows that

$$\begin{aligned}
p\left(\mathbf{x}|M_q\right) &= \int_{-\infty}^{0} \int_{r_4}^{\infty} \int_{\max\left\{z_g - \epsilon,\ z_{[h]}\right\} - \frac{r}{2}}^{\min\left\{z_g - \epsilon,\ z_{[h]}\right\} + \frac{r}{2}} \frac{1}{r^n} \frac{\alpha\beta}{r} \lambda e^{\lambda\epsilon} dm dr d\epsilon \\
&= \int_{-\infty}^{0} \frac{\alpha\beta\lambda e^{\lambda\epsilon}}{n\left(n-1\right)\left(\max\left\{z_g - \epsilon,\ z_{[h]}\right\} - \min\left\{z_g - \epsilon,\ z_{[h]}\right\}\right)^{n-1}} d\epsilon.
\end{aligned}$$

Let z_i and $z_{i'}$ denote the smallest and largest of the z_g respectively, where it is assumed that $z_{i'}$ differs more from the sample median than $\max\left\{z_{[h]}\right\}$ does because otherwise we will have potential upper outliers in the sample. It can be shown that the Bayes factor for comparing the models M_0 and M_q is

$$B_{0,q} = \frac{1}{\tau_3\left(\max\left\{z_{[h]}\right\} - z_i\right)^{n-1}} \text{ when } z_{i'} - \max\left\{z_{[h]}\right\} \leq z_i - \min\left\{z_{[h]}\right\}$$

and

$$B_{0,q} = \frac{1}{\tau_4\left(\max\left\{z_{[h]}\right\} - z_i\right)^{n-1}} \text{ when } z_{i'} - \max\left\{z_{[h]}\right\} \geq z_i - \min\left\{z_{[h]}\right\},$$

where

$$\begin{aligned}
\tau_3 &= \frac{e^{\lambda\left(z_i - \min\left\{z_{[h]}\right\}\right)} - e^{\lambda\left(z_{i'} - \max\left\{z_{[h]}\right\}\right)}}{\left(\max\left\{z_{[h]}\right\} - \min\left\{z_{[h]}\right\}\right)^{n-1}} \\
&+ \int_{z_i - \min\left\{z_{[h]}\right\}}^{0} \frac{1}{\left(\max\left\{z_{[h]}\right\} - z_i + \epsilon\right)^{n-1}} \lambda e^{\lambda\epsilon} d\epsilon \\
&+ \int_{-\infty}^{z_{i'} - \max\left\{z_{[h]}\right\}} \frac{1}{\left(z_{i'} - \min\left\{z_{[h]}\right\} - \epsilon\right)^{n-1}} \lambda e^{\lambda\epsilon} d\epsilon
\end{aligned}$$

and

$$\begin{aligned}
\tau_4 &= \frac{e^{\lambda\left(z_{i'} - \max\left\{z_{[h]}\right\}\right)} - e^{\lambda\left(z_i - \min\left\{z_{[h]}\right\}\right)}}{\left(z_{i'} - z_i\right)^{n-1}} \\
&+ \int_{z_{i'} - \max\left\{z_{[h]}\right\}}^{0} \frac{1}{\left(\max\left\{z_{[h]}\right\} - z_i + \epsilon\right)^{n-1}} \lambda e^{\lambda\epsilon} d\epsilon \\
&+ \int_{-\infty}^{z_i - \min\left\{z_{[h]}\right\}} \frac{1}{\left(z_{i'} - \min\left\{z_{[h]}\right\} - \epsilon\right)^{n-1}} \lambda e^{\lambda\epsilon} d\epsilon.
\end{aligned}$$

All of the integrals can be evaluated numerically because they have the same form as the integrals I^l and J^l.

If we were to assume that ϵ is known, then the corresponding Bayes factor is equal to

$$B_{0,q} = \left(\frac{\max\left\{z_i - \epsilon,\ z_{[h]}\right\} - \min\left\{z_i - \epsilon,\ z_{[h]}\right\}}{\max\left\{x_j\right\} - \min\left\{x_j\right\}} \right)^{n-1}.$$

Also we use the same procedure as before to deal with masking and swamping, based on adding the next largest observation in the case when we have upper outliers and the next smallest observation in the case when we have lower outliers into the set of extreme observations on each iteration.

If ϵ_1 and ϵ_2 are both assumed to be known and $\epsilon_1 > \epsilon_2$, then it can be shown that the Bayes factor for testing whether the model M_0 that all of the X_j have a uniform(θ_1, θ_2) distribution or the model M_{q+q^*} that all of the $Z_{[h]}$ have a uniform(θ_1, θ_2) distribution, all of the Z_g have a uniform$(\theta_1 + \epsilon_1, \theta_2 + \epsilon_1)$ distribution and all of the Z_{g^*} have a uniform$(\theta_1 + \epsilon_2, \theta_2 + \epsilon_2)$ distribution, is more appropriate is

$$B_{0,q+q^*} = \left(\frac{\max\left\{z_i - \epsilon_1,\ z_{i^*} - \epsilon_2,\ z_{[h]}\right\} - \min\left\{z_i - \epsilon_1,\ z_{i^*} - \epsilon_2,\ z_{[h]}\right\}}{\max\left\{x_j\right\} - \min\left\{x_j\right\}} \right)^{n-1}.$$

Similar Bayes factors can be derived for case when we have any number of sets of outliers, but when this number of sets is large it is questionable as to what extreme means. We use the same procedure as before to deal with masking and swamping, based on adding the next most extreme observation into the set of observations suspected of being least contaminated and then considering all of the reasonable outlier models on each iteration.

When the amounts of contamination are unknown, we use the following method for the case of testing whether extreme observations are outliers generated by different probability distributions:

(i) Consider the least extreme observation that could possibly be an outlier and delete all of the more extreme observations from the sample. Then for this new sample test whether it is an outlier. If we conclude that it is, then all of the other extreme observations are outliers. Otherwise we conclude that this observation is not an outlier and repeat the procedure until either we conclude that an observation is an outlier or that none of the observations are outliers.

If we are only interested in declaring whether or not extreme observations are outliers, then this is sufficient, otherwise we continue by using the following method to see which outliers are generated by the same probability distribution.

(ii) Consider the two least extreme outliers and delete the rest of them. If we find out that these two outliers are generated by the same probability distribution (noting that this will not be the case if one is an upper outlier and the other is a lower outlier), then we consider adding in a third outlier. Otherwise we conclude that the first and second least extreme outliers are generated by different probability distributions, but then consider if the first and third or second and third least extreme outliers are generated by the same probability distribution using the same approach. This is done until we have found out which probability distribution the most extreme outlier has been generated by relative to the other outliers.

Example 2.6 As an example to illustrate the methods in this subsection, we return to our uniform$(-1, 1)$ data which was used in **Example 2.5**.

(i) We have added three to observations one and five so that they are now 2.775 and 3.561 respectively. When $\lambda = \frac{1}{3}$ it can be shown that $B_{0,1} = 0.265$

and therefore by comparing this to our answer from part (ii) of **Example 2.5**, we can see that masking has definitely occurred. We know that $z_i = 3.561$, $z_{i'} = 2.775$, $\max\left\{z_{[h]}\right\} = 0.972$ and $\min\left\{z_{[h]}\right\} = -0.981$, hence $2.775 + 0.981 > 3.561 - 0.972$ and so we calculate $B_{1,2}$ based on τ_1. Therefore it can be shown that $B_{1,2} = 9.71 \times 10^{-3}$ and we should certainly select M_2 over M_1. We do not have to calculate any more Bayes factors because we can clearly see that 0.972 is not extreme, therefore we select M_2 as our final model and conclude that 3.561 and 2.775 are upper outliers generated by the same probability distribution. Note that the $B_{\gamma,\gamma+1}$ are hardly at all sensitive to any reasonable choice of λ, but very sensitive to the sample size. This is because of the same reasons that were given for the additive multiple outlier model when ϵ is unknown in the one parameter case.

(ii) We have added three to observation five and subtracted three from observation two so that they are now 3.561 and -3.745 respectively. We can clearly see that -0.981 is not extreme and therefore start by deleting observation two from the sample and testing whether 3.561 is an upper outlier. When $\lambda = \frac{1}{3}$ it can be shown that $B_{0,1} = 4.42 \times 10^{-3}$, hence 3.561 is an upper outlier and therefore we definitely conclude that 3.561 is an upper outlier and -3.745 is a lower outlier.

Having built on the Bayesian methods discussed in **Section 1.3**, we come to the following conclusions:

For a sample from a one parameter uniform distribution we have shown that the largest observation in the sample has the smallest conditional predictive ordinate. Hence we have derived the Bayes factor for testing whether it is an outlier when the amount of contamination is known and unknown us-

ing two different outlier models. We then investigated this problem when we had multiple outliers, assuming that our outliers are generated by the same probability distribution or by different probability distributions. Similarly for two parameter uniform samples we have shown that the most extreme observation in the sample has the smallest conditional predictive ordinate. Hence we derived the Bayes factors for testing whether extreme observations are outliers using an additive outlier model. In practice, outlier detection and declaration is only a secondary task of the analysis of a data set. Therefore our methods for uniform samples could be built into statistical software packages in a similar way to existing outlier tests and goodness of fit tests, so that they can be used without much additional effort.

3. MODELLING OUTLIERS IN PARETO SAMPLES

3.1 Modelling a single outlier in a Pareto sample

Suppose that $\{X_1, ..., X_n\}$ are independent Pareto(θ, k) random variables, then the joint probability density function of $\{X_1, ..., X_n\}$ not including X_i is

$$p\left(\mathbf{x_{(i)}}|\theta, k\right) = \theta^{n-1} e^{-\theta \sum_{j \neq i} \log\left(\frac{x_j}{k}\right)} \prod_{j \neq i} \frac{1}{x_j},$$

for $\theta > 0$, $0 < k < s$ and $s = \min\{x_j : j \neq i\}$. The conjugate prior for θ is a gamma prior and for this problem θ shall be given a gamma(α, β) prior, so that $p(\theta) = \frac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)}$, for $\theta > 0$ and where both α and β are assumed to be known. When k is assumed to be known, we use the transformation $Y_j = \log\left(\frac{X_j}{k}\right)$ to transform our data to an exponential(θ) sample so that the methods in Pettit (1988) can be used to model outliers. We shall assume that k is unknown and is given an improper prior such that $p(k) = \frac{a}{k}$, where a is an unknown constant whose exact value does not matter in what follows. We can find the joint posterior density function of θ and k given $\{x_1, ..., x_n\}$ not including x_i, which is written as $p\left(\theta, k|\mathbf{x_{(i)}}\right)$ and is such that

$$\begin{aligned} p\left(\theta, k|\mathbf{x_{(i)}}\right) &\propto p\left(\mathbf{x_{(i)}}|\theta, k\right) p(\theta)\, p(k) \\ &\propto \theta^{\alpha+n-2} e^{-\theta\left(\beta + \sum_{j \neq i} \log\left(\frac{x_j}{k}\right)\right)} \frac{1}{k}. \end{aligned}$$

The constant C is such that

$$\begin{aligned} & C\int_0^s \int_0^\infty \theta^{\alpha+n-2} e^{-\theta\left(\beta+\sum_{j\neq i}\log\left(\frac{x_j}{k}\right)\right)} \frac{1}{k} d\theta dk \\ = \; & C\int_0^s \frac{\Gamma(\alpha+n-1)}{\left(\beta+\sum_{j\neq i}\log\left(\frac{x_j}{k}\right)\right)^{\alpha+n-1} k} dk \\ = \; & \frac{C\Gamma(\alpha+n-1)}{(n-1)(\alpha+n-2)\left(\beta+\sum_{j\neq i}\log\left(\frac{x_j}{s}\right)\right)^{\alpha+n-2}} \\ = \; & 1, \end{aligned}$$

thus

$$C = \frac{(n-1)(\alpha+n-2)\left(\beta+\sum_{j\neq i}\log\left(\frac{x_j}{s}\right)\right)^{\alpha+n-2}}{\Gamma(\alpha+n-1)}$$

and so

$$p\left(\theta,k|\mathbf{x_{(i)}}\right) = \frac{(n-1)(\alpha+n-2)\left(\beta+\sum_{j\neq i}\log\left(\frac{x_j}{s}\right)\right)^{\alpha+n-2}\theta^{\alpha+n-2}e^{-\theta\left(\beta+\sum_{j\neq i}\log\left(\frac{x_j}{k}\right)\right)}}{\Gamma(\alpha+n-1)\,k}.$$

The conditional predictive ordinate is then given by

$$\begin{aligned} p\left(x_i|\mathbf{x_{(i)}}\right) &= \int_0^{s'}\int_0^\infty p(x_i|\theta,k)\,p\left(\theta,k|\mathbf{x_{(i)}}\right)d\theta dk \\ &= \int_0^{s'}\int_0^\infty C\theta^{\alpha+n-2}e^{-\theta\left(\beta+\sum_{j\neq i}\log\left(\frac{x_j}{k}\right)\right)}\frac{1}{k}\frac{\theta}{x_i}e^{-\theta\log\left(\frac{x_i}{k}\right)}d\theta dk \\ &= \frac{C\Gamma(\alpha+n)}{n(\alpha+n-1)\,x_i\left(\beta+\sum_{j=1}^n\log\left(\frac{x_j}{s'}\right)\right)^{\alpha+n-1}}, \end{aligned}$$

where $s' = \min\{x_j\}$. We can clearly see that the largest observation in the sample has the smallest conditional predictive ordinate and therefore will not consider small observations to be possible lower outliers. From hereafter an upper outlier is referred to as an outlier.

Suppose that x_i is the largest observation in the sample and it is suspected of being an outlier. We can derive the Bayes factor to test whether the model

M_0 that all of the X_j have a Pareto(θ, k) distribution or the model M_1 that all of the X_j except for X_i have a Pareto(θ, k) distribution and X_i has a Pareto$(\theta, \delta k)$ distribution, is more appropriate, where $\delta > 1$ and is known. The Bayes factor is $B_{0,1} = \frac{p(\mathbf{x}|M_0)}{p(\mathbf{x}|M_1)}$, where

$$\begin{aligned} p(\mathbf{x}|M_0) &= \int_0^s \int_0^\infty \frac{\theta^n e^{-\theta \sum_{j=1}^n \log\left(\frac{x_j}{k}\right)} \beta^\alpha \theta^{\alpha-1} e^{-\beta\theta} a}{\left(\prod_{j=1}^n x_j\right) \Gamma(\alpha) k} d\theta dk \\ &= \frac{a\beta^\alpha \Gamma(\alpha+n)}{n(\alpha+n-1)\Gamma(\alpha)\left(\prod_{j=1}^n x_j\right)\left(\beta + \sum_{j=1}^n \log\left(\frac{x_j}{s}\right)\right)^{\alpha+n-1}}, \\ p(\mathbf{x}|M_1) &= \int_0^{s^*} \int_0^\infty \frac{\theta^{n-1} e^{-\theta \sum_{j\neq i} \log\left(\frac{x_j}{k}\right)} \theta e^{-\theta \log\left(\frac{x_i}{\delta k}\right)} \beta^\alpha \theta^{\alpha-1} e^{-\beta\theta} a}{\left(\prod_{j=1}^n x_j\right) \Gamma(\alpha) k} d\theta dk \\ &= \frac{a\beta^\alpha \Gamma(\alpha+n)}{n(\alpha+n-1)\Gamma(\alpha)\left(\prod_{j=1}^n x_j\right)\left(\beta + \sum_{j=1}^n \log\left(\frac{x_j}{s^*}\right) - \log(\delta)\right)^{\alpha+n-1}} \end{aligned}$$

and

$$s^* = \min\left\{\frac{x_i}{\delta},\ x_j : j \neq i\right\}.$$

Therefore

$$B_{0,1} = \left(\frac{\beta + \sum_{j=1}^n \log\left(\frac{x_j}{s^*}\right) - \log(\delta)}{\beta + \sum_{j=1}^n \log\left(\frac{x_j}{s}\right)}\right)^{\alpha+n-1},$$

which is minimized when x_i is very large compared to the other observations for any reasonable choice of δ and any sensible sample size such as $n \geq 5$.

Now consider the previous testing problem when δ is unknown. We shall give δ an improper prior such that $p(\delta) = \frac{b_1}{\delta}$, for $\delta > 1$ and where b_1 is an unknown constant to be determined. In order for it to make sense to test for outliers we assume that $n \geq 5$. The Bayes factor is $B_{0,1} = \frac{p(\mathbf{x}|M_0)}{p(\mathbf{x}|M_1)}$, where $p(\mathbf{x}|M_0)$ is given by the same expression as we had for this problem when δ

was known and

$$\begin{aligned} p\left(\mathbf{x}|M_1\right) &= \int_1^{\infty}\int_0^{s^*}\int_0^{\infty}\frac{\theta^{n-1}e^{-\theta\sum_{j\neq i}\log\left(\frac{x_j}{k}\right)}\theta e^{-\theta\log\left(\frac{x_i}{\delta k}\right)}\beta^{\alpha}\theta^{\alpha-1}e^{-\beta\theta}ab_1}{\left(\prod_{j=1}^{n}x_j\right)\Gamma(\alpha)k\delta}d\theta dk d\delta \\ &= \int_1^{\infty}\frac{ab_1\beta^{\alpha}\Gamma\left(\alpha+n\right)}{n\left(\alpha+n-1\right)\Gamma\left(\alpha\right)\left(\prod_{j=1}^{n}x_j\right)\left(\beta+\sum_{j=1}^{n}\log\left(\frac{x_j}{s^*}\right)-\log\left(\delta\right)\right)^{\alpha+n-1}\delta}d\delta. \end{aligned}$$

By splitting the previous integral up:

If $s^* = s$, we have

$$\begin{aligned} &\int_1^{\frac{x_i}{s}}\frac{ab_1\beta^{\alpha}\Gamma\left(\alpha+n\right)}{n\left(\alpha+n-1\right)\Gamma\left(\alpha\right)\left(\prod_{j=1}^{n}x_j\right)\left(\beta+\sum_{j=1}^{n}\log\left(\frac{x_j}{s}\right)-\log\left(\delta\right)\right)^{\alpha+n-1}\delta}d\delta \\ = &\frac{ab_1\beta^{\alpha}\Gamma\left(\alpha+n\right)}{n\left(\alpha+n-1\right)\left(\alpha+n-2\right)\Gamma\left(\alpha\right)\left(\prod_{j=1}^{n}x_j\right)\left(\beta+\sum_{j\neq i}\log\left(\frac{x_j}{s}\right)\right)^{\alpha+n-2}} \\ - &\frac{ab_1\beta^{\alpha}\Gamma\left(\alpha+n\right)}{n\left(\alpha+n-1\right)\left(\alpha+n-2\right)\Gamma\left(\alpha\right)\left(\prod_{j=1}^{n}x_j\right)\left(\beta+\sum_{j=1}^{n}\log\left(\frac{x_j}{s}\right)\right)^{\alpha+n-2}} \\ = &\frac{a\beta^{\alpha}\Gamma\left(\alpha+n\right)}{n\left(\alpha+n-1\right)\Gamma\left(\alpha\right)\prod_{j=1}^{n}x_j}\phi_a; \end{aligned}$$

If $s^* = \frac{x_i}{\delta}$, we have

$$\begin{aligned} &\int_{\frac{x_i}{s}}^{\infty}\frac{ab_1\beta^{\alpha}\Gamma\left(\alpha+n\right)}{n\left(\alpha+n-1\right)\Gamma\left(\alpha\right)\left(\prod_{j=1}^{n}x_j\right)\left(\beta+\sum_{j=1}^{n}\log\left(\frac{x_j}{x_i}\right)+\left(n-1\right)\log\left(\delta\right)\right)^{\alpha+n-1}\delta}d\delta \\ = &\frac{ab_1\beta^{\alpha}\Gamma\left(\alpha+n\right)}{n\left(n-1\right)\left(\alpha+n-1\right)\left(\alpha+n-2\right)\Gamma\left(\alpha\right)\left(\prod_{j=1}^{n}x_j\right)\left(\beta+\sum_{j\neq i}\log\left(\frac{x_j}{s}\right)\right)^{\alpha+n-2}} \\ = &\frac{a\beta^{\alpha}\Gamma\left(\alpha+n\right)}{n\left(\alpha+n-1\right)\Gamma\left(\alpha\right)\prod_{j=1}^{n}x_j}\phi_b. \end{aligned}$$

Therefore

$$p\left(\mathbf{x}|M_1\right) = \frac{a\beta^{\alpha}\Gamma\left(\alpha+n\right)}{n\left(\alpha+n-1\right)\Gamma\left(\alpha\right)\prod_{j=1}^{n}x_j}\left(\phi_a+\phi_b\right)$$

and hence

$$B_{0,1} = \frac{1}{\left(\beta+\sum_{j=1}^{n}\log\left(\frac{x_j}{s}\right)\right)^{\alpha+n-1}\left(\phi_a+\phi_b\right)}.$$

When x_i is very large compared to the other observations, the Bayes factor will be close to zero because $\lim_{x_i \to \infty} (B_{0,1}) = 0$ and in such cases we should conclude that x_i is an outlier. To find the constant b_1, we use the method of imaginary observations described in Spiegelhalter and Smith (1982). The smallest possible experiment to distinguish between the models M_0 and M_1 would have two observations and gives maximal support to M_0 if they are equal. We shall denote this observation by x and so b_1 can be found by solving the equation $B_{0,1} = 1$ for b_1. When we have a sample of two equal observations $\log\left(\frac{x_j}{s}\right) = 0$, so that $\phi_a = 0$ and $\phi_b = \frac{b_1}{\alpha\beta^\alpha}$, hence $B_{0,1} = \frac{\alpha}{\beta b_1}$ and therefore $b_1 = \frac{\alpha}{\beta}$.

Example 3.1 As an example to illustrate the previous methods, we shall consider the data displayed in **Table 3.1** (Ryland, 1841) which was taken from Barnett and Lewis (1995). This shows the annual incomes (in order of magnitude to the nearest pound) of the top 69 of 91 scientific and literary societies in England in 1840. Barnett and Lewis (1995) argue that a Pareto model is appropriate for the original data set of 91 observations and show that this argument is unaffected by the fact that they have truncated the lower values (less than 75 pounds) in the original data set. They suspect that the observation 7000 is an outlier.

Figure 3.1 shows the quantile plot for this data set, which suggests that there is little evidence to suspect that the observation 7000 is an outlier because the corresponding point lies roughly on the straight line connecting all of the other points. **Figure 3.2** again shows the quantile plot for this data set with the observation 7000 replaced by 12000, which suggests that 12000 may possibly be an outlier. We shall use the previous methods to

Tab. 3.1: (Ryland, 1841) showing the annual incomes (in order of magnitude to the nearest pound) of the top 69 of 91 scientific and literary societies in England in 1840 and taken from Barnett and Lewis (1995)

77	77	79	80	80	84	87	90	90	90
92	100	102	110	112	115	120	120	120	125
130	135	136	138	140	147	150	150	169	170
170	190	200	200	200	200	201	206	208	230
230	237	249	290	300	309	335	350	400	404
431	445	456	500	650	650	700	800	844	900
900	1050	1300	1400	1878	2000	2363	3000	7000	

formally test whether 7000 is an outlier for the cases when both δ is known and unknown. This shall be done using a variety of different combinations of α and β, noting that the maximum likelihood estimate of k is equal to 77 and the maximum likelihood estimate of θ is close to 0.8.

(i) When δ is known and $Y \sim \text{Pareto}(0.8, 77\delta)$, we can find the value of δ such that $p(Y > 7000) = \pi$, where π is some fixed probability. For $\pi = 0.90$, 0.95 or 0.99 it follows that $\delta = 80$, 85 or 90 respectively. In all of these cases $s^* = s = 77$, hence $\sum_{j=1}^{69} \log\left(\frac{x_j}{s}\right) = 83.7829$ and so the Bayes factors using our various different combinations of the known parameters are given in **Table 3.2**, **Table 3.3** and **Table 3.4**. We can see from these three tables that the Bayes factor is hardly at all sensitive to any reasonable choice of δ. Also we see that the Bayes factor is not sensitive to our choices of α and β when the prior mean of θ is equal to 0.8, otherwise the size of the Bayes factor gets smaller as the prior mean of θ gets larger. The reason for this is

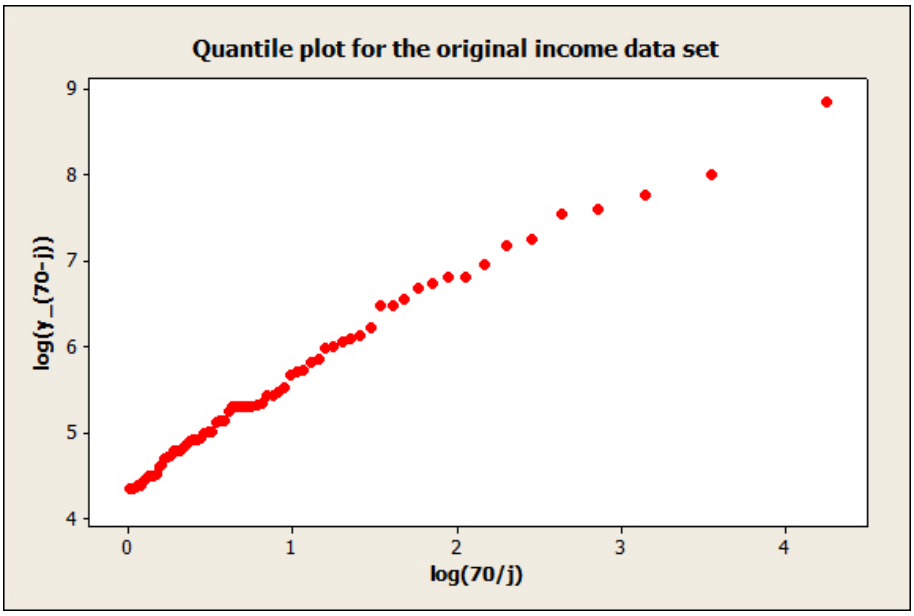

Fig. 3.1: Quantile plot for the original income data set, where $\{y_1, ..., y_{69}\}$ denotes the ascending ordered sample.

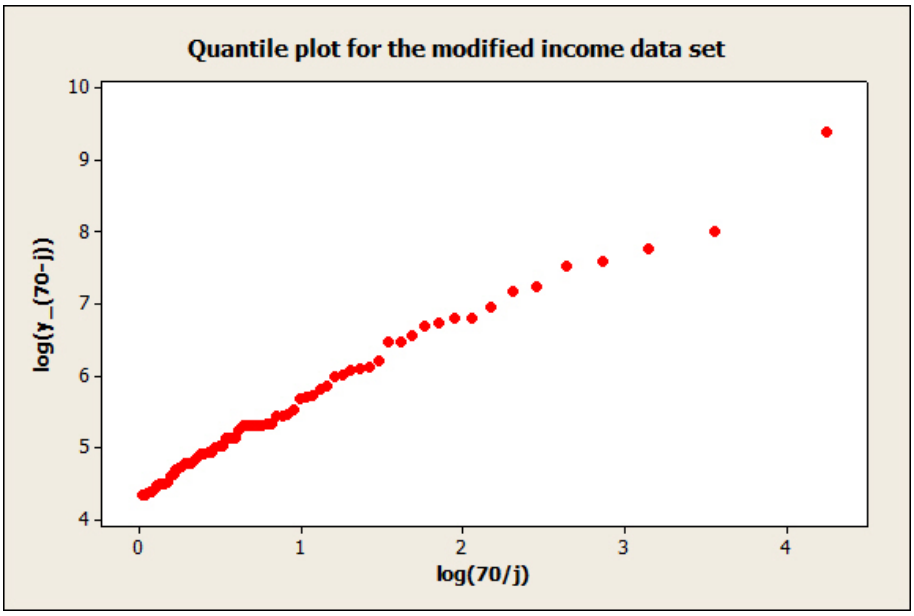

Fig. 3.2: Quantile plot for the modified income data set, where $\{y_1, ..., y_{69}\}$ denotes the ascending ordered sample.

because for M_0 we know $p\left(X_i > x_i\right) = \left(\frac{k}{x_i}\right)^{\theta}$ and so large observations are less likely as θ gets larger. Therefore as $B_{0,1}$ is greater than 0.015 for most of these combinations of the known parameters there is not enough evidence to conclude that 7000 is an outlier. Note that in general, increasing the sample size has a similar effect to increasing α, which is shown by our formula for $B_{0,1}$.

(ii) When δ is unknown, $\sum_{j=1}^{69} \log\left(\frac{x_j}{s}\right) = 83.7829$ and $\sum_{j\neq i} \log\left(\frac{x_j}{s}\right) = 79.2730$, hence the Bayes factors using exactly the same combinations of α and β as before are given in **Table 3.5**. We can see from **Table 3.5** that the Bayes factor is not sensitive to our choices of α and β when the prior mean of θ is equal to 0.8 and has similar values to the case when δ is known. Otherwise the larger we initially believe the mean of θ is, the more likely we are to conclude that 7000 is an outlier for the same reason as before. Again in general, increasing the sample size has a similar effect to increasing α. Note that for $\alpha = 4$ and $\beta = 5$ it can be shown that the critical value for this sample is equal to 12970 and agrees with our quantile plots.

Tab. 3.2: Bayes factors when $\delta = 80$

	α				
β	1	2	4	8	16
1.25	0.0260	0.0246	0.0222	0.0179	0.0117
2.5	0.0274	0.0260	0.0235	0.0190	0.0125
5	0.0304	0.0289	0.0261	0.0213	0.0142
10	0.0368	0.0351	0.0319	0.0263	0.0180
20	0.0510	0.0488	0.0448	0.0377	0.0267

Tab. 3.3: Bayes factors when $\delta = 85$

	α				
β	1	2	4	8	16
1.25	0.0247	0.0234	0.0210	0.0169	0.0110
2.5	0.0261	0.0247	0.0222	0.0180	0.0118
5	0.0290	0.0275	0.0248	0.0202	0.0134
10	0.0351	0.0335	0.0304	0.0250	0.0170
20	0.0489	0.0468	0.0429	0.0360	0.0253

Tab. 3.4: Bayes factors when $\delta = 90$

	α				
β	1	2	4	8	16
1.25	0.0235	0.0222	0.0199	0.0160	0.0104
2.5	0.0248	0.0235	0.0211	0.0171	0.0111
5	0.0276	0.0262	0.0236	0.0192	0.0127
10	0.0336	0.0320	0.0290	0.0238	0.0161
20	0.0470	0.0449	0.0411	0.0344	0.0242

Tab. 3.5: Bayes factors when δ is unknown

	α				
β	1	2	4	8	16
1.25	0.0248	0.0119	0.0055	0.0023	0.0008
2.5	0.0518	0.0249	0.0115	0.0049	0.0017
5	0.1121	0.0539	0.0249	0.0106	0.0039
10	0.2593	0.1250	0.0581	0.0251	0.0093
20	0.6613	0.3202	0.1501	0.0659	0.0253

3.2 Modelling multiple outliers in a Pareto sample

We now consider the problem when it is believed that more than one observation in the sample is an outlier. Suppose that $\{z_1, ..., z_q\}$ are the q largest observations in the sample and are suspected of being outliers generated by the same probability distribution, where q is the number of observations

that we suspect of being outliers, $q < n$, $\{Z_1, ..., Z_q\} \subset \{X_1, ..., X_n\}$ and $\left\{Z_{[1]}, ..., Z_{[n-q]}\right\}$ denote the random variables corresponding to the observations not suspected of being outliers. We can derive the Bayes factor to test whether the model M_0 that all of the X_j have a Pareto(θ, k) distribution or the model M_q that all of the $Z_{[h]}$ have a Pareto(θ, k) distribution and all of the Z_g have a Pareto$(\theta, \delta k)$ distribution, is more appropriate. It is assumed that δ is unknown, so that $p(\delta) = \frac{b_q}{\delta}$, for $\delta > 1$ and where b_q is an unknown constant to be determined. In what follows we again have $n \geq 5$. It is assumed that z_i is the smallest of the z_g and that $q < \frac{n}{2}$, as if $q \geq \frac{n}{2}$ it might imply that the $Z_{[h]}$ should be suspected of being contaminants rather than the Z_g. We write the Bayes factor as $B_{0,q} = \frac{p(\mathbf{x}|M_0)}{p(\mathbf{x}|M_q)}$ to compare these models. Therefore

$$\begin{aligned} p(\mathbf{x}|M_q) &= \int_1^\infty \int_0^t \int_0^\infty \frac{\theta^{n-q} e^{-\theta \sum_{h=1}^{n-q} \log\left(\frac{z_{[h]}}{k}\right)} \theta^q e^{-\theta \sum_{g=1}^{q} \log\left(\frac{z_g}{\delta k}\right)} \beta^\alpha \theta^{\alpha-1} e^{-\beta\theta} a b_q}{\left(\prod_{j=1}^n x_j\right) \Gamma(\alpha) k\delta} d\theta dk d\delta \\ &= \int_1^\infty \frac{a b_q \beta^\alpha \Gamma(\alpha+n)}{n(\alpha+n-1)\Gamma(\alpha)\left(\prod_{j=1}^n x_j\right)\left(\beta + \sum_{j=1}^n \log\left(\frac{x_j}{t}\right) - q\log(\delta)\right)^{\alpha+n-1} \delta} d\delta \\ &= \frac{a\beta^\alpha \Gamma(\alpha+n)}{n(\alpha+n-1)\Gamma(\alpha)\prod_{j=1}^n x_j} (\phi_c + \phi_d), \end{aligned}$$

where

$$\begin{aligned} \phi_c &= \frac{b_q}{q(\alpha+n-2)\left(\beta + \sum_{j=1}^n \log\left(\frac{x_j}{s}\right) - q\log\left(\frac{z_i}{s}\right)\right)^{\alpha+n-2}} \\ &- \frac{b_q}{q(\alpha+n-2)\left(\beta + \sum_{j=1}^n \log\left(\frac{x_j}{s}\right)\right)^{\alpha+n-2}}, \\ \phi_d &= \frac{b_q}{(n-q)(\alpha+n-2)\left(\beta + \sum_{j=1}^n \log\left(\frac{x_j}{s}\right) - q\log\left(\frac{z_i}{s}\right)\right)^{\alpha+n-2}} \end{aligned}$$

and

$$t = \min\left\{\frac{z_i}{\delta},\ z_{[h]}\right\}.$$

Hence

$$B_{0,q} = \frac{1}{\left(\beta + \sum_{j=1}^{n} \log\left(\frac{x_j}{s}\right)\right)^{\alpha+n-1} (\phi_c + \phi_d)}.$$

When the z_g do not hugely differ and are very large compared to the $z_{[h]}$, the Bayes factor will be close to zero because $\lim_{z_i \to \infty} (B_{0,q}) = 0$ and in such cases we should conclude that $\{z_1, ..., z_q\}$ are outliers generated by the same probability distribution. We again use the method of imaginary observations to find the constant b_q. The smallest possible experiment to distinguish between the models M_0 and M_q would have $q + 1$ observations and gives maximal support to M_0 if they are equal. We shall denote this observation by x and so it can be shown that $b_q = \frac{\alpha+q-1}{\beta}$. Note that if we were to assume that δ is known, then the corresponding Bayes factor is equal to

$$B_{0,q} = \left(\frac{\beta + \sum_{j=1}^{n} \log\left(\frac{x_j}{t}\right) - q \log(\delta)}{\beta + \sum_{j=1}^{n} \log\left(\frac{x_j}{s}\right)}\right)^{\alpha+n-1}.$$

We can see that the previous methods are sensitive to both masking and swamping and so it creates problems. To overcome these problems, we repeatedly calculate Bayes factors of the form $B_{\gamma,\gamma+1} = \frac{p(\mathbf{x}|M_\gamma)}{p(\mathbf{x}|M_{\gamma+1})}$, starting with $\gamma = 0$, until we can see that γ is equal to some value such that the $(\gamma + 1)^{\text{th}}$ largest observation in the sample is not extreme. If on all of the iterations we select the model M_γ over the model $M_{\gamma+1}$, then we should select M_0 as our final model. Otherwise our final model is $M_{\eta+1}$, which is the last model that we selected over the model suspected of having one less outlier. Using a similar derivation as we did for finding $B_{0,q}$ it can be shown that the Bayes factor for comparing the models M_γ and $M_{\gamma+1}$ is

$$B_{\gamma,\gamma+1} = \frac{\phi_{c,\gamma} + \phi_{d,\gamma}}{\phi_{c,\gamma+1} + \phi_{d,\gamma+1}},$$

where $\gamma \geq 1$. Also $\phi_{c,\gamma}$, $\phi_{c,\gamma+1}$, $\phi_{d,\gamma}$ and $\phi_{d,\gamma+1}$ have the same form as ϕ_c and ϕ_d respectively, except that we replace q by γ or $\gamma+1$ and z_i by the γ^{th} or $(\gamma+1)^{\text{th}}$ largest observation in the sample. Note that the constant b_q cancels out in all of the Bayes factors used in this procedure except for $B_{0,1}$. If we strongly believe that there are q outliers in the sample, we could use the previous method by putting γ equal to $q-1$ instead of zero on the first iteration, so that we can save time from not having to perform as many iterations to arrive at the final model. Also if we were to assume that δ is known, then

$$B_{\gamma,\gamma+1} = \left(\frac{\beta + \sum_{j=1}^{n} \log\left(\frac{x_j}{t_{\gamma+1}}\right) - (\gamma+1)\log(\delta)}{\beta + \sum_{j=1}^{n} \log\left(\frac{x_j}{t_\gamma}\right) - \gamma \log(\delta)} \right)^{\alpha+n-1},$$

where $t_\gamma = \min\left\{\frac{z_i}{\delta},\ z_{[1]}, ..., z_{[n-\gamma]}\right\}$, $t_{\gamma+1} = \min\left\{\frac{\tilde{z}_i}{\delta},\ \tilde{z}_{[1]}, ..., \tilde{z}_{[n-\gamma-1]}\right\}$, the sample $\left\{\tilde{z}_{[1]}, ..., \tilde{z}_{[n-\gamma-1]}\right\}$ is the same as the sample $\left\{z_{[1]}, ..., z_{[n-\gamma]}\right\}$ with $\max\left\{z_{[h]}\right\}$ removed from it and $\tilde{z}_i = \max\left\{z_{[h]}\right\}$.

If δ_1 and δ_2 are both assumed to be known and $\delta_1 > \delta_2$, then it can be shown that the Bayes factor for testing whether the model M_0 that all of the X_j have a Pareto(θ, k) distribution or the model M_{q+q^*} that all of the $Z_{[h]}$ have a Pareto(θ, k) distribution, all of the Z_g have a Pareto$(\theta, \delta_1 k)$ distribution and all of the Z_{g^*} have a Pareto$(\theta, \delta_2 k)$ distribution, is more appropriate is

$$B_{0,q+q^*} = \left(\frac{\beta + \sum_{j=1}^{n} \log\left(\frac{x_j}{u}\right) - q_1 \log(\delta_1) - q_2 \log(\delta_2)}{\beta + \sum_{j=1}^{n} \log\left(\frac{x_j}{s}\right)} \right)^{\alpha+n-1},$$

where z_i and z_{i^*} are the most extreme of the z_g and z_{g^*} respectively and

$$u = \min\left\{ \frac{z_i}{\delta_1},\ \frac{z_{i^*}}{\delta_2},\ z_{[h]} \right\}.$$

Similar Bayes factors can be derived for case when we have any number of sets of outliers, but when this number of sets is large it is questionable as to what extreme means. To deal with masking and swamping, we use the same procedure that we did for the case of comparing the models M_0 and M_q, except that on each iteration we compare the current model with all of the reasonable models containing one more outlier and use the one which gives the smallest Bayes factor as the current model for the next iteration. If on all of the iterations we select the current model over the best model with one more outlier, then we should select M_0 as our final model. Otherwise our final model is the last one that we selected over the model suspected of having one less outlier.

When δ is unknown and we do not conclude that a set of extreme observations are outliers generated by the same probability distribution, but still suspect that they are outliers, we use the following method:

(i) Consider the smallest observation that could possibly be an outlier and delete all of the more extreme observations from the sample. Then for this new sample test whether it is an outlier. If we conclude that it is, then all of the other extreme observations are outliers. Otherwise we conclude that this observation is not an outlier and repeat the procedure until either we conclude that an observation is an outlier or that none of the observations are outliers.

If we are only interested in declaring whether or not extreme observations are outliers, then this is sufficient, otherwise we continue by using the following method to see which outliers are generated by the same probability distribution.

(ii) Consider the two smallest outliers and delete the rest of them. If we find out that these two outliers are generated by the same probability distribution, then we consider adding in a third outlier. Otherwise we conclude that the first and second smallest outliers are generated by different probability distributions, but then consider if the second and third smallest outliers are generated by the same probability distribution while deleting the first smallest outlier as well as the fourth smallest to the largest outliers. This is done until we have found out which probability distribution the largest outlier has been generated by relative to the other outliers.

Example 3.2 As an example to illustrate the methods in this subsection, we return to our Income data which was used in **Example 3.1**. We have replaced the observation 7000 by 15000 and have also added an observation of 20000 to the sample. For $\frac{\alpha}{\beta} = 0.8$, we know from part (ii) of **Example 3.1** that 15000 is an outlier when deleting 20000 from the sample, hence 15000 and 20000 are definitely both outliers in this case. **Table 3.6** shows the Bayes factors for comparing the models M_1 and M_2 using the same combinations of α and β as before. We see that $B_{1,2}$ is not sensitive to our choices of α and β when the prior mean of θ is equal to 0.8, but otherwise the larger we initially believe the mean of θ is, the more likely we are to conclude that 15000 and 20000 are outliers generated by the same probability distribution. Therefore we generally conclude that 15000 and 20000 are outliers generated by different probability distributions.

Tab. 3.6: Bayes factors for comparing M_1 and M_2

	α				
β	1	2	4	8	16
1.25	0.0313	0.0295	0.0262	0.0207	0.0129
2.5	0.0332	0.0314	0.0279	0.0221	0.0138
5	0.0373	0.0353	0.0315	0.0251	0.0159
10	0.0461	0.0437	0.0393	0.0317	0.0207
20	0.0659	0.0629	0.0572	0.0472	0.0321

3.3 Modelling outliers in a multivariate Pareto sample

When we have N independent samples of independent Pareto random variables, all of our results for the univariate Pareto distribution can be extended to the multivariate case by performing our tests on each of the marginal Pareto samples individually. When these N samples are not independent ways of defining the multivariate Pareto distribution are discussed in Mardia (1962). For modelling outliers we use "Multivariate Pareto Type 1". This has probability density function

$$p\left(x_1,...,x_N|\theta,k_1,...,k_N\right)=\frac{\theta\left(\theta+1\right)...\left(\theta+N-1\right)}{\left(\prod_{l=1}^{N}k_l\right)\left\{\left(\sum_{l=1}^{N}{k_l}^{-1}x_l\right)-N+1\right\}^{\theta+N}},$$

for $x_l > k_l > 0$, $\theta > 0$, where the marginal distribution of X_l is Pareto(θ, k_l). When $\theta > 2$, all the correlations of zero order are equal to $\frac{1}{\theta}$, every partial correlation coefficient of the m^{th} order is $\frac{1}{\theta+m}$ and the multiple correlation of a variate with the other $N-1$ variates is $\left[\frac{N-1}{\theta(\theta+N-2)}\right]^{\frac{1}{2}}$. For the bivariate

case it follows that

$$p\left(x_1, x_2|\theta, k_1, k_2\right) = \frac{\theta\left(\theta+1\right)}{k_1 k_2\left(\frac{x_1}{k_1}+\frac{x_2}{k_2}-1\right)^{\theta+2}}$$

and therefore the joint probability density function of $\{(X_{11}, X_{21}), ..., (X_{1n}, X_{2n})\}$ not including (X_{1i}, X_{2i}) is given by

$$p\left(\mathbf{x_{1(i)}}, \mathbf{x_{2(i)}}|\theta, k_1, k_2\right) = \frac{\left(\theta^{2n-2}+\theta^{n-1}\right)e^{-(\theta+2)\sum_{j\neq i}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)}}{\left(k_1 k_2\right)^{n-1}},$$

for $\theta > 0$, $0 < k_1 < s_1$ and $0 < k_2 < s_2$, where $s_1 = \min\{x_{1j} : j \neq i\}$ and $s_2 = \min\{x_{2j} : j \neq i\}$. By letting $p(k_1) = \frac{a_1}{k_1}$, $p(k_2) = \frac{a_2}{k_2}$ and giving θ a gamma(α, β) prior it follows that

$$p\left(\theta, k_1, k_2|\mathbf{x_{1(i)}}, \mathbf{x_{2(i)}}\right) \propto \frac{\left(\theta^{\alpha+2n-3}+\theta^{\alpha+n-2}\right)e^{-\theta\left(\beta+\sum_{j\neq i}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)\right)}}{\left(k_1 k_2\right)^n e^{2\sum_{j\neq i}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)}}.$$

The constant C is such that

$$\begin{aligned}
& C\int_0^{s_2}\int_0^{s_1}\int_0^{\infty}\frac{\left(\theta^{\alpha+2n-3}+\theta^{\alpha+n-2}\right)e^{-\theta\left(\beta+\sum_{j\neq i}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)\right)}}{\left(k_1 k_2\right)^n e^{2\sum_{j\neq i}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)}}d\theta dk_1 dk_2 \\
= \; & C\int_0^{s_2}\int_0^{s_1}\frac{\Gamma\left(\alpha+2n-2\right)}{\left(k_1 k_2\right)^n e^{2\sum_{j\neq i}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)}\left(\beta+\sum_{j\neq i}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)\right)^{\alpha+2n-2}}dk_1 dk_2 \\
+ \; & C\int_0^{s_2}\int_0^{s_1}\frac{\Gamma\left(\alpha+n-1\right)}{\left(k_1 k_2\right)^n e^{2\sum_{j\neq i}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)}\left(\beta+\sum_{j\neq i}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)\right)^{\alpha+n-1}}dk_1 dk_2 \\
= \; & 1.
\end{aligned}$$

At this early stage the integral cannot be evaluated exactly and therefore we assume that both k_1 and k_2 are known.

When k_1 and k_2 are both assumed to be known and θ is unknown it follows that $C = \xi^{-1}$, where

$$\xi = \frac{\Gamma\left(\alpha+2n-2\right)}{\left(k_1 k_2\right)^{n-1} e^{2\sum_{j\neq i}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)}\left(\beta+\sum_{j\neq i}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)\right)^{\alpha+2n-2}}$$

$$+ \quad \frac{\Gamma(\alpha+n-1)}{(k_1k_2)^{n-1}e^{2\sum_{j\neq i}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)}\left(\beta+\sum_{j\neq i}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)\right)^{\alpha+n-1}}.$$

It can be shown that the conditional predictive ordinate is then given by $p\left(x_{1i}, x_{2i}|\mathbf{x_{1(i)}}, \mathbf{x_{2(i)}}\right) = \xi'$, where

$$\begin{aligned}
\xi' &= \frac{\Gamma(\alpha+2n)}{\xi(k_1k_2)^{n}e^{2\sum_{j=1}^{n}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)}\left(\beta+\sum_{j=1}^{n}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)\right)^{\alpha+2n}} \\
&+ \frac{\Gamma(\alpha+2n-1)}{\xi(k_1k_2)^{n}e^{2\sum_{j=1}^{n}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)}\left(\beta+\sum_{j=1}^{n}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)\right)^{\alpha+2n-1}} \\
&+ \frac{\Gamma(\alpha+n+1)}{\xi(k_1k_2)^{n}e^{2\sum_{j=1}^{n}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)}\left(\beta+\sum_{j=1}^{n}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)\right)^{\alpha+n+1}} \\
&+ \frac{\Gamma(\alpha+n)}{\xi(k_1k_2)^{n}e^{2\sum_{j=1}^{n}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)}\left(\beta+\sum_{j=1}^{n}\log\left(\frac{x_{1j}}{k_1}+\frac{x_{2j}}{k_2}-1\right)\right)^{\alpha+n}}.
\end{aligned}$$

We clearly see that the point in the sample with the smallest conditional predictive ordinate has the largest value of $\frac{x_{1i}}{k_1}+\frac{x_{2i}}{k_2}-1$. This is in agreement with Barnett (1979) and so we could use his approach for testing whether extreme points in the sample are outliers, as it is not even possible to derive Bayes factors for the case when both k_1 and k_2 are assumed to be known.

Having built on the Bayesian methods discussed in **Section 1.3**, we come to the following conclusions:

For a sample from a univariate Pareto distribution we have shown that the largest observation in the sample has the smallest conditional predictive ordinate and derived the Bayes factor for testing whether it is an outlier when the amount of contamination is known and unknown. We then investigated this problem when we had multiple outliers, assuming that our outliers are generated by the same probability distribution or by different probability distributions. Finally we extended these ideas to the multivariate case both

when the marginal samples are independent of one another and when there are correlations/partial correlations. In practice, outlier detection and declaration is only a secondary task of the analysis of a data set. Therefore our methods for Pareto samples could be built into statistical software packages in a similar way to existing outlier tests and goodness of fit tests, so that they can be used without much additional effort. A further research problem may be to use a copula based approach for modelling outliers in multivariate Pareto samples, as it could be used to overcome the difficulties that we had in **Section 3.3**.

4. REFERENCES

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

Barnett, V. (1979). Some outlier tests for multivariate samples. *South African Statistical Journal,* **13,** 29–52.

Barnett, V. and Lewis, T. (1995). *Outliers in Statistical Data,* 3rd edition. Chichester: Wiley.

Barnett, V. and Roberts, D. (1993). The problem of outlier tests in sample surveys. *Communications in Statistics – Theory and Methods,* **22,** 2703–2721.

Beckman, R.J. and Cook, R.D. (1983). Outlier.....s (with Discussion). *Technometrics,* **25,** 119–163.

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

David, H.A., Hartley, H.O. and Pearson, E.S. (1954). The distribution of the ratio, in a single normal sample, of range to standard deviation. *Biometrika,* **41,** 482–493.

Dixon, W.J. (1951). Ratios involving extreme values. *The Annals of Mathematical Statistics,* **22,** 68–78.

Ferguson, T.S. (1961). On the rejection of outliers. In *Proceedings of the*

Fourth Berkeley Symposium on Mathematical Statistics and Probability, **1,** 253–287.

Freeman, P.R. (1980). On the number of outliers in data from a linear model (with Discussion). In *Bayesian Statistics,* (eds. Bernardo, J.M. et al.) 349–365. University Press, Valencia.

Geisser, S. (1980). In discussion of G.E.P, Box. *Journal of the Royal Statistical Society, B,* **42,** 416–417.

Grubbs, F.E. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics,* **21,** 27–58.

Guttman, I. Dutter, R. 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.

Irwin, J.O. (1925). On a criterian for the rejection of outlying observations. *Biometrika,* **17,** 238–250.

Jeffreys, H. (1961). *Theory of Probability,* Oxford, U.K.: Oxford University Press.

Justel, A. and Pena, D. (1996). Gibbs sampling will fail in outlier problems with strong masking. *Journal of Computational and Graphical Statistics,* **5,** 176–189.

Kimber, A.C. (1982). Tests for many outliers in an exponential sample. *Journal of Applied statistics,* **31,** 263–271.

Kimber, A.C. and Stevens, H.J. (1981). The null distribution of a test for two upper outliers in an exponential sample. *Journal of Applied statistics,* **30,** 153–157.

Lewis, T. and Fieller, N.R.J. (1979). A recursive algorithm for null distributions for outliers: I. Gamma samples. *Technometrics,* **21,** 371–376.

Likes, J. (1966). Distribution of Dixon's statistics in the case of an exponential population. *Metrika,* **11,** 46–54.

Mardia, K.V. (1962). Multivariate Pareto Distributions. *The Annals of Mathematical Statistics,* **33,** 1008–1015.

O'Hagan, A. (1995). Fractional Bayes factors for models comparison (with discussion). *Journal of the Royal Statistical Society, B,* **57,** 99–138.

Pearson, E.S. and Stephens, M.A. (1964). The ratio of range to standard deviation in the same normal sample. *Biometrika,* **51,** 484–487.

Pettit, L.I. (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. (1994). Bayesian approaches to the detection of outliers in Poisson samples. *Communications in Statistics – Theory and Methods,* **23,** 1785–1795.

Pettit, L.I. (1995). Contribution to the discussion of O'Hagan, A. (1995). *Journal of the Royal Statistical Society, B,* **57,** 124–126.

Pettit, L.I. and Smith, A.F.M. (1983). Bayesian model comparisons in the presence of outliers. *Bulletin of the International Statistical Institute,* **50,** 292–306.

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

Ryland, A. (1841). Income of scientific and literary societies in England. *Journal of the Statistical Society,* **4,** 264–267.

Shapiro, S.S. and Wilk, M.B. (1965). An analysis of variance test for normality (complete samples) *Biometrika,* **52,** 591–611.

Shapiro, S.S. and Wilk, M.B. (1972). An analysis of variance test for the exponential distribution (complete samples) *Technometrics,* **14,** 335–370.

Shapiro, S.S., Wilk, M.B. and Chen M.J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association,* **63,** 1343–1372.

Sothinathan, N. and Pettit, L.I. (2005). Bayes methods for outliers in binomial samples. *Communications in Statistics – Theory and Methods,* **34,** 351–366.

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

Stephens, M.A. (1978). On the W-test for exponentiality with origin known. *Technometrics,* **20,** 33–35.

Verdinelli, I. and Wasserman, L. (1990). Bayesian analysis of outlier problems using the Gibbs sampler. *Statistics and Computing,* **1,** 105–117.